ASD
Big Data & Aplication

12th EDITION

02
03
MAY
2018

CONFERENCE ON
Advances of Decisional
Systems

Marrakech

Editors

Azedine BOULMAKOUL
Omar.BOUJSAID
Hassan.BADIR

Conference proceedings

rajouter S à proceedings

ASD'2018

12th Edition of the conference on Advances of Decisional Systems

Relever l'image, elle est écrase le texte d'en dessous

Marrakech-Morocco

May 2-3, 2018

Editors

Azedine BOULMAKOUL
Omar BOUSSAID
Hassan BADIR

Publisher: FST Mohammedia/IOSIS

Publisher by : FST Mohammedia / IOSIS

Partners


FST
FACULTÉ DES SCIENCES ET TECHNIQUES
DE MOHAMMEDIA
UNIVERSITÉ HASSAN II DE CASABLANCA


UNIVERSITE DE SFAX


UNIVERSITÉ
LUMIÈRE
LYON 2
UNIVERSITÉ DE LYON


EMSI
RECHERCHE
INNOVATION
EMSI
ECOLE MAROCAINE DES
SCIENCES DE L'INGENIEUR
Membre de
HONORIS UNITED UNIVERSITIES


ENSA


INPT


ANRT


i
s

Preface

The importance of Big Data for the scientific community, as well as for professionals, raises new interests and opens up new research tracks. The central focus of Data really attracts different communities around new research issues, and then creates new synergies. These communities are called upon to collaborate and cogitate together to understand the complexity of Data. They have challenges to develop new approaches, techniques and software tools, capable of generating value from Big data.

The pressure issued from the professional world creates regularly new scientific and technological challenges that academic researchers have to identify. Among these challenges, one encourages academic researchers and professionals to work together: thus, many advances in Data processing within different disciplines must be combined together in order to make Big Data a key resource for both organizations and scientific research.

ASD 2018 aims to consolidate the experiences of researchers, professionals and users from communities working on decision-making systems, Big Data and IoT (Internet of Things). The Decision Systems Advance Conference (ASD), extended to Big Data and IoT from this edition, is part of this process. The aim of this twelfth edition of the conference, especially after the success of previous editions, is to contribute to further boosting research in these fields and to create synergy between researchers, mainly but not exclusively North African, working in their country. Or in research laboratories abroad. On the other hand, it aims to strengthen existing links and build new relationships in order to bring out a thematic community decision-making systems and Big Data at the Maghreb level.

Replace North African by Maghrebians

This twelfth edition, with a strong connotation in Big Data and IoT (Internet des Objets),, will allow researchers who have already participated in previous editions to find themselves in the same context for presenting their latest scientific works. The ASD: Big Data & Applications conference addresses also professionals who are involved or interested in the decision-making and Big Data domains, who, through their pragmatic vision, will contribute by expressing new requirements or by valuing existing solutions. ASD: Big Data & Applications is also an opportunity for Ph.D. students to expose and reveal their work and to make themselves part of this community. This event in Big Data and IoT sectors will allow discovering specialized works in these fields, as well as receiving many feedbacks.

proceedings

The present acts regroup the articles accepted and presented during this new edition organized in the form of 8 tracks. ASD 2018 received 96 submissions from different countries (Algeria, France, Morocco, Tunisia, Palestine, Saudi Arabia, Greece). After evaluation by the scientific committee of each track, bringing together 81 international expert researchers in these fields, **60** long articles were selected. These papers cover different themes of research and application on decision-making systems, Big Data and IoT.

A supprimer, ne pas oublier la doubl e virgule

supprimer le gras

ASD 2018, entitled: ASD: Big Data & Applications will take place in Marrakech (Morocco) from May 2nd to 3rd, 2018. The Faculty of Science and Technology of Mohammedia, Hassan II Univ. of Casablanca and the Moroccan School of Engineering Sciences jointly organized this edition. It has received their support as well as those of various public institutions of teaching and research that we wish to thank: The “Centre National pour la Recherche Scientifique et Technique” (CNRST) and “Agence Nationale de Réglementation des Télécommunications” (ANRT); as well as international institutions: the Institute of Communication (ICOM) and the ERIC Laboratory of Lyon 2 Univ. (France), HASSAN University of casablanca. (Morocco), the Faculty of Sciences and Techniques of Mohammedia (Morocco), the Moroccan School of Engineering Sciences (Morocco), the Faculty of Economic Sciences and Management of Sfax (Tunisia), the Research Center in Computer Science, Multimedia and Digital Data Processing of Sfax (Tunisia), as well as all other institutions that have helped far or near to the success of this event.

supports

University

The success of this new edition of ASD: Big Data & Applications would not have been achieved without the close cooperation of the three committees: Steering Committee, Scientific Committee and Organizational Committee, which we would like to thank very warmly.

We are very grateful for their support.

We would like to thank all the authors who submitted to this edition of ASD: Big Data & Applications. We congratulate those whose articles have been accepted. We encourage other authors of unsuccessful papers to persevere and continue their efforts.

Editors

Azedine BOULMAKOUL, Omar BOUSSAID and Hassan BADIR

Committee

Conference General Chairs

- Azedine BOULMAKOUL, Université Hassan II, Maroc
- Kamal DAISSAOUI, EMSI, Morocco

Steering Committee

- BADIR Hassan (ENSAT, Morocco)
- BEN ABDALLAH Hanêne (MIRACL, University of Sfax, Tunisia)
- BENTAYEB Fadila (ERIC, University Lumière Lyon 2, France)
- BOULMAKOUL Azedine (University of Hassan II, Morocco)
- BOUSSAID Omar (ERIC, University of Lumière Lyon 2, France)
- FEKI Jamel (MIRACL, University of Sfax, Tunisia)
- GARGOURI Faiez (MIRACL, University of Sfax, Tunisia)
- HARBI Nouria (ERIC, University Lumière Lyon 2, France)

Organization Committee

- Karim ALAMI, EMSI, Morocco
- Zineb BESRI, Université Abdelmalek Essaadi Morocco
- Adil El BOUZIRI, Université Hassan II de Casablanca, Morocco
- Abdelfattah IDRI, Université Hassan II de Casablanca, Morocco
- Meriem MANDAR, Université Hassan I, Morocco
- Mohamed Sahbi MOALLA, ISET Sfax - Tunisia
- Mariyem OUKARFI, Université Hassan II de Casablanca, Morocco
- Aziz Mabrouk, Université Abdelmalek Essaadi Morocco
- Azedine BOULMAKOUL, Université Hassan II de Casablanca, Morocco
- Mohamed Tabaa , EMSI, Morocco
- Lamia KARIM, Université Hassan I, Morocco
- Rabia MARGHOUBI, INPT Rabat, Morocco
- Ghyzlane CHERRADI, Université Hassan II de Casablanca, Morocco

Program Committee

- ABDELMALEK Amine, Saida University, Algeria
- ABDI Mustapha K., Université d'Oran, Algérie
- AHMED OUAMER Rachid, Université Tizi Ouzou, Algérie
- ALIANE Hassina, CERIST, Algeria
- ALIMAZIGHI Zaia, USTHB University, Algeria
- AMAROUCHE Amine Idir, USTHB University, Algeria
- ASFARI Ounas, Université Lyon2, France
- ATMANI Baghdad, Université d'Oran, Algérie
- BAAZIZ Abdelhalim, Université Badji Mokhtar, Annaba, Algérie
- BADACHE Nadjib, CERIST Alger, Algérie
- BADARD Thierry, Université Laval, Canada
- BADIR Hassan, ENSAT, Maroc
- BADRI Abdelmajid, Université Hassan II, Maroc
- BELLAFKIH Mostafa, INPT Rabat, Maroc
- BELLATRECHE Ladjel, ENSMA Poitiers, France
- BEN ABDALLAH Hanene, Université de Sfax, Tunisie
- BENBLIDIA Nadjia, Université de Blida Algérie
- BENHARKAT Nabila, INSA de Lyon, France
- BENKHLIFA Elhadj, STAFFORDSHIRE University, UK
- BENKRID Soumia, Higher National School of Computer Science, Algeria
- BENMAISSA Yann , INPT Rabat , Maroc
- BENSLIMANEl Djamel, Université de Lyon1, France
- BENTAYEB Fadila, Université Lumière Lyon 2, France
- BOUCHEBOUT Khoutir, USTHB University, Algeria
- BOUFAIDA Mahmoud, Université de Constantine, Algérie
- BOUFAIDA Zizette, Université de Constantine, Algérie
- BOUFARES Faouzi, LIPN Paris France
- BOUKHALFA Kamel, USTHB, Alger, Algérie
- BOUKRAA Doulkifli, Univesrité de Jijel, Algérie
- BOULMALKOUL Azedine, Université Hassan II, Maroc
- BOURAMAOUL Ramzi Abdelkrim, Université de Constantine, Algérie
- BOUSSAID Omar, Université Lumière Lyon 2, France
- BOUSTIA Narhimène, Saad Dahla University - Blida1, Algeria
- DAHCHOUR Mohamed, INPT Rabat, Maroc
- DARMONT Jérôme, Université Lumière Lyon 2, France
- DERRAR Hacene , Université de Blida, Algérie
- EL AKKAOUI Zineb, INPT Rabat, Maroc
- EN-NOUAARI Abdesslam , INPT Rabat , Maroc
- FAVRE Cécile, Université Lumière Lyon 2, France
- FEKKI Jamel, Université de Sfax, Tunisie
- FERRAG Mohamed Amine, Université du 8 mai 1945, Guelma, Algérie
- GARGOURI Faiez, Université de Sfax, Tunisie
- GHOZZI Faiza, Université de Sfax, Tunisie
- HAFFIDI Hatim , INPT Rabat , Maroc
- HARBI Nouria, Université Lumière Lyon 2, France
- HEMAM Sofiane, Khenchela University, Algeria
- HIDOUCI Walid, ESI Alger, Algérie
- HIOUAL Ouassila, Khenchela University, Algeria
- IDRISSI Abdellah, Université Mohammed V, Rabat, Maroc
- JARARWEH Yaser, University of Science and Technology, Jordan
- KABACHI Nadia, Université Lyon1, France
- KAZAR Okba, Université Biskra, Algérie
- KHOURI Selma, Higher National School of Computer Science, Algeria
- LEMIRE Daniel, Université du Québec à Montréal, Canada
- MAHDAOUI Latifa, USTHB University, Algeria
- MARGHOUBI Rabia, INPT Rabat, Maroc
- MELIT Ali, Université de Jijel, Algérie
- MEROUANI Hayet Farida, Université Badji Mokhtar, Annaba, Algérie
- MEZRIOUI Abdellatif, INPT Rabat, Maroc
- MEZIANE Abdelkrim, CERIST, Algérie
- MOUSSA Rim, Université de Carthage, Tunisie
- NABLI Ahlem, Université de Sfax, Tunisie
- NAFAA Jabeur, German University of Technology in Oman, Muscat, Oman
- NAFAA Mehdi, Université Badji Mokhtar, Annaba, Algérie
- NAÏT BAHLOUL Safia, Oran University, Algeria
- OUKID Saliha, Université de Blida, Algérie
- OULAD HAJ THAMI Rachid, ENSIAS Rabat, Maroc
- RAVAT Frank, Université de Toulouse, France
- REGUIEG F Zohra, Université de Blida, Algérie
- S. Sid Ali, Blida University, Algeria
- SEKHRI Larbi, Université d'Oran, Algérie
- SERIDI Hassina, Université Badji Mokhtar, Annaba, Algérie
- SIDHOM Sahbi, Université de Nancy, France
- TAHARI Karim, Laghouat University, Algeria
- TESTE Olivier, Université de Toulouse, France
- ZAIDOUNI Dounia, INPT Rabat, Maroc
- ZERAOULIA Khaled, USTHB University, Algeria
- ZAROUR Nasreddine, Université de Constantine, Algérie
- ZEGOUR Djamel Eddine, ESI Alger, Algérie
- ZGHAL Sami, Université de Jendouba, Tunisie
- ZIYATI Houssaine - ESTC Casablanca
- ZURFLUH Gille, Université Toulouse Capit

Content

Chapter 5: Big data & Business Intelligence 377

Chapter 1

Internet of Things (IoT) & Healthcare

ASD'2018

Content

Computational IoT-Framework based on Smart and Low-cost Devices for Medical Image Segmentation

Hassna Bensag*, Mohamed Youssfi*, Omar Bouattane*, Fatima Ezzahra Ezzrhari*

*LSSDIA, Hassan II University of Casablanca, Mohammedia, Morocco
h.bensag@gmail.com
med@youssfi.ent
o.bouattane@gmail.com
ezzrhari.fz@gmail.com

Abstract. The exponential growth of information technologies is revolutionizing our lives. They have a great influence on economy, medicine and other areas of society. In this approach, healthcare has been highly connected to technology. In particular, the internet of things (IoT) has been used to interconnect medical resources to ensure healthcare services to patient. Iot platforms are also helping imaging workstations to process their intensive imaging algorithms on the cloud. IoT could also revolutionize medical imaging. In this paper, a distributed framework based on the internet of things is proposed for MRI image segmentation. The main advantages of the proposed system is the intelligence in segmentation image by using low cost resources. This proposed framework can be applied to any applications area, especially those where intensive task and high processing needs take place.

1 Introduction

Human beings have always been fascinated by new technologies: more powerful computers, music players, TVs, Smartphones, etc. This trend gave birth to miraculous inventions we never thought possible. Besides hardware advancements, another significant and more revolutionary shift has taken place through data processing. This new approach known as Internet of Things (IoT) is based upon online digitalization of our physical world and promise to solve many inefficiencies and dangerous modern life practices. The IoT stands at the center of interest of both consumers and companies, for example it is at the top of Gartner' S 2016 Hype Cycles (Haubenwaller and al.,2015), which involves the development of numerous platforms intended specifically for IoT such as SicsthSense , Xively , SensorCloud (Diaz and al., 2016) and IoT architecture (Razzaque and al., 2016). (Fig.1.).

IoT makes use of smart devices easily available at a low cost and allowing us to gather data from our environment via integrated sensors and to share it on the Internet. IoT allows a wide variety of physical devices and machines (e.g. home appliances, surveillance camera, vehicles, and plants) to interact across networks, fostering the development of applications in many fields including house automation, industry, medical field, intelligent networks and traffic management.

The connected objects on the Internet of the things are frequently considered as producers of data and tasks, whereas much of them also have processing capacity, we might cite, as examples: Raspberry Pi, Arduinos and Intel Edison. Thus, we should take advantage of those devices potential in order to decrease the quantity of inputs data (Wadhwa an al., 2015). The execution time could also be reduced if the devices on the same network can treat data (Duttagupta and al. , 2016). This leads to conceiving distributed programs executing precise and independent parallel tasks. In order to simplify the distribution of those tasks, a middleware to which all the objects are connected is necessary (Razzaque and al., 2016). Such a framework should receive comprehensible and interpretable information and automatically transmit them to available devices (Maciej and al., 2014). These requirements constitute the rationale for adopting an information distribution model according to "push-based" approach, (Akkermans and al, 2016). That can be very well supported by the paradigm publish/subscribe. This model of interaction consists of a set of producers and consumers. The producers of information publish events on the system and the consumers of information subscribe to these events (Happ and al ,2017). This approach ensures the availability, real-time performances, and scalability (Hongyan and al, 2014)(Rostanski and al., 2014).

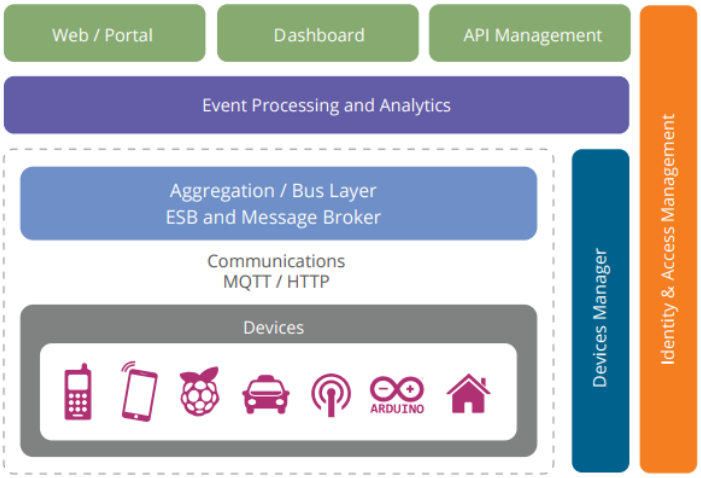


Fig.1. IoT Architecture by WOS2

In this paper, we propose a framework of task distribution for high performance computing. This model is based on the asynchronous communication by using AMQP protocol (Jorge and al., 2015). The objective of this approach is to treat the data by IoT devices. The tasks will be deployed in such a way that execution time and the quantity of data that circulates in the network are minimized. The framework will be developed using publish/subscribe pattern where a task producer (publisher) deploy the byte codes of the task to all the consumers (subscriber) according to a queue. Thereafter, the producer will be able to ask one of the consumers to execute the code of a task already deployed by sending the data to be treated in another queue dedicated to the execution of the requests

The article will be presented as follows: in part 2 we present the proposed model. We discuss the implementation and results in part3. Finally a conclusion and perspectives.

2 Proposed model

2.1 Model Overview

The proposed framework is designed to distribute task in a publish-subscribe IoT architecture, the particularity for the framework is the ability to execute any data intensive task. From there, separate distribution layer from the task execution layer is necessary. The figure 2 shows an application architecture based on the proposed framework.

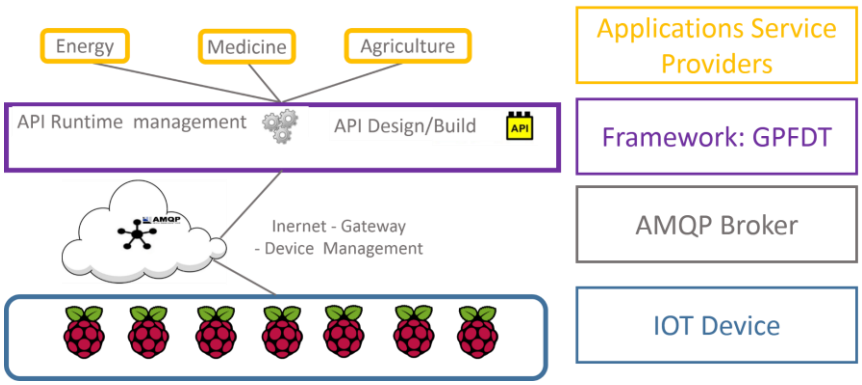


Fig.2. Architecture based on proposed framework GFDT

Task are the main interest in the proposed framework. Hence, modeling task execution process is fundamental. Task is modeled by a process that needs input data to start, executes a sequence of operations sequentially and then produces output data. There are four steps to performing task (shows Fig.3.):

1. Preparing task by defining the operation and input data
2. Initializing task is generally providing data to be performed
3. Execution and finally
4. Post-Processing task by recording and displaying results.

We conclude that to execute task, you have to prepare it first (Bensag and al, 2017). For this step, the proposed framework recommends a model to follow in order to carry out this preparation flexibly. The other three steps are supported by the framework. The developer should be careful to process most of the data intensive operation in step 3 (execution task), this step is distributed in remote nodes.

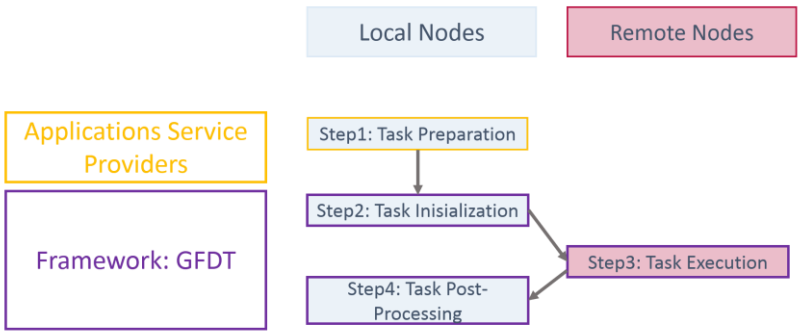


Fig.3. GFDT Task Flow

To ensure Task performing the framework must comprised the following component (Fig.4.):

- ✓ Task Exchange (TE) is responsible for or distributing tasks prepared by the user application.
- ✓ Task Producer Workers (TPW) in the local node. They are responsible for performing step 2 and 4 of the task execution process.
- ✓ Task Remote Workers (TRW) in the remote nodes. They are responsible for performing step 3 of the task execution process.

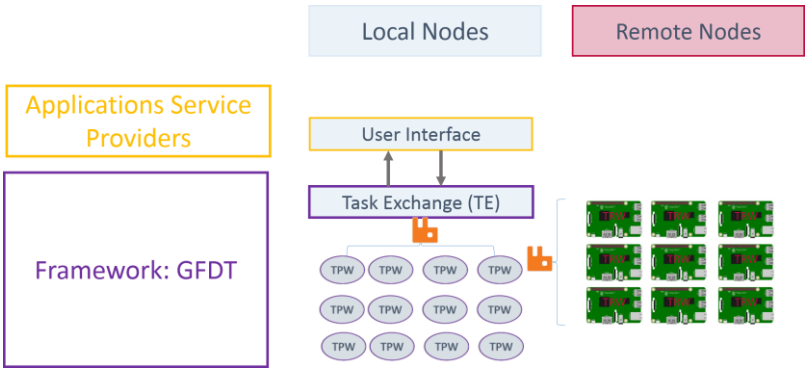


Fig.4. GFDT Componnents

2.2 Task Distribution Process Model

The data intensive task consists of the data and instructions to perform. The data can be any user defined structure. The instructions are only represented by the names of the instruction classes. These are created by the user in accordance with the model imposed by the proposed

framework (Figure 5). Multiple tasks of different types can be executed simultaneously. The tasks to be performed are added as they arrive in a queue managed by the TaskExchange. To be processed, a task requires two available workers: a local producer worker Task Producer Worker and remote worker Task Remote Worker. Task Exchange maintains a list of local workers and another of the available remote workers. When a local worker and a remote worker are available, and a task is on hold, the task distribution process starts. It is carried out in three principal phases:

2.2.1 Task Pre-processing

In this phase, we must define instructions and data to be performed. To define instructions we must implement AbstractProducerPostTask() , AbstractProducerPreTask(), and AbstractWorkerRemoteTask() and define the algorithm to be executed in local for task initialization and finalization and in remote nodes for task execution. This will not an impact on latency, only the names of these three classes are communicated to the workers and instantiation will be done through reflection. For data definition we must implement AbstractDataObject() and define the specific data structure. Nevertheless, we should keep in mind that these data will be transmitted through the network, so they must be serializable.

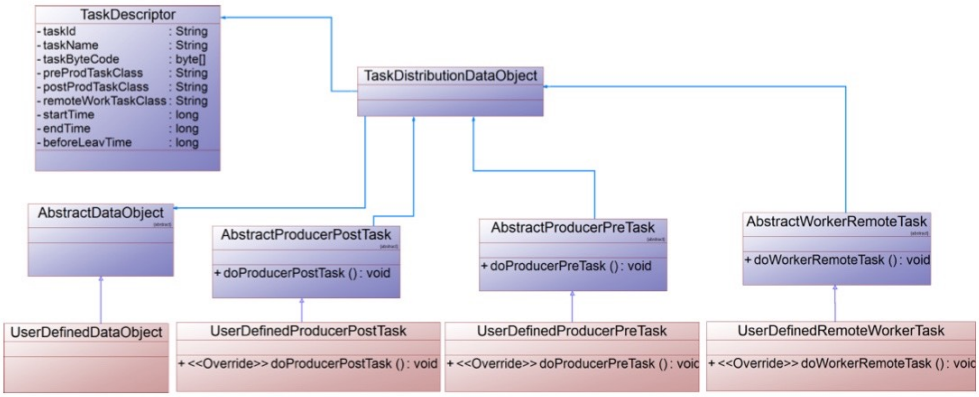


Fig.5. Preparing Task Class Diagram

2.2.2 Task deployment

Once the tasks are added to the TE queue and there are available Workers (TPW and TRW). The TE sends task in a direct exchange to TPW, this latter load the task bytes code and sends the latter in an AMQP message. The message that contains task are not published directly to TRW queues, yet the TPW sends message to a fanout exchange. Fanout exchange main responsibility is pushing the received message to all TRW queues that bound to it. The message is held in each queue until it is handled by a TRW.

2.2.3 Task Execution

When the task is deployed, the TPW will ask exchange for available workers to execute the task. The request is pushed in a unique queue shared with all TRWs. The TPW forward data to the appropriate worker node in an AMQP Message. Sometimes, it is not sufficient to perform real-time processing and scalability with the single producer. The TE comes to the solution here. If a multiple TPW deploys tasks to be processed, a group of TRWs behaves like a distributed system. Executing more than one task fetches task messages from TE queue using round robin distribution, balancing the load across all TRWs, sending each data to the next TRW in sequence. Each task remote node receives and executes the same number of tasks, guaranteeing load balancing.

3 Simulations and results

To prove the reliability of the proposed framework, a case study was developed for k-means segmentation using Breast Magnetic Resonance Image (MRI). In order to perform k-means segmentation as distributed program. Firstly, we must define the actions to be carried out at each step of the task execution process, then we explain how to deploy the application in a cluster consisting of raspberry pi (version B2).

3.1 Fuzzy K-means Algorithm

Clustering is a method of splitting a data set into a specific number of groups. One of the most popular methods is K-means clustering (see Dhanachandra and al, 2015) (Bensag and al, 2015). In k-means, it divides the input data into k cluster disjoint. The K means algorithm is composed of two different phases. In the first phase, it computes the centroid k and in the second phase, it takes each point to the cluster that has the closest centroid to the respective data point. To determine the nearest centroid distance, one of the most frequently used methods is Euclidean distance.

Let us consider an image with resolution of $x * y$ and the image has to be cluster into *k* number of cluster. Let $u(x, y)$ be an input pixels to be cluster and v_k be the cluster centers. The algorithm for *k*-means clustering is achieved according to the following steps:

1. Initialize number of cluster k and center
2. For each pixel, compute the Euclidean distance *d* between then center and each pixel of an image using the given equation:

$$d = ||p(x, y) - v_k|| \quad (1)$$

3. Assign all pixels to the nearest center based en equation (1).
4. Compute the new centroids using the equation given below:

$$\frac{1}{k} \sum_{y \epsilon c_k} \sum_{x \epsilon c_k} p(x, y)$$

5. Repeat the process until it satisfies the error value.
6. Reshape the cluster pixels into image.

3.2 K-means Execution Process

As presented in section 2.1, the model recommends the following steps:

3.2.1 Preparing Task

The Execute Process task has as input and output an object of type TaskDistributionDataObject(). The later encapsulates both task data in an AbstractDataObject () object and task information in a TaskDescriptor() object. TaskPreparation() consists of:
Defining of the task metadata. Required fields are: the names of the classes (see Figure 5) extending abstract classes : ProducerPreTask(), ProducerPostTask(), WorkerRemoteTask().
Adding image data through DataObject class. At this step, we give the name, the path of the image to classify and the number of classes desired within a new instance of DataObject class.

Fig.5. Class Diagram for K-means classification

3.2.2 Initializing Task

This step is achieved by doProducerPreTask() method, which loads the image file into TaskDistributionDataObject as a byte array. To do this, it needs the file path and the dataObject object that encapsulates the task data. The latter is already prepared in step 1(Preparing Task).

3.2.3 Execution Task

This is where the image classification will be done by doWorkerRemoteTask() method. Which will have to load the image file into memory, convert the image to grayscale levels using getGrayScaleImageData() and then run classification by calling the doClassification () method and finally compress the result using the compress () method and put it back into the TaskDistributionDataObject.

3.2.4 Post-Processing Task

This is the last step in task execution process, it is performed by doPostProducrTask() method. The classification result is first decompressed and is then used to generate c output segmented images where c corresponds to the class number.

3.3 Distributed Clustering Communication

A distributed computing environment, as illustrated in Figure 6, presents how GFDT can perform the k-mean classification as a distributed program. To illustrate the main idea of this application, we present in Figure 7 the k-means program implemented according to SPMD architecture over a cluster of three raspberry pi 2 model B. In this model each Raspberry, is asked to perform the k-means program using its assigned image data. This distributed k-means classification is performed according to three global phases:

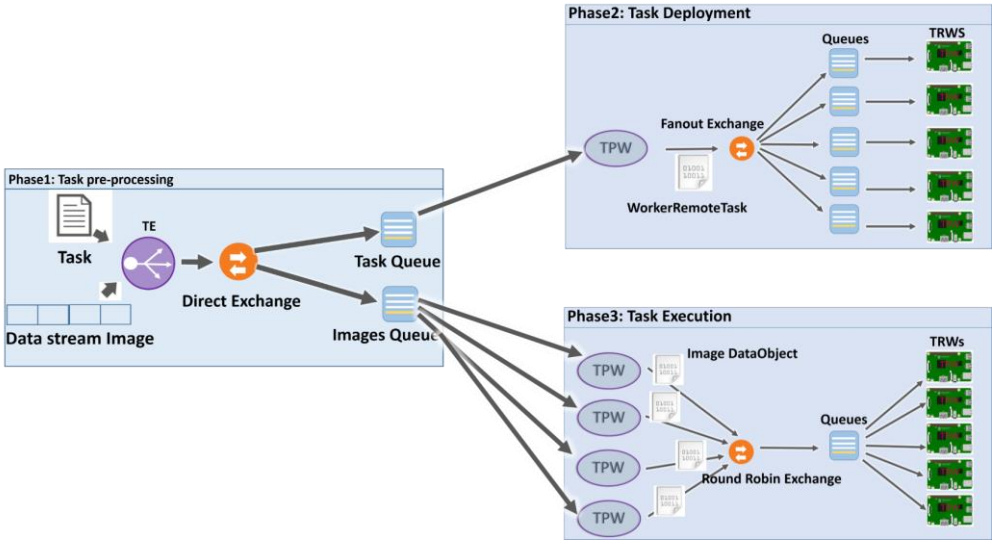


Fig.6. GFDT communication mechanism

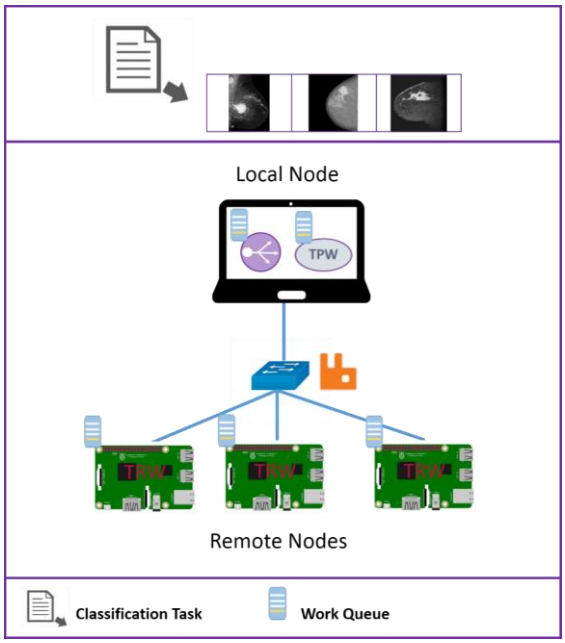


Fig.7. K-means classification cluster

3.3.1 Task Pre-Processing

In this phase, The TE is initialized by the data stream image and the name of classification classes: ProducerPreTask, ProducerPostTask, WorkerRemoteTask.

3.3.2 Task deployment

In This phases the TE deploy just WorkerRemotTask, when the task are added to TE queue and there are available TPWs and TRWs. The TE sends an AMQP message that contains the task to a direct Exchange. The later takes the message and routes in into TPW queue. The TPW load the task byte code and broadcast the latter in a fanout exchange to all TRW queues. The message remains in each queue until it is processed by a TRW.

3.3.3 Task execution

When the task is deployed, Each TPW loads is Image Data into an object that will be transported in an AMQP message. Then, it will ask the broker to select available TRW to perform the classification task (WorkerRemoteTask) by sending the image data object to it. The message is sent to a single queue shared by all TRWs and each image data is send to the next TRW in sequence. Every TRW, performs Executing Task Step and return their classification results to TPW that provided the image to be processed. Each TPW carries Post-Processing step and displays the segmented output image.

3.4 K-means classification results

The proposed distributed K-means algorithm is implemented in this model for MRI breast image. To do so, we choose three breast MRI images: (Img1), (Img2) and (Img3). Each Raspberry pi performs the classification progam using its assigned image data. At the end of execution process these image will be segmented into c output images where c corresponds to the class number (Figure 8 (a)-(b)-(c)).

Table 1 - Total running time of the three different images segmentation

Image Id	Image size (octet)	Execution Time (ms)
Img1	18 366	187
Img2	296 033	2429
Img3	53 036	944

The distributed K-Means classification time in Table 1, shows clearly that the classification time of the three images achieve minimum values of 187ms for Img1, 2429ms for Img2 and 944ms for Img3.

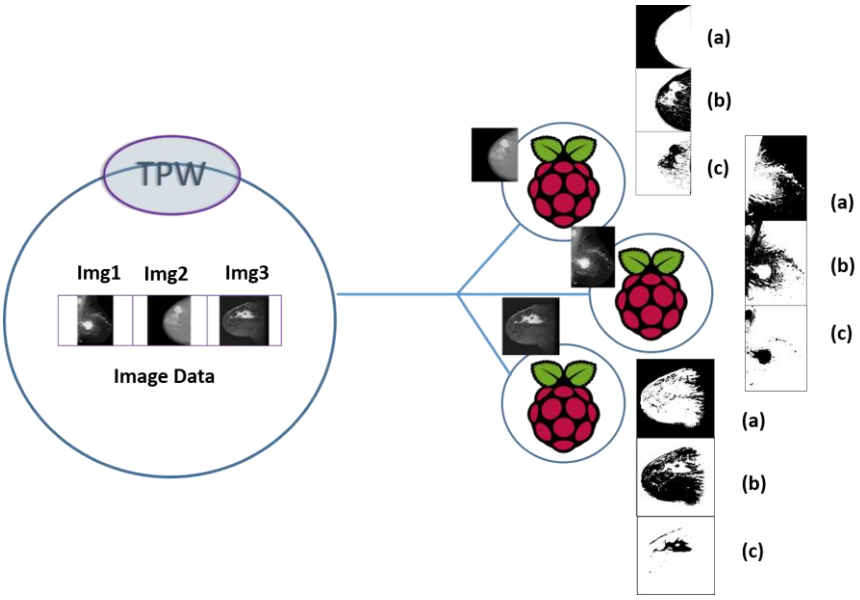


Fig.7. K-means classification results

4 Conclusion and perspectives

In this paper, we presented a distributed framework (GFDT) based on IoT devices for MRI image classification. The framework is implemented on a SPMD model based on IoT architecture. GFDT uses the asynchronous communication mechanism, which is based on

exchanging AMQP messages between GFDT components. The proposed framework is not only designed for image classification, but it can be used for any compute intensive task. The GFDT adds an abstraction layer to RabbitMQ, and makes the framework generic.

The first version of the GFDT framework shows encouraging results both in terms of ease use and performance. However, improvements can be made at several levels. One of the most important will be to automate event detection. The generic task model can be improved also by adding priority indicator. The results allow us to confirm the possibility of using a large amount of data and IoT devices to take more advantage of the framework performance and opening the possibility of designing more scalable HPC models.

References

Akkermans, S., Bachiller, R., Matthys, N., Joosen, W., Hughes, D., & Vučinić, M. (2016). Towards efficient publish-subscribe middleware in the IoT with IPv6 multicast. IEEE International Conference on, 1-6.

Bensag, H., Youssfi, M., & Bouattane, O. (2015). Embedded agent for medical image segmentation. IEEE International Conference on Microelectronics, 190-193.

Bensag, H., Youssfi, M., & Bouattane, O. (2017). Efficient Model for Distributed Computing based on Smart Embedded Agent. INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS, 8(2), 102-109.

Dhanachandra, N., Manglem, K., & Chanu, Y. J. (2015). Image segmentation using K-means clustering algorithm and subtractive clustering algorithm. Procedia Computer Science, 54, 764-771.

Duttagupta, S., Kumar, M., Ranjan, R., & Nambiar, M. (2016). Performance prediction of iot application: An experimental analysis. International Conference on the Internet of Things, ACM, 43-51.

Diaz Manuel, Cristian Martin, Bartolome Rubio, (2016), State-of-the-art, challenges, and open issues in the integration of Internet of Things and Cloud Computing, Journal of Network and Computer Applications, 99-117.

Happ, D., Niels Karowski1, Thomas Menzel1, Vlado Handziski, Adam Wolisz,(2017) Meeting IoT platform requirements with open pub/sub solutions, Annals Telecommunications,72: 41-52

Hongyan Mao, Li yuan, Zhengwei Qi, (2014), A Load Balancing and Overload Controlling Architecture in Clouding Computing, International Conference on Computational Science and Engineering, IEEE.

Haubenwaller, A. M., & Vandikas, K. (2015). Computations on the edge in the internet of things. Procedia Computer Science, 52, 29-34.

Jorge E. Luzuriaga, Miguel Perezy, Pablo Boronaty, Juan Carlos Cano, Carlos Calafate, Pietro Manzoni, (2015), A comparative evaluation of AMQP and MQTT protocols over unstable and mobile networks, Consumer Communications and Networking Conference, IEEE.

Maciej ROSTA_SKI , Krzysztof GROCHLA, Aleksander SEMAN, (2014), Highly available and fault-tolerant architecture guidelines for clustered middleware servers, Theoretical and Applied Informatics, 69 – 85.

Razzaque, Mohammad Abdur, et al. (2016) Middleware for internet of things: a survey. IEEE Internet of Things Journal 3.1: 70-95.

Rostanski, M., Grochla, K., & Seman, A. (2014). Evaluation of highly available and fault-tolerant middleware clustered architectures using RabbitMQ. Computer Science and Information Systems , IEEE, 879-884.

Wadhwa R., Mehra A., Singh P., Singh M.(2015). A pub/sub based architecture to support public healthcare data exchange. In Communication Systems and Networks, IEEE, 1-6.

Résumé

La croissance exponentielle des technologies de l'information révolutionne notre mode de vie. Elles ont une grande influence sur l'économie, la médecine et d'autres domaines de la société. Dans cette approche, les soins de santé ont été fortement liés à la technologie. En particulier, l'Internet des objets (IoT) a été utilisé pour interconnecter les ressources médicales afin d'assurer des services de santé aux patients. Les plates-formes Iot aident également les stations de travail à traiter leurs algorithmes d'imagerie intensive sur le nuage. L'IoT pourrait également révolutionner l'imagerie médicale. Dans cet article, un Framework distribué basé sur l'internet des choses est proposé pour la segmentation de l'image IRM. Les principaux avantages du système proposé sont l'intelligence de la segmentation d'image en utilisant des ressources moins cher. Ce Framework proposé peut s'appliquer à n'importe quel domaine d'application, en particulier à ceux qui nécessitent des travaux intensifs et un traitement intensif.

A pipeline approach for automatic segmentation of free-text medical reports

Walid Zeghdaoui*,** Frederik Joly**
Omar Boussaid* Fadila Bentayeb*

*Université de Lyon, Université Lyon 2, ERIC EA 3083
5, Av. Pierre Mendès-France, 69676 Bron, France
{walid.zeghdaoui, omar.boussaid, fadila.bentayeb}@univ-lyon2.fr
**Sword Group - 9, Av. Charles de Gaulle, 69370 St-Didier-au-Mont-d'Or, France
{walid.zeghdaoui, frederik.joly}@sword-group.com

Abstract. One of the major challenges in precision medicine is to guide the research of specific therapeutic solutions through the extraction of knowledge from medical reports. This could help physicians identify the most appropriate diagnosis among many possible choices. These data are often unstructured and maintained in free-text form. Therefore, the use of Natural Language Processing techniques combined to text segmentation methods became obvious to identify the parts that are of great interest. In this article, we propose and implement our pipeline approach for automatic section segmentation of medical reports which consists of two components. First, a rule-based algorithm to detect sections using our titles identification method. The second part focuses on sentence classification into pre-defined categories. This last task is based on machine learning algorithms since it is considered complex to be apprehended only by rules. This system was evaluated on 500 reports and achieved more than 94% classification accuracy.

Keywords: Medical informatics, Clinical reports segmentation, Text classification, Machine learning, Natural language processing.

1 Introduction

The vast majority of clinical reports in hospitals is still maintained in free-text form, and the amount of these data has grown dramatically over the last decade. These documents contain a wealth of information and knowledge that must be used to improve the health care process. Indeed, the knowledge discovered could assist medical staff in a myriad of ways, particularly in medical making decision, for example identifying the most appropriate diagnosis for a patient, but not only in direct patient care, but also for secondary purposes like clinical research. Actually, the medical community is constantly striving for new means to conduct research in the battle against diseases. In order to facilitate their task, the next step consists of making these informations both usable and useful. Thus, the access to this knowledge using information retrieval and information extraction techniques has become more judicious than ever before, as it helps greatly understanding clinical report content.

However, the unstructured format of clinical reports makes it difficult to retrieve meaningful information, all the more when they are long and only some portions of the text are interesting to specific users. For example a physician may wish only to see surgical history or conclusion section of a report, while a computer scientist may be interested in another section of the report, especially the one containing information that must be anonymized such as personal data (names, addresses, phone numbers, etc.) of patients that are not allowed to be disclosed due to patient privacy and regulation issues concerning medical data.

Section segmentation of these clinical reports plays actually a key role not only because it allows us to divide them into meaningful units, so we can return a specific part of a report corresponding to a query as a result, but also because it helps us to better understand knowledge contained on them. For instance, *SNOMED CT*[1] are often made up of 5 digits likewise French postal codes. So using only classical Named Entity Recognition (NER) techniques is not enough, since these codes could have a multiple meanings. This problem is also-known as a Named Entity Disambiguation (NED). As we know that postal codes are usually used in section dedicated to personnel informations in the beginning of the reports, unlike *SNOMED CT* which are much used in conclusions at the end of the reports, it is therefore easy for us to distinguish between the two terms and so overcome this kind of issues. Similarly, several automated tasks could be set up by automatically identifying section boundaries.

As part of this work, we deal with clinical reports of four institutes and a center among the 20 French Comprehensive Cancer Centers (FCCC). These are Institut Paoli-Calmettes (IPC) in Marseille, Institut Curie (IC) in Paris, Institut du Cancer de Montpellier (ICM) and Centre Léon-Bérard (CLB) in Lyon respectively. Furthermore, we also aim to deploy our segmentation system in other centers. As a result, the identification of report types becomes very complicated since there is presently no universal format for written medical reports in France.

Since each center has its own computer system and uses one or more document formats to store medical reports, such as PDF, Word, HTML or Plain text file, the processing performed on these documents may give rise to additional challenges like dealing with dirty data. Indeed, sentences extracted from ocerized PDF are sometimes poorly cut or represent a meaningless sequence of special characters.

In view of this issues, we propose in this paper our work towards building a scalable automatic segmentation system of medical reports into predefined categories. We report performance results on 500 reports from four member institutions of French Comprehensive Cancer Centers.

In section II we discuss the related work. In section III, the data features are described. In section IV, we present the developed methods. In section V, we describe our experimentations to evaluate our segmentation system and we discuss the results. Finally, conclusion and future research directions are outlined.

1. *SNOMED CT* is the most international comprehensive and precise clinical health terminology, enabling healthcare professionals and researchers to adopt a common language.

2 Related work

The problem of text segmentation (i.e., the process of dividing written text into meaningful units) has been widely studied. Various segmentation methods have emerged during the last decade for different kinds of language and applications. However, less effort has been devoted to its application in the clinical domain. Segmentation of medical reports is a difficult issue, because each physician has his own writing style to structure these reports. Furthermore, since there is no universal template for written medical reports in France, physicians do not follow neither a strict section naming conventions nor a defined structuration format.

Ganesan and Subotin (2015) proposed a supervised model using L1-regularized logistic regression with a constraint combination approach that is capable of recognizing the header, footer, and all of the top-level sections of a clinical texts. This method operates at the line-level rather than the sentence-level, which could generate a label sequence that does not make any sense.

Tepper et al. (2012) trained a supervised statistical machine learning model using discharge summaries and radiology reports and proved that the two-step approach which first identifies the section headings followed by their categorization outperform the one-step approach. However, they report low adaptability when their model is applied to unseen documents. In fact, their segmentation model had accuracy on the same dataset but significantly lower accuracy on a separate data set. There could be several reasons for this including the types of documents used for training along with the features, preventing the model from generalizing sufficiently well.

Apostolova et al. (2009) developed a document segmenter based on an SVM classifier that divides text segments into eight semantic units from radiology reports which are outpatient notes and have a fairly consistent format and a very concise structure compared to other clinical reports which represents the limitation of this approach.

Cho et al. (2003) used rule-based filters based on string, phrase, lexical and statistical analysis to automatically partition the text within radiology and urology reports into topically cohesive sections. However, this approach is not practical because, firstly, it is effective due to inherently structured nature of these documents, and secondly, we have to train and maintain as many models as report types.

While the above-mentioned approaches have achieved a good accuracy, their main limitation is that they rely on the nature of some report types in Electronic Medical Record (EMR). In this paper, we propose a pipeline approach to automatically segment free-text medical reports into semantic sections. This one is used to develop our robust and scalable medical report segmentation system and consists of two components. The first step is based on our titles detection algorithm for identify titles and thus some section boundaries. Then the results of the first step are passed to the second step, where a separate sentence classifier is called upon to assign for each sentence an appropriate section category.

3 Task definition and dataset

3.1 Dataset and section category

We used 500 reports randomly selected of 417 patients from four member institutions of French Comprehensive Cancer Centers. All of these reports were then used to evaluate and test the performance of our system. One hundred randomly selected reports from the dataset were used for preliminary analysis and construct our section categorization. With the help of an expert, 7 categories were identified covering all sections of the reports (Table 1).

Section category	Description	Count
Personal History	Past medical history	461 (10.9%)
Family history	Family medical history	206 (4.8%)
Personal data	Any information relating to an identified natural person	814 (19.2%)
Noisy data	Low medical value information	524 (12.4%)
Recommendation	Recommendations for additional studies and follow up	264 (6.2%)
Conclusion	Conclusion	167 (4%)
Content	Otherwise	1802 (42.5%)
Total	-	4238 (100%)

TAB. 1 – *Section category for 500 reports and their count.*

3.2 Challenges in segmenting clinical reports

Loose text formatting is commonly used to structure medical reports. It is customary for physicians to preface some paragraphs with an appropriate titles. The identification of these key titles could therefore help us in our segmentation task since each title marks the beginning of a new paragraph, or even a new section. We have manually listed all the titles contained in our subset of reports and found that more than 95% share some common features. For example: A maximum length of 6 words, beginning with a line break followed by a capital letter and ending with a colon and/or a line break. (Figure 1).

Antécédent personnel: Antécédent de carcinome endométrioïde.

Résultats du bilan:
Confrontation au précédent scanner du 12 décembre 2012.

CONCLUSION
L'ensemble du bilan actuel n'a retrouvé aucune lésion évolutive.
80903

FIG. 1 – *Example of titles in a report.*

The intuition to use rules to detect titles and then label all the sentences grouped in the following paragraph might seem rational, however use only regular expressions is far from sufficient for the following reasons:

1. Since physicians have complete freedom to structure their reports, sentences belonging to different categories may be grouped in the same paragraph, in which case, this method will mislead us.
2. In the same way, we can find sentences of a single category dispatched on several paragraphs. This can also mislead our segmentation task.
3. Titles are prone to misspellings due to human error, and may not be detected.
4. And finally, most reports do not contain explicit titles.

Fait le 15/06/2004.
Cs Dr DUPONT

Résultast:
La cytoponction est en faveur d'une cytostéatonécrose.

A fait une échographie EV: pas d'anomalie particulière.

FIG. 2 – *Sample report with some of the challenges such a report.*

The above sample report (Figure 2) illustrates the different difficulties listed previously. Indeed, the first sentence is not preceded by a title, belongs to the *Noisy data* section and followed by a sentence that belongs to *Personal data* section. The third sentence of the sample report represents a title that refers to the *Content* section. However, using regular expressions is not sufficient to match misspelled strings («**Résultats**» instead of «**Résultast**»). Finally, the last two sentences of this example belong to two separate paragraphs, are part of the same section and only the first one is preceded by a title.

To overcome these rules-based system limitations, we have adopted a sentence classification solution based on machine learning algorithm. Indeed, these algorithms have the ability to create generalizations automatically on theses reports during the training process regardless the number of variations of these sentences. Therefore, we need a labeled data to train our model.

3.3 Annotation task

The quality of the learning sentences is important to avoid generating a biased model. Furthermore, these sentences must be representative of the global corpus to best cover the various possible cases. We used a graphical interface to manually annotate all sentences of our dataset. The user must select manually each sentence and indicate an appropriate category among those we have defined.

4 Materials and Methods

We have divided our segmentation system into two steps. First, a rule-based algorithm is developed to automatically detect titles within reports. Each title is then used as a reliable rule

to identify the category of all sentences that form the next paragraph within the report. The second step is modeled as a text classification task, involving assigning each report sentence to one of seven categories we have defined.

4.1 A rule-based algorithm

The first step of our segmentation system performs a first pass in the reports in order to extract reliable and useful information for the next step. This first step is exclusively based on title detection. The algorithm assigns each identified title within the report one of the seven categories we have defined. However, the rules used for titles detection have a high probability of yielding false positives (Sentences that meet the potential title definition). To overcome this issue, we suggest a two-phase algorithm:

In the first phase, all potential titles that are preselected must respect the following constraints:

1. Containing at least one word and at most 6 words.
2. Holding on a single line.
3. Starting with a line break followed by a capital letter.
4. Ending with a colon and/or a line break.

In the second phase, regular expressions are used to detect titles among those preselected in the first phase, based on a title inventory. This inventory serves as a mapping between the section categories and the titles compiled on our reports. For example, *Last name* and *First name* would correspond to *personal data* section. To construct this mapping, the categories were manually assigned during the preliminary analysis. This allows us to choose only true positives. When matching a candidate label, care must be taken to consider possible variants that have not yet been identified during the preliminary analysis. To solve this problem, a normalization process is applied to the selected titles being compared:

1. Remove special characters (e.g., hyphen, asterisk, etc.);
2. Remove stopwords;
3. Replace accented characters (e.g., «é» by «e»);
4. Expand abbreviations;
5. Convert to all lowercase.

At this point, titles detection is done. Then, all sentences in the following paragraph are labelled with the same label which is one assigned to the title, since we consider that the narratives following a recognized section title should belong to this corresponding section.

Analysis of the preliminary reports revealed that some sections share several common features, including the format of the titles used. It is customary practice for physicians to use a single term for two titles of two different sections. For example, the term *History* could be used as a title for both *Personal history* and *Family history* sections. Furthermore, *Personal data* and *Noisy data* sections tend to be grouped into a single paragraph (Figure 2). We therefore decided to temporarily group some categories as shown in table below. (Table 2).

Temporary sections	Old sections
Data	Personal data
	Noisy data
History	Personal history
	Family history
Content	Content
Conclusion	Conclusion

TAB. 2 – *Section categories of 500 reports and their count.*

This choice was motivated mainly by two reasons:

1. Reduce the risk of a misallocation of a category to a title.
2. Reduce the number of rules to maintain.

Finally, the preliminary analysis also revealed that almost all conclusions are prefaced by titles. Identifying these sections becomes an easy task based solely on rules. Especially since the sentences of these sections have no features that could distinguish them from those of the *content* section.

4.2 Machine learning text classification

The second step of our segmentation system was modeled as a text classification task assigning each sentence within the report to one of the six categories used to train our model. In fact, *Conclusion* sections are detected at the first phase and do not require further processing. For this, we tested the performance of several machine learning algorithms of sentence classification using our manually annotated reports as an input.

To build our sentence classification model, there is several techniques, such as Decision Trees, Support Vector Machines, and Naive Bayes algorithms. Each technique adopts a learning algorithm to identify a model that best fits the relationship between sentences and labels. Therefore, the key objective is to build predictive model that accurately predicts the labels of previously unknown sentences. In our experiment, we compared the performance of the four following algorithms:

4.2.1 Decision Tree Classifier

The decision tree algorithm tries to solve the problem by using tree representation. This method classifies a population, sentences for example, into branch-like segments that construct an inverted tree with a root node, internal nodes, and leaf nodes. Each internal node of the tree corresponds to an attribute, and each leaf node corresponds to a label. Decision trees are widely used for establishing classification systems based on multiple covariates or for developing prediction algorithms for a target variable. (Venkatasubramaniam et al., 2017).

4.2.2 Support Vector Machine

Support vector machine is a powerful machine learning method in data classification and can be employed mainly for both classification and regression purposes. SVMs are based on the idea of finding hyperplanes that best divide a dataset. This leads to good generalization accuracy on unseen data and supports specialized optimization methods that allow SVM to learn from a large amount of data. (Cortes and VAPNIK, 2009).

4.2.3 Naive Bayes Classifier

The naïve bayes classifier is one of the simplest approaches that is still capable of providing reasonable accuracy in classification tasks and represents a supervised learning method as well as a statistical method for classification. It assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature (naive independence assumptions between the features). Bayesian classification calculates explicit probabilities for hypothesis and it is robust to noise in input data. In statistical classification the bayes classifier minimises the probability of misclassification. (Raschka, 2014).

4.2.4 FastText Classifier

FastText is a library created by the Facebook Research Team for efficient learning of word representations and sentence classification. It combines some of the most successful natural language processing techniques of the last few years and machine learning. These include representing sentences with bag of words and bag of n-grams, as well as using subword information, and sharing information across classes through a hidden representation. The real motivation behined fastText is using shallow neural network to overcome some limatations of deep neural networks, since even if these models achieve very good performance in classification tasks, they can be slow to train and test. (Joulin et al., 2017).

4.3 Evaluation

The values of precision (i.e., measure of a classifiers exactness) and recall (i.e., measure of a classifiers completeness) determine the accuracy of each classification algorithm. We evaluate our results using precision (formula 1), recall (formula 2), and F-measure (formula 3, with $\beta = 1$).

$$Precision = \frac{True\ positive}{True\ positive + False\ positive} \tag{1}$$

$$Recall = \frac{True\ positive}{True\ positive + False\ negative} \tag{2}$$

$$F - measure = \frac{(1+\beta)^2 \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{3}$$

We used the scikit-learn and fastText libraries to carry out our experiments. The classification performance are shown in the table below.

Algorithm	Precision	Recall	F-measure
DecisionTree	73%	76%	72%
LinearSVCl2 (SVM)	89%	77%	81%
BernoulliNB	72%	86%	77%
FastText	84%	84%	84%

TAB. 3 – *Accuracy comparaison across classifiers.*

We tested these different algorithms on the same test data. A normalization process was first applied on all sentences and consists of:

1. Remove special characters (e.g., hyphen, asterisk, etc.);
2. Replace accented characters;
3. Convert to all lowercase.

We used each algorithm to train several models by varying the input hyperparameters at each execution. In fact, in machine learning, the same kind of model can require different hyperparameters such as, weights or learning rate. These measures have to be tuned to yield an optimal model which minimizes a predefined loss function and thus can optimally solve the problem. We were thus able to compare the best models of each algorithm.

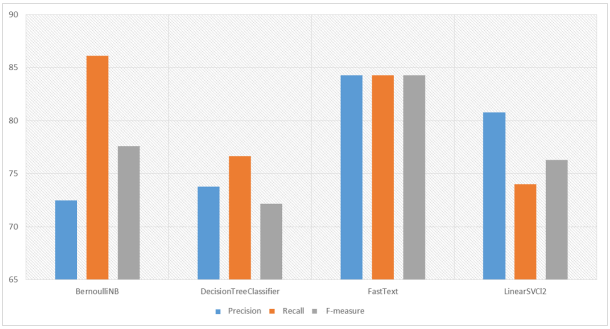


FIG. 3 – *Sentence classification algorithms performance.*

We chose to keep *fastText* for our sentence classification task, for several reasons. First of all, based on the results presented above, *fastText* provides the best classification accuracy using our dataset. Actually, we used *F-measure* metric to evalute our classification models because we deal with imbalanced class distribution problem where more than 40% of sentences belong to the *Content* section. Furthermore *fastText* word vectors are built from vectors of substrings of characters contained in it. This allows to build vectors even for misspelled words or concatenation of words.

5 Experimentation and results

In our experimentations, we used all the manually annotated sentences within the 500 reports of our dataset. We used only 200 reports to build the segmentation system and use the remaining 300 for independent testing. For the evaluation, we have combined the rule-based algorithm with the fastText classification model as shown in the figure below.

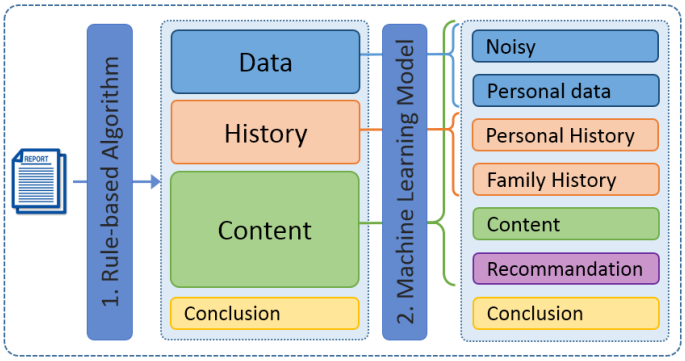


FIG. 4 – *Segmentation system.*

The trained model in the second step of our approach is used in different ways depending on the temporary section assigned to sentences at the end of the rule-based algorithm. In fact, the used fastText model predicts for each sentence a list of our six labels ordered by decreasing probability. For instance, if the model predicts the following label sequence for the sentence «*Name: Last name*» : *Content, Personal data, Noisy data, Family history, Conclusion, Personal history*, we will only be interested in *Personal data* and *Noisy data* labels. Since it concerns the temporary *Data* section. Thus, the first one (*Personal data*) will be chosen.

For the temporary *Data* and *History* sections, only labels of *Personal data*, *Noisy data* and of *Personal history*, *Family history* are concerned respectively. The results obtained are shown in the following table.

Section category	Count	Accuracy
Personal History	254	90,11 %
Family history	133	93,9%
Personal data	279	97,21
Noisy data	271	94,7%
Recommendation	141	97,1%
Conclusion	79	99,2%
Content	979	96 %
Total	2136	94,5 %

TAB. 4 – *Accuracy of our segmentation system on 300 reports.*

The accuracy rate ranged from 90.11% for *Personal history* section to 99.2% for *Conclusions* which are detected using only our rule-based algorithm. This results show that combining natural language processing techniques and machine learning text classification methods could be successfully applied to solving the automatic segmentation of free-text medical reports.

The experiments also reveal some challenges to the task. In fact, even after carring out hyperparameter tuning for each model used in our text classification task, it turned out that a high bias problem (an error from erroneous assumptions in the learning algorithm) still remains. This confirms the importance of our rule-based algorithm and how we combined the two phases of our segmentation system.

6 Conclusion and Future Work

Section segmentation of free-text medical reports provides important contextual information for other automated information extraction tasks. This could help to improve the health care process and to advance clinical research. In this paper, we proposed a fully automatic segmentation system of clinical reports. The system consists of two components. First a rule-based algorithm is developed to identify some sections using our two-phase titles detection approach. This first step plays an elementary role in our segmentation system. Indeed, the algorithm makes a first pass on the reports to extract the maximum of informations based on features that look at both shape and the content of these reports. This informations are eventualy used during the second step except for the conclusions which are identified at the end of the first step. The second part consists of training a machine learning sentence classification model using *fastText* which assigns a section category to each sentence within the reports. Depending of the results of the our rule-based algorithm, this model is used in a different way. The system was tested on 500 manually annotated reports of four member institutions of French Comprehensive Cancer Centers, and achieved a good performance. This segmentation system is used to facilitate information retrieval and extraction of knowledge from clinical reports.

There is still room for improvement, particularly to ensure the robustness of our system, since it is based on features defined during the preliminary analysis of a random sample drawn from a larger corpus. As future work, we plan to deploy our system in all institutions member of FCCC. We will also explore other techniques for text classification such as Deep Neural Network text classification algorithms. In fact, deep neural networks have revolutionized the field of natural language processing (NLP). The dominant approaches are recurrent neural networks (Elman, 1990), in particular LSTMs, and convolutional neural networks (Lecun et al., 1998). Parallel to that, we look further into semi-automatic annotation methods to alleviate the tedious work of annotation.

References

Apostolova, E., D. Channin, D. Demner-Fushman, J. Furst, S. Lytinen, and D. Raicu (2009). Automatic segmentation of clinical texts. *2009*, 5905–8.

Cho, P., R. Taira, and H. Kangarloo (2003). Automatic section segmentation of medical reports. *2003*, 155–9.

Cortes, C. and V. VAPNIK (2009). Support-vector networks. *297*, 273–297.

Elman, J. (1990). Finding structure in time. *14*, 179–211.

Ganesan, K. and M. Subotin (2015). A general supervised approach to segmentation of clinical texts. pp. 33–40.

Joulin, A., E. Grave, P. Bojanowski, and T. Mikolov (2017). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 427–431.

Lecun, Y., L. Bottou, Y. Bengio, and P. Haffner (1998). Gradient-based learning applied to document recognition. *86*, 2278 – 2324.

Raschka, S. (2014). Naive bayes and text classification i - introduction and theory.

Tepper, M., D. Capurro, F. Xia, L. Vanderwende, and M. Yetisgen (2012). Statistical section segmentation in free-text clinical records.

Venkatasubramaniam, A., J. Wolfson, N. Mitchell, T. Barnes, M. JaKa, and S. French (2017). Decision trees in epidemiological research. *14*.

On the performance of NoSQL stores for managing proteomics data

Chaimaa Messaoudi, Rachida Fissoune and
Hassan Badir

National School of Applied Sciences, ENSA,
Abdelmalek Essaadi University, Tangier, 90 000, Morocco
messaoudi.chaimaa@gmail.com
ensat.fissoune@gmail.com
hbadir@ensat.ac.ma

Abstract. Genome sequencing instruments are capable of sequencing thousands of samples in parallel, leading to the production of several terabytes of raw genome sequence data in one day, which creates the challenges of storage of biomedical big data. NoSQL data stores can serve as an alternative to traditional relational database systems, particularly for handling biomedical big data applications. Applications that model their data using two or more simple NoSQL models are known as applications with polyglot persistence. Recently, a family of multi-model data stores is introduced, integrating simple NoSQL data models into a single unique system. In this paper, we evaluate the storage, deletion and query efficiencies of a polyglot persistence approach and two multi-model systems (OrientDB, ArangoDB). The polyglot persistence approach combines two NoSQL stores: Graph-oriented database (Neo4j) and Document-oriented database (MongoDB). In order to evaluate the performance of the integration of proteomics data, we used two species datasets: *Homosapiens* as a large dataset and *Lactobacillus Rhamnosus* as a small dataset.

1 Introduction

New technologies have emerged and revolutionized biological and biomedical research such as Next Generation Sequencing (Mardis, 2008) and Mass Spectrometry techniques advances (Yergey et al., 2013). The development of these methods led to the exponential growth of biological data: DNA, protein sequences, microarrays, and metabolic pathways that need to be stored, retrieved and analyzed. Biological data are eligible for the name "Big data" (Marx, 2013). Storing and extracting useful information from these biological big data is one of the main endeavors for bioinformatics community. Moreover, biological data sources are distributed and heterogeneous: each source has its own data format and structure, and it is common that the scientific terms used to describe the data differ from one source to another. These challenges are needed to be addressed because the current relational database technologies have insufficient resources to handle them (Atzeni et al., 2013).

In order to manage these large and heterogeneous data, the database research community has developed new technologies to overcome the insufficiency of the relational databases technologies. New data stores systems called non-relational data stores systems have emerged under the name of *NoSQL systems*. These systems support different types of data models that are efficiently scalable and distributed. We distinguish four NoSQL categories with examples, each one having its own specificities and facilitating the management of some particular kind of data: key-value stores (DynamoDB), column family database (Cassandra, HBase), document-based storage (MongoDB, OrientDB) and graph database (AllegroGraph, OrientDB, Neo4j) (Moniruzzaman and Hossain, 2013). The use of a NoSQL store mainly relies on the application context and the data model (e.g. graph). Some applications require more than one NoSQL stores. For example, in a proteomics application, a protein-protein interaction dataset should be modeled as a graph, and a protein sequence information dataset is more appropriate to be stored in a document database. Those types of applications that simultaneously use different models and data stores are called applications with polyglot persistence(Sadalage and Fowler, 2012).

The polyglot persistence approach requires the understanding of more than one query language and user interface, addition to managing the communication between the different data stores used in the application. There have been advances to provide a unique NoSQL system that contains multiple data models. These systems are called NoSQL multi-model, and they simplify the process of application development because they use only one store, but they could decrease the performance of applications (Wiese, 2015).

Several research studies have been conducted to evaluate the performance of NoSQL stores such as MongoDB, Cassandra, ArangoDB, CouchDB and OrientDB for the management of large biomedical data sets (Shao and Conrad, 2015; Guimaraes et al., 2015; Wang et al., 2014; Have and Jensen, 2013; Schulz et al., 2016; Aniceto et al., 2015; Lee et al., 2013).

To the best of our knowledge, there is no publication on the evaluation of a polyglot approach using biomedical real data. This paper presents a performance study of NoSQL stores for the integration of proteomics data. We compare the performance of NoSQL multi-model (OrientDB and ArangoDB) to the polyglot persistence approach. OrientDB and ArangoDB manage both document and graph data models, and they were chosen because they are an open source data models stores. The polyglot persistence consists of combining the document database (MongoDB) with a graph database (Neo4j). The comparisons are made from the following aspects: Insertion, deletion, importation and query performance.

This paper is structured as follows. Section 2 covers a brief introduction to NoSQL databases and the main features of polyglot persistence. It discusses some related works that evaluate NoSQL data store performance. Section 3 presents the evaluation study, the datasets used and discusses the practical results obtained. Section 4 concludes and suggests future works.

2 NoSQL Stores and Data Models

NoSQL databases have appeared as a solution for storage scalability, management of large volumes of unstructured data and parallelism. We aim to provide a brief overview of NoSQL store models as well as polyglot systems particularly polyglot persistence and multi-model system. In the same section, we discuss some related works on NoSQL stores for the integration

of biomedical data.
Key-value: similar to maps or dictionaries where data are associated to a unique key. This makes the system accessible and available at runtime anytime without conflicting with any other stored data. Values are isolated and independent from each other.
Document: it designed to manage and store documents. These documents are encoded in a standard data exchange format such as XML, JSON (Javascript Option Notation) or BSON (Binary JSON).
Column: stores data tables as columns rather than rows offering a more precise access to data, especially in very large datasets.
Graph: this model has three basic components: nodes, relationships, and properties of nodes and relationships. The graph is directed, nodes are connected by edges. This model is opportune for applications requiring queries traversing several levels of relationships between data.

In the last decade, much attention has been given to storing biomedical data using NoSQL data models. ONDEX (Köhler et al., 2006) and Biozon (Birkland and Yona, 2006), collected the data from various sources under a single data store, using a graph data schema centered around the non-redundant set of biological objects shared by each data source. The data model in both systems is a graph with typed nodes and edges, allowing for the incorporation of arbitrary data sources. In addition to curated data derived from the source databases, both ONDEX and Biozon include in-house data such as similarity links computed from sequence similarity of proteins and predicted links derived by text mining. A Similar approach is presented in (Eronen and Toivonen, 2012), called Biomine which is a system that integrates cross-references from several biological databases into a graph model with multiple types of edges, such as protein interactions, gene-disease associations and gene ontology annotations. They also formulate a protein interaction prediction and disease gene prioritization tasks as instances of link prediction.

In Lioni et al. (2010), the authors present SeqWare Query Engine, which has been created using modern cloud computing technologies and designed to support databasing information from thousands of genomes. Their backend implementation was built using the highly scalable, NoSQL HBase database from the Hadoop project. This software is open source and freely available from the SeqWare project (http://seqware.sourceforge.net). The paper Messina (2015) presents an integrated database structured as a NoSQL graph database based on OrientDB, which allows the integration of different types of data sources (Gene, miRBase, mirCancer), facilitating the performance of bioinformatics analysis using only one system. The authors in (Bonnici et al., 2014) presented ncRNA-DB, a NoSQL database based on the OrientDB platform that put together many biological resources that deal with several classes of non-coding RNA (ncRNA) such as miRNA, long-non-coding RNA (lncRNA), circular RNA (circRNA) and their interactions with genes and diseases. More recently, Bio4j (Pareja-Tobes et al., 2015) and BioGraphDB (Fiannaca et al., 2016b,a), have been developed. Bio4j is based on a Java library that allows building an integrated cloud-based data platform upon a graph structure, focused on the analysis of proteomic data. It, in fact, integrates data about protein sequences and annotations, GO terms, enzymes. Since Bio4j has fewer resources rather than BioGraphDB, an integrative database structured as a NoSQL graph database, based on the OrientDB platform, collecting data related to genes, microRNA (miRNA), proteins, pathways and diseases from ten online public resources. Moreover, In (Lysenko et al., 2016), they suggest

that graph databases provide a flexible solution for the integration of multiple types of biological data and facilitate exploratory data mining to support hypothesis generation. In (Gundla and Chen, 2016), they described a methodology of building two NoSQL application databases (MongoDB and AllegroGraph) using GO ontology, and then discuss how to achieve query relaxation through GO ontology NoSQL with ontologies.

Another application of NoSQL stores in bioinformatics is BigQ Gabetta et al. (2015), as an extension of the i2b2 framework, which integrates patient clinical phenotypes with genomic variant profiles generated by Next Generation Sequencing. The i2b2 web service is composed of an efficient and scalable document-based database that manages annotations of genomic variants and of a visual programming plug-in designed to dynamically perform queries on clinical and genetic data. The system is based on CouchDB. In Manyam et al. (2013), TargetHub a CouchDB based database used for storing miRNA-gene interactions for integration into high-throughput genomic analysis is developed. It integrates data from multiple miRNA repositories and allows users to systematically integrate data from multiple sources. In addition, CouchDB has been used to build three new bioinformatics resources (Manyam et al., 2012). GeneSmash as a database that collects data from various bioinformatics resources and provides automated gene-centric annotations needed and used in large scale projects such as the Cancer Genome Atlas (TCGA). The drugBase database used for storage of drug-target interactions and the HapMapCN drug-target database which provides an interface to query the copy number variations identified using the HapMap datasets. One can conclude from this overview that NoSQL stores have been used and implemented to address many challenges such as storing unstructured datasets and processing a huge volume of data, in addition to their usefulness to visualize biological networks and process distributed resources. However, much work is needed in order to compare different architectures such as polyglot persistence and multi-model.

2.1 Polyglot databases architecture

Polyglot database architectures are classified into three main types, Lambda architecture, Polyglot persistence, and Multi-Model databases. Lambda architecture is a combination of a slower batch processing layer and a speedier stream processing layer when real-time data processing is a requirement. In this paper, we will take interest in only two Polyglot persistence and Multi-Model systems.

2.1.1 Polyglot Persistence

The term polyglot persistence refers to using different data stores in different circumstances (Sadalage and Fowler, 2012), instead of choosing just one single database management system to store the entire data. Different kinds of data are best dealt with different data stores. Polyglot persistence makes it possible to choose as many databases as needed since they are built for different purposes. Figure 1 shows the polyglot persistence and the multi-model concept applied to proteomics data.

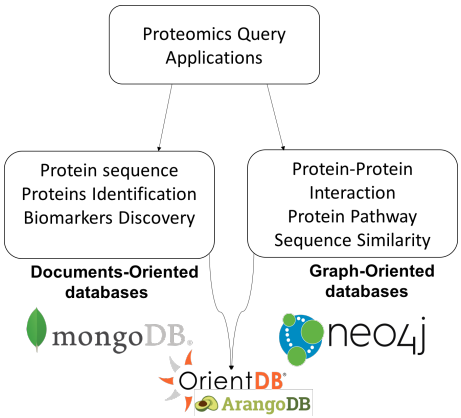


FIG. 1 – *Application of polyglot persistence and multi-model to proteomics data*

2.1.2 Multi-Model systems

Multi-model systems provide a database system that stores data in a single store but accesses the data with different APIs according to different data models. Indeed, multi-model databases are relying on different storage backends, which increases the overall complexity of the system and raises concerns like inter-database consistency, inter-database transactions, and interoperability as well as version compatibility and security. They either support different data models directly inside the store engine or they offer layers for additional data models on top of a single-model engine, see Figure 1.

Two open source multi-model databases are OrientDB and ArangoDB. OrientDB has a document API, an object API, and a graph API; it offers extensions of the SQL standard to interact will all three APIs. ArangoDB is a system that implements a data model integrating document, graph, and key-value models. It supports transactions, partitioning (sharding) and replication, and the Foxx language to develop components on the server side. It supports the query language AQL, which allows joins, operations on graphs, iterations, filters, projections, ordering, grouping, aggregate, functions, union, and intersection.

A comparison of different NoSQL stores (e.g. MongoDB, HBase) on storing and querying mass spectrometry datasets is presented in Shao and Conrad (2015). In another study (Have and Jensen, 2013), an experimental comparison of PostgreSQL and Neo4j using data imported from STRING v9.1: protein-protein interaction networks containing 20,140 proteins and 2.2 million interactions is given. It showed that speedup in Neo4j could be hundred or thousand times higher. The authors concluded that graph databases allow for efficient queries and give advantages in scalability with respect to any relational database. In (Wang et al., 2014), a study of high dimensional biological data retrieval optimization with NoSQL technology using a key-value model is presented. In (Guimaraes et al., 2015), they authors presented a proposal to store provenance data generated during the execution of biological workflows in a NoSQL document-oriented database system, MongoDB. The authors in Oliveira and del Val Cura (2016) presented a performance evaluation of multi-model data stores using polyglot persistence. They implemented a synthetic data generator to create the hybrid datasets.

3 Experimental Study

3.1 Datasets and Material

We conducted an experimental approach to compare the latencies of MongoDB combined with Neo4j and OrientDB, on storing interactions and protein sequence information of five datasets. For the graph, the data are available in the IntAct Molecular Interaction database [1] in PSI-mitab 2.5 format. For the documents datasets, we provide protein sequence and functional information from UniProt (Universal Protein Resource) [2] in CSV format. We use the following two data files:

— Homo sapiens as a large dataset of 159743 proteins with 11.5 million Interactions.
— Lactobacillus rhamnosus as a small dataset of 11707 proteins with 1.8 million interactions.

These experiments were performed using a virtual machine accessing one server running the NoSQL stores with Intel Xeon CPU E5-1650 3.20GHz processor, 16GB RAM running (Red Hat 4.8.5-11) and 500GB of storage. The versions of systems used in the experiments are OrientDB 2.2.20, MongoDB 3.4.1, and Neo4j 3.2.1 community version. The load and query operations used the web interface provided by the Neo4j and OrientDB data stores. Viewing the collections and documents created in MongoDB can be done using a command prompt. MongoDB does not offer a complete Web interface, then we used the (Robomongo) software which provides the user interface in order to access, view, create, add, edit and delete the existing or new collections and documents.

3.2 Data Modeling

The publicly available datasets listed in the previous section give us a huge amount of information, that we have to integrate in a harmonious and consistent way. Evaluating the loading, deletion, and querying of these data is the goal. Moreover, the datasets are available for download in several different formats, such as tab-delimited plain-text, structured XMLs, FASTA. In the graph model, each biological entity (protein) and its properties have been mapped respectively into a vertex and its attributes, and each relationship between two biological entities(protein) has been mapped into an edge. If a relationship has some properties, they are also saved as edge's attributes. Vertices and edges are grouped into classes, according to the nature of the entities. For example, all the proteins imported from Uniprot become instances of the protein vertex class. The latest available release of OrientDB has a powerful tool to move data from and to a database by executing an Extract-Transformer-Loader (ETL) process, described by a JSON configuration file. For ArangoDB, the data has been imported using the command-line tool utility named "arangoimp".

3.3 Results and Discussion

The performance study consists of comparing NoSQL stores in terms of i) data storing, deletion, and ii) queries latencies. Two real datasets are used to perform the comparison. OrientDB, ArangoDB, MongoDB, and Neo4j are considered for this study. The number of

1. http://www.ebi.ac.uk/intact/downloads
2. http://www.uniprot.org/

seconds taken to complete each operation is calculated 30 times, and the average is given to compare different stores. Smaller values of the average time indicate better performance.

3.3.1 Data Storing and Deletion

Data storing concerns with two operations i) the importation of the dataset into the NoSQL stores and ii) the insertion of a single data record into the NoSQL stores. The importation consists of loading the whole dataset in the stores while the insertion is done for a single data record. Moreover, the deletion concerns with deleting the whole dataset from the stores. These operations are applied to the small and large datasets.

Figure 2 shows the importation and deletion performance for document stores (MongoDB, OrientDB and ArangoDB) using the small dataset while Figure 3 displays result regarding the large dataset. The results for the importation reveal that MongoDB has better performance than OrientDB and ArangoDB in the small dataset. The results for the large dataset show that ArangoDB has better performance than MongoDB and OrientDB. The same conclusion is given for the deletion operation. We can conclude that for large datasets ArangoDB is most likely the most efficient system for importing and deleting Document-oriented data. This is also demonstrated in Figure 4, where MongoDB performs better than ArangoDB and OrientDB on the insertion of a single record. In all cases, OrientDB shows the worst performance.

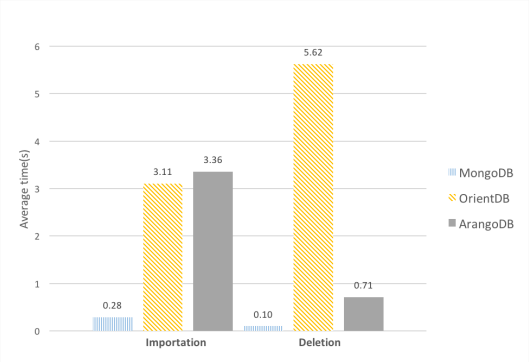


FIG. 2 – *Document operation for Small dataset*

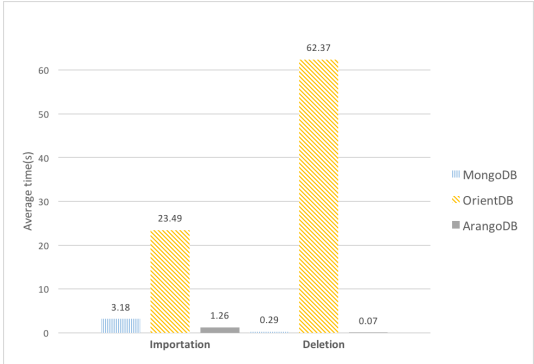


FIG. 3 – *Document operation for Large dataset*

Figure 5 shows the importation and deletion performance for graph stores (Neo4j, OrientDB and ArangoDB) using the small dataset while Figure 6 displays result regarding the large dataset. The results for the importation reveal that ArangoDB has the best performance in loading the graph than OrientDB and Neo4j in both cases, small and large dataset. The same conclusion is given for the deletion operations. There is a significant performance gain for Neo4j compared only to OrientDB when the importation is conducted in the large dataset. For larger network, Neo4j is better than OrientDB. ArangoDB is the most efficient in Graph-oriented data.

Notice that Neo4j includes a 'LOAD CSV' Cypher clause for data import, which is a powerful ETL tool. It can load a CSV file from the local filesystem or from a remote URI (i.e.

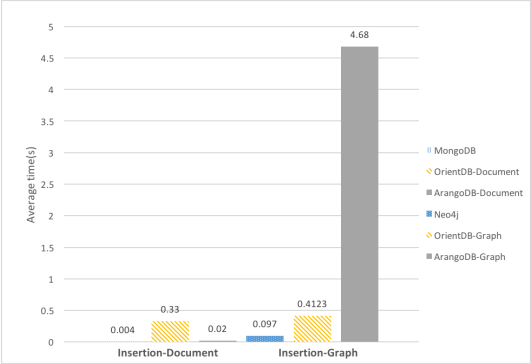


FIG. 4 – *Insertion of a single record*

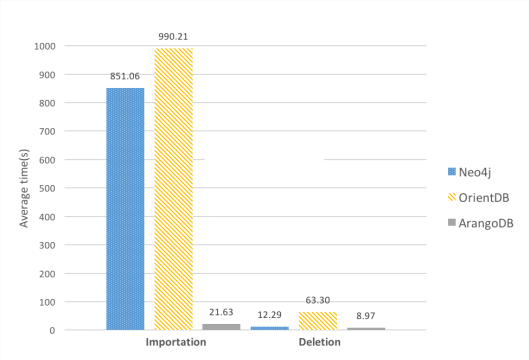


FIG. 5 – *Graph operation for Small dataset*

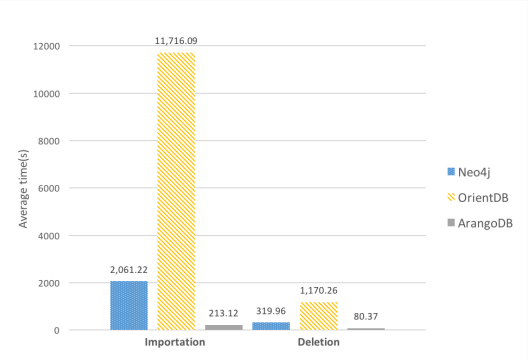


FIG. 6 – *Graph operation for Large dataset*

Dropbox, Github, etc.) and can be combined with USING PERIODIC COMMIT to group the operations on multiple rows in transactions to load large amounts of data. This can explain the superior performance of Neo4j compared to OrientDB.

In Figure 4, we present the performance results for the insertion of a single record in two data models: graph (Neo4j, OrientDB, and ArangoDB) and document (MongoDB, OrientDB and ArangoDB). In document data store OrientDB has the lower performance compared to MongoDB and ArangoDB, but for graph data store ArangoDB has the worst performance.

3.3.2 Query performance

We evaluate the performance of the multi-model and the polyglot persistence approaches using a query that retrieves a document and its network. The document key is randomly selected from the document-oriented database, then the network of the selected document is extracted from the graph-oriented database with a traversal through the graph up to a fixed depth level from 1 to 5. For example, using polyglot persistence data stores, each query was

run in two steps. In the first step, a key is randomly selected and the matched Uniprot-ID is retrieved from MongoDB. In the second step, the set of nodes connected with the selected Uniprot-ID is retrieved from Neo4j. The total elapsed time of the query is computed as a sum of both the Neo4j and MongoDB elapsed query times. In multi-model data stores, each query returns the matched documents and their connected documents in the graph.

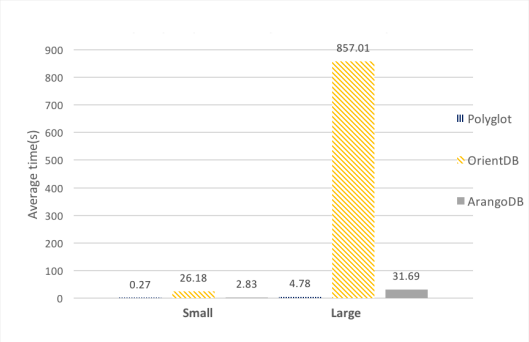


FIG. 7 – *Graph query retrieving documents with depth level 1*

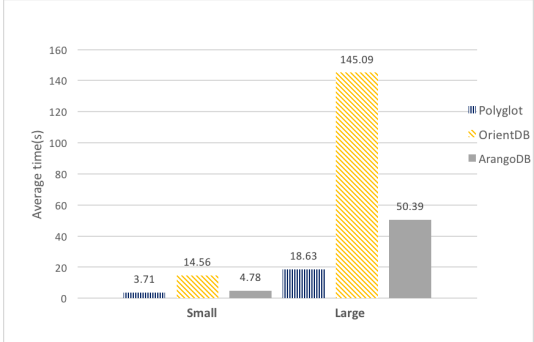


FIG. 8 – *Graph query retrieving documents with depth level 2*

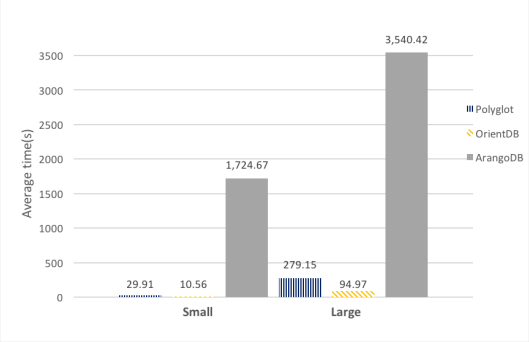


FIG. 9 – *Graph query retrieving documents with depth level 3*

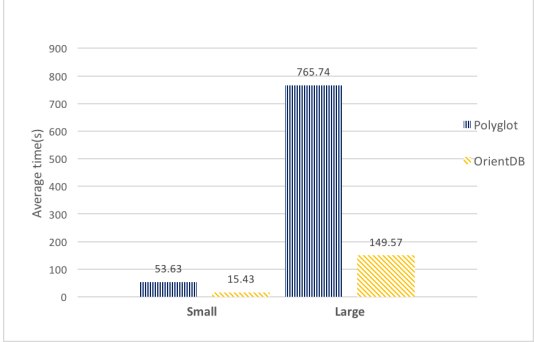


FIG. 10 – *Graph query retrieving documents with depth level 4*

Figures 7 and 8 show the performance results for querying the small and large datasets in depth levels 1 and 2. The results show that polyglot persistence (Neo4j and MongoDB) has the best performance for queries with graph traversal up to a depth level 2. Figure 9 shows that the performance of ArangoDB decreases while OrientDB reaches the best performance for queries that require graph traversal of depth level 3 probably because the graph engine implementation in ArangoDB is based on documents. Figure 10 and 11 show that OrientDB is still the multi-model data store that reaches the best performance for graph traversal depth levels 4 and 5. We conclude that when an application requires deeper levels of graph traversal, the best performance is reached by OrientDB. The same conclusions are made when querying the large

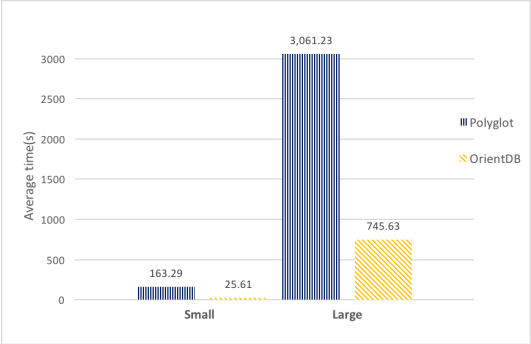


FIG. 11 – *Graph query retrieving documents with depth level 5*

dataset. There is a significant gain in the performance of polyglot persistence approach when large datasets are used and query for graph traversal up to a depth level 2 (see Figures 7 and 8). For instance, for a depth level 2 polyglot persistence approach has an average time much better (18.63s) than the multi-model (145.09s).

The results of our experiments show that both the size of the dataset and the depth level of graph query may impact the performance of both single and multi-model stores. We conclude the following statements regarding the use of *NoSQL* stores to manage proteomics data:

— MongoDB is faster than OrientDB and ArangoDB for importing and deleting protein information in small datasets.
— For large datasets, ArangoDB is the most efficient system for importing and deleting Document-oriented data.
— ArangoDB has the best performance in loading and deleting the graph than OrientDB and Neo4j in both cases, small and large dataset.
— OrientDB shows slower average time than MongoDB, Neo4j, and ArangoDB for single record insertion.
— Polyglot persistence (MongoDB+Neo4j) has faster average time than OrientDB and ArangoDB for query retrieval with depth level up to 2.
— ArangoDB becomes very slow when the depth level is greater than 3.
— OrientDB shows the best performance compared to Polyglot persistence for query retrieval with deeper depth level, greater than 2.

These results can be used as guidance to select a NoSQL system in order to store proteomics data.

4 Conclusion

In this paper, a performance study is provided to evaluate the time needed for storing, deleting and querying data using a polyglot persistence approach and a multi-model system. We found out that both the graph depth levels queries and the size of the graph influence the performance of both polyglot persistence and multi-model data stores. We conclude that for importing, inserting and deleting biomedical data as illustrated in this paper, ArangoDB is

faster than OrientDB and MongoDB regarding large document-oriented datasets. For small document-oriented datasets, MongoDB is the best. In the case of graph oriented datasets, ArangoDB shows better performance than OrientDB and Neo4j. In the query performance, we found out that when the application requires deeper levels of graph traversal, the best performance is reached by OrientDB.

References

Aniceto, R., R. Xavier, V. Guimarães, F. Hondo, M. Holanda, M. E. Walter, and S. Lifschitz (2015). Evaluating the cassandra nosql database approach for genomic data persistency. *International journal of genomics 2015*.

Atzeni, P., C. S. Jensen, G. Orsi, S. Ram, L. Tanca, and R. Torlone (2013). The relational model is dead, sql is dead, and i don't feel so good myself. *ACM SIGMOD Record 42*(2), 64–68.

Birkland, A. and G. Yona (2006). Biozon: a system for unification, management and analysis of heterogeneous biological data. *BMC bioinformatics 7*(1), 70.

Bonnici, V., F. Russo, N. Bombieri, A. Pulvirenti, and R. Giugno (2014). Comprehensive reconstruction and visualization of non-coding regulatory networks in human. *Frontiers in bioengineering and biotechnology 2*, 69.

Eronen, L. and H. Toivonen (2012). Biomine: predicting links between biological entities using network models of heterogeneous databases. *BMC bioinformatics 13*(1), 119.

Fiannaca, A., L. La Paglia, M. La Rosa, A. Messina, P. Storniolo, and A. Urso (2016a). Integrated db for bioinformatics: A case study on analysis of functional effect of mirna snps in cancer. In *International Conference on Information Technology in Bio-and Medical Informatics*, pp. 214–222. Springer.

Fiannaca, A., M. La Rosa, L. La Paglia, A. Messina, and A. Urso (2016b). Biographdb: a new graphdb collecting heterogeneous data for bioinformatics analysis. *Proceedings of BIOTECHNO*.

Gabetta, M., I. Limongelli, E. Rizzo, A. Riva, D. Segagni, and R. Bellazzi (2015). Bigq: a nosql based framework to handle genomic variants in i2b2. *BMC bioinformatics 16*(1), 415.

Guimaraes, V., F. Hondo, R. Almeida, H. Vera, M. Holanda, A. Araujo, M. E. Walter, and S. Lifschitz (2015). A study of genomic data provenance in nosql document-oriented database systems. In *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, pp. 1525–1531. IEEE.

Gundla, N. K. and Z. Chen (2016). Creating nosql biological databases with ontologies for query relaxation. *Procedia Computer Science 91*, 460–469.

Have, C. T. and L. J. Jensen (2013). Are graph databases ready for bioinformatics? *Bioinformatics 29*(24), 3107.

Köhler, J., J. Baumbach, J. Taubert, M. Specht, A. Skusa, A. Rüegg, C. Rawlings, P. Verrier, and S. Philippi (2006). Graph-based analysis and visualization of experimental results with ondex. *Bioinformatics 22*(11), 1383–1390.

Lee, K. K.-Y., W.-C. Tang, and K.-S. Choi (2013). Alternatives to relational database: comparison of nosql and xml approaches for clinical data storage. *Computer methods and programs in biomedicine 110*(1), 99–109.

Lioni, A., C. Sauwens, G. Theraulaz, and J.-L. Deneubourg (2010). Seqware query engine: storing and searching sequence data in the cloud. *BMC bioinformatics 11*, S2.

Lysenko, A., I. A. Roznovăţ, M. Saqi, A. Mazein, C. J. Rawlings, and C. Auffray (2016). Representing and querying disease networks using graph databases. *BioData mining 9*(1), 23.

Manyam, G., C. Ivan, G. A. Calin, and K. R. Coombes (2013). targethub: a programmable interface for mirna-gene interactions. *Bioinformatics 29*(20), 2657–2658.

Manyam, G., M. A. Payton, J. A. Roth, L. V. Abruzzo, and K. R. Coombes (2012). Relax with couchdb into the non-relational dbms era of bioinformatics. *Genomics 100*(1), 1–7.

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in genetics 24*(3), 133–141.

Marx, V. (2013). Biology: The big challenges of big data. *Nature 498*(7453), 255–260.

Messina, A. (2015). Etls for importing ncbi entrez gene, mirbase, mircancer and microrna into a bioinformatics graph database.

Moniruzzaman, A. and S. A. Hossain (2013). Nosql database: New era of databases for big data analytics-classification, characteristics and comparison. *arXiv preprint arXiv:1307.0191*.

Oliveira, F. R. and L. del Val Cura (2016). Performance evaluation of nosql multi-model data stores in polyglot persistence applications. In *Proceedings of the 20th International Database Engineering & Applications Symposium*, pp. 230–235. ACM.

Pareja-Tobes, P., R. Tobes, M. Manrique, E. Pareja, and E. Pareja-Tobes (2015). Bio4j: a high-performance cloud-enabled graph-based data platform. *bioRxiv*, 016758.

Robomongo. The web api for mongodb retrieved january 23, 2017 https://robomongo.org/.

Sadalage, P. J. and M. Fowler (2012). *NoSQL distilled: a brief guide to the emerging world of polyglot persistence*. Pearson Education.

Schulz, W. L., B. G. Nelson, D. K. Felker, T. J. Durant, and R. Torres (2016). Evaluation of relational and nosql database architectures to manage genomic annotations. *Journal of Biomedical Informatics 64*, 288–295.

Shao, B. and T. Conrad (2015). Are nosql data stores useful for bioinformatics researchers? *International Journal on Recent and Innovation Trends in Computing and Communication 3*(3), 1704–1708.

Wang, S., I. Pandis, C. Wu, S. He, D. Johnson, I. Emam, F. Guitton, and Y. Guo (2014). High dimensional biological data retrieval optimization with nosql technology. *BMC genomics 15*(8), S3.

Wiese, L. (2015). Polyglot database architectures= polyglot challenges. In *LWA*, pp. 422–426.

Yergey, A. L., C. G. Edmonds, I. A. Lewis, and M. L. Vestal (2013). *Liquid chromatography/mass spectrometry: techniques and applications*. Springer Science & Business Media.

Etude comparative sur les différentes attaques Iot : La couche perception

Kawtar Aarika*, Meriem Bouhlal*
El habib Benlhmar**, Sanaa El fillali***

*Avenue idriss el harti my rachid Gr 04 Casa
Aarika.kawtar@gmail.com
*Jamila 5 rue 28 N°22 Casa
bouhlalmeriem@gmail.com
**Av Driss El Harti Sidi Othmane B.P 7955، 20700
h.benlahmer@gmail.com
***Av Driss El Harti Sidi Othmane B.P 7955، 20700
elfilalis@gmail.com

Résumé. Le paradigme de l'Internet des objets (IoT) est l'une des innovations les plus passionnantes de ces dernières années. Cependant avec chaque objet connecté le nombre d'attaques possibles va augmenter de manière exponentielle. Cet article présente une étude comparative entre les différentes solutions contre les attaques de l'Iot, nous allons d'abord présenter une architecture de sécurité qu'on peut la diviser en 3 couches principales couche de perception, couche de réseau, et couche d'application. Chaque couche est divisée par des hiérarchies qui permettent de garantir la sécurité à savoir la confidentialité l'intégrité, l'authenticité, ainsi que l'acquisition des données, ensuite nous nous sommes focalisés sur les attaques de la couche perception qui est extrêmement vulnérable aux attaques de sécurité tout en classant ces attaques selon les deux technologies sans fil (WSN, RFID). Finalement Nous discutons sur les différentes approches procurées contre ces attaques.

1 Introduction

Le paradigme de l'Internet des objets (IoT) est l'une des innovations les plus passionnantes de ces dernières années.

L'exploitation de l'espace d'adressage IPv6, ainsi que la miniaturisation des dispositifs électroniques d'émission-réception ont ouvert la voie à la fourniture d'une adresse Internet à chaque objet sur Terre et du support technologique pour le transformer en objet communicant. Une fois que l'objet possède des capacités de communication, le nombre d'applications possibles devient potentiellement infini, cependant avec chaque objet connecté le nombre d'attaques possibles va augmenter de manière exponentielle. La preuve provient d'une méta-étude récente qui découvre que les appareils intelligents utilisés dans les soins de santé, les maisons et les bâtiments intelligents posent des risques considérables.

Les chercheurs quantifient les risques liés aux appareils de l'Internet des objets (Iot): **[1]**

- 90% des appareils collectaient au moins certaines informations via l'appareil
- 80% des appareils, ainsi que leurs composants Cloud et mobiles, n'ont pas exigé un

mot de passe assez complexe

- 70% des appareils, ainsi que leurs composants Cloud et mobiles, ont permis l'attaquant d'identifier les comptes d'utilisateurs via l'énumération.

Réduire le nombre des attaques pour l'internet des objets est une tache assez complexe ça nécessite une connaissance architecturale de la chaine de valeur qui relie les objets au Cloud. Dans la chaine de valeur des objets connectés, il faut commencer par les objets eux-mêmes, de nombreuse entreprise proposent des outils permettant de sécuriser telle ou telle partie de la chaine de valeur mais elles se sont positionnées pour une courte période sur les objets connectés **[3].**

Dans cet article, nous allons d'abord présenter un aperçu architecture de sécurité qui est divisé en 3 couche principale : la couche perception, la couche réseau et la couche application, chaque couche est divisé en hiérarchies qui permet de garantir la sécurité ,la confidentialité l'intégrité, l'authenticité ainsi que l'acquisition des donnée ensuite nous allons mis l'accent principalement sur quelques attaques de la couche perception parce qu'elle est extrêmement vulnérable aux attaques de sécurité, finalement nous allons définir les contres measure de chaque attaque cité qui utilisent principalement des techniques de détection et de réponse aux intrusions pour résister efficacement aux attaques illégales .

2 Architecture de sécurité pour l'Iot

La plupart des équipements de l'Iot n'ont pas d'environnement d'exécution uniforme ni puissance de calcul élevée, ce qui rendre l'implémentation d'une stratégie de sécurité unifié basée sur l'environnement fondamentale de l'Iot très difficile par conséquence il influencera la sécurité de l'Iot.

Les auteurs ont proposé une architecture de sécurité hiérarchique pour se protéger contre l'ouverture inhérente, l'hétérogénéité et la vulnérabilité du terminal. L'architecture proposée vise à améliorer l'efficacité, la fiabilité et la contrôlabilité de l'ensemble du système de sécurité.

La structure en réseau de l'Internet des objets est divisée en trois hiérarchies: la hiérarchie inférieure est l'équipement de détection pour l'acquisition d'informations; la hiérarchie intermédiaire est le réseau de transmission de données, tandis que la hiérarchie supérieure est conçue pour les applications et les middlewares, comme le montre la figure 1**[16].**

-Hiérarchie de sécurité de l'acquisition de l'information dans l'Internet des Objets permet de garantir L'intégrité et La fiabilité des données.

-Hiérarchie de sécurité de la transmission de l'information dans l'Internet des Objets afin de garantir la confidentialité, l'intégralité, l'authenticité et l'instantanéité des données **[16].**

-Hiérarchie de sécurité du traitement de l'information. Permet d'assurer la confidentialité, ainsi que le stockage sécurisé des informations, implique principalement dans la protection de la vie privée, la sécurité middleware, etc, et correspond à la sécurité de la hiérarchie des applications dans l'Internet des objets.

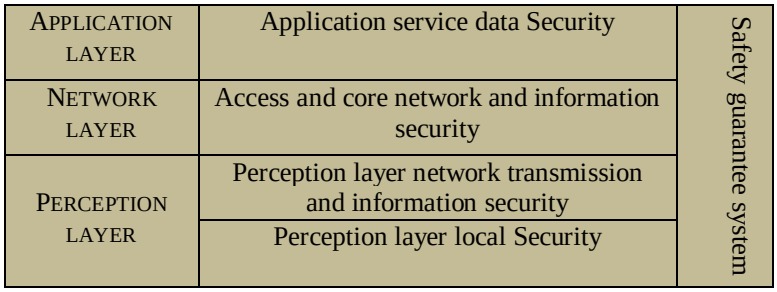


FIG. 1 –*Architecture de sécurité d'IoT [16]*

3 Les attaques de l'Iot pour la couche perception.

3.1 La différence entre WSN et RFID.

WSN sont des structures de nœuds indépendants dont la communication sans fil se fait sur une bande passante et fréquence limitée. Les nœuds des réseaux de capteurs sans fil sont constitués des éléments suivants, tels que Capteur, Microcontrôleur, Batterie, radio émetteur-récepteur et mémoire. En raison de la portée de communication limitée de chaque nœud de capteur WSN, un relais d'information multi-sauts a lieu entre la source et la station de base. Les réseaux de communication sont formés dynamiquement par l'utilisation d'émetteurs-récepteurs radio sans fil qui facilite la transmission de données entre les nœuds.

Le system RFID est un système d'identification par radiofréquence c'est une technologie basé sur la communication sans fil dans il se compose sur 2 composant principale à savoir Les étiquettes RFID (les tags RFID), un lecteur RFID ces deux technologie ce communique entre eux grâce à des antennes RFID intégrée dans chacun des 2 composants. Cette communication permet de transmettre le signal radiofréquence entre les deux composants **[2]**

3.2 La classification des attaques par Technologie.

Cet article tente de capturer un spectre plus large des failles de sécurité et des attaques de la couche perception dans les systèmes IoT. Notre classification est classifié selon les deux technologies WSN et RFID. Un résumé de la classification des attaques est présenté dans le tableau 1 ci-dessous.

Les attaques de la couche perception selon la technologie	
Wireless Sensor Network (WSN)	Radiqo Frequency Identification (RFID)
Jamming	Permanently Disabling Tags
Tampering	Temporarily Disabling Tags
Exhaustion	Relay Attacks
Collision	
Unfairness	

TAB. 1 – CLASSIFICATION DES ATTAQUES SELON LA TECHNOLOGIE

1. Jamming (Brouilleur) :

   Un brouilleur est un dispositif qui peut perturber partiellement ou entièrement le signal d'un nœud. Les réseaux de capteurs sans fil sont construits sur un support partagé qui permet aux adversaires de commettre facilement des interférences radio, ou brouillage, des attaques qui provoquent effectivement un déni de service des fonctionnalités de transmission ou de réception. Ces attaques peuvent facilement être accomplies par un adversaire en contournant les protocoles de la couche physique ou en émettant un signal radio visant à brouiller un canal particulier **[13-8]**. Il existe plusieurs types de brouilleurs qui sont classés en tant que brouilleurs constants, brouilleurs trompeurs, brouilleurs aléatoires et brouilleurs réactifs.

   (a) brouillage Proactive

   L'objectif d'un brouilleur proactive est de rendre tous les nœuds fonctionnels non réactifs, il transmettant des signaux indépendamment de toute communication de données dans le réseau en plaçant tous les nœuds dans un seul canal l jusqu'à ce que son énergie soit épuisée **[7-8-6].**

   (b) brouilleur réactif

   Ce type de brouilleur bloque le signal lorsque il observe une activité réseau sur un canal données un brouilleur réactif vise à compromettre la réception d'un message. Il peut perturber les paquets de petite et de grande taille **[6-4].** Voici deux façons différentes de mettre en œuvre un brouilleur réactif.

2. Tampering :

   visent généralement à attaquer le composant physique des appareils. Dans cette attaque, l'attaquant obtient un accès direct au composant matériel des nœuds tel que le microcontrôleur. Depuis, les nœuds WSN sont généralement utilisés dans un champ et laissés sans surveillance, ils sont vulnérables aux attaques de trempe **[5]**.

3. Collision :

   Dans l'attaque par collision, l'adversaire envoie son propre signal lorsqu'il entend qu'un nœud légitime va transmettre un message afin de faire des interférences. Les paquets entrent en collision lorsque deux nœuds tentent de transmettre simultanément sur la même fréquence, Cette attaque peut causer beaucoup de perturbations au fonctionnement du réseau **[5].**

4. Exhaustion

   Attaque par épuisement des ressources: l'opération consiste en des collisions répétées dans les trames et des retransmissions multiples jusqu'à la mort du nœud. Un

nœud malveillant demande ou transmet en permanence sur le canal **[5].**

5. Unfairness:

Certains algorithmes de couche MAC utilisent "aMacBattLifeExt". Cette technique donnera la priorité aux nœuds qui épuisent leur batterie. En d'autres termes, les nœuds proches de la mort ont la priorité pour envoyer les paquets. Un nœud adverse peut prendre l'avantage et mettre son "aMacBattLifeExt" à true. En mettant ce bit à true, le nœud adverse aura une priorité plus élevée pour l'envoi de données et rendra injuste .L'application de ce type d'attaque dépend du protocole de couche MAC implémenté dans les réseaux de capteurs **[20].**

6. Permanently Disabling Tags :

La désactivation permanente des étiquettes RFID englobe toutes les menaces possibles pouvant résulter de la destruction totale d'une étiquette. Les moyens possibles pour rendre une étiquette RFID définitivement inutilisable sont Tag Removal, Tag Destruction, KILL Command

(a) Tag Removal; les étiquettes qui sont pas intégré dans les éléments peuvent facilement être supprimés d'un élément

(b) Tag Destruction ; l'étiquette peut être détruite physiquement par des conditions environnementales extrêmes telles que les pressions ou des charges de tension. exposition chimique abrasion brusque, température ect **[19].**

(c) KILL Command ; chaque étiquette RFID a un mot de passe unique qui est défini par le fabricant de l'étiquette et son utilisation peut rendre une étiquette RFID définitivement inutilisable. Bien que cette fonctionnalité puisse être utilisée pour des raisons de confidentialité, il est évident qu'elle peut être exploitée par des adversaires malveillants afin de saboter les communications RFID **[19].**

7. Temporarily Disabling Tags:

Les étiquettes RFID comportent également le risque d'une désactivation temporaire involontaire causée par des conditions environnementales (par exemple, une étiquette recouverte de glace). La désactivation temporaire des étiquettes peut également être le résultat d'interférences radio passives ou actives.

(a) interférence passive: Considérant le fait que les réseaux RFID fonctionnent souvent dans un environnement intrinsèquement instable et bruyant, leur communication est rendue vulnérable aux interférences et aux collisions possibles de toute source d'interférence radio telle que les générateurs électroniques bruyants et les alimentations électriques. l'eau ou des billes de ferrite peuvent également perturber ou même bloquer le signal radio et conduire à un désaccord de fréquence radio. Cette interférence empêche une communication précise et efficace **[19].**

(b) Brouilleur active: un adversaire peut provoquer un brouillage électromagnétique en créant un signal dans la même plage que le lecteur afin de bloquer la communication entre les étiquettes **[19].**

8. Relay Attacks :

Une attaque par relais, également appelée attaque de l'intercepteur, consiste à placer un dispositif illégal entre le lecteur et l'étiquette de manière à intercepter les informations entre les deux nœuds, puis à les modifier ou à les transmettre directement au système. Les informations transmises par des dispositifs illégaux rencontreront un certain retard, et par conséquent, ces attaques sont appelées attaques relais **[12].**

4 Les contremesures des attaques Iot pour la couche perception.

Afin de sécurisé les données échangées de l'internet des objets nous proposons quelque contremesure des attaque cité dans le deuxième axe.

Attaque	Contremesure
Jamming	FHSS,DSSS, Regulated transmitted power,
Tampering	Raising Alarm
Exhaustion	Rate Limiting
Collision	Error -Correction Code
Unfairness	Small Frames Transmission
Permanently Disabling Tags	cryptographie à clé publique
Temporarily Disabling Tags	cryptographie à clé publique
Relay Attacks	Distance Limiting

TAB. 2 – LES CONTREMESURES DES ATTAQUES IOT

1. Regulated transmitted power

Consiste à diminuer la probabilité de découverte d'un attaquant Une puissance transmise plus élevée implique une plus grande résistance au brouillage car un signal de brouillage plus fort est nécessaire pour surmonter le signal d'origine [**14-15]**

2. Frequency-Hopping Spread Spectrum

La technique d'étalement de spectre par saut de fréquence consisté a changé la fréquence reçue en utilisant un algorithme partagé reconnue compréhensible par l'émetteur et le récepteur afin d'éviter des interférences. Elle apporte plusieurs avantages les environnements WSN **[14-15]**:
-Minimisation des interceptions non autorisé et le brouillage entre les nœuds et aussi le rapport signal sur bruit requis pour le porteur.

3. Direct-Sequence Spread Spectrum

Les systèmes à séquence directe est une méthode antibrouillages prometteuse il consiste à redonder les d'information à chaque envoi d'une séquence de bits en multipliant le signal RF entrant par le même porteur modulé PN(Le porteur PN permet de Ce signal numérique PN est une séquence pseudo-aléatoire) afin de trouver retrouver la donnée adéquate, sans rémission du paquet **[14-15].**

4. Raising d'alarme

Une solution propose contre l'attaque Tampering c'est de déclencher une alarme chaque fois qu'un nœud est touché par une partie non autorisée. Cependant, cela va venir avec des frais généraux **[9].**

5. Error-Correction Code

Consiste à incorporés des codes de correction d'erreur afin de tolérer des niveaux variables de corruptions dans les messages à n'importe quelle couche pour se défendre contre une collision. Mais cette solution a un coût plus élevé en termes de complexité de calcul et de consommation d'énergie. **[10- 9].**

6. Rate limiting

Afin éviter l'épuisement une méthode contre l'attaque qu'on l'appelle la limitation du taux de contrôle d'admission qui permet d'ignorer les demandes excessives et

éviter ainsi de transmettre des transmissions radio coûteuses**[9].**

7. Small Frames Transmission

Cette technique consiste à utiliser des petites trames afin de réduire le temps dont dispose un attaquant pour capturer le canal de communication **[11].**

8. cryptographie à clé publique

Cette méthode consiste à attribuer une clé d'identification privée permanente (non effaçable) à chaque étiquette en partageant avec le serveur Ensuite, lorsqu'une étiquette est contestée par un lecteur, elle génère une réponse en utilisant cette clé privée **[12].**

9. Distance Limiting

Afin de défendre contre le relais cette méthode utilise le temps d'aller-retour des messages pour mesurer la distance physique entre les deux côtés légaux. Cette méthode est assez difficile à résister à une attaque par relais **[21].**

5 Comparaison entre les différentes attaques IOT.

Ce tableau représente un résumé d'une étude comparative entre les différentes attaques en se basant sur les différents critères à savoir le type d'attaque qui permet de détecter l'aspect confidentialité et l'aspect intégrité, ensuite nous avons traité le critère du niveau d' endommagement qui permet de classifier le niveau de danger d'attaque IOT , et quant à la chance de détection de cette attaque que nous avons présenté dans ce tableau, nous montre la difficulté de de prévention d'attaque , finalement les différents algorithmes proposés par des auteurs se sont présentés dans ce tableau traitant les contres attaquent IOT finalement NOYS AVONS PRÉSENTÉ quelques algorithmes proposés par les auteurs concernant les contre-mesure utilisés pour chaque attaque.

Attaques	Les critères de classification				
	Type d'attaque	Niveau d'endommagement	Chance de détection	Possibilité de prévention	Contremesure
Jamming	Active - rendre tous les nœuds fonctionnels non réactifs	Elevé Epuisement d'énergie	Facile a détecté	Oui classification des paquets en temps réel sur la couche physique **[18]**	FHSS,DSSS, Regulated transmitted power,
Tampering	Active - Accès direct au composant matériel des nœuds	Elevé endommager ou remplacer un capteur	Difficile	Non impossible de contrôler l'accès aux capteurs dispersés sur plusieurs distances	Raising Alarm
Exhaustion	Active Collision répété	Elevé La mort des nœuds	Difficile	Oui limiter le taux de contrôle d'admission MAC	Rate Limiting
Collision	Active - Une modification dans une portion de données	Elevé erreur de vérification au niveau du récepteur	Difficile	Oui mécanismes de détection de collisions	Error - Correction Code
Unfairness	Active -dégrader le service	Moyen augmentation des délais de protocole MAC utilisées	Difficile	Non Techniques de détection de mauvais comportement	Small Frames Transmission
Permanently Disabling Tags and	Active même bloquer le signal radio	Elevé la destruction totale d'une étiquette	Difficile	Non (pas de moyen pour prévenir)	cryptographie à clé publique
Temporarily Disabling Tags	Active même bloquer le signal radio	Moyen empêche une communication précise et efficace	Difficile	Non	cryptographie à clé publique
Relay attack	Active –passive les modifier ou à les transmettre directement au système	Moyen Augmenter un certain retard au niveau de transmission des données	Très difficile	Oui Techniques de détection de mauvais comportement	Distance Limiting

TAB. 3 – COMPARAISON ENTRE LES ATTAQUES IOT

6 Conclusion

Avec la popularisation progressive de l'Internet des objets dans la vie quotidienne, la sécurité de l'IoT fait face à de plus en plus de défis.

Dans cet article, nous avons analysé différentes dimensions de la sécurité dans la couche perception pour un réseau IOT qui représente une grande variété d'attaques. Nous avons présenté et identifié les attaques les plus importants et ce, après les avoir classées selon la technologie utilisée à savoir : WSN, RFID. A la fin, nous avons traité les différentes techniques de détection et de défense associées à ces attaques pour les gérer et tous cela afin de retracer les futures activités de nos recherches dans ce domaine.

Le but de cet article est de mettre en évidence l'importance de la sécurité dans l'IoT, ainsi de fournir des bases à nos futurs études dont le but est de prévoir, de concevoir et de mettre en œuvre une nouvelle approche permettant la détection des failles de la couche perception et ce qui nécessite également des études supplémentaires.

Références

[1] The Internet of Hackable Things Nicola Dragoni, Alberto Giaretta and Manuel Mazzara Springer International Publishing AG 2018

[2] Security In the Internet of Things Based on RFID: Issues and Current Countermeasures Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013) Published by Atlantis Press, Paris, France. Xiao Nie , Xiong Zhong

[3] Internet of Things Security Internet of Things Security Yassine Chahid*, Mohamed Benabdellah, Abdelmalek Azizi 978-1-5090-6681-0/17/$31.00 2017 IEEE

[4] Jamming and Anti-jamming Techniques in Wireless Networks: A Survey Int. J. Ad Hoc and Ubiquitous Computing, Vol. 17, No. 4, 2014 197

[5] An Ontology for Attacks in Wireless Sensor Networks An Ontology for Attacks in Wireless Sensor Networks Wassim Znaidi, Marine Minier, Jean-Philippe Babau Submitted on 24 Oct 2008

[6] Literature Survey on Jamming Attack in Wireless Adhoc Network, Literature Survey on Jamming Attack in Wireless Adhoc Network Pinal Rupani Prof. Naren Tada 2017 IJEDR | Volume 5, Issue 2 | ISSN: 2321-9939

[7] Security Vulnerabilities and Countermeasures against Jamming Attacks in Wireless Sensor Networks 2017 International Conference on Computer, Communications and Electronics (Comptelix) Manipal University Jaipur, Malaviya National Institute of Technology Jaipur & IRISWORLD, July 01-02, 2017

[8] A Survey Jamming Sensor Networks: Attack and Defense Strategies Wenyuan Xu, Ke Ma, Wade Trappe, and Yanyong Zhang, Rutgers University 0890-8044/06/$20.00 2006 IEEE

[9] Denial of Service Attacks and Their Countermeasures in WSN IRACST – International Journal of Computer Networks and Wireless Communications (IJCNWC), ISSN: 2250-3501 Vol.3, No2, April 2013

[10] Security Threats in Wireless Sensor Networks, Hiren Kumar Deva Avijit Kar 1-4244-01 74-7/06/$20.00 ©2006 IEEE

[11] Denial of Service in Sensor Networks Anthony D. Wood John A. Stankovic University of Virginia

[12] RFID Security: Attacks, Countermeasures and Challenges Mike Burmester and Breno de Medeiros Computer Science Department Florida State University Tallahassee, FL 32306

[13] Jamming Attack Detection and Countermeasures In Wireless Sensor Network Using Ant System

[14] Internet of Things Security: Layered classification of attacks and possible Countermeasures

Summary

The paradigm of the Internet of Things (IoT) is one of the most exciting innovations of our last years. However with each connected object the number of possible attacks will increase exponentially. This article presents a comparative study between the different solutions and the attacks of the Iot. First, we will present a security architecture that can be divided into 3 main layers: Perception layer, network layer, and layer of application. Each layer is divided by hierarchies that guarantee security, namely confidentiality, integrity, authenticity, and data acquisition. Next, we will focus on the attacks of the perception layer, which is extremely vulnerable to security attacks while classifying these attacks on both wireless technologies (WSN, RFID). Finally we will discuss about the different approaches taken against these attacks.

Indicateurs de risque d'accidents piétons : Vers une décision floue intuitionniste

Meriem MANDAR*, Azedine BOULMAKOUL**
Lamia KARIM ***

*École Nationale des Sciences Appliquées

Bd Béni Amir, BP 77 Khouribga, Maroc

meriem.mandar@gmail.com,

**Faculté des sciences et techniques de Mohammedia,

Département d'informatique, Laboratoire LIM / IOS, Maroc

azedine.boulmakoul@gmail.com

***Ecole Supérieure de Technologie

Quartier Taqadoum Route aljazair BP 218 Berrechid Maroc

lkarim.lkarim@gmail.com

Résumé. Nous nous trouvons tous dans la situation d'un piéton quotidiennement au moins pour une courte période de temps. Dépourvu de protection, nous nous exposons aux risques d'accidents. Ces derniers ne devraient pas être acceptés comme inévitables car ils sont le produit d'un système global combinant le comportement individuel, système de transport et d'environnement dans des conditions qui rendent le dysfonctionnement à la fois prévisible et évitable. Par conséquent, la réduction ou l'élimination des risques pour les piétons est un objectif important et réalisable. Nous récapitulons dans ce papier, différentes approches que nous avons développé pour mesurer l'exposition des piétons aux risques d'accidents. Ces approches sont basées d'une part sur la modélisation floue et d'autre part sur la logique floue intuitionniste qui semble prometteuse car elle permet d'aborder la psychologie comportementale des piétons et des conducteurs en se basant sur les ensembles flous. La théorie intuitionniste floue a la flexibilité d'étudier l'hésitation et l'indécision. Une solution logicielle est développée pour cette instance en réutilisant les modèles de simulation piétons développés dans nos travaux précédents.

1 Introduction

Pratiquement chaque parcours commence et se termine avec une marche. Ce mode de déplacement naturel, bénéfique pour la santé et écologique, n'est pas aussi sain qu'il ne devrait l'être. Dépourvu de protection, les piétons sont des usagers du réseau routier particulièrement vulnérables. La conception des rues repose principalement sur des caractéristiques spécifiques aux voitures, sans prendre en considération celles des piétons. Et ce car la marche n'est pas encore considérée comme une part essentielle et nécessaire du système global de transport.
Les accidents des piétons, comme d'autres accidents de la circulation routière, ne devrait pas être acceptés comme étant inévitables parce qu'ils sont, le produit d'un système associant comportement individuel, outils de transport et environnement, dans des conditions qui rendent le dysfonctionnement à la fois prévisible et évitable. La réduction ou l'élimination des risques encourus par les piétons est un objectif important et réalisable. Il existe une association étroite entre l'environnement de la marche et la sécurité des piétons. Marcher dans un environnement qui manque d'infrastructures dédiées aux piétons augmente le risque d'accident des piétons. La capacité de répondre à la sécurité des piétons est une composante importante des efforts visant à prévenir les accidents de la circulation routière. Outre les attitudes à l'égard des risques et leur évaluation, tout usage de la route doit également prendre en compte un ensemble de règles gouvernant les interactions de l'environnement routier. La transgression de ces règles est à l'origine des comportements dangereux pouvant induire des accidents. Cet acte est autant plus observé chez les hommes que les femmes, qu'ils soient conducteurs ou piétons. Ceci peut s'expliquer par le fait que dans une telle situation, les femmes sont plus sensibles aux potentielles pertes, alors que les hommes le sont par rapport aux gains. Les accidents de piétons se produisent majoritairement en milieu urbain, du fait d'une exposition au risque supérieure en ville. Le taux de gravité exprimé par le ratio du nombre de personnes tuées et le nombre de celles accidentées est plus important en dehors des agglomérations sur les autoroutes et les routes nationales. Et ce à cause d'une part de l'importance des vitesses et densités de circulation et que les conducteurs s'attendent moins à trouver des piétons. Par ailleurs, beaucoup d'accidents de moindre gravité se produisent aux passages des piétons où les vitesses de circulation sont plus faibles et les conducteurs sont avertis de la présence des piétons. Les conducteurs sont plus susceptibles à s'arrêter lorsque les piétons semblent déterminés à traverser, contrairement à ceux qui attendent passivement leur tour pour traverser. Cependant des différences existent selon les catégories de piétons : les enfants sont percutés au début de la traversée puisqu'ils s'élancent sans regarder, tandis que les piétons âgés sont percutés en fin de traversée vu qu'ils n'ont pas le temps de traverser alors que la circulation a déjà repris Weidmann et al. (2012).

Facteurs de risque pour les piétons Les principaux risques pour les piétons sont relatifs à un large éventail de facteurs : Les véhicules en termes de vitesse de déplacement et de conception surtout aux niveaux des fronts solides. L'effet sur le risque d'accidents provient principalement de la relation entre la vitesse et la distance d'arrêt. Le manque d'expérience de la circulation routière et des aspects relatifs à la sécurité des piétons, permettent d'expliquer la sur-implication des enfants et les personnes âgées.

Tandis que Les piétons, qui sont également conducteurs, possèdent une meilleure perception des risques d'accidents. Ils se servent à la fois de leur expérience de la conduite et de leur point de vue de piétons pour évaluer les dangers potentiels. Le risque d'accidents des piétons augmente lorsque la conception des routes et de l'aménagement du territoire ne parvient pas à planifier et fournir des installations comme les trottoirs, les passages, les points de refuges ou les médianes soulevées, ou un examen adéquat des accès des piétons aux intersections. D'autres facteurs liés aux traits de comportement des piétons et des conducteurs contribuent également aux accidents des piétons.
Dans ce contexte, nous avons développé différentes approches pour mesurer l'exposition des piétons aux risques d'accidents. Et ce en se basant d'une part sur la modélisation floue et d'autre part sur la logique floue intuitionniste. Cette dernière permet d'aborder la psychologie comportementale des piétons et des conducteurs en se basant sur les ensembles flous, tout en ayant la flexibilité d'étudier l'hésitation et l'indécision. Une solution logicielle est développée pour cette instance en réutilisant les modèles de simulation piétons développés dans nos travaux précédents Boulmakoul et Mandar (2011) Mandar et Boulmakoul (2016). Ce papier récapitule les différentes approches développées pour mesurer l'exposition des piétons aux risques d'accidents. Pour en avoir plus de détails et notamment sur les résultats obtenus, voir les références de nos travaux précédents. Après l'introduction, le reste de l'article est organisé comme suit. Dans la deuxième section, nous présentons le concept d'exposition des piétons aux risque d'accidents. Alors que dans la section 3, nous présentons trois approches que nous avons développés dans nos travaux précédents pour mesurer cette exposition. Une conclusion est donnée dans la section suivant e.

2 Exposition aux risques d'accidents

Le domaine de l'épidémiologie définit l'exposition comme une situation dans laquelle un agent est soumis à une substance potentiellement dangereuse. Le risque est une fonction de l'exposition et du danger. Il se réfère à la probabilité d'occurrence d'un événement dangereux après un ensemble d'essais représentant des unités d'exposition. Une «exposition» peut être définie comme un événement se produisant à un endroit donné. Ainsi, l'exposition de l'organisme i situé à la position (x, y, z) est donnée par :

$$E_i\left(t\right) = \int_0^T c\left(x, y, z, t\right) dt \tag{1}$$

Dans nos travaux, l'exposition des piétons est donc définie comme un taux de contact des piétons avec un trafic de véhicules potentiellement dangereux. Il est important de comprendre le concept de l'exposition des piétons et leur relation avec le risque pour les piétons. Il n'y a pas de meilleure mesure pour l'exposition des piétons. Cependant, selon les besoins et les objectifs spécifiques, certaines mesures sont mieux adaptées que d'autres. Nous supposons que les piétons sont exposés à des risques lorsqu'ils traversent la route. Cette hypothèse est presque réaliste compte tenu du faible taux d'accidents des routes hors traversée. En outre, les formes des trajectoires de traversées des piétons peuvent avoir des formes différentes. Le choix de la trajectoire est habituellement un compromis entre la perception du risque par le piéton et sa capacité

à traverser avec le plus grand confort possible. Sur la zone de traversée, le chemin prend généralement la forme d'une ligne perpendiculaire à la route. Et hors zone de traversée, les piétons ont tendance ces derniers ont tendance à arrondir les angles et choisir des lignes obliques. Indépendamment du chemin et de la section de route choisit, les piétons tentent d'ajuster leur vitesse en fonction de leurs situations.

3 Mesures d'exposition des piétons aux risques d'accidents

3.1 Première approche

Dans notre contexte, les piétons sont exposés aux risques d'accidents dans un segment de route, impliqués par un flux de véhicules pendant leur temps de traversée. Cette exposition est définie comme suit Mandar et Boulmakoul (2014) :

$$E\left(t\right) = \int_0^{t_c} q_v dt = q_V \cdot t_C \tag{2}$$

Où q_V et t_C désignent respectivement le flux des véhicules et le temps de traversée des piétons. Nous supposons que le piéton traverse une route ayant une largeur donnée D dans une ligne rectiligne, avec une vitesse donnée ν_p. En outre, le flux de véhicules peut être exprimé en termes de leurs densités et vitesses selon l'équation suivante :

$$\upsilon = \upsilon_{\max}\left(1 - \frac{\rho}{\rho_{\max}}\right) \tag{3}$$

Par conséquent l'exposition des piétons aux risques d'accidents devient :

$$Exp_{p/V} = \rho_V \cdot \upsilon_{\max} \cdot t_p - \frac{\rho_V^2 \cdot \upsilon_{\max} \cdot t_p}{\rho_{\max}} \tag{4}$$

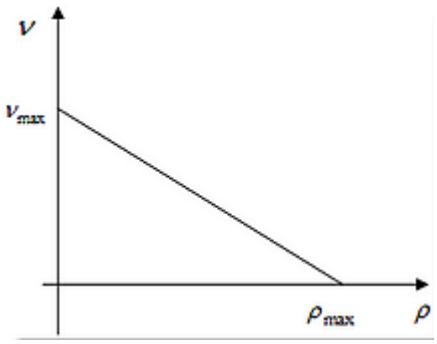


FIG. 1 – *Variation de la vitesse en fonction de la densité.*

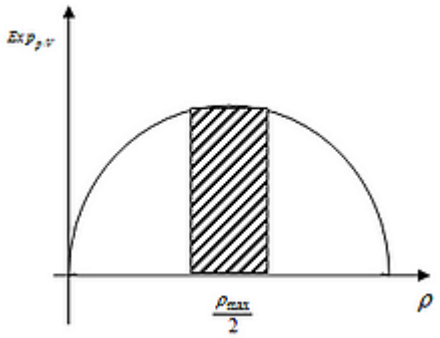


FIG. 2 – *Variation de l'indicateur de risques d'accidents des piétons selon la densité des véhicules.*

Cette exposition aux risques d'accidents, suit une courbe parabolique. En effet, si la densité du véhicule atteint sa valeur maximale $\rho_V = \rho_{\max}$, le risque d'accidents s'annule $Exp_{p/V} = 0$ et les piétons peuvent traverser inter véhicules. Alors que si la densité des véhicules est nulle $\rho_V = 0$, les piétons ne courent aucun risque pendant leur traversée, puisqu'ils peuvent l'accomplir en absence de véhicules et on a alors $Exp_{p/V} = 0$. Le problème se pose alors pour des valeurs intermédiaires de la densité des véhicules, dans un intervalle de centre $\rho_V = \rho_{\max}/2$

Par ailleurs, nous pouvons penser qu'il existe un risque d'accidents pour les véhicules, imposé cette fois par les piétons. En effet, les piétons transgressent leurs règles de passage sur la route généralement en formant des groupes. Et ce car, quand un piéton réussi à traverser dans l'écart entre deux véhicules, d'autres vont accélérer et le suivre, imposant ainsi aux véhicules de leurs céder le passage. Cette exposition des véhicules aux risques d'accidents prend la même forme que dans le cas des piétons :

$$Exp_{V/p} = q_p \cdot t_V \tag{5}$$

Où cette fois q_P et t_V désignent le débit des piétons et le temps de traversée des véhicules respectivement. Cette exposition suit également une courbe parabolique. Lorsque la densité des piétons atteint sa valeur maximale $\rho_p = \rho_{\max}$, le risque d'accidents s'annule pour les véhicules $Exp_{V/p} = 0$ puisqu'ils doivent complètement s'arrêter et attendre le passage de la foule de piétons. Alors que si la densité des piétons est nulle $\rho_p = 0$ les véhicules ne courent aucun risque pendant leur passage, puisqu'ils peuvent l'accomplir en absence de piétons et on a alors $Exp_{V/p} = 0$. Le problème se pose alors également pour des valeurs intermédiaires de la densité des piétons, dans un intervalle de centre $\rho_p = \rho_{\max}/2$. C'est le cas où les véhicules se voient dans l'obligation de freiner pour éviter un accident potentiel. Cette réaction se propage vers tous les véhicules suiveurs sur la voie en question. Ces variations de la valeur d'exposition des véhicules aux risques d'accidents en fonction de la densité des piétons sont schématisées de la même manière que pour le cas des piétons.

3.2 Deuxième approche : modélisation floue et distance de sécurité

Le processus de traversée de route repose sur la théorie de la maximisation de l'utilité. D'autre part, les véhicules ne peuvent s'arrêter immédiatement après la perception des piétons en raison de leurs vitesses. Ils ont besoin de ce qu'on appelle une distance de sécurité afin de pouvoir s'arrêter sans collision avec les piétons ou d'autres obstacles dans leurs itinéraires. Cette distance peut être exprimée par la formule suivante :

$$S = L + T_r\nu + {}^{\nu^2}/_{2\gamma} \tag{6}$$

Où : T_r est le temps de réaction du conducteur ;L est la longueur du véhicule ; ν est la vitesse du véhicule ; γ est l'accélération du véhicule. De l'équation précédente, nous pouvons obtenir ce que nous appelons le temps d'arrêt de sécurité pour un véhicule, qui est donné par :

$$T_s = {}^{S}/_{\nu} = {}^{L}/_{\nu} + T_r + {}^{\nu}/_{2\gamma} \tag{7}$$

Les études statistiques en France définissent la distance de sécurité pour un véhicule par la formule suivante : $S = 8+0.2\nu+0.03\nu^2 \cong 0.2\nu+0.03\nu^2$. Ainsi le temps d'arrêt de sécurité pour le véhicule devient : $T_s = 0.2+0.03\nu$. En outre, le débit du véhicules peut être exprimé en termes de densité et de vitesse, qui atteignent leurs valeurs maximales si la densité est nulle et s'annulent dans le cas de la densité maximale, selon l'équation suivante : $\nu = \nu_{\max}\left(1 - \rho/\rho_{\max}\right)$

Où : ν et $\nu_{\max}$ représentent la vitesse du véhicule et sa valeur maximale respectivement ; ρ et $\rho_{\max}$ représentent la densité des véhicules et sa valeur maximale respectivement ; Par conséquent, le temps d'arrêt de sécurité pour un véhicule devient :

$$T_s = T_r + \frac{\nu_{\max}}{2\gamma}\left(1 - \rho/\rho_{\max}\right) \tag{8}$$

Sachant que le flux d'un organisme est le produit de sa densité et de sa vitesse, l'exposition des piétons aux risques d'accidents devient :

$$E = T_P \cdot \rho \cdot \nu \tag{9}$$

Par ailleurs, le débit du véhicule peut être exprimé en termes de densité et de vitesse, qui atteint sa valeur maximale si la densité est nulle et s'annule dans le cas d'une densité maximale. Par conséquent, l'exposition des piétons aux risques d'accidents devient :

$$\mathrm{E\,(t)} = \mathrm{T_P} \cdot \rho \cdot \nu_{\max}\left(1 - \rho/\rho_{\max}\right) = \mathrm{T_P} \cdot \rho \cdot \nu_{\max} - \mathrm{T_P} \cdot \rho^2 \nu_{\max}/\rho_{\max} \tag{10}$$

Nous considérons que le temps d'arrêt de sécurité pour un véhicule et le temps de traversée des piétons sont des nombres flous triangulaires $\tilde{T}_P = tfn\,(T_P, \alpha_l, \alpha_r)$ et $\tilde{T}_S = tfn\,(T_S, \beta_l, \beta_r)$ respectivement. Et ce car leur connaissance n'est pas déterministe et reste imprécise autant pour les piétons que pour les véhicules Mandar et al. (2017a).

Par conséquent, l'exposition des piétons aux risques d'accidents devient alors :

$$\tilde{E}' = \left(\tilde{T}_{\mathrm{S}} - \tilde{T}_{\mathrm{P}}\right) \cdot \rho \cdot \nu = \tilde{T}_{\mathrm{S}} \cdot \mathrm{q} - \tilde{E} \tag{11}$$

De plus, en référant le temps d'arrêt de sécurité pour un véhicule dans la nouvelle formulation de l'exposition des piétons aux risques d'accidents, nous obtenons :

$$\tilde{E}' = \left(\mathrm{T_r} + \frac{\nu_{\max}}{2\gamma}\left(1 - \rho/\rho_{\max}\right)\right) \cdot \mathrm{q} - \tilde{T}_{\mathrm{P}} \cdot \rho \cdot \nu = (\tilde{\alpha} - \beta\nu) \cdot \mathrm{q} \tag{12}$$

où $\tilde{\alpha} = T_r - \tilde{T}_P$ et $\beta = 1/2\gamma$.

Par conséquent, les nouveaux et l'ancien indicateurs de risque d'accident sont liés par la formule suivante :

$$\tilde{E}' + \tilde{E} = T_r\nu - \beta\nu^2\rho \tag{13}$$

Lors d'une situation de traversée, deux cas peuvent être discutés. Le premier est quand $\tilde{T}_P > \tilde{T}_S$, ce qui signifie que les piétons ont suffisamment de temps pour traverser la route avant que le véhicule ne s'arrête. Dans ce cas le risque d'accidents pour les piétons tend vers zéro. Soient $P\left(\tilde{T}_S < \tilde{T}_P\right)$ et $P\left(\tilde{T}_P < \tilde{T}_S\right)$ la probabilité que le temps de passage des piétons soit plus grand, ou moins élevé, que le temps d'arrêt de

sécurité d'un véhicule respectivement. Les piétons traversent la route si et seulement si $P\left(\tilde{T}_S < \tilde{T}_P\right) \gg P\left(\tilde{T}_P < \tilde{T}_S\right)$.
Le deuxième cas est quand $\tilde{T}_P < \tilde{T}_S$. Dans ce cas le temps d'arrêt de sécurité du véhicule est plus grand que le temps de traversée du piéton. Par conséquent ce dernier ne peut traverser la route en sécurité et son risque d'être heurté par un véhicule augmente. Par conséquent, le véhicule se déplace si et seulement si $P\left(\tilde{T}_S < \tilde{T}_P\right) \ll P\left(\tilde{T}_P < \tilde{T}_S\right)$.

3.3 Troisième approche : modélisation floue intuitionniste

3.3.1 logique intuitionniste floue

Dans la théorie des ensembles flous, l'appartenance d'un élément à un ensemble flou est une valeur unique dans l'intervalle [0,1] Zadeh (1975). Néanmoins, en réalité il n'est peut-être pas toujours vrai que le degré de non-appartenance d'un élément dans un ensemble flou est égal à 1 moins le degré d'appartenance. Et ce, car il peut y avoir un degré d'incertitude. Par conséquent, Atanassov (1986, 1999) a proposé une généralisation d'ensembles flous, comme des ensembles flous intuitionnistes qui intègrent le degré d'incertitude ou d'hésitation appelé marge d'hésitation Atanassov (1986) Atanassov (1999). Cette marge étant définit comme 1 moins la somme des degrés d'appartenance et de non-appartenance respectivement.

Pourquoi le risque intuitionniste flou ? Dans la modélisation du risque des piétons en interaction avec les véhicules, seuls les paramètres cinématiques du véhicule et du piéton ainsi que certains critères géométriques de la route sont souvent considérés. L'intégration des facteurs comportementaux liés à la perception de l'espace et à la décision des deux usagers de la route reste totalement absente. Pour surmonter ce problème, nous utilisons l'approche intuitionniste qui permet de relier les deux réalités perçues par les piétons et les conducteurs. L'approche intuitionniste repose à la fois sur l'information objective et subjective du conducteur et du piéton. D'une part, nous avons intégré les deux perceptions antagonistes dans les nombres flou intuitionnistes et, d'autre part, développé une méthode de classement relatif pour dériver les indicateurs d'exposition au risque.

3.3.2 Deuxième formulation du modèle intuitionniste d'exposition aux risques

Dans ce travail, nous avons utilisé les nombres flou intuitionnistes. Le concept du nombre flou Intuitionniste triangulaire (TIFN) est une généralisation de celle du nombre flou triangulaire Mandar et al. (2017b).
Définition : Un nombre flou intuitionniste triangulaire $\tilde{u} = (\underline{u}, u, \overline{u}, \alpha_{\tilde{u}}, \beta_{\tilde{u}})$ est un ensemble flou intuitionniste spécial sur l'ensemble des nombres réels R, dont les fonctions d'appartenance et de non-appartenance peuvent être données comme suit :

$$\mu_{\tilde{u}}(x) = \begin{cases} \frac{(x-\underline{u})\times\alpha_{\tilde{u}}}{(u-\underline{u})} & if\ \underline{u} \le x < u \\ \alpha_{\tilde{u}} & if \quad x = u \\ \frac{(\overline{u}-x)\times\alpha_{\tilde{u}}}{(\overline{u}-u)} & if\ u < x \le \overline{u} \\ 0 & if \quad \overline{u} < x \ or \ x < \underline{u} \end{cases} \tag{14}$$

$$\vartheta_{\tilde{u}}(x) = \begin{cases} \frac{(u-x)+(x-\underline{u})\times\beta_{\tilde{u}}}{(u-\underline{u})} & if \quad \underline{u} \le x < u \\ \beta_{\tilde{u}} & if \quad x = u \\ \frac{(x-u)+(\overline{u}-x)\times\beta_{\tilde{u}}}{(\overline{u}-u)} & if \quad u < x \le \overline{u} \\ 1 & if \quad \overline{u} < x \ or \ x < \underline{u} \end{cases} \tag{15}$$

les valeurs $\alpha_{\tilde{u}}$ et $\beta_{\tilde{u}}$ représentent le degré maximum d'appartenance et le degré minimum de non-appartenance, respectivement, en vérifiant les inéquations suivantes $0 \le \alpha_{\tilde{u}} \le 1$, $0 \le \beta_{\tilde{u}} \le 1$ et $0 \le (\alpha_{\tilde{u}} + \beta_{\tilde{u}}) \le 1$. Ces paramètres sont introduit pour refléter le niveau de confiance et le niveau de non-confiance du TIFN $\tilde{u} = (\underline{u}, u, \overline{u}, \alpha_{\tilde{u}}, \beta_{\tilde{u}})$

Soient $\tilde{u}_i = (_i, \mathrm{u_i}, \overline{\mathrm{u_i}}, \alpha_{\tilde{u}_i}, \beta_{\tilde{u}_i}) \quad i = 1 \cdot \cdot N$ des nombres flous intuitionnistes.

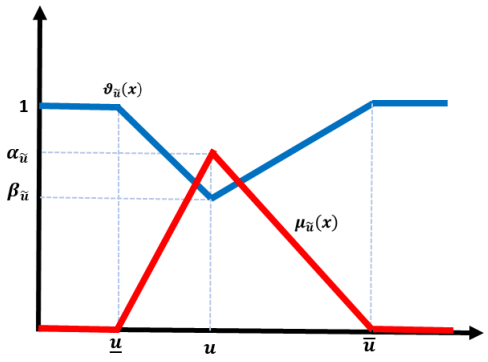


FIG. 3 – *Représentation d'un nombre flou intuitionniste* $\tilde{u} = (\underline{u}, u, \overline{u}, \alpha_{\tilde{u}}, \beta_{\tilde{u}})$.

Nous adoptons un méthode de tri des nombres flous intuitionnistes qui se base sur un rapport entre l'indice de valeur et l'indice d'ambiguïté pour un TIFN qui est défini comme suit Li (2010) :

$$\mathrm{R}(\tilde{u}_i) = \frac{[_i + 4 \times \mathrm{u_i} + \overline{\mathrm{u_i}}]\,[(\lambda - 1) \times \alpha_{\tilde{u}_i} + \lambda \times (1 - \beta_{\tilde{u}_i})]}{2 \times [3 + (\overline{\mathrm{u_i}} - _i) \times [(1 - \lambda) \times (1 - \beta_{\tilde{u}_i}) + \lambda \times \alpha_{\tilde{u}_i}]]} \tag{16}$$

Où $\lambda \in [0, 1]$.

3.3.3 Modèle flou intuitionniste d'exposition aux risques

L'idée fondamentale de ce modèle est de définir la prise de décision d'un acteur (piéton ou véhicule) par deux schémas. Pour un acteur donné, le premier schéma suppose que l'acteur antagoniste prenne la bonne décision qui reste cohérente avec ses connaissances pour prendre la bonne ou la mauvaise décision. Dans le deuxième schéma, l'un des acteurs suppose que son acteur antagoniste prenne la mauvaise décision qui reste cohérente avec ses connaissances pour prendre la bonne ou la mauvaise décision. Précisément, cette indécision hypothétique peut être modélisée par la théorie des nombres flous intuitionnistes.

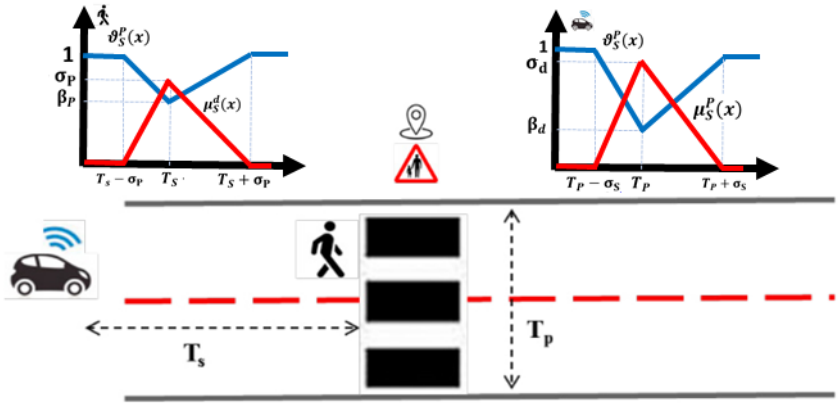


FIG. 4 – *Représentation des décision et indécision intuitionnistes.*

Soient (σ_s, σ_p) deux paramètres correspondant à la déviation du temps d'arrêt de sécurité pour un véhicule T_s le , et T_p du temps de traversée des piétons, respectivement. Nous définissons deux nombres flous intuitionnistes triangulaires pour modéliser les estimations du risque d'accident du point de vue des deux utilisateurs de la route piétons et conducteurs.

— Soit $\mathcal{R}^{\mathrm{d}} = \left(\mu_{\mathrm{S}}^{\mathrm{d}}(\mathrm{x}), \vartheta_{\mathrm{S}}^{\mathrm{d}}(\mathrm{x})\right) = (\mathrm{T_p} - \sigma_{\mathrm{d}}, \mathrm{T_p}, \mathrm{T_p} + \sigma_{\mathrm{d}}, \alpha_{\mathrm{d}}, \beta_{\mathrm{d}})$ qui dénote une évaluation incorrecte du risque du point de vue du conducteur. Ce nombre permet de modéliser l'estimation du temps de traversée du piéton du point de vue du conducteur.

— Soit $\mathcal{R}^p = (\vartheta_S^p(x), \vartheta_S^p(x)) = (T_s - \sigma_p, T_s, T_s + \sigma_p, \alpha_p, \beta_p)$ qui dénote une évaluation incorrecte du risque du point de vue du conducteur. Ce nombre permet de modéliser l'estimation du temps d'arrêt de sécurité du véhicule du point de vue du piéton.

Où

— $\mu_{\mathrm{S}}^{\mathrm{d}}(\mathrm{x})$ désigne le degré d'indécision dont dispose le conducteur du véhicule pour estimer correctement le temps de traversée des piétons.

— $\vartheta_S^{\mathrm{d}}(\mathrm{x})$ désigne le degré d'incertitude dont dispose le conducteur pour ne pas estimer correctement le temps de traversée du piéton.
— $\vartheta_S^p(x)$ désigne le degré d'indécision dont dispose le piéton pour estimer correctement le temps d'arrêt de sécurité du véhicule.
— $\vartheta_S^p(x)$ désigne le degré d'incertitude dont dispose le piéton pour ne pas estimer correctement le temps d'arrêt de sécurité du véhicule.

Avec ces considérations, nous obtenons les scénarios de risque présentés dans le tableau suivant qui présente les clauses d'occurrence d'accidents ou des conflits dangereux entre piétons et véhicules. D'autres situations sont plausibles et seront prises en considération dans des travaux futurs. Les niveaux de risque sont classés comme suit : $\mu^{+} < \mu\vartheta^{++} < \mu\vartheta^{+++} < \vartheta^{++++}$

	Décision du conducteur	
Décision du piéton	$\boldsymbol{\mu_S^d(x)}$	$\boldsymbol{\vartheta_S^d(x)}$
$\boldsymbol{\mu_S^p(x)}$	$\boldsymbol{\mu^{+}}$	$\boldsymbol{\mu\vartheta^{+++}}$
$\boldsymbol{\vartheta_S^p(x)}$	$\boldsymbol{\mu\vartheta^{++}}$	$\boldsymbol{\vartheta^{++++}}$

Tab. 1 – *Les prémisses du Hasard selon la théorie des ensembles flous intuitionnistes*

Si $\tilde{T}_P < \tilde{T}_S$ alors les piétons ne peuvent pas traverser la route en sécurité parce qu'ils seront percutés par un véhicule. Par conséquent, nous devons calculer $\mathcal{F} = \frac{\mathrm{R}(\tilde{T}_P)}{\mathrm{R}(\tilde{T}_S)}$. Si $\mathcal{F} \ll 1$ alors le risque d'accident augmente considérablement, et par conséquent, nous adopterons indicateur intuitionniste floue du risque. Certes, l'idée est intéressante puisqu'elle permet de modéliser des prises de décision hypothétiques incertaines pour les conducteurs et les piétons. Évidemment, d'autres facteurs psychologiques doivent être inclus dans notre approche. Dans tous les cas, l'approche est forte par sa simplicité et par sa capacité à intégrer à la fois le raisonnement abductif et l'incertitude, grâce à la théorie des ensembles flous intuitionnistes. Un système logiciel est en cours de développement à cette fin et réutilise des modèles de simulation de piétons développés dans nos travaux précédents. Les développements futurs de ce travail considéreront le problème de préférence multi-attributs et le processus de classement des nombres flous intuitionnistes.

Dans le même besoin d'étudier les interactions entre conducteurs des véhicules et piétons pour améliorer le processus de traversée de ces derniers et le rendre aussi sain qu'il devrait l'être. Nous avons développé un système de transport flou intelligent temps réel basé sur des variables temporelles floues gaussiennes pour la génération des messages d'alerte concernant la dangerosité de la conduite face aux piétons sur route et pour l'analyse du comportement des conducteurs vis-à-vis de la signalisation routière. Ce système est optimisé pour l'aide à la prévention et à la protection des usagers vulnérables de la route de type piéton. Le système permet la génération des alertes pour les profils de conduite dangereuse à l'approche des zones piétonnes, des

zones génératrices de piétons, des zones accidentogènes pour les piétons. Et ce par le biais du calcul d'indicateurs de risque fondé sur des variables de trafic temporelles floues gaussiennes et l'usage des marqueurs virtuels stockés dans une base de données spatiale temps réel. De plus le système proposé donne aussi la possibilité d'évaluer les comportements des usagers de la route en considération de la signalisation verticale, sur la base d'un calcul simple moyennant les marqueurs virtuels. La présente invention tire profit des technologies des bases de données spatiales temps réel et des technologies de géolocalisation et de l'environnement ubiquitaire. Cette invention sera d'un grand intérêt pour le processus d'amélioration de la sécurité des piétons et du respect de la réglementation routière.

4 Conclusion

Nous résumons dans ce papier différentes approches que nous avons développées pour mesurer l'exposition des piétons aux risques d'accidents. La première approche présente une mesure de l'indicateur de risque d'accidents mutuels des véhicules et piétons virtuels, en fonction de leurs flux et de leurs densités. La deuxième approche présente une nouvelle formulation d'un indicateur de risque d'accidents pour les piétons. Cet indicateur est basé sur le temps de traversée des piétons, le temps d'arrêt de sécurité pour les véhicules, la densité et la vitesse de ces derniers. Le temps de traversée des piétons et le temps d'arrêt de la sécurité des véhicules sont modélisés comme des nombres flous en raison de leur imprécision. Dans la troisième approche, nous avons aligné la théorie du flou intuitionniste aux objectifs de modélisation du risque d'accidents pour les piétons. Nous avons proposé de nouveaux indicateurs pour modéliser l'exposition des piétons aux accidents. Cette approche permet d'aborder la psychologie comportementale des piétons et des conducteurs avec des méthodes intuitives basées sur la théorie des ensembles flous. Les piétons en tant que véhicules sont considérés comme des objets mobiles se déplaçant selon un modèle comportemental donné. La dynamique des piétons est modélisée à l'aide du modèle de base de fourmis floues Boulmakoul et Mandar (2011) Mandar et Boulmakoul (2016), auquel nous avons intégré des champs de potentiels artificiels. La dynamique des véhicules est modélisée à l'aide du modèle de conducteur intelligent IDM pour les déplacements longitudinaux Treiber et al. (1999) et du modèle MOBIL pour le changement de voies Kesting et al. (2007). Cependant, la solution logicielle est en cours de développement et peut inclure d'autres modèles de mouvement pour ces entités. Les résultats de la simulation confirment les prédictions données par la théorie du flux de trafic de premier ordre. La validation du modèle de simulation par rapport aux données du monde réel est recommandée pour un complément d'étude.

Références

Atanassov, K. (1986). Intuitionistic fuzzy sets. *Fuzzy Sets and Systems*, 87–96.

Atanassov, K. (1999). Intuitionistic fuzzy sets : Theory and applications. *Physica-Verlag, Heidelberg.*

Boulmakoul, A. et M. Mandar (2011). Fuzzy pheromone potential fields for virtual pedestrian simulation. *The Open Operational Research Journal*, 19–29. DOI : 10.2174/1874243201105010019.

Kesting, A., M. Treiber, et D. Helbing (2007). Mobil – a general lane-changing model for car-following models. *Transportation Research Record : Journal of the Transportation Research Board*, 86–94.

Li, D. F. (2010). A ratio ranking method of triangular intuitionistic fuzzy numbers and its application to madm problems. *Computers and Mathematics with Applications, Elsevier 60*, 1557–1570.

Mandar, M. et A. Boulmakoul (2014). Virtual pedestrians' risk modeling. *International Journal of Civil Engineering and Technology 5*, 32–42.

Mandar, M. et A. Boulmakoul (2016). Fuzzy pheromone potential fields for virtual pedestrian simulation. *Adv. Fuzzy Systems*. 4027687:1-4027687:11.

Mandar, M., A. Boulmakoul, et A. Lbath (2017a). Pedestrian fuzzy risk exposure indicator. *Transportation Research Procedia 22*, 124 – 133.

Mandar, M., L. Karim, A. Boulmakoul, et A. Lbath (2017b). Triangular intuitionistic fuzzy number theory for driver-pedestrians interactions and risk exposure modeling. *Procedia Computer Science 109*, 148 – 155.

Treiber, M., A. Hennecke, et D. Helbing (1999). Congested traffic states in empirical observations and microscopic simulations. *Physical Review E 62*, 1805–1824.

Weidmann, U., U. Kirsch, et M. Schreckenberg (2012). *Pedestrian and Evacuation Dynamics 2012, Springer*.

Zadeh, L. (1975). *Information Sciences 8*, 199–249.

Summary

We are all in the situation of a pedestrian daily for at least a short period of time. Without protection, we expose ourselves to the risk of accidents. These should not be accepted as inevitable because they are the product of a global system that combines individual behavior, transport and environment under conditions that make the dysfunction both predictable and avoidable. Therefore, reducing or eliminating pedestrian risks is an important and achievable goal. We summarize in this paper different approaches that we have developed to measure the exposure of pedestrians to the risk of accidents. These approaches are based on fuzzy modeling on the one hand, and intuitionistic fuzzy logic on the other, which seems promising because it allows the behavioral psychology of pedestrians and drivers to be approached on the basis of fuzzy sets. Fuzzy intuitionistic theory has the flexibility to study hesitation and indecision. A software solution is developed for this instance by reusing pedestrian simulation models developed in our previous work.

Chapter

2

Intelligent Transportation Systems: Big Data, Machine Learning Cloud Computing

ASD'2018

Content

Proposition d'une méthode hybride pour la sélection des services Cloud

Ouassila Hioual*, Hakim Amamiche**
Sofiane Mounine Hemam***, Redha Zidane**

*Abbes Laghrour University of Khenchela / LIRE Laboratory of Constantine, Algeria
Ouassila.hioual@gmail.com
**Abbes Laghrour University of Khenchela, Algeria
{hakimamamiche, redhain}@gmail.com
***Abbes Laghrour University, Khenchela/ICOSI Laboratory of Khenchela, Algeria
sofiane.hemam@gmail.com

Résumé. Avec l'augmentation croissante de l'utilisation du Cloud Computing, de nombreux nouveaux besoins ont émergé, parmi lesquels : la nécessité d'avoir des systèmes de recherche et de sélection des services Cloud qui correspondent aux exigences des utilisateurs. La contribution de ce papier est de proposer une méthode basée sur la classification, le Pareto optimal et la méthode TOPSIS. Notre méthode (CloudOptimizer) permet aux utilisateurs de spécifier les exigences de qualité des services Cloud qu'ils souhaitent utiliser. CloudOptimizer est composée de trois étapes : dans la première étape, nous utilisons la classification, plus précisément la méthode K-means, afin de minimiser le nombre très volumineux des services Cloud. Dans la deuxième étape, nous appliquons l'algorithme de front de Pareto afin de sélectionner les classes non dominées. Et enfin, dans la troisième étape, on utilise les poids fournis par l'utilisateur afin de sélectionner la classe de services Cloud la plus adapté à ces exigences.

1 Introduction

Aujourd'hui, le Cloud Computing a gagné de plus en plus de popularité dans la communauté de recherche et le monde du commerce. Beaucoup d'utilisateurs finaux et d'entreprises utilisent des services Cloud pour sauvegarder leurs données ou pour obtenir plus de puissance informatique.

L'utilisation d'un service Cloud présente de nombreux avantages pour les utilisateurs finaux. Tout d'abord, elle permet une réduction significative des coûts, puisque les utilisateurs n'achètent que les ressources dont ils ont besoin, sans surplus ni besoin d'investir dans l'infrastructure ou la maintenance. Il y a aussi la garantie d'accès instantané et ininterrompu à l'informatique (le Computing) et aux ressources de stockage pour tout utilisateur qui a une machine connectée au réseau Internet. En plus, les utilisateurs peuvent facile-

ment adapter des ressources à leurs besoins spécifiques et peuvent ajouter des ressources à la demande.

Tous ces avantages ont conduit à une augmentation de l'utilisation du Cloud Computing. Avec cette augmentation, de nombreux nouveaux besoins ont émergé, parmi lesquels : la nécessité d'avoir des systèmes de recherche et de sélection des services Cloud qui correspondent aux exigences (besoins) des utilisateurs finaux. Notre contribution est dans cet axe de recherche et consiste à proposer une méthode hybride basée sur la classification, le Pareto optimal et la méthode TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) (Hioual et al., 2017). Notre méthode (CloudOptimizer) permet aux utilisateurs de spécifier les exigences de qualité des services Cloud qu'ils souhaitent utiliser.

Pour se faire, CloudOptimizer se connecte à une base de services Cloud et sélectionne ceux qui correspondent aux exigences des utilisateurs tout en leur donnant la possibilité d'obtenir la valeur optimale de certains de ces exigences, à savoir le coût et le temps de réponse.

CloudOptimizer est composée de trois étapes : dans la première étape, nous utilisons la classification, plus précisément la méthode K-means (Lloyd, 1982), afin de minimiser le nombre très volumineux des services Cloud sur le Net. Dans la deuxième étape, nous appliquons l'algorithme de front de Pareto (Hemam et Hioual, 2017) (Pareto Optimal) afin de sélectionner les classes non dominées. Et enfin, dans la troisième étapes, nous appliquons la technique de TOPSIS. Cette dernière utilise les poids fournis par l'utilisateur afin de sélectionner la classe des services Cloud la plus adapté à ces exigences.

Ce papier est organisé comme suit : dans la section 2, nous présentons quelques travaux traitant le problème de la sélection de services Cloud ; dans un deuxième temps, nous présentons la problématique et les objectifs visés par notre recherche ; puis nous introduisons l'architecture générale à base d'agents et un exemple introductif, qui sert de fil conducteur à la présentation de notre architecture et le méthode de sélection des services Cloud proposée, avant de détailler son fonctionnement. Nous terminons par une conclusion.

2 Travaux connexes

L'utilisation accrue de Cloud Computing a entraîné l'émergence de nouveaux besoins, tels que la nécessité d'avoir des systèmes de recherche et de sélection des services Cloud qui répondent aux exigences des utilisateurs.

De nombreux travaux ont été réalisés pour offrir de nouvelles solutions qui aideront les utilisateurs à choisir les services Cloud répondant à leurs besoins. Comme déjà mentionné ci-dessus, notre objectif principal est de trouver les services Cloud qui correspondent le mieux aux besoins des utilisateurs. Dans cette section, on va citer quelques travaux qui s'articulent autour du problème de la sélection des services Cloud. Ces travaux sont classés en deux grandes classes : ceux qui se basent sur la similarité et ceux qui se basent sur les méthodes MCDA (Muli-Criteria Decision Analysis).

Dans Zeng et al (2009), les auteurs ont présenté un algorithme de sélection de services Cloud. L'algorithme détermine le coût et les gains de la disponibilité des services Cloud pouvant être atteints par des Proxy et retourne comme résultat ceux qui maximisent les gains et minimisent le coût. Cet algorithme se déroule en deux étapes. Dans la première étape, le proxy sélectionne les services Cloud disponibles suite à la demande envoyée par l'utilisateur.

Dans la deuxième étape, l'algorithme calcule les gains et le coût des services Cloud Sélectionnés et renvoie ceux qui optimisent les deux critères Coût et Gains.

Kang et Sim ont présenté un portail Cloud avec un Moteur de recherche de service Cloud dans Kang et Sim (2010, a). Ce système utilise le concept de la similarité (Resnik, 1999) et consulte l'ontologie de Cloud adoptée pour sélectionner les services Cloud qui correspondent aux exigences spécifiées par l'utilisateur. Ces auteurs ont aussi proposé Cloudle Kang et Sim**.** (2010, b) qui est un moteur de recherche de services Cloud dont les fonctionnalités principales sont : le traitement des requêtes, le raisonnement de similarité et la notation. Comme le portail présenté dans Kang et Sim (2010, a), Cloudle consulte une ontologie Cloud pour calculer la similitude entre les services Cloud et retourne une liste de résultats triés par similarité agrégée.

Les auteurs dans (Yoo, et al, 2009), ont présenté dans un service de sélection de ressources basé sur l'ontologie Cloud. Cela génère des ontologies virtuelles (Vons) basé sur des ressources virtuelles et les combine en de nouvelles ressources. Ensuite, il calcule la similitude entre ces nouvelles ressources pour déterminer ceux qui répondent le mieux aux exigences et besoins de l'utilisateur.

Dans Zang et al. (2009), les auteurs ont présenté un algorithme de correspondance de service et un algorithme de composition de service. Ces algorithmes recherchent à travers les services Cloud et calcule la similarité sémantique (Resnik, 1999) entre eux pour déterminer si les deux services Cloud sont interopérables.

Comme on peut le constater, ces différents travaux se basent principalement sur la similarité pour déterminer lesquels des services répondent plus aux exigences d'un utilisateur. Ainsi, ils seraient mieux adaptés aux utilisateurs qui veulent trouver des services Cloud similaires à ceux qu'ils connaissent ou qu'ils l'ont déjà utilisé.

Il existe de nombreux travaux qui ont utilisé les méthodes MCDA pour traiter le problème de la sélection des services Cloud. L. Sun et al. conduit une étude approfondie des techniques de sélection de service Cloud dans (Sun et al., 2014), y compris les techniques basées sur les MCDA telles que la méthode AHP (Analytic Hierarchy Process) Saaty (1980), la méthode ANP (Analytic Network Process) (Saaty, 1996), MAUT (Multiple Attribute valUe Theory) (Churchman, 1957) et les méthodes de dépassement (outranking methods) comme par exemple la méthode ELECTRE (Roy, 1991).

Les chercheurs ont proposé dans (Godse et Mulik, 2009). une approche basée sur l'AHP pour la sélection des services Cloud de produits SaaS. Les critères utilisés sont la fonctionnalité, l'architecture, la convivialité, la réputation du fournisseur et le coût. Parmi les limites de cette méthode : elle n'a été testé qu'à l'aide de trois services Cloud seulement (produits SaaS). De plus, elle est utilisée uniquement pour comparer les services Cloud de type SaaS, en laissant de côté les autres catégories de services Cloud.

Les auteurs ont présenté dans (Rubayet et al., 2013) une approche de mapping basée Qualité de Services (QoS) pour combiner les produits SaaS et IaaS, puis classer les services Cloud combinés pour les utilisateurs finaux. Les auteurs ont utilisé la méthode AHP pour le classement des différents services Cloud.

A notre connaissance, ils n'existent pas encore de travaux qui combinent à la fois une méthode de classification, le front de Pareto et la méthode MCDA pour la recherche et la sélection de services Cloud. Notre motivation est de minimiser le temps de réponse de la requête utilisateur. Dans la section qui suit, nous présentons la problématique et les objectifs de notre travail.

3 Notre contribution

3.1 Problématique et objectifs

Dans un environnement multi Clouds, il est plus chanceux de trouver une séquence de services, qui composent un service composé, dans un cloud individuel. Ainsi, dans de nombreux cas, nous avons besoin de trouver des services qui seront déployés dans plusieurs Clouds dans le cas où on ne peut pas les avoir tous au niveau d'un seul cloud. Plusieurs approches ont été proposées pour résoudre le problème de sélection de service, y compris l'Analyse Multi-Critère d'Aide à la Décision (MCDA : Multi Criteria Decision Analysis). Le problème majeur qui se pose au moment de la sélection est le temps important nécessaire pour répondre aux besoins de l'utilisateur. Afin d'optimiser le temps de recherche et de sélection d'un service, il est nécessaire d'organiser, dans une première étape, les services cloud sous forme de classes en utilisant une des méthodes de classification pour avoir une présélection d'un sous ensemble de services. Puis, dans une deuxième étape, d'appliquer les algorithmes de front de Pareto et de TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) pour la sélection finale du ou des meilleurs services répondant aux exigences de l'utilisateur.

3.2 Aperçu global de la méthode proposée et techniques utilisées

Notre travail s'articule autour du problème de la sélection de services Cloud. Notre objectif consiste à proposer une méthode permettant de sélectionner le meilleur service qui répond au mieux aux exigences des utilisateurs, à savoir le coût et le temps de réponse du service demandé. Afin de répondre à cet objectif, la méthode proposée sera composée de trois étapes :

Dans la première étape, nous utilisons la classification afin de minimiser le nombre très volumineux des services Cloud sur le Net. En effet, le regroupement des services cloud qui se rapprochent en classes permet de minimiser encore plus le temps de réponse puisque au lieu de sélectionner un service Cloud parmi des milliers de services, nous sélectionnons la classe des services Cloud la plus appropriée aux besoins de l'utilisateur. Parmi les méthodes de classification qui existent nous utilisons k-moyennes (k-means) pour plusieurs raisons parmi lesquelles :

- C'est une méthode très utilisée
- Une méthode de regroupement rapide et très facile à comprendre et à mettre en œuvre.
- Elle fournit des classes à travers le nombre initial K clusters que nous le choisissons au début, et chaque classe (ou cluster) contient les services cloud similaires selon des critères bien définies à savoir le cout et le temps de réponse.
- Applicable à des données de grandes tailles telles que : le Text-Mining, le Web-Mining, la bio-informatique, etc.

Dans la deuxième étape, nous appliquons l'algorithme de front de Pareto (Hemam et Hioual, 2017) afin de sélectionner les classes non dominées. Les classes sélectionnées sont considéré comme étant les meilleures classes qui répondent aux exigences de l'utilisateur. Donc cette étape va minimiser l'espace de recherche puisque elle va sélectionner que quelques classes, ce qui permet d'optimiser encore plus le temps de réponse.

Dans la troisième étape, nous appliquons la technique de TOPSIS. Cette dernière utilise les poids fournis par l'utilisateur afin de sélectionner la classe des services cloud la plus adapté à ces exigences. Une fois la classe des services est choisie, nous appliquons une deuxième fois la technique de TOPSIS mais cette fois ci sur les services Cloud de la classe sélectionnée ce qui permet de choisir le service Cloud nécessaire.

3.2.1 Exemple de déroulement du front de Pareto dans notre contexte

Soient six centres de classes sous forme ***Nom_classe= (Coût, temps)*** : la sélection des classes optimum consiste à ***extraire les classes qui ont le coûts et le temps de réponse le plus petit (***Problème min-min***).***

Nous avons : a= (a1, a2), b= (b1, b 2), c= (c1, c2), d= (d1, d2), e= (e1, e2), f= (f1, f 2).

Soit la fonction objective F (x) [f1 (x), f2 (x)] La représentation graphique de l'image de ces six classes par la fonction F est fournie dans la figure 1.

Dans cet exemple on peut écrire :

f : est dominé par tous les autres classes.

b : est dominé uniquement par c.

d : est dominé par c est e.

a, c et e : ne sont dominés par aucune classe.

Par conséquent, les solutions (a, c, e) sont Pareto optimaux. Donc, le front de Pareto est l'ensemble de classe : (a, c, e).

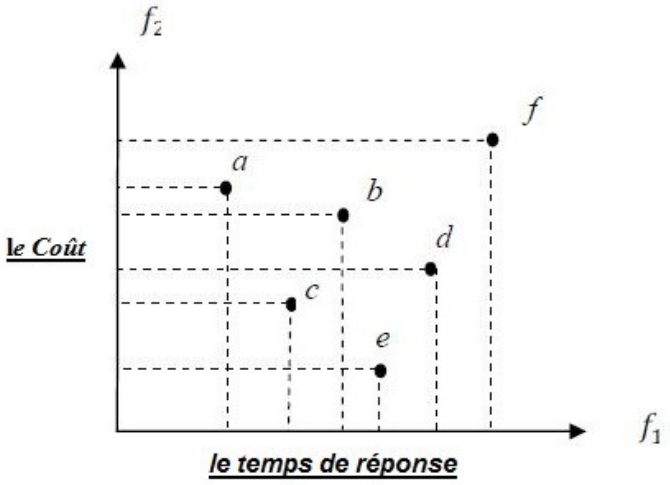


FIG. 1 – *Exemple de dominance et d'optimalité au sens de Pareto*

3.2.2 TOPSIS (Technique for Order Preference by Similarity to Ideal Solution)

TOPSIS est une méthode dont le but est de pouvoir classer par ordre de choix un certain nombre d'alternative sur la base d'un ensemble de critères favorables ou défavorables. Cette méthode s'inscrit dans les techniques utilisées dans le domaine d'aide à la décision multicritère (MCDM (Multiple Criteria Decision Making)). Elle a été développée par Hwang et Yoon en 1981 (Hwang and Yoon, 1981). Son principe consiste à déterminer pour chaque alternative un coefficient compris entre 0 et 1 sur la base des distances (euclidiennes) entre chaque alternative d'une part et les solutions idéales favorable et défavorable.

Une alternative est dite ***idéale favorable*** si elle est plus loin de la pire alternative et la plus proche de la meilleur alternative.

Une alternative est dite ***idéal défavorable*** si elle est la plus proche de la pire alternative et la plus loin de la meilleur alternative (Yezza, 2017).

Pour l'application de la méthode TOPSIS classique, les poids des critères sont connus avec précision. Toutefois, dans la pratique, une majorité des données n'est pas connue avec précision.(Majumdara et al., 2010) (Rubayet, et Karmake, 2016).

Le processus général de TOPSIS se compose de sept étapes et qui sont :

Etape 1. Choisir une échelle de mesure des valeurs des critères selon le tableau ci-dessous :

Valeur numérique	Valeur linguistique
1	Pas intéressent du tout
2	Pas intéressent
3	Très peu intéressent
4	Moyennement intéressent
5	Intéressent
6	Très intéressent
7	super intéressant
8	Parfaitement intéressent

TAB. 1 – *Choisir une échelle de mesure des valeurs des critères*

Etape 2. On applique la formule indiquée ci-dessous pour obtenir les nouvelles entrées r_{ij} de la matrice. Notons que : nous avons opté pour une normalisation euclidienne, car les métriques qui seront calculées dans la suite sont basées sur la distance euclidienne ce qui garantit des résultats cohérents (voir. TAB. 2 et TAB. 3). Donc, la formule à appliquer est :

$$r_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} \sqrt{x_{ij}^{2}}} \qquad (1)$$

Etape 03. Construire la matrice de décision pondérée normalisée par multiplication de la matrice de décision normalisée par son Poids associés. La v_{ij} valeur normalisée (Voir. TAB. 4) pondérée est calculée comme suit : $v_{ij}=w_{ij} X r_{ij}$

	le coût (Poids 0.75)	**Le temps** (Poids 0.25)
Classe 1	6	8
Classe 2	8	5
Classe 3	3	1
Classe 4	1	4
Classe 5	2	3
La formule	$\sqrt{(6^2 + 8^2 + 3^2 + 1^2 + 2^2)}$	$\sqrt{(8^2 + 5^2 + 1^2 + 4^2 + 3^2)}$

TAB. 2 – *Normaliser tous les scores de la matrice des niveaux attribués aux critères*

	le coût (Poids 0.75)	**Le temps** (Poids 0.25)
Classe 1	0.561	0.746
Classe 2	0.749	0.466
Classe 3	0.280	0.093
Classe 4	0.093	0.373
Classe 5	0.187	0.279

TAB. 3 – *Normaliser tous les scores de la matrice finale*

	le coût	**Le temps**
wij	**0.75**	**0.25**
Classe 1	0.420	0.186
Classe 2	0.561	0.116
Classe 3	0.210	0.023
Classe 4	0.069	0.093
Classe 5	0.140	0.069

TAB. 4 – *La matrice de décision pondérée normalisée*

Etape 04. Déterminer les solutions idéales positive et négative (Voir. TAB. 5) selon les formules suivantes :

$$A^* = \{(\max v_{ij}/j \in J), (\min v_{ij}/j \in J')\ for\ i = 1,2,3 \dots M\} = \{v_1^* + v_2^* + \cdots + v_N^*\}$$

$$A^- = \{(\min v_{ij}/j \in J), (\max v_{ij}/j \in J')\ for\ i = 1,2,3 \dots M\} = \{v_1^- + v_2^- + \cdots + v_N^-\}$$

Tel que : *J*= {*j*=1, 2,3…, *N*/*j* associé à profit de critères positifs}, et *J'*= { *j*=1, 2,3…, *N*/*j* associé à profit de critères négatifs}

	le coût	Le temps
w_{ij}	**0.75**	**0.25**
Classe 1	0.420	0.186
Classe 2	0.561	0.116
Classe 3	0.210	0.023
Classe 4	0.069	0.093
Classe 5	0.140	0.069
A*	0.561	0.186
A-	0.069	0.023

TAB. 5 – *La solution idéale positive A* et solution idéale négative A-*

Etape 05. Calculer la mesure de séparation (cf. TAB 6). La séparation de chaque variante de l'idéal positive est donnée par :

$$S_i^* = \left\{\sum_{j=1}^{N}(v_{ij} - v_j^*)^2\right\}^{(\frac{1}{2})}, i = 1,2,\dots,M$$

	le coût	Le temps	S*
Classe 1	(0.420-0.561)² = 0.019	(0.186-0.186)² = 0	0.137
Classe 2	(0.561-0.561)² = 0	(0.116-0.186)² = 0.004	0.063
Classe 3	(0.210-0.561)² =0.123	(0.023-0.186)² =0.026	0.386
Classe 4	(0.069-0.561)² =0.242	(0.093-0.186)² =0.008	0.500
Classe 5	(0.140-0.561)² = 0.242	(0.069-0.186)² =0.013	0.504
A*	0.561	0.186	
A-	0.069	0.023	

TAB. 6– *La solution idéale positive s**

Etape 6. De même, la séparation de chaque variante de l'idéal négative (cf. TAB. 7) est donnée par la formule suivante :

$$S_i^- = \left\{\sum_{j=1}^{N}(v_{ij} - v_j^-)^2\right\}^{(\frac{1}{2})}, i = 1,2,\dots,M$$

	le coût	Le temps	S^-
Classe 1	(0.420-0.069)² = 0.123	(0.186-0.023)² = 0.026	0.386
Classe 2	(0.561-0.069)² = 0.242	(0.116-0.023)² = 0.008	0.5
Classe 3	(0.210-0.069)² =0.019	(0.023-0.023)² =0	0.137
Classe 4	(0.069-0.069)² =0	(0.093-0.023)² =0.004	0.063
Classe 5	(0.140-0.069)² = 0.005	(0.069-0.023)² =0.002	0.083
A*	0.561	0.186	
A-	0.069	0.023	

TAB. 7– *La solution idéale négative* s^-

Etape 07. Calculer la proximité par rapport à la solution idéale. La proximité relative par rapport à A* est défini comme suit:

$$C^* = \frac{S_i^-}{(S_i^* + S_i^-)}, 0 < c^* < 1$$

	S^-	S*	C*	L'ordre de choix
Classe 1	0.386	0.137	0.738	2
Classe 2	0.5	0.063	0.888	1
Classe 3	0.137	0.386	0.261	3
Classe 4	0.063	0.5	0.111	4
Classe 5	0.083	0.504	0.141	5

TAB. 8– *Ordre des classes*

3.3 Fonctionnement de la méthode CloudOptimizer

Rappelons que notre travail se situe dans une intersection de trois domaines à savoir, la sélection des services Cloud, la classification et l'optimisation de critères.

Nous allons, tout d'abord, modéliser notre méthode à travers des diagrammes UML. Ensuite, nous présentons l'algorithme de fonctionnement de la méthode CloudOptimizer.

3.3.1 Modélisation avec UML

UML (Unified Modeling Language) est un langage de modélisation graphique et textuel permettant de décrire et comprendre des besoins, spécifier, concevoir des solutions possibles et communiquer des points de vue (Roques, 2008).

UML unifie à la fois les notations et les concepts orientés objet. Il ne s'agit pas d'une simple notation, mais les concepts transmis par un diagramme ont une sémantique précise et sont porteurs de sens au même titre que les mots d'un langage. UML permet de modéliser de manière claire et précise la structure et le comportement d'un système indépendamment de toute méthode où de tout langage de programmation (Roques et Vallée, 2004).

La figure 2 représente le diagramme de séquence du scénario de la sélection de services Cloud, et la figure 3 représente le diagramme d'activité de la méthode CloudOptimizer :

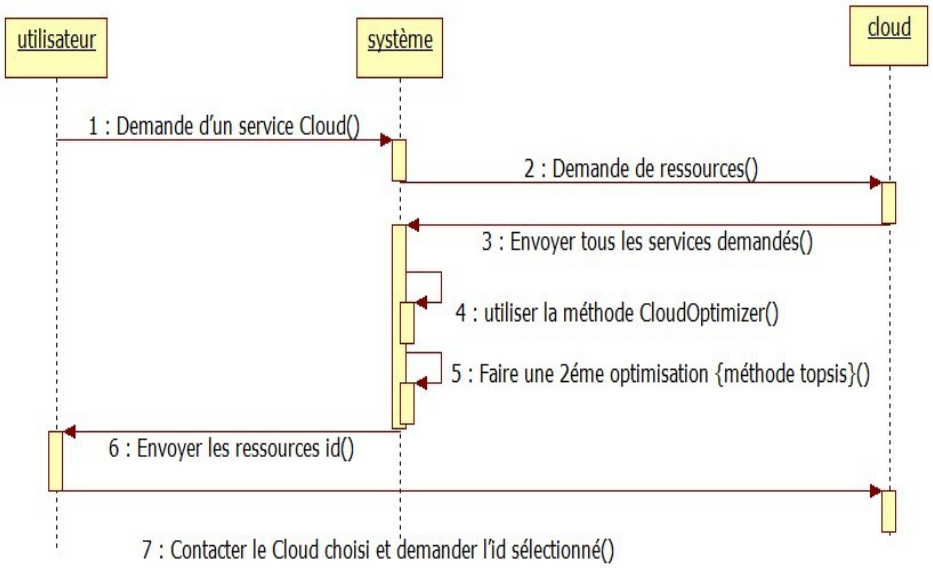


FIG. 2 – *Diagramme de séquence du scénario « Sélection de services Cloud »*

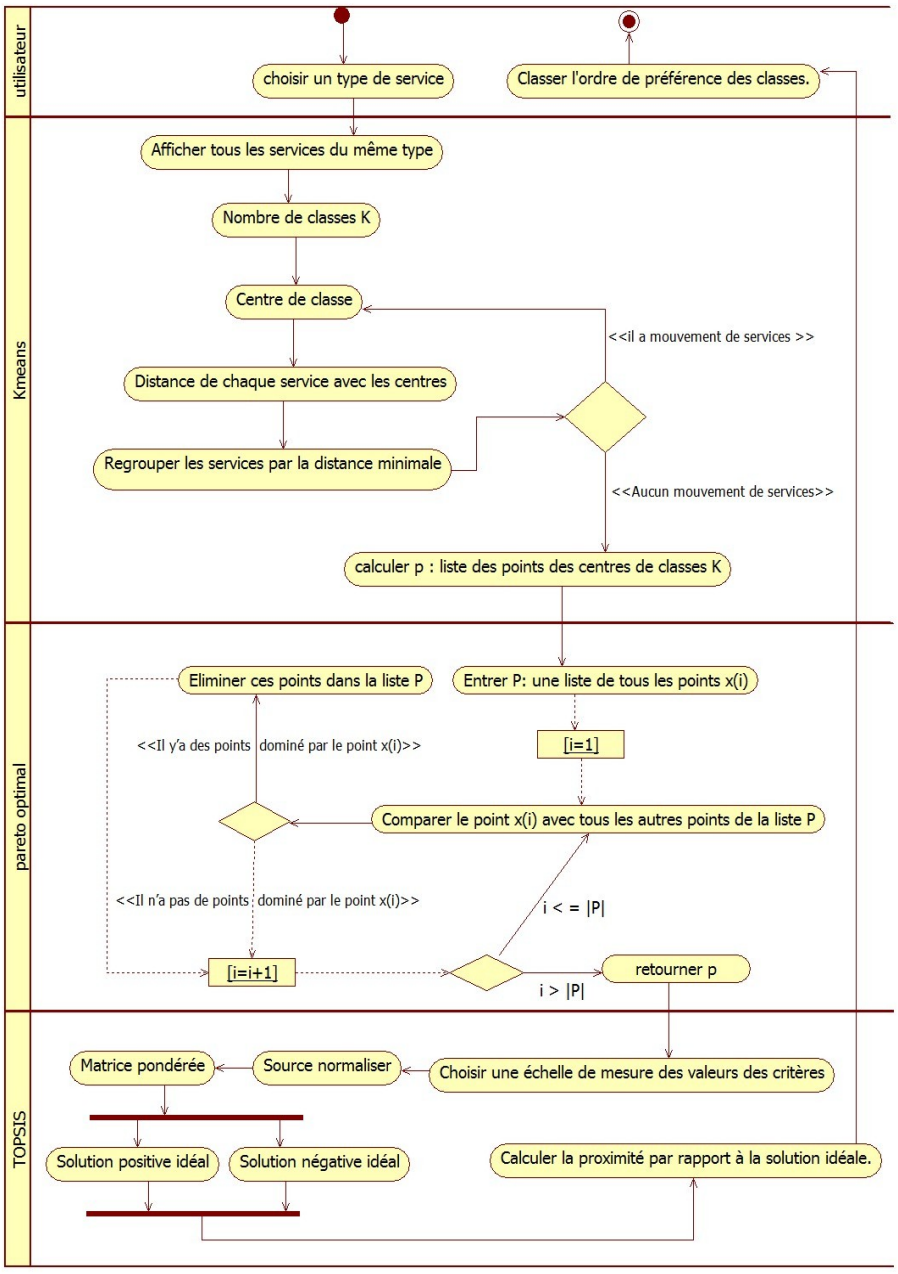


FIG. 3 – *Diagramme d'activité de la méthode CloudOptimizer*

4 Conclusion

La méthode proposée est composée de trois étapes : dans un premier lieu, nous avons utilisé la classification afin de minimiser le nombre très volumineux des services Cloud sur le Net. En effet, le regroupement des services Cloud qui se rapprochent en classes permet de minimiser encore plus le temps de réponse puisque au lieu de sélectionner un services Cloud parmi les milliers de services, nous sélectionnons la classe des services Cloud la plus appropriée aux besoins de l'utilisateur.

Dans un deuxième, nous avons appliqué l'algorithme du front de Pareto afin de sélectionner les classes non dominées. Les classes sélectionnées sont considéré comme étant les meilleures classes qui répondent aux exigences de l'utilisateur. Donc, cette étape a minimisé l'espace de recherche puisque elle a permis de sélectionner que quelques classes, ce qui a permis d'optimiser encore plus le temps de réponse.

Et, dans un troisième lieu, nous avons appliqué la technique de TOPSIS. Cette dernière utilise les poids fournis par l'utilisateur afin de sélectionner la classe des services Cloud la plus adapté à ces exigences. Une fois la classe des services est choisie, nous avons appliqué une deuxième fois la technique de TOPSIS mais cette fois ci sur les services Cloud de la classe sélectionnée ce qui a permet de choisir le service Cloud nécessaire.

Dans notre travail, nous nous sommes basé sur deux exigences de l'utilisateur, à savoir l'optimisation du coût et temps de réponse. Ainsi, ce travail peut être amélioré en prenant en considération d'autres critères (exigences) de l'utilisateur d'un coté et l'amélioration des performances du système d'un autre côté.

Références

Churchman, C. W., R. L. Ackoff and E.L. Arnoff (1957). Introduction to Operations Research. *New York: Wiley.*

Godse M. and S. Mulik (2009). An approach for selecting software-as-a-service (SaaS) product. In: the IEEE international conference on Cloud Computing (CLOUD), Bangalore.

Hemam, S.M., and O., Hioual (2017). A Hybrid Load Balancing Algorithm for P2P-Cloud System Aware of Constraints Optimization of Cost and Reliability Criteria. *International Journal of Internet Protocol Technology*: 10(2), 99-114.

Hioual, O. Z., Boufaida and S.M., Hemam (2017). Load balancing, Cost and Response Time Minimization Issues in Agent Based Multi Cloud Service Composition. *International Journal of Internet Protocol Technology*: 10(2),73-88.

Hwang, C.L. and K., Yoon (1981). *Multiple Attribute Decision Making: Methods and Applications*. Berlin, Springer-Verlag.

Kang, J. and K. M. Sim (2010, a). *Cloudle: A Multi-criteria Cloud Service Search Engine*. In: IEEE Asia-Pacific Services Computing Conference, Hangzhou, China.

Kang, J. and K. M. Sim (2010, b). Cloudle : An Agent-based Cloud Search Engine that consults a Cloud Ontology. In: Annual International Conference on Cloud Computing and Virtualization, Singapore Management University, Singapore.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Trans. Inform. Theory*: 28,129–137. Originally as an unpublished Bell laboratories Technical Note (1957).

Majumdara, A., R., Mangla and A., Gupta (2010). Developing a decision support system software for cotton fibre grading and selection. *Indian Journal of Fibre and Textile Technology*: 35(3), 195-200.

Resnik, P. (1999). Semantic similarity in a taxonomy: an information-based measure and its application to problem of ambiguity in natural language. *Journal of Artificial Intelligence Research* : 11, 95-130.

Roques, P. and F., Vallée (2004). *UML 2 en action : de l'analyse des besoins à la conception Java*. Eyrolles.

Roques, P. (2008). *UML2 : Modéliser une application Web*. Eyrolles.

Roy, B. (1991). The outranking approach and the foundation of the ELECTRE methods. *Theory and decision*: 31 (1), 49-73.

Rubayet, K., C. Ding and A. Miri (2013). An end-to-end QoS mapping approach for Cloudservice selection. In: The IEEE 9th world congress on services (SER-VICES), Santa Clara Marriott, CA.

Rubayet, K, and C. L. Karmake (2016). Machine Selection by AHP and TOPSIS Methods. *American Journal of Industrial Engineering* : 4(1), 7-13.

Saaty, T.L. (1980). The Analytic Hierarchy Process for Decision in a Complex World. *Pittsburgh, PA: RWS Publications*.

Saaty, T. L. (1996). Decisions with the analytic network process (ANP). *University of Pitts urgh (USA), ISAHP 96*.

Sun, L., H. Dong, F. K. Hussain, O. K. Hussain, and E. Chang (2014). Cloud service selection: State-of-the-art and future research directions. *Journal of Network and Computer Applications* : 45, 134-150.

Yezza, A. (2017). La méthode TOPSIS expliqué pas à pas. Rapport proposé Mais 2015, mis à jour Avril 2017.

Yoo, H., C. Hur, S. Kim, and Y. Kim (2009). An Ontology-based Resource Selection Service on Science Cloud. *Communications in Computer and Information Science*: 63, 221-228.

Zang, C., X. Guo, W. Ou and D. Han (2009). Cloud Computing Service Composition and Search Based on Semantic. *Cloud Computing*: 5931, 290-300.

Zeng, W., Y. Zhao and J. Zeng (2009). Cloud Service and Service Selection Algorithm Research. In: The first ACM/SIGEVO Summit on Genetic and Evolutionary Computation, Shanghai, China.

Proposition d'une méthode hybride pour la sélection des services Cloud

Summary

With the increasing use of cloud computing, many new needs have emerged, including: the need for cloud-based search and selection systems that meet end-user requirements (needs). The contribution of this paper is to propose a method based on the classification, the Pareto optimal and the TOPSIS method. Our method (CloudOptimizer) allows users to specify the quality of requirements of the cloud services they want to use. CloudOptimizer consists of three steps: in the first step, we use the classification, more precisely the K-means method, to minimize the very large number of cloud services on the Net. In the second step, we apply the Pareto (Pareto Optimal) algorithm to select the non-dominated classes. And finally, in the third step, we use the weights provided by the user to select the most appropriate cloud service class for these requirements.

Computing Shortest Paths in Large Scale Multimodal Graphs

Mariyem Oukarfi*, Abdelfettah Idri**
Azedine Boulmakoul***

* LIM Lab. IOS, Computer Sciences Department, Faculty of Sciences and Technology Mohammedia, Mohammedia, Morocco
oukarfi.mariyem@gmail.com
**National School of business and Management, Casablanca, Morocco
abdelfattah.id@gmail.com
*** LIM Lab. IOS, Computer Sciences Department, Faculty of Sciences and Technology Mohammedia, Mohammedia, Morocco
azedine.boulmakoul@gmail.com

Abstract. Intelligent Transport Systems (ITS) contribute to the control of mobility in all its forms and constraints and recently has generated a strong interest in multimodality. Multimodal trip planner (MTP) is one of the products of ITS that helps to plan a dynamic day trip in a complex network where the traveler can commute through private and public modes, while taking into account the variability of travel times and transfer times. A relevant part of creating an MTP is preparing the large scale multimodal network. For this purpose, the article focuses on techniques of abstracting a multimodal network passing through a first step of preparing the geographic data, then connecting the network layers in transit stations and finally creating the multimodal network and displaying it in geographic information system (GIS). After the construction of the multimodal network it comes the routing part where we implement a shortest path algorithm based on an optimization approach presented in a previous work.

1 Introduction

Route planning in highly developed transportation networks is becoming increasingly important in light of changing transportation and the emergence of advanced intelligent transportation systems. Nowadays, the mobility of goods and people becomes a major challenge, especially in critical areas such as the transport of dangerous goods.

Travelers are demanding efficient routing methods to reach their destinations via large-scale networks involving different modes of scheduled and unplanned transport. For this reason, many algorithm applications that compute an optimal route from a source to a destination in a multimodal network have been proposed to address this combinatorial optimization problem under additional constraints such as changing dynamic displacement data.

As a result, the shortest single-source tracking algorithms of a single source in the dynamic multimodal transport network have seen more and more improvements by adopting different optimization approaches. On the one hand, with regard to the issue of time dependence, multiple searches have been carried out. In addition to conventional solutions for routing in static networks such as acceleration techniques (Bastet al 2014), Cooke and Halsey (1966) have introduced the dynamic aspect to cope with the temporal dependence of modes of public transport. Pyrga et al. (2008) detailed the realistic version of the model dedicated to time and expanded over time for the public transport network. Bakalov et al. (2015) described the time-dependent network model adapted to the need for transport network modeling. On the other hand, we can find relevant work in case of temporal dependency problem. Schultes et al. (2008) and Pajor et al. (2009) have done extensive research to extend single-mode to multimodal networks. Liu et al. (2009) proposed a switching point approach for model multimodal transport networks. Peng et al. (2008) proposed a distributed solution for travel planning in a larger transportation system. Ayed et al. (2008) proposed a graphical transfer approach for multimodal transport problems. Lozano et al. (2002) adopted the concept of hypergraphs as Bielli et al. (2006) worked on a hierarchical graph. Zhang et al. (2014) investigates the multimodal network design problem that optimizes the automatic network expansion scheme and the bus network design scheme. Ziliaskopoulos and Wardell (2000) presented an optimal intermodal time travel algorithm for multimodal transport networks that explains the delays in mode and arc change points.

2 Multimodal network model

2.1 OSM and GTFS data

A multimodal network used on a trip planner should takes into account the dynamic transportation data as public transport schedules, delays, stop times and the transit data as the transit stations, point of interest (POI). In order to prepare these two kinds of data, we chose to use Open Street Map (OSM) and the General Transit Feed Specification (GTFS).

OSM is a collaborative project to create a free editable map of the world. It provides data into a special format (nodes, ways, relations and tags) that can be imported as a shapefile.

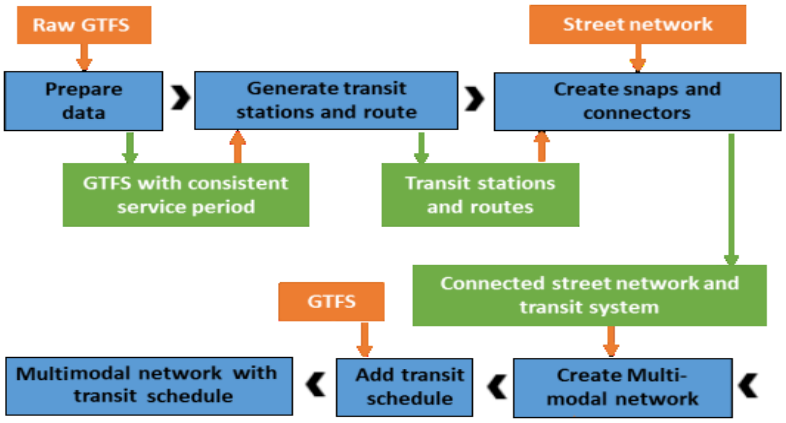


FIG. 1 - *Preparing network data*

The GTFS has become the most popular world-wide data format to describe fixed-route transit services. Many transit agencies have created and published GTFS data with the primary purpose being integration with Google Maps. However, GTFS data can power many other different types of transit and multimodal software applications, including multimodal trip planning, timetable creation, mobile apps, visualization, accessibility, analysis tools for planning, real-time information, and interactive voice response.

The different steps for integrating the network data are explained as follow (fig.1) :

1. Collect, clear-up, and create GTFS and street networks. GTFS come from different sources; during this step, it is important to ensure that all GTFS share the same service period.
2. Generate transit routes and stations. The latitude/longitude information of transit stations in GTFS is read in QGIS (Quantum Geographic Information System) and point shapefiles are created to store the spatial information and other fields. Straight lines are generated to connect two adjacent stations; lines are converted to line shapefiles as the transit routes.
3. Create connectors between transit stations to street networks. Due to the different sources, transit stations are not always mapped to the street network. Therefore, connectors, short straight lines which are perpendicular to streets, are generated to connect transit system and street network. This step is very vital. By creating snaps and connectors, layers are connected and only connected at stops, which prevents walking along transit lines.
4. Create a multimodal transit network. With “creating a multimodal network dataset” toolkit provided in ArcGIS Network Analyst, a multimodal transit network is created.
5. Add GTFS transit schedule to network. Convert GTFS to QGIS recognizable transit schedule table, and attach the table to the multimodal network.

The data model of the dynamic transit is designed as belowin figure 2:

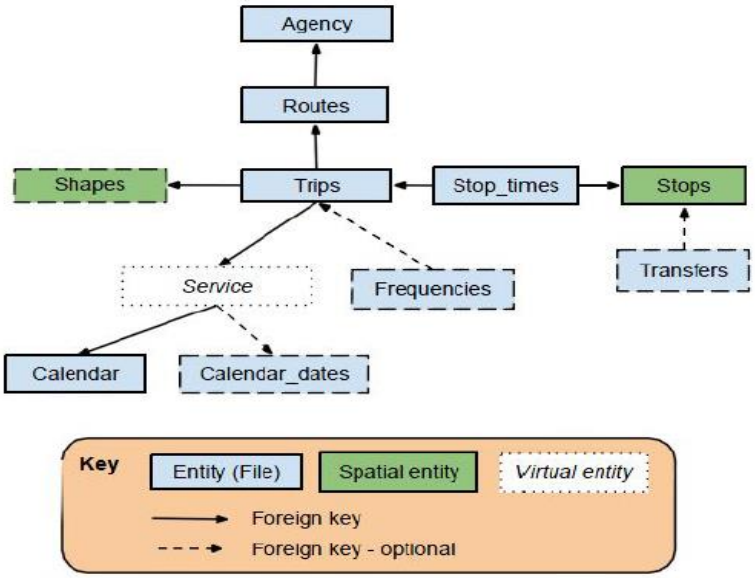


FIG. 2 - *Data model for transit*

2.2 Graph connection

The main graph, the support graph, is the road graph. The vertices of the other graphs (public transport, POI) are paired with the road graph at its arcs. It is possible to architect the final network so that road graphs and transport graphs can be used both in isolation and when they form together a single multimodal graph.

The multimodal graph is based on the data of its road graph and its transport graphs to make a graph as such.

A node of a multimodal graph is either a road vertex or the node of a given transport graph or a POI, an arc of a multimodal graph is a pair of multimodal vertices (fig.3).

The adjacency of a multimodal graph is thus defined as follows:

The adjacent arcs of a multimodal vertex v are:

If v is a road vertex:

- the road arcs adjacent to v ;
- the arcs connecting v to tv (tv is a transport vertex accessible by one of the road arcs adjacent to v) ;
- the arcs connecting v to pv (pv is a POI accessible by one of the road arcs adjacent to v) ;

If v is a transport vertex:

- the transport arcs adjacent to v for the same network ;
- the two arcs linking v to sv and tv (sv and tv are the road vertices of the road arc where the vertex v is paired) ;

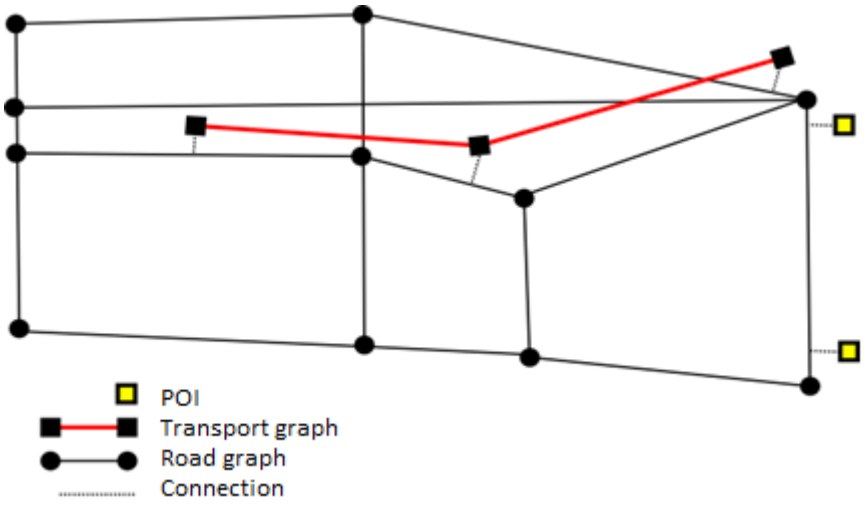


FIG. 3 - *Multimodal graph*

Below are the table responsible for linking the POIs and road arcs with dynamic data.

#	Nom	Type	Longueur	Null	Défaut
1	id_	int4	4	N	
2	poi_type	int4	4	Y	
3	name	varchar		Y	
4	parking_transport_modes	_int4		N	
5	road_section_id	int8	8	Y	
6	abscissa_road_section	float8	8	N	
7	geom	geometry (PointZ,2154)		Y	

FIG. 4 - *POI table*

#	Nom	Type	Longueur	Null	Défaut
1	id_	int4	4	N	nextval('pt_stop_time_id_seq'::regclass)
2	trip_id	int8	8	N	
3	arrival_time	time	8	N	
4	departure_time	time	8	N	
5	stop_id	int4	4	N	
6	stop_sequence	int4	4	N	
7	stop_headsign	varchar		Y	
8	pickup_type	int4	4	Y	0
9	drop_off_type	int4	4	Y	0
10	shape_dist_traveled	float8	8	Y	

FIG. 5 - *Trip table*

3 Routing algorithm

This section formally describes our approach by specifying the algorithm and its input parametres (Idri et al., 2017).

The proposed algorithm is a target-oriented algorithm based on the concept of "proximity" to the target node as a heuristic to drive the search towards its destination. It is mainly

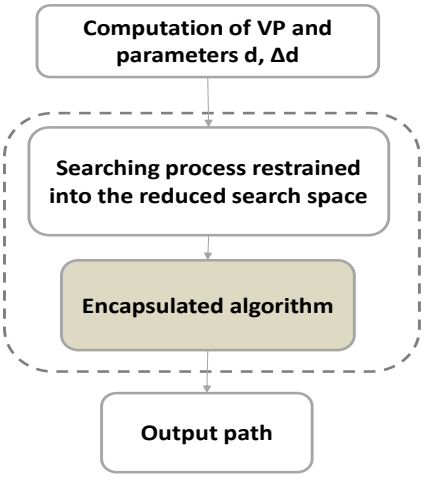


FIG. 6 - *Algorithm process*

based on calculating a virtual path that is the Euclidean distance from the source node to the target node, then conducting the search to the nodes connected near this virtual path by a parameter d (see fig.6). The idea of the proposed algorithm comes from a logical and obvious procedure of finding the shortest cut towards our destination, which obviously stays as close as possible to the destination and avoids the detours that could take us away from it. . Since the known algorithms A* (A star) and ALT (A star Landmark and Triangular inequality) use the concept of lower bounds in the triangular inequality to prioritize the nodes that are supposed to be closer to the target, our algorithm uses the virtual path as the driver of the search space and the parameter d and Δd. This virtual path is considered as an ideal path from the source node s to the target node t and used as a reference path to carry out the search in a restricted space. The algorithm can navigate backwards in the graph as it can move forward, to ensure that it explores all possible paths until the shortest path is constructed.

The approach is applicable for any shortest path algorithm as well as it can work on itself without the integration of another algorithm.

4 System overview

The trip planner is responsible for searching the Real time transit network and generates the desired travel plan based on the user's preferences. Plan search algorithm searches reachable nodes from the source node. Then, all links will be recorded in an order. After that, it will visit each of these nodes one by one. Again from these nodes, it will look for all possible reachable nodes and will record them in order. This process will be continued till the destination node is found. In case, real time information is not available due to technical reason then dynamic network is used to search the travel plans.

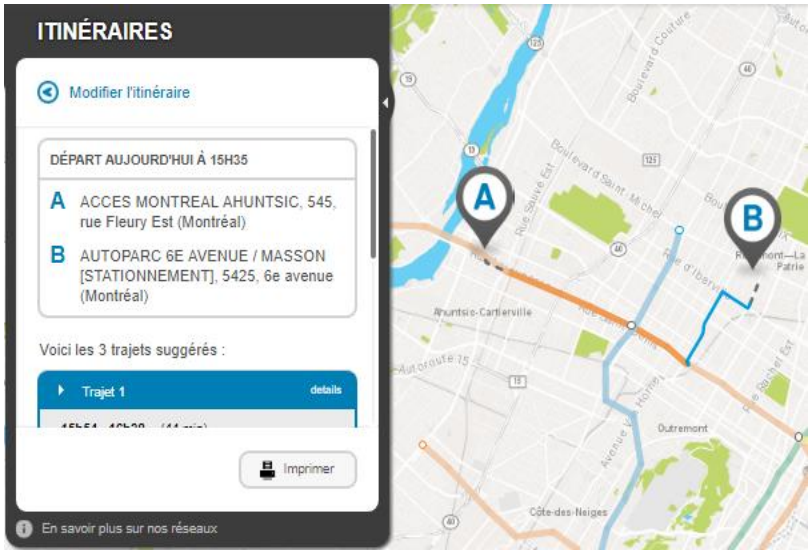


FIG.7 - *Application interface*

A web interface (fig.7) has been implemented to access real-time transit trip planner system. Travelers need to give input parameters such as: source stop, destination stop of journey, departure time from source stop or arrival time at destination stop, transportation mode (can be only buses or buses and trains). Suggested travel plans include the instructions to reach at destination stop from source stop along with transit service details, total covered distance and expected fare of trip. Route for each travel plan along with intermediate waypoints is displayed on map using Google map Application Programming Interface (API).

5 Conclusion

In this paper we treated some design and implementation aspects regarding our approach dealing with the time-dependant multimodal transport problem expressed as finding the shortest path algorithm. Especially, we focused on the design techniques to solve the problem in the context of the different search dimensions including the user defined constraint

that influences the search process and the final results. We intend in our future work to investigate deeply the behaviour of existing algorithms when embedding this approach within these algorithms in a parallel distributed architecture.

Références

Ayed, H., & Khadraoui, D., 2008. "Transfer graph approach for multimodal transport problems". Second International Conference MCO, Metz, France – Luxembourg.

Bakalov, P., Hoel, E., & Heng, W. L. 2015. Time dependent transportation network models. In Data Engineering (ICDE), 2015 IEEE 31st International Conference on (pp. 1364-1375). IEEE.

Bast, H. et al., 2009. "Route planning in transportation networks." Technical Report MSR-TR-2014-4. Microsoft Research, Microsoft Co.

Bielli, M., Boulmakoul, A., and Mouncif, H., 2006. "Object modeling and path computation for multimodal travel systems". European Journal of Operational Research pp. 175, 1705–1730.

Cooke, K. & Halsey, E., 1966. "The Shortest Route through a Network with Time-Dependent Intermodal Transit Times". Journal of Mathematical Analysis and Applications, (14):493{498.

Idri, A., Oukarfi, M., Boulmakoul, A., Zeitouni, K. & Masri, A., 2017. "A new time-dependent shortest path algorithm for multimodal transportation network". Procedia Computer Science. volume 109, 2017, Pages 692–697.

Liu, L. & Meng, L., 2009. "Algorithms of multi-modal route planning based on the concept of switch point". 6th International Symposium on LBS & TeleCartography, Nottingham, UK.

Lozano, A. & Storchi, G., 2002. "Shortest viable hyperpath in multimodal networks". Transportation Research Part B, 36, 853–874.

Pajor, T., 2009. "Multi-modal route planning". Master thesis of Karlsruhe Institute of Technology.

Peng, Z. R. & Kim, E., 2008. "A standard-based integration framework of distributed transit trip planning systems". Journal of the Intelligent Transportation Systems, 12(1), pp. 13-19.

Pyrga, E., Schulz, F., Wagner, D., & Zaroliagis, C. 2008. "Efficient models for timetable information in public transportation systems". Journal of Experimental Algorithmics (JEA), 12, 2-4.

Schultes, D., 2008. "Route planning in road networks". Ph.D Thesis of Karlsruhe Institute of Technology.

Zhang, L., Yang, H., Wu, D., & Wang, D. 2014. "Solving a discrete multimodal transportation network design problem". Transportation Research Part C: Emerging Technologies, 49, 73-86.

Ziliaskopoulos, A., & Wardell, W. 2000. “An intermodal optimum path algorithm for multimodal networks with dynamic arc travel times and switching delays”. European Journal of Operational Research, 125(3), 486-502.

Résumé

Les systèmes de transport intelligents (STI) contribuent au contrôle de la mobilité sous toutes ses formes et contraintes et ont récemment suscité un fort intérêt pour la multimodalité. Le planificateur de voyage multimodal est l'un des produits de STI qui aide à planifier une excursion d'une journée dans un réseau complexe où le voyageur peut se déplacer en mode privé et public tout en tenant compte de la variabilité des temps du voyage et de transfert. Une partie non pertinente de la création de ce logiciel est de préparer le réseau multimodal à grande échelle. A cet effet, l'article s'intéresse aux techniques d'abstraction d'un réseau multimodal passant par une première étape de préparation des données géographiques, puis la liaison des couches réseaux dans les stations de transit et enfin la création du réseau multimodal et son affichage dans le système d'information géographique. Après la construction du réseau multimodal vient la partie routage où nous implémentons un algorithme de chemin le plus court basé sur une approche d'optimisation présentée dans un travail précédent.

Optimization of a controlled trajectory using artificial neural networks for a mobile robot

Meryem Khouil*, Mohammed Mestari**

*meryem.khouil@gmail.com
**mestari@enset-media.ac.ma

Abstract. We apply in this paper a new optimization method for planning a controlled trajectory in an environment with dynamic obstacles. The algorithm behind this method has an ability to optimize, control and plan a free collision path for a four wheel mobile engine. The Artificial Neural Networks techniques are used here to overcome the time and computation complexity using its parallelism and structured architecture.

1 Introduction

We focus in this paper on the navigation of autonomous agents such as mobile robots in virtual three-dimensional environment with the presence of dynamic obstacles. Path planning is an acute function for these agents since it will automatically affect their movement's autonomy and their ability to act in these worlds. The target here is about elaborating an optimal and non colliding path. The computational complexity Stentz (1994) is one of the main reason slowing down the advancement of robotic systems. This complexity comes from the high number of potential cases that may face the robot. The problem is even more difficult when the movements are important and especially when there is a risk of collisions. Optimization problems considering several criteria in conflicting situations Eichfelder (2008) are indeed the subject of different fields such as robotics and engineering.

The movement of the mobile robot of four wheels here studied can be modeled by a kinematic model. The latter is formulated as a nonlinear constrained multicriterion optimization problem (NECMOP), where the constraints and the objective functions are nonlinear functions assembling the variables considered.The classic NECMOP's application are known of being complex to solve due to the complicated mathematical tools chosen and also the computation's cost. In the scientific literature, there are many new approaches proposing to solve NECMOP's applications Aggelogiannaki and Sarimveis (2007); Zitzler et al. (2000); Agrawal et al. (2008); taking advantage of genetic algorithms. The multiobjective genetic algorithm hardly approach to optimal pareto front Abraham and Jain (2005); Ahn (2006) when the problem's constraints becomes more ponderous or when the objective space is non convex. However there is still some interesting researches for general algorithms discussing the pareto optimal solutions in multiobjective optimization problems which is computationaly fast and simultaneously capable of finding a well converged and well distributed set of solutions

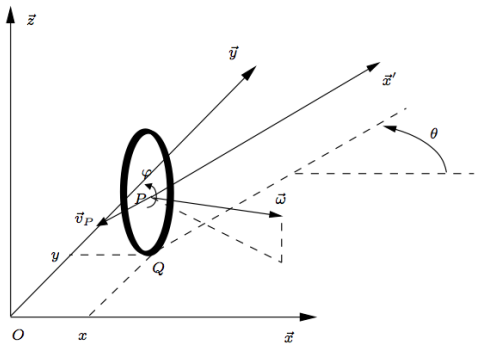


FIG. 1 – *Characterization of rolling without slipping*

Deb et al. (2005).

Our main objective is to reach an optimal and effective set of control inputs that will help the robot attain at time N the desired path avoiding the dynamic obstacles. In this case, we take use of decomposition coordination principle so that the nonlinearity can be treated at a local level and the coordination concluded via lagrange multipliers Mestari et al. (2001).
The scientific input in this project is the implementation of a different way of NECMOP's resolution with a view to plan an optimal trajectory avoiding dynamic obstacles. The remainder will be arranged along these lines: In the second fraction 2, we discuss the modelization of robot's motion. Then in 3, we resolve the NECMOP related to robot's kinematic model. In 4, we experiment the use of artificial neural networks applied to the case of planning an optimal and collision free path for a vehicle of four wheels.

2 Modelization of the Robot's Motion

2.1 Analysis of Constraints

We consider that the robot is a rigid vehicule moving on a plane surface, also we suppose that the wheel is respecting the constraint of rolling without sliding. We note P as the point of the wheel's center, Q the contact point between the wheel and the ground, ψ the wheel's clean rotation angle and θ the angle linking the wheel's plane and the ground's one (See Fig. 1).

As seen above, the relative speed $\overrightarrow{V}_Q$(wheel/ground) equals zero at the contact point which lead to an equation associating the vector speed $\overrightarrow{V}_P$ and $\overrightarrow{W}$ the vector speed of the wheel rotation :

$$\overrightarrow{V}_Q = \overrightarrow{V}_P + \overrightarrow{W} \wedge \overrightarrow{PQ} = \overrightarrow{0}. \tag{1}$$

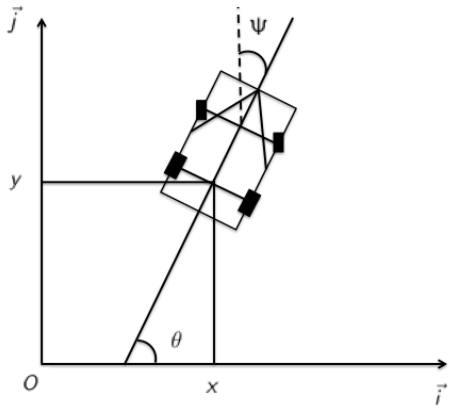


FIG. 2 – *A robot of four wheels (car type)*

2.2 Establishment of the Appropriate Kinematic Model

The kinematic model of each dynamic object is simply the description of how its state changes. The two major commands for any type of a mobile robot are acceleration and direction.

We describe in this section the mobile robot's kinematic model with all the details concerning the mechanics law governing the engine's movement. The methodology of the kinematic model is to elaborate a model of the engine's motion as a function of time according to the physics laws. In the following system, the velocity of the front wheel and the initial body pose can fix the exact localisation of the robot. We consider the reference point (x, y) as the mid-axis related to the rear axis wheels where matrices wheels are located (see Fig. 2). The complexity of the problem usually increases with its dimension. Therefore we consider the simple Newton law: $F = ma$. Besides, we also consider the following relation : $F = C \times U$ where C is the electro-mechanical transmission coefficient.

We then take by convention as explained in Egerstedt (2013) the following equation:

$$\dot{q} = \frac{C}{m} \times u \tag{2}$$

The model (2) above is a different representation of the robot which take into account the dynamics of the robot having on the left-hand side the speed and acceleration of the robot.

With a view to stabilize the equation and assure a minimal error. We obtain:

$$\dot{q} = \frac{C}{m}u - \gamma q \tag{3}$$

The transition from (2) to (3) was depicted in Egerstedt (2013), where tested in a mobile robot Khouil et al. (2016) the model (2) gave less satisfying results.

The corresponding nonlinear model for a four wheel mobile robot which takes into account both kinematic and dynamic constraints is:

$$\begin{cases} \dot{x} = \dfrac{C}{m} \times V \times \cos\theta - \gamma \times x \\ \dot{y} = \dfrac{C}{m} \times V \times \sin\theta - \gamma \times y \\ \dot{\theta} = \dfrac{C}{m} \times \dfrac{V}{L} \times \tan\psi - \gamma \times \theta \\ \dot{\psi} = \dfrac{C}{m} \times W - \gamma \times \psi. \end{cases} \tag{4}$$

where $(\dot{x}, \dot{y}, \dot{\theta}, \dot{\psi})$ are the first derivatives respective to time. The coefficient C designate the electro-mechanical transmission coefficient, m is the robot's mass whereas L represents the distance separating the front's axes and rear wheels. Ψ denotes the steering angle (the angle between the front wheels and the main axis), V refers to linear velocity, W the angular one and γ the wind resistance coefficient.

2.3 Nonlinear and Discrete Time Kinematic Model

The general purpose of this study is to optimize and planify the robot's path using an algorithm that handles the resolution of nonlinear systems with several constraints. The said algorithm will be applied in robotics for the very first time. Many nonlinear control techniques have a sufficient level of maturity to deal effectively with problems related to monitoring and/or control of industrial systems within a nonlinear modeling. However, we note a certain reluctance to adopt nonlinear control methods often considered difficult to understand, complicated to implement and which systematic performance analysis is complex Benalia (2004).

In order to apply the algorithm properly, we have to transform the robot's kinematic model above (4) from continuous time to discrete time with respect to a specific format of the NECMOP seen in Mestari et al. (2015):

$$\begin{cases} \dfrac{\delta x}{\delta t} = \dfrac{C}{m} \times V \times \cos\theta - \gamma \times x \\ \dfrac{\delta y}{\delta t} = \dfrac{C}{m} \times V \times \sin\theta - \gamma \times y \\ \dfrac{\delta \theta}{\delta t} = \dfrac{C}{m} \times \dfrac{V}{L} \times \tan\psi - \gamma \times \theta \\ \dfrac{\delta \psi}{\delta t} = \dfrac{C}{m} \times W - \gamma \times \psi. \end{cases} \tag{5}$$

According to the forward Euler rule, we convert the differential equations system above into a difference equations system:

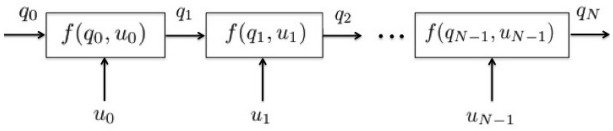


FIG. 3 – *N interconnected subsystems*

$$\begin{cases} x_{k+1} = \dfrac{C}{m} \times V_k \times \cos\theta_k \times \delta t + (1 - \gamma \times \delta t) \times x_k \\ y_{k+1} = \dfrac{C}{m} \times V_k \times \sin\theta_k \times \delta t + (1 - \gamma \times \delta t) \times y_k \\ \theta_{k+1} = \dfrac{C}{m} \times \dfrac{V_k}{L} \times \tan\psi_k \times \delta t + (1 - \gamma \times \delta t) \times \theta_k \\ \psi_{k+1} = \dfrac{C}{m} \times W_k \times \delta t + (1 - \gamma \times \delta t) \times \psi_k. \end{cases} \tag{6}$$

This is the matrix form:

$$\begin{pmatrix} x_{k+1} \\ y_{k+1} \\ \theta_{k+1} \\ \psi_{k+1} \end{pmatrix} = \frac{C}{m} \times \delta t \times \begin{pmatrix} \cos\theta_k & 0 \\ \sin\theta_k & 0 \\ \tan\psi_k / L & 0 \\ 0 & 1 \end{pmatrix} \times \begin{pmatrix} V_k \\ W_k \end{pmatrix} + (1 - \gamma \times \delta t) \times \begin{pmatrix} x_k \\ y_k \\ \theta_k \\ \psi_k \end{pmatrix}. \tag{7}$$

To simplify we note:

$$q_{k+1} = f(q_k, u_k). \tag{8}$$

with $q_k = \begin{pmatrix} x_k \\ y_k \\ \theta_k \\ \psi_k \end{pmatrix}$ and $u_k = \begin{pmatrix} V_k \\ W_k \end{pmatrix}$.

U_k represents the control function characterized by the velocity relative to the robot, which is the input of the system and q_k designate the system's position at time k.

We observe in Fig. 3 a representative diagram to visualize the sequence of planning the trajectory of the studied robot.

3 Necmop Resolution

3.1 Analysis of the Problem

At this stage of the study, we are going to apply the decomposition-coordination method seen in Mestari et al. (2015) with a view to obtain an optimal path free of

collisions through the kinematic model constructed above. As a start, we set the following system:

$$\begin{cases} \min E(q,u) \\ q_{k+1} = f(q_k, u_k) \\ \quad q_0 = q(0). \end{cases} \tag{9}$$

The studied problem consists over all to determine the control inputs specialized in optimizing many objective functions considered in conflicting situations so that we can afford a compromise between them.

The key step of solving the problem (9) is to divide the above system into numerous subsystems. The decomposition suggested help the transformation from a dynamic nonlinear system to a group of N interconnected subsystems assembled respective to a specific structure (Fig. 3). Z_k refers to the subsystem's output, u_k the input where q_k is the state:

$$Z_k = f(q_k, u_k), \quad k = 0, 1, 2, ..., N-2. \tag{10}$$

and

$$q_k = Z_{k-1}, \quad k = 1, 2, ..., N-1. \tag{11}$$

We define L as the ordinary Lagrange function:

$$L = \sum_{k=0}^{N-1} L_k. \tag{12}$$

For $k = 0$

$$L_0 = \frac{1}{N} E(q,u) + \mu_0^T (f(q_0, u_0) - Z_0). \tag{13}$$

For $1 \leq k \leq N-2$

$$L_k = \frac{1}{N} E(q,u) + \mu_k^T (f(q_k, u_k) - Z_k) + \beta_k^T (q_k - Z_{k-1}). \tag{14}$$

For $k = N-1$

$$L_{N-1} = \frac{1}{N} E(q,u) + \mu_{N-1}^T (f(q_{N-1}, u_{N-1}) - q_d) + \beta_{N-1}^T (q_{N-1} - Z_{N-2}) \tag{15}$$

μ_k and β_k are the lagrange multiplier vectors, they are in charge of the quality constraints (10) , (11) and are composed of four components. we have four components in q_k and $f(q_k, u_k)$ whereas in u_k we have only two components:

$$q_k^T = (x_k, y_k, \theta_k, \psi_k). \tag{16}$$

and

$$u_k^T = (V_k, W_k). \tag{17}$$

and

$$f(q_k, u_k) = (f_1(q_k, u_k), f_2(q_k, u_k), f_3(q_k, u_k), f_4(q_k, u_k)). \tag{18}$$

The equality constrained minimization problem (**??**) is being transformed into differential equations system with the help of ordinary Lagrange function derivations (12). We note the equilibrium point $(q_k^*, u_k^*, \mu_k^*, \beta_k^*, Z_k^*)$ satisfying the equations below according to the KKT rules **?**:

$$\begin{aligned} \nabla_{q_k} L = & \frac{1}{N} \times \frac{\delta E}{\delta q_k} + \frac{\delta f^T}{\delta q_k} \times \mu_k^* + \beta_k^* = 0. \\ & \text{For} \quad 1 \leq k \leq N-1 \end{aligned} \tag{19}$$

$$\begin{aligned} \nabla_{u_k} L = & \frac{1}{N} \times \frac{\delta E}{\delta u_k} + \frac{\delta f^T}{\delta u_k} \times \mu_k^* = 0. \\ & \text{For} \quad 0 \leq k \leq N-1 \end{aligned} \tag{20}$$

$$\begin{aligned} \nabla_{z_k} L = & -\mu_k^* - \beta_k^* = 0. \\ & \text{For} \quad 0 \leq k \leq N-2 \end{aligned} \tag{21}$$

$$\begin{aligned} \nabla_{\mu_k} L = & f(q_k^*, u_k^*) - Z_k^* = 0. \\ & \text{For} \quad 0 \leq k \leq N-1 \end{aligned} \tag{22}$$

$$\begin{aligned} \nabla_{\beta_k} L = & q_k^* - Z_{k-1}^* = 0. \\ & \text{For} \quad 1 \leq k \leq N-1 \end{aligned} \tag{23}$$

These five differential equations above (19)-(23) help us solve the problem (**??**).

3.2 The Method of Decomposition-Coordination

With a view to solve efficiently the equality constrained minimization problem (**??**), we've decomposed the treatment of the differential equations associated system into two separated levels. At the lower level we have a separated form in a way that every subproblem k can only handle unknown variables with indices $'k'$ for $k \in [0, N-1]$. The processing of equations system (19)-(23) is divided between two levels, the upper level is responsible of fixing $Z_k, k \in [0, N-2]$ and $\beta_k, k \in [1, N-1]$ taking into account equations (21) and (23). Once Z_k and β_k are fixed, the upper level transmit these informations to the lower level.

In consequence, the result of any subproblem equals the treatment of equations (19), (20), (22) for Z_k, $k \in [0, N-2]$ and β_k, $k \in [1, N-1]$ given by the upper level. Therefore, when applying the method of gradient we get the following differential equations system:

$$\frac{\delta q_k}{\delta t} = -\lambda_q \times \nabla_{q_k} L. \tag{24}$$

$$\frac{\delta u_k}{\delta t} = -\lambda_u \times \nabla_{u_k} L. \tag{25}$$

$$\frac{\delta \mu_k}{\delta t} = -\lambda_\mu \times \nabla_{\mu_k} L. \tag{26}$$

where $\lambda_q, \lambda_u, \lambda_\mu$ are all strictly positive. The equations (24)-(26) are transformed into the current scalar form:

$$\frac{\delta q_{ks}}{\delta t} = -\lambda_q \left(\frac{1}{N} \times \frac{\delta E}{\delta q_{ks}} + \sum_{i=1}^{4} \frac{\delta f_i}{\delta q_{ks}} \times \mu_{ki} + \beta_{ks} \right). \quad k \in [1, N-1] \quad \text{and} \quad s \in [1, 4] \tag{27}$$

$$\frac{\delta u_{ks}}{\delta t} = -\lambda_u \left(\frac{1}{N} \times \frac{\delta E}{\delta u_{ks}} + \sum_{i=1}^{4} \frac{\delta f_i}{\delta u_{ks}} \times \mu_{ki} \right). \quad k \in [1, N-1] \quad \text{and} \quad s \in [1, 2] \tag{28}$$

$$\frac{\delta \mu_{ks}}{\delta t} = -\lambda_\mu \left(f_s(q_k, u_k) - Z_{ks} \right). \quad k \in [0, N-1] \quad \text{and} \quad s \in [1, 4] \tag{29}$$

In pursuance of realizing the discrete time network using the forward Euler rule, we're able to remodel the differential equations system (27)-(29) to difference equations system.

$$q_{ks}^{(l+1)} = q_{ks}^{(l)} - \lambda_q \left(\frac{1}{N} \times \frac{\delta E^{(l)}}{\delta q_{ks}} + \sum_{i=1}^{4} \frac{\delta f_i^{(l)}}{\delta q_{ks}} \times \mu_{ki}^{(l)} + \beta_{ks}^{(j)} \right). \quad k \in [1, N-1] \quad \text{and} \quad s \in [1, 4] \tag{30}$$

$$u_{ks}^{(l+1)} = u_{ks}^{(l)} - \lambda_u \left(\frac{1}{N} \times \frac{\delta E^{(l)}}{\delta u_{ks}} + \sum_{i=1}^{4} \frac{\delta f_i^{(l)}}{\delta u_{ks}} \times \mu_{ki}^{(l)} \right). \quad k \in [1, N-1] \quad \text{and} \quad s \in [1, 2] \tag{31}$$

$$\mu_{ks}^{(l+1)} = \mu_{ks}^{(l)} + \lambda_\mu \left(f_s(q_k^{(l)}, u_k^{(l)}) - Z_{ks}^{(j)} \right). \quad k \in [0, N-1] \quad \text{and} \quad s \in [1, 4] \tag{32}$$

Once the upper level gives the information about $\beta_k^{(j)}, k \in [1, N-1]$ and $Z_k^{(j)}, k \in [0, N-2]$, then we have the ability to solve this difference equations system (30)-(32). The coordination between the two levels is quiet important for the transmission of necessary information for the functionning of lower level.

The coordination parameters are made up through the simultaneous work on $\beta_k^{(j)}, k \in$

$[1, N-1]$ and $Z_k^{(j)}, k \in [0, N-2]$ executed by the upper level.
In the lower level, this coordination allow the local resolution of the difference equations system (30)-(32) and also to determine $q_k^*(Z_k^{(j)}, \beta_k^{(j)})$, $u_k^*(Z_k^{(j)}, \beta_k^{(j)})$ and $\mu_k^*(Z_k^{(j)}, \beta_k^{(j)})$ with respect to equations (19), (20) and (22).
The resulted variables $q_k^*(Z_k^{(j)}, \beta_k^{(j)})$ and $\mu_k^*(Z_k^{(j)}, \beta_k^{(j)})$ are given to the upper level in order to verify if the previous information are correct and make the correction if not. Then the lower level can restart the process with valid variables.
The latter correction is needed for equations satisfaction (21) and (23).
The main goal of the upper level is to convert to the satisfaction of coordination parameters (19) and (21) by making the exact corrections. In fact, the said parameters at iteration $j+1$ are corrections of the ones at iteration j (see Fig. 4).

$$\begin{aligned} &Z_{ks}^{(j+1)} = Z_{ks}^{(j)} - \lambda_Z \times (-\mu_{is}^*(Z_k^{(j)}, \beta_k^{(j)}) - \beta_{k+1,s}^{(j)}). \\ &k \in [0, N-2] \qquad \text{and} \qquad s \in [1, 4] \end{aligned} \tag{33}$$

$$\begin{aligned} \beta_{ks}^{(j+1)} = \beta_{ks}^{(j)} + \lambda_\beta \times (q_{ks}^*(Z_k^{(j)}, \beta_k^{(j)}) - Z_{k-1,s}^{(j)}). \\ k \in [1, N-1] \qquad \text{and} \qquad s \in [1, 4] \end{aligned} \tag{34}$$

λ_Z, λ_β positive.

We repeat the system of solutions (30)-(32) till to the satisfaction of coordination, which means to satisfy equations (21) and (23).

The choice of λ at each iteration impacts the convergence celerity of the algorithm. Undeniably, if we choose a fixed coefficient λ, the convergence to an optimum point will be tardy and will not offer a great interest for the algorithm. In consequence, we apply the theorem below so that we can reach the adequate convergence celerity.

4 Modelization Using Artificial Neural Networks

The target here is using artificial neural networks for modeling and controlling process. The tasks which those networks are intended for, are effectively those of predictors or simulation models for process control, as well as regulators or markers. A neural network is an interconnected system of nonlinear operators, receiving outside signals through its inputs, and providing output signals which are the activities of certain neurons. A neural network is designed to perform a task defined by a sequence of entries, and a corresponding sequence of desired values for the activities of some of the neural network which are the output neurons.

4.1 Useful Neurons

Every single network is using some common neurons Mestari (2004) Mestari et al. (2001); Namir et al. (2008) such as the linear and threshold logic neuron.

The said neurons are in charge of summing the n weighted inputs and transfer the result as follows:

$$y = \Phi\left(\sum_{i=1}^{n} w_i x_i - \theta\right). \tag{35}$$

Φ is proposed as an activation function, it is a nonlinear or limiting transfer function. θ represents the external threshold, w_i designates the synaptic weights, x_i the inputs, n is the number of inputs and y the output.

The linear activation function:

$$\Phi_L(x) = x. \tag{36}$$

The threshold logic activation function:

$$\Phi_T(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{37}$$

4.2 Networks Used

The WN (weighting network) seen in Mestari et al. (2015) is in charge of the construction and calculation of the energy function $E(q, u)$. The energy function is formed through a linear neuron and an adaptive nonlinear building blocks. The aim of WN is to convert a multiobjective problem into a single objective by gathering all the objectives underneath the form of a weighted sum.

The two JNN (Jacobian neural network) seen in Mestari et al. (2015) are important for the computation of J_F and J_E related to the described dynamic system (10) and energy function (**??**). It is also a significant part of the final architecture network. The two JNN are also fundamental to resolve the difference equations (30)-(32).

The input network is realized by using the switched capacitor techniques using simple and basic elements as in Mestari et al. (2015) and Mestari et al. (2001).

The decomposition-coordination method illustrated in Fig. 4 includes the upper and lower level networks as seen in Mestari et al. (2015). It can be implemented using SC techniques Mestari et al. (2001). The obstacle detection network ODN is in charge of avoiding obstacles appearing at the optimal trajectory, once the obstacle is detected, the robot changes his path so that he can bypass the obstacle and then follow his path on an other optimal trajectory by using the decomposition coordination method each time we face an obstacle in order to regenerate a new optimal trajectory. The ODN has as inputs: x_k, y_k(the two first coordinate defyning the state of the robot), $x_{\text{obs}}, y_{\text{obs}}$(the two first coordinate defyning the state of the dynamic obstacles), $\Delta x, \Delta y, \Delta q$ (The three thresholds are fixed to help define whether there is a collision between the robot and the obstacles or not) and as outputs a binary result which tells us about the appearance of a near obstacle next to the state with the following coordinate (x_k, y_k). This network is only composed (see Fig. 5) of the absolute value network and the linear and threshold logic neuron Saber et al. (2016); Khouil et al. (2014b); Khouil et al. (2014c); Khouil et al. (2014a); Saber et al. (2014) which makes the time computation and complexity very interesting.

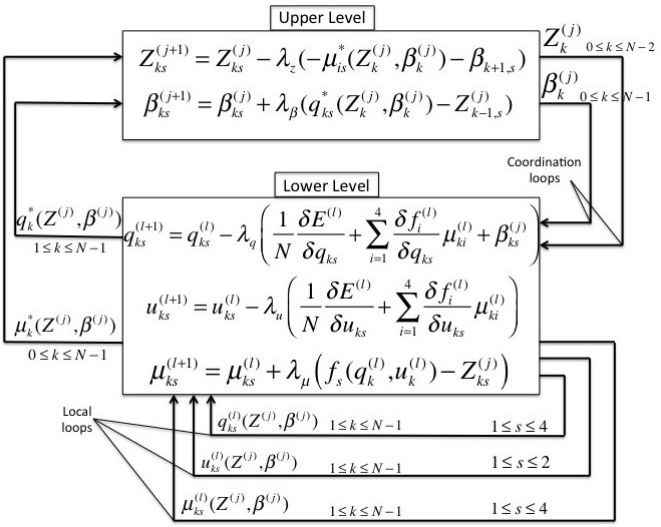


FIG. 4 – *Coordination between Upper and Lower Level*

The function corresponding to the ODN network is as follows:

$$S_{\text{ODN}} = (\Phi_T(\Delta x - |x_k - x_{\text{obs}}|) + \Phi_T(\Delta y - |y_k - y_{\text{obs}}|)) \times \Delta q + q_k. \tag{38}$$

The Absolute value network is indeed providing the distance between a specific state and a potential obstacle with a view to compare it to the fixed threshold $(\Delta x, \Delta y)$) (see Fig. 6).

The latter has as inputs x_k and x_{obs} or y_k and y_{obs}. Using the linear and threshold neuron in a very simple and structured architecture, we get as outputs the distance between two states as said before. The function corresponding to the AVN network is as follows:

$$S_{\text{AVN}} = |x - x'| = \Phi_T(x - x') \times (x - x') + \Phi_T(x' - x) \times (x' - x). \tag{39}$$

4.3 Implementation of the Final Network

The path planification algorithm detailed in the sections before is being implemented inside a simple and structured neural network architecture using switched capacitor (SC) techniques Mestari et al. (2001). All the networks and neurons defined in the subsection before are indeed a part of this final network, each network has a specific function. It is composed of an iteration of NECMOP network which is related to the decomposition coordination method as seen in Fig.7. On every iteration, there is a connexion with the obstacle detection method related to the ODN network and some commonly used neurons. Hence, the main objective of this final network is to

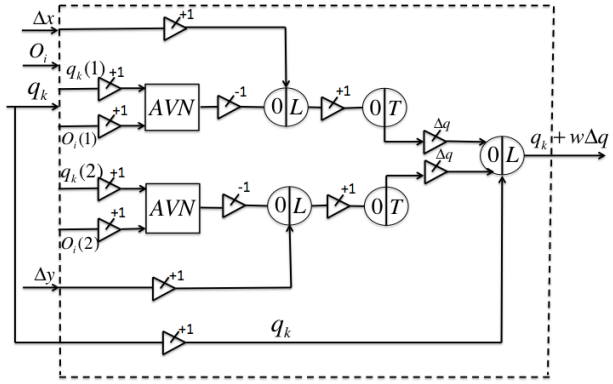


FIG. 5 – *Architecture of the Obstacle Detection Network*

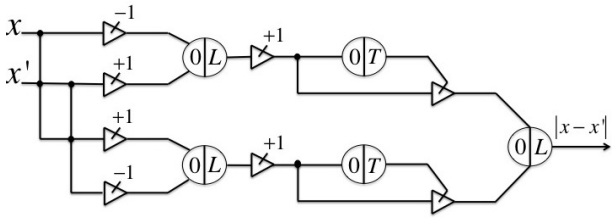


FIG. 6 – *Architecture of the Absolute Value Network*

find an optimal trajectory for a robot of 4 wheels avoiding dynamic obstacles.

In Fig.7, the inputs are q_0 and q_d respectively the initial and final state, then just after being processed through the NECMOP Network seen in Mestari et al. (2015), we obtain the optimal trajectory which will be the input of the Obstacle Detection Network along with the dynamic obstacle's position that will appear in the environment. As an output of the ODN network, we get the state $q_i + \delta q$ (δq designate the constant added to the robot's position subject to collision in order to avoid the obstacle), which represent the position of the robot after avoiding the i^{th} obstacle and at the same time an other input of the NECMOP network as the new initial robot's state.
The operation is repeated iteratively until obtaining the optimal trajectory q_k^* free of obstacles.

5 Conclusion

The optimization techniques are commonly established on a general processing inserted in an iterative functionning. The modelization and resolution of a NECMOP

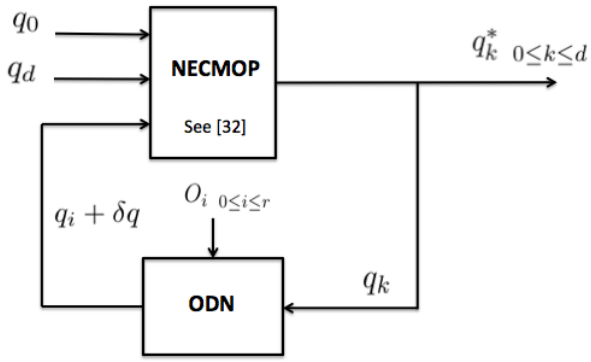


FIG. 7 – *Architecture of the Final Network*

problem is usually heavy in time and memory space.

This research is bound for the theoretical implementation of the decomposition-coordination algorithm and its application on a four wheel engine. In addition to a collision detection method that helps the robot to avoid the existing obstacles in the environment and then continue its path with a view to plan an optimal and collision free path. The implementation is executed after studying the constraints movement of the robot and developed its kinematic model in the form of differential equations system.

Using the parallelism of the artificial neural networks, we were able to challenge the time and memory space constraints. The architecture of neural network is composed of elements with a simple circuit based on VLSI techniques.

The novelty of this work relies in the overall planning of different trajectories using the decolposition-coordination method allowing the robot enough freedom to deviate the eventual obstacles. Another features of the studied method is that it is easily implemented on an Analog Neural Network.

References

Abraham, A. and L. Jain (2005). Evolutionary multiobjective optimization. In *Evolutionary Multiobjective Optimization*, pp. 1–6. Springer.

Aggelogiannaki, E. and H. Sarimveis (2007). A simulated annealing algorithm for prioritized multiobjective optimization—implementation in an adaptive model predictive control configuration. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 37*(4), 902–915.

Agrawal, S., B. Panigrahi, and M. K. Tiwari (2008). Multiobjective particle swarm algorithm with fuzzy clustering for electrical power dispatch. *IEEE Transactions on Evolutionary Computation 12*(5), 529–541.

Ahn, C. W. (2006). *Advances in evolutionary algorithms*. Springer.

Benalia, A. (2004). *Contribution à la modélisation et la commande robuste du confort thermique au sein d'un habitacle automobile*. Ph. D. thesis, Paris 11.

Deb, K., M. Mohan, and S. Mishra (2005). Evaluating the ε-domination based multi-objective evolutionary algorithm for a quick computation of pareto-optimal solutions. *Evolutionary computation 13*(4), 501–525.

Egerstedt, M. (2013). Controls for the masses [focus on education]. *IEEE Control Systems 33*(4), 40–44.

Eichfelder, G. (2008). *Adaptive scalarization methods in multiobjective optimization*. Springer.

Khouil, M., H. I. ENSET, M. I. Sanou, M. M. Mestari, M. A. Aitelmahjoub, and E. Casablanca (2016). Planification of an optimal path for a mobile robot using neural networks. *Applied Mathematical Sciences 10*(13), 637–652.

Khouil, M., N. Saber, and M. Mestari (2014a). Collision detection for three dimension objects in a fixed time. In *Information Science and Technology (CIST), 2014 Third IEEE International Colloquium in*, pp. 235–240. IEEE.

Khouil, M., N. Saber, and M. Mestari (2014b). Neural network in fixed time for collision detection between two convex polyhedra. *network 1*, 6.

Khouil, M., N. Saber, and M. Mestari (2014c). Réseaux de neurones pour la détection de collision et localisation de contacts des polyèdres convexes. *Revue Méditerranéenne des Télécommunications 4*(2), 104–108.

Mestari, M. (2004). An analog neural network implementation in fixed time of adjustable-order statistic filters and applications. *IEEE transactions on Neural Networks 15*(3), 766–785.

Mestari, M., M. Benzirar, N. Saber, and M. Khouil (2015). Solving nonlinear equality constrained multiobjective optimization problems using neural networks. *IEEE transactions on neural networks and learning systems 26*(10), 2500–2520.

Mestari, M., A. Namir, and J. Abouir (2001). Switched capacitor neural networks for optimal control of nonlinear dynamic systems: Design and stability analysis. *Systems Analysis-Modelling-Simulation 41*(3), 559.

Namir, A., M. Mestari, K. Akodadi, and A. Badi (2008). θ (1) time neural network minimum distance classifier and its application to optical character recognition problem. *Applied Mathematical Sciences 2*(26), 1253–1282.

Saber, N., M. ENSET, Mestari, and A. Ait El Mahjoub (2016). The multi-constrained least-cost multicast problem with neural networks in fixed time. *Applied Mathematical Sciences 10*(19), 931–945.

Saber, N., M. Khouil, and M. Mestari (2014). Neural networks based on adjustable-order statistic filters for multimedia multicast routing. In *Information Science and Technology (CIST), 2014 Third IEEE International Colloquium in*, pp. 435–439. IEEE.

Stentz, A. (1994). Optimal and efficient path planning for partially-known environments. In *Robotics and Automation, 1994. Proceedings., 1994 IEEE International Conference on*, pp. 3310–3317. IEEE.

Zitzler, E., K. Deb, and L. Thiele (2000). Comparison of multiobjective evolutionary algorithms: Empirical results. *Evolutionary computation 8*(2), 173–195.

Résumé

Dans ce présent document, une nouvelle méthode d'optimisation est appliqué pour planifier une trajectoire optimale et contrôlée dans un environnement à obstacles dynamiques. L'algorithme derrière cette méthode sert en effet à optimiser, contrôler et planifier une trajectoire sans collisions pour un robot mobile à quatre roues. Les réseaux de neurones artificiels sont utilisées afin de réduire la complexité de temps et calcul grâce au parallélisme et à la simplicité d'architecture qu'offrent ces techniques.

Processus de calcul parallèle des réseaux spatiaux de Voronoï basé sur une architecture distribuée

Aziz Mabrouk*, Hafssa Aggour*, Azedine Boulmakoul**

*Université Abdelmalek Essaadi, ER Ingénieurie des Systèmes d'Information
Route de Martil – Tétouan, Maroc.
Aziz.mabrouk@gmail.com, hafssaaggour@gmail.com
**Université Hassan II, FST de Mohammedia, LIM/IOS
Mohammedia, Maroc
azedine.boulmakoul@gmail.com

Résumé. Dans le contexte des villes intelligentes, des flux de données sémantiques et spatiales en temps réel sont produits en continu et analysés par des algorithmes efficaces et performants capables de gérer les complexités liées aux Big Data afin d'obtenir des connaissances exploitables permettant les fonctions de base des systèmes d'aide à la décision. En outre, les diagrammes spatiaux de Voronoï sont très utilisés pour modéliser et analyser des réseaux spatiaux. Aussi, le recours au calcul distribué reste l'un des défis actuels de la recherche en sciences de l'information géographique. Dans ce sens, nous proposons un processus de calcul des réseaux spatiaux de Voronoï basé sur une architecture distribuée adaptés au contexte et aux objectifs de l'étude de phénomènes géographiques réels. Nous exploitons ce processus distribué dans un prochain travail pour analyser spatialement les villes intelligentes et ce en profitant de sa capacité de géo-traiter des données spatiales massives.

1 Introduction

Aujourd'hui, le grand défi et la préoccupation portent sur les problèmes des villes intelligentes qui reposent sur des infrastructures de télécommunications performantes dont l'utilisation des données est perçue comme un levier de pilotage et d'action. Cependant, la collecte, le traitement, le stockage, la gestion, l'analyse, l'interprétation et l'affichage des données spatiales massives et volumineuses en toute efficacité deviennent de temps en temps difficiles, délicats et ennuyeux, sachant que ces données sont caractérisées par leurs natures différentes (temporelle, spatiale, hybride, etc.). En outre, l'analyse spatiale de ces données massives impose forcément la mise en place des algorithmes qui respectent les normes de l'efficacité, l'efficience et la pertinence et qui garantissent l'optimisation du temps de traitement et la prise de décision en temps réel. Des algorithmes qui permettent un traitement rapide et fiable qualitativement et quantitativement. Dans le même sens, l'architecture distribuée permet l'accélération des processus de calcul aussi bien l'optimisation du temps de traitement et d'interprétation des données massives (Carmen De Maio, 2017). En effet, l'objectif de cet article est de proposer une nouvelle approche de calcul des diagrammes de

Voronoï en se basant sur une architecture distribuée. En fait, ces diagrammes jouent un rôle important dans les calculs géométriques dans différents domaines. Dans la Science de l'Information Géographique, après les premiers prototypes, ces structures géométriques ont été appliquées dans le marketing et l'analyse urbaine en tant que des outils de modélisation géométrique et mathématique. Le Diagramme de Voronoï est entré ensuite dans le monde des SIG grâce au développement étendu du calcul géométrique. Dans ce papier, nous présentons dans un premier temps les différentes méthodes classiques de calcul des Diagrammes de Voronoï Spatiaux et ensuite nous présentons notre processus de calcul parallèle. Ce processus de calcul reçoit les données spatiales du réseau sous forme d'un ensemble des données spatiales distribuées résilients (RDD spatiales). Ensuite il partitionne spatialement le réseau en se basant sur le calcul distribué de la distance euclidienne. Et puis il procède au calcul distribué des arbres des plus courts chemins.

2 Les Diagrammes de Voronoï Spatiaux

Le Diagramme de Voronoï de type Réseau (DVR) est défini par la division du réseau en sous réseau de Voronoï dont chacune contient les points les plus proches à chaque générateur de Voronoï en parcourant le plus court chemin entre ces composantes (Okabe et al., 2008). Si le réseau analysé est un réseau spatial réel (réseau routier, réseau de transport, etc.…), ce diagramme s'appelle le Diagramme de Voronoï du Réseau Spatial ($DVR_{Spatial}$) (Mabrouk et Boulmakoul, 2012). La figure 1 présente une partie du DVR spatial généré par les hôpitaux de la ville de Tétouan.

FIG. 1 – *Une partie du DVR spatial généré par les hôpitaux de la ville de Tétouan*.

Soit $N(S, A)$ un réseau de sommets S et d'arcs A et P un ensemble de sommets.
$G = \{g_1, \dots, g_n\}$ avec$G \subseteq S$. G représentent les générateurs de Voronoï dans le DVR.

Chaque poids valuant un arc, est représenté par un nombre réel. Considérons v et w deux sommets qui appartiennent au S. Nous allons utiliser $P(v, w)$ pour représenter le poids du plus court chemin de v à w dans le réseau N. le DVR pour G divise le réseau N en n sous réseaux de Voronoï $Vor(1), \dots, Vor(n)$ avec :

$$Vor(i) = \{\forall p \in P \,/\, P(pi, p) \leq P(pj, p), 1 \leq \forall j \leq n, i \neq j\} \quad (1)$$

2.1 Méthodes de calcul du DVR

Nous présentons dans cette section deux méthodes de calcul existantes dans la littérature :

Une méthode utilisant Dijkstra parallèle, considèrent les générateurs de Voronoï comme des sources multiples et puis cherchent en parallèle les nœuds les plus proches à chaque générateur de Voronoï, et ce en se basant sur les poids des chemins parcourus.

Une méthode qui Construit d'abord les arbres des plus courts chemins$ACC(g_i)$enraciné à $g_i,\ i = 1 \dots, n_g$; et prolonge en seconde lieu ces arbres à $APCC(\ g_1), \dots, APCC(g_{n_g})$ pour choisir le générateur le plus proche à chaque point sur le réseau.

2.1.1 Méthodes utilisant Dijkstra parallèle

Le calcul du Diagramme de Voronoï de type réseau selon Okabe (2008), se fait par la division du réseau en nœuds et arcs, En principe il y a deux étapes pour construire le diagramme de Voronoï pour un réseau spatial :

Construire **Le diagramme de Voronoï des nœuds du réseau**: Ce diagramme assigne chaque nœud du réseau spatial à un générateur de Voronoï dont la distance du plus court chemin du nœud à ce générateur est la plus petite de toutes les distances des plus courts chemins parcourus pour atteindre les autres générateurs de Voronoï.

Construire **Le diagramme de Voronoï des arcs du réseau**: Il assigne chaque arc, dans le réseau spatial, à un générateur de Voronoï. Il coupe éventuellement des arcs en deux parts dont les points sur une part de l'arc coupé sont assignés à un générateur (qui est le plus proche de ces points), et les points sur l'autre partie de l'arc coupé sont assignés à un autre générateur (qui est le plus près d'eux).

Le temps de calcul du Diagramme de Voronoï de type réseau en utilisant Dijkstra parallèle est de $O(\ n_a + \ n_s log n_s)$ avec n_s et n_a sont respectivement le nombre des nœuds et le nombre des arcs constituant le réseau.

Pour construire le diagramme de Voronoï de type réseau, Erwig(2000) et Okabe et al. (2008) ont utilisé l'algorithme Dijkstra. C'est un algorithme qui calcule dans un réseau le chemin le plus court d'un nœud choisi à n'importe quel autre nœud (Dijkstra 1959). Dans ce contexte, il doit être modifié pour calculer en parallèle les chemins les plus courts de plusieurs générateurs. Chaque nœud dans le réseau est assigné par la distance la plus courte seulement au prochain générateur. Cet algorithme, qui est basé sur l'algorithme Dijkstra, utilise une file prioritaire d'attente où les opérations inserer(), Extraire_Min() et Mettre_à_jour() sont disponibles. Il considère les générateurs de Voronoï comme des sources multiples et cherche en parallèle les nœuds les plus proches à chaque générateur en se basant sur les poids des chemins parcourus.

2.1.2 Méthode utilisant l'arbre prolongé du plus court chemin (APCC)

Okabe et al.(2008) proposent une méthode naïve pour construire le DVR orienté c'est d'abord, construire n_g APCCs enraciné aux n_g générateurs; et en second lieu, choisir le générateur le plus proche à chaque point sur le réseau en utilisant lesn_g APCCs (Voir la figure 2).

L'astuce est :

d'abord, ajouter un sommet factice g_0 et n_g arcs factices qui joignent g_0 et les générateurs g_i, $i = 1, \dots, n_g$ avec le poids zéro $P(g_0, g_i) = 0$ au réseau original ;

en second lieu, construire $APCC(g_0)$ enracinée au sommet factice g_0;

enfin, supprimer le sommet factice et les arcs factices $APCC(g_0)$.

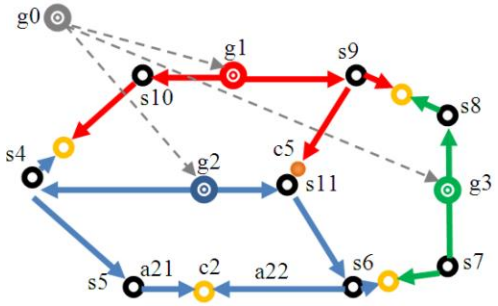


FIG.2 – *DVR orienté se compose de trois APCCs : APCC(g1), APCC(g2), et APCC(g3).*

Alors, $APCC(g_0)$ est décomposé en n_g sous-arbres, $APCC(g_1) \dots, APCC(g_{n_g})$.Sur la 0, $APCC(g_1), APCC(g_2)\ et\ APCC(g_3)$ sont indiqués par les segments rouges, les segments bleus et les segments verts. $g_0$est un sommet factice et les lignes discrètes sont les arcs factices.$g1, g2, et\ g3$ sont des points générateurs. Le sous-réseau de Voronoï $Vor(i)$ est donné par $APCC(gi)$, et le $DVR_{Orienté}$"Vor" est écrit comme $Vor = \{APCC(g1) \dots, APCC(g_{n_g})\}$. Considérant un réseau, $R(S, A)$ contenant l'ensemble de sommets, $S = \{s_1 \dots, s_{n_s}\}$, et l'ensemble d'arcs, $A = \{a_1 \dots, a_{n_a}\}$. Un arc a_i peut être à sens unique ou bidirectionnel.

Pour construire un DVR orienté, (Okabe, 2008) propose deux phases :

La première phase : Construire $ACC(g_1), \dots, ACC(g_{n_g})$, où $ACC(g_i)$ est l'arbre des plus courts chemins enraciné à g_i, $i = 1 \dots, n_g$;

La deuxième phase : Prolonger les arbres à $APCC(g_1), \dots, APCC(g_{n_g})$.

2.2 Discussion

Les diagrammes de Voronoï de type réseau sont discutés par Erwig(2000), Okabe et al (2008) et bien d'autres auteurs. Les méthodes de calcul proposées utilisent l'algorithme Dijkstra pour chercher en parallèle les nœuds les plus proches à chaque générateur de Voronoï en considérant ces générateurs comme des sources multiples. En outre, Mabrouk et Boulmakoul (2012) ont utilisé cet algorithme pour calculer le Diagramme de Voronoï pour les réseaux spatiaux (réseau routier, réseau de transport, etc.…).

Le temps de construction du Diagramme de Voronoï est calculé principalement en fonction de nombre des nœuds n_s et des arcs n_a constituant le graphe modélisant le réseau. En

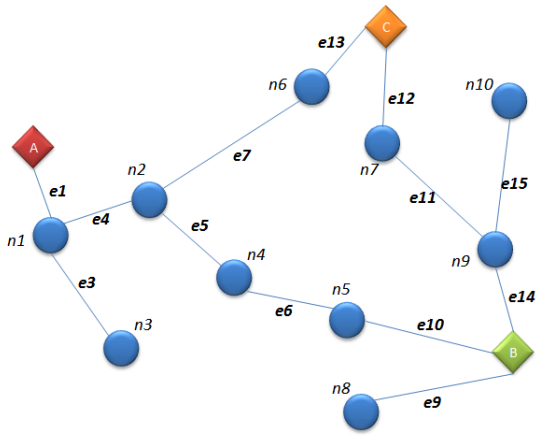
effet, le temps de calcul de ce diagramme en utilisant Dijkstra parallèle est de $(n_a + n_s log n_s)$. Cependant, ce parallélisme de calcul n'est pas effectif au niveau du processeur, mais il s'agit en fait d'un calcul des plus courts chemins « à tour de rôle » entre chaque générateur de Voronoï et l'ensemble des nœuds du réseau.

Ainsi, la collecte, le traitement, le stockage, la gestion, l'analyse, l'interprétation et l'affichage des données spatiales massives et volumineuses en toute efficacité deviennent de temps en temps difficiles, délicats et ennuyeux, sachant que ces données sont caractérisées par leurs natures différentes (temporelle, spatiale, hybride, etc.). En outre, l'analyse spatiale de ces données massives impose forcément la mise en place des algorithmes qui respectent les normes de l'efficacité, l'efficience et la pertinence et qui garantissent l'optimisation du temps de traitement et la prise de décision en temps réel.

3 Notre approche : Réseaux spatiaux de Voronoï distribués

3.1 Représentation des données spatiales du réseau dans une plateforme distribuée

Soit $R(N, E)$ un réseau spatial de k nœuds $N = \{n_1, \dots, n_k\}$ et de q arcs $E = \{e_1, \dots, e_q\}$
Soit G un ensemble de *m* nœuds représentant les générateurs de Voronoï
$G = \{g_1, \dots, g_m\}$ avec $G \subseteq N$.
Chaque poids We_j valuant un arc e_j est représenté par un nombre réel. Chaque nœud n_i et générateur de voronoï g_j sont spatialement référencés respectivement avec les coordonnées $n_i(x_{ni},y_{ni})$ *et* $g_j(x_{Gi},y_{Gi})$



E	N1	N2
e1	A	n1
e2	n1	n2
e3	n1	n3
e4	n2	n4
e5	n6	C
e6	n4	n5
e7	n2	n6
e8	n5	B
e9	B	n5
e10	B	n9
e11	n9	n7
e12	C	n7
e13	n9	n10

N	X	Y
n1	12	56
n2	48	52
n3	21	75
n4	88	63
n5	23	47
n6	36	95
n7	89	73
n8	29	47
n9	55	37
n10	81	22

FIG.3 – Modélisation du réseau spatial en graphe

Les nœuds et les arcs sont chargés en tant que un ensemble de données distribué résilient (RDD), l'abstraction de base dans une plateforme distribuée. Il représente une collection partitionnée d'éléments pouvant être manipulés en parallèle.

3.2 Stratégie de partition du réseau spatial basé sur le calcul distribué de la distance euclidienne

Pour permettre un traitement parallèle, les données spatiales "distribuées" relatives au réseau sont représentées sous forme d'un ensemble des données spatiales distribuées résilients (RDD spatiales). Ceci dit, elles doivent être partitionnées.

L'idée est de décomposer ces données selon la proximité euclidienne des nœuds aux différents générateurs de Voronoï (une sorte de diagramme de Voronoï euclidien) (voir la figure 4). Notre processus de calcul découpe alors le réseau spatial en n_g sous-réseaux Sub_{Net} avec n_g est le nombre des générateurs de voronoï. Chaque sous-réseau Sub_{Net} est composé d'un sous réseau de nœuds $N_{Sub_{Net}}$et un sous réseau d'arcs $E\_Sub_{Net_i}$ avec :

$$N\_Sub_{Net_i} = \{\forall n_i \in N | dist_{Euc}(n_i, g_i) \leq dist_{Euc}(n_i, g_j), j \neq i, j = 1, \dots, m\}$$
$$E\_Sub_{Net_i} = \{\forall e_i \in E | GV(nd_i) = GV(nf_i), j \neq i, j = 1, \dots, q\} \quad (2)$$

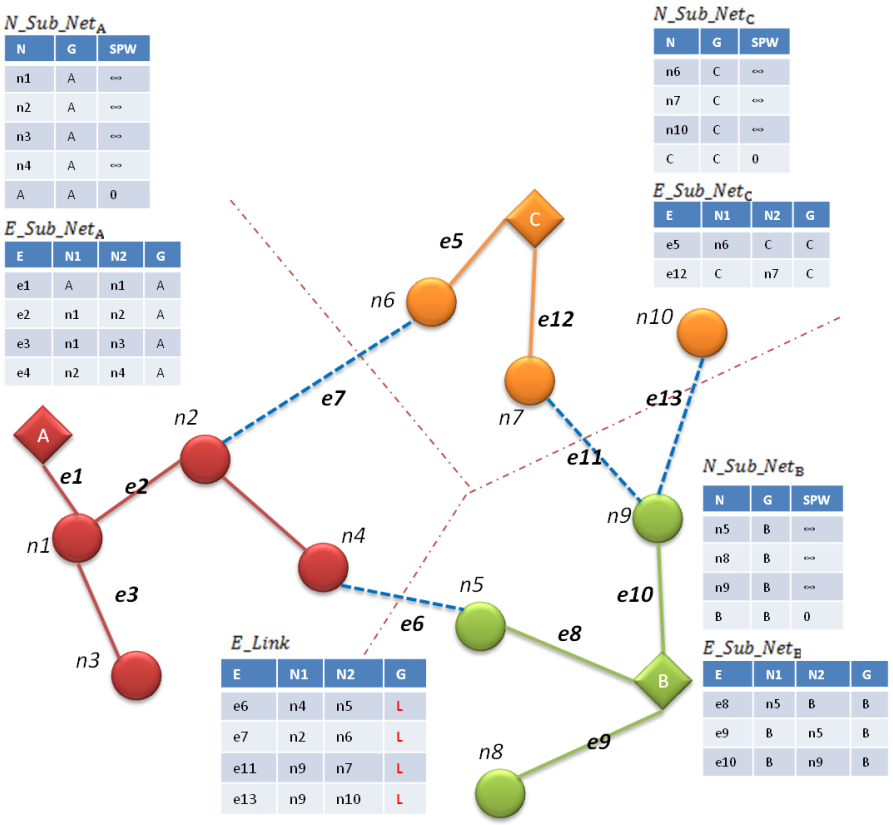


FIG.4 – Partitionnement spatial du réseau basé sur le calcul distribué de la distance euclidienne

En effet, l'ensemble des données spatiales distribuées résilients (RDD spatiales) est composé de n_g partitions. Chaque partition comporte les données spatiales d'un sous réseau de nœuds $N_{Sub_{Net}}$et un sous réseau d'arcs $E\_Sub_{Net_i}$.

Lors de la décomposition du réseau spatial, on constate des arcs dont les deux extrémités appartiennent à deux sous réseaux différents. Cependant ces arcs constituent des liaisons entres ces sous réseaux. Ceci dit, ils ne peuvent être intégrer dans aucune partition. Ces arcs sont rangés dans l'ensemble E_Link qui servira par la suite une table de routage (voir la section suivante).

$$E\_Link = \{ \forall e_i \in E | GV(nd_i) \neq GV(nf_i), j \neq i, j = 1, \dots, q\} \quad (3)$$

Cette étape du processus est aussi exécutée dans un environnement distribué et elle est concrétisée par ces deux pseudos code :

3.2.1 Pseudo code : Partition spatiales des nœuds du réseau spatial

---

Pour *chaque gj* ***faire*** *SPW(gi)=0*

Pour *chaque nœud n_i faire*
  Pour *chaque g_j faire*

$$dist_{Eucl}\ (n_i, g_j) = \sqrt{(x_{Gj} - x_{ni})^2 + (y_{Gj} - y_{ni})^2}$$

Si $dist_{Eucl}\ (n_i, g_j) < dist_{Eucl}\ (n_i, g_{j-1})$ ***Alors***
Ajouter n_i au $N\_Sub\_Net_{g_j}$
$GV(n_i) = g_j$
$SPW(n_i) = \infty$
Fin Si

  Fin pour
Fin pour

3.2.2 Pseudo code : Partition spatiales des arcs des sous-réseaux et sauvegarde des arcs de liaison

---

Pour *chaque arc e_i faire*

Si $GV(nd_i) == GV(nf_i)$ ***Alors***
Ajouter e_i au $E\_Sub\_Net_{GV(nd_i)}$
$GV(e_i) = GV(nd_i)$
Sinon
Ajouter e_i au E_Link
$GV(e_i) = "L"$
Fin Si

Fin pour

3.3 Calcul distribué des arbres des plus courts chemins

Les données spatiales sont réparties dans $n_g$partirions. Ceci dit elles seront à la charge de n_g executors lors des traitements. En parallèle, au niveau de chaque partition j le processus de calcul suit l'algorithme Dijkstra pour chercher les nœuds les plus proches à la racine g_j de l'arbre $Sub_{Net\,j}$ en se basant sur les poids des chemins parcourus (voir la figure 5). Il marque le prédécesseur $Pr(n_i)$ et le poids $SPW[n_i]$du plus courts chemin pour atteindre chaque nœud n_i . L'arbre des plus courts chemins $ACC(g_j)$ est alors construit par listes d'adjacence. Le résultat est un ensemble des arbres des plus courts chemins *ACC(gi)* enraciné à *gi, i=1…,ng*;

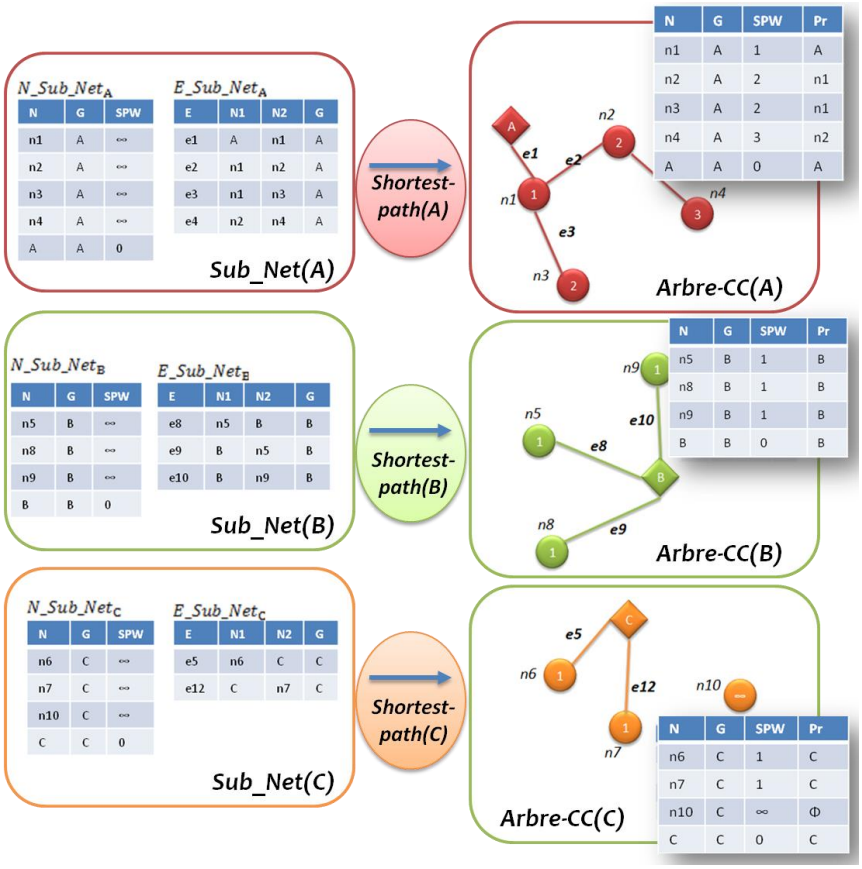


FIG.5 – Calcul distribué des arbres des plus courts cheminsACC(A) , ACC(B) et ACC(C) enracinés à A, B et C

3.4 Mise à jour des affectations des générateurs aux sommets et aux arrêtes

Les arbres des plus courts chemins $ACC(g_i)$ enraciné à g_i, $i = 1 \ldots, n_g$; sont calculés en parallèle et d'une manière indépendante dont chaque $ACC(g_j)$ comporte l'ensemble des sommets de points de repère, en constituant un graphe où chaque sommet comporte l'information sur le poids du plus court chemin pour atteindre la racine *gi*. Cependant, ce poids peut être supérieur au poids du plus court chemin pour atteindre un autre générateur de Voronoï. Ceci dit, la mise à jour de l'affectation des sommets aux plus proches générateurs est nécessaire. En fait, chaque deux sommets constituant les extrémités des arcs de liaison (Voir la section précédente) n'appartiennent pas au même arbre ACC. Par conséquence ce sont les sommets avec lesquels on doit commencer la comparaison des poids de leurs chemins plus courts.

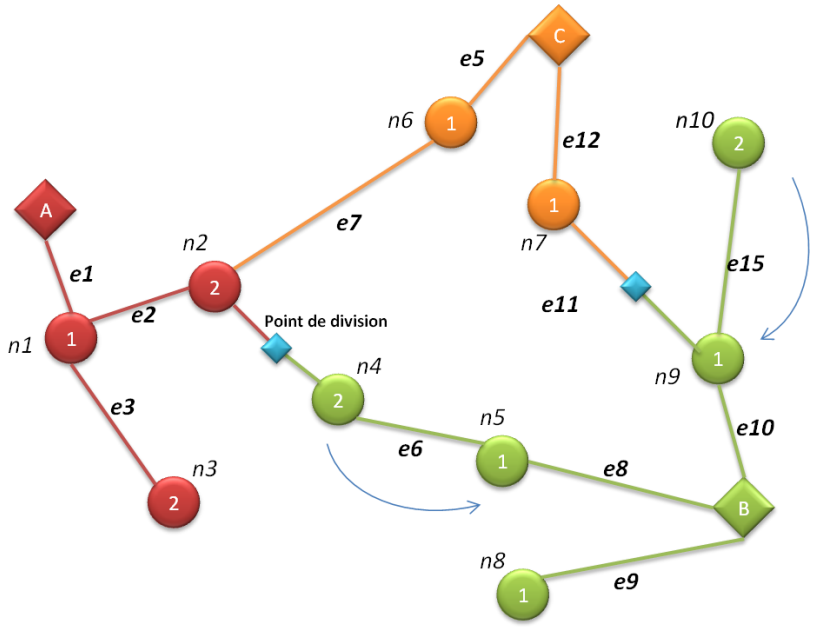


FIG.6 – Mise à jour des affectations des générateurs aux sommets et aux arrêtes

Soit deux arbre *ACC(gi) et ACC(gj)* enraciné à g_i et g_j. Le processus de calcul, que nous proposons avec le pseudo code ci-dessous, lit les extrémités début et fin de chaque arc liant ces deux arbres et puis compare leurs poids des plus courts chemins pour atteindre *gi* et *gj*. Tant que le poids du sommet de début est inférieur strictement au poids du sommet de fin, le processus de calcul ajoute ce dernier à l'arbre où il appartient le sommet de début et met à jour leurs poids des plus courts chemins (Voir la figure 6).

Pseudo code

Pour *chaque arrête* $e_i \in$ *au E_Link* ***faire***

Lire *ni_d et ni_f // les extrimités début et fin de l'arrête ei*

Lire *SPW(ni_d) et SPW(ni_f)*

Tanque *SPW(ni_d) < SPW(ni_f)* ***faire***

GV (ni_f) = GV(ni_d)
SPW(ni_f) = SPW(ni_d) + W(e(ni_d, ni_f))

GV ***(e****(ni_d, ni_f))= GV(ni_d)*

ni_d=ni_f
ni_f = Pr(ni_f)

Si *SPW(ni_d)* $\geq$ *SPW(ni_f)* ***Alors***

Partitionner l'arrête (ni_d, ni_f) en deux arrêtes a1(ni_d, pt_div) et a2(pt_div, ni_f) Avec SPW(pt_div, GV(ni_d)) = SPW(pt_div, GV(ni_f))

Ajouter a1(ni_d, pt_div) au $E\_Sub\_Net_{GV(nd_i)}$
Ajouter a2(pt_div, ni_f) au $E\_Sub\_Net_{GV(nf_i)}$

Fin Si
Fin Tanque

Fin Pour

3.4.1 Mettre à jour de l'affectation des arrêtes au plus proche générateurs

Le processus de calcul que nous proposons (voir le pseudo code précédent) met à jour aussi l'affectation des arrêtes aux bons arbres des plus courts chemins (voir figure 7). D'après Margot Graf et Stephan Winter (2003), chaque arc est marqué selon l'affectation de son nœud de début et celle de fin aux générateurs de Voronoï. Ils distinguent quatre cas différents qui peuvent se produire. Parmi ces cas si Le nœud de début et celui de la fin appartiennent à différents générateurs. L'arc est bidirectionnel et n'est pas nécessairement symétrique. Il est accessible par n'importe quel point qui lui appartient. Dans ce cas l'arc doit être divisé. Le point de division est le point auquel les coûts de déplacement au générateur du nœud de début sont égaux aux coûts de déplacement au générateur du nœud de fin. La première partie XA (nœud de début au point de division) sera assignée à l'arbre des plus courts

chemins où il appartient le nœud de début, et la deuxième partie, XE, du point de division au nœud de fin, sera assignée à l'arbre des plus courts chemins où il appartient le nœud de fin. Au point de division on présentera un nœud qui n'est assigné à aucun arbre des plus courts chemins. Il représente la frontière entre deux sous réseaux de Voronoï (Mabrouk, A et al, 2017).

FIG.7 – Mise à jour des affectations des générateurs aux sommets et aux arrêtes

4 Conclusion et perspective

Le processus de calcul, proposé dans ce papier, est basé sur une architecture distribuée. Il permet de calculer les diagrammes spatiaux de voronoï en traitant par partie et en parallèle des données spatiales massives qui deviennent aujourd'hui primordiales, et constituent les sous bassement de l'analyse spatiale des zones géographiques. En fait, les fluctuations et la production continue des données spatiales imposent donc, le traitement et l'analyse en temps réel afin d'intervenir rapidement à prendre des décisions spatiales. La perspective de ce travail est l'implémentation de ce processus de calcul avec une plateforme distribuée Spark qui propose un Framework complet et unifié pour répondre aux besoins de traitements Big Data pour divers jeux de données, divers par leur nature (texte, graphe, etc.) aussi bien que par le type de source (batch ou flux temps-réel). Nous exploitons ce processus distribué dans un prochain travail pour analyser spatialement les villes intelligentes et ce en profitant de sa capacité de géo-traiter des données spatiales massives.

Références

Carmen De Maio et al., Distributed Online Temporal Fuzzy Concept Analysis for stream processing in smart cities. J. Parallel Distrib Comput. (2017).

Erwig, M., 2000: The Graph Voronoi Diagram with Applications. Networks, 36 (3): 156-163.

Okabe, A., Satoh, T., Furuta, T., Suzuki, A. and Okano, K. (2008), Generalized network Voronoï diagrams: Concepts, computational methods, and applications', International Journal of Geographical Information Science.

Okabe, Atsuyuki and Suzuki, Atsuo, (1997), Locational optimization problems solved through Voronoi dia-grams. European Journal of Operational Research, Elsevier, Vol. 98, pp.445-456

Margot Graf et Stephan Winter , Network Voronoï Diagrams, Institut für Geoinformation, Technische Universität Wien,2003

Mabrouk, A., Boulmakoul, A. Karim, L. and Lbath, A. (2017). Safest and shortest itineraries for transporting hazardous materials using split points of Voronoï spatial diagrams based on spatial modeling of vulnerable zones, Procedia Computer Science, Vol. 109C, pp. 156-163.

Mabrouk, A., Boulmakoul A., (2012). Modèle spatial objet base sur les diagrammes spatiaux de voronoï pour la geo-gouvernance des espaces urbains, INTIS 2012, ISBN: 2168/2008 978-9981-1-3000-1, pp. 105-116

Summary

In the context of smart cities, real-time semantic and spatial data flows are continually generated and analyzed by efficient, high-performance algorithms able to handle the complexities associated with Big Data in order to obtain exploitable knowledge for the functions basic decision support systems. In addition, Voronoi spatial diagrams are widely used to model and analyze spatial networks. Also, the use of distributed computing remains one of the current challenges in geographic information science research. In this sense, we propose a distributed architecture of a process of computation of the Voronoï spatial diagrams adapted to the context and the objectives of the study of real geographical phenomena. We use this distributed process in a future work to spatially analyze smart cities, taking advantage of its ability to geo-process massive spatial data.

Supervised Learning and Multi Agent Systems for Fault Tolerance in Cloud Computing

Rayen Derbal*, Amira Hassad*, Ouassila Hioual**

* Abbes Laghrour University of Khenchela, Algeria
{derbalrayen, hassad.amira9}@gmail.com
**Abbes Laghrour University of Khenchela, Algeria,
**LIRE Laboratory of Constantine, Algeria
ouassila.hioual@gmail.com

Abstract. Cloud Computing is composed of a set of services distributed over a global communication network. These services offer a real-time system and a virtual environment with immense computing capacities. Most of real-time systems are critical to security and require high reliability and a very high level of fault tolerance for their execution. In this work, we propose a fault-tolerance model that allows making an appropriate decision in the event of a breakdown. The proposed model is based on supervised learning techniques and multi-agent systems (MAS). Thus, our system is composed of three types of agents: the Supervisor Agent (SA) which is responsible for the calculation of the reliability threshold of the nodes, using the "decision tree" technique. The Replicator Agent (RA) has as a main role the manager of the provisional replication of Cloud Service alternatives. The Monitoring Agent (MA) which can detect changes in the reliability of different nodes.

1 Introduction

Cloud computing is widely treated as a promising information technique (IT) infrastructure because of its powerful functionalities (Ying-Si and Qing-An, 2018). It is an emerging platform which provides computing and storage to the end-users as a service. This paradigm has gained more and more popularity, every day, in the research community and commercial world. There are three main service layers of cloud computing namely, software as a service (SaaS), platform as a service (PaaS) and infrastructure as a service (IaaS) (Gerardo and de Assis López-Fuentes, 2014).

Most of real-time systems are critical to security and require high reliability and a very high level of fault tolerance for their execution. The primary reasons that can accelerate failures and faults in this type of systems are dynamic execution environments, frequent updates and upgrades, online repairs, etc.

As the cloud computing systems continue to grow in scale and complexity, it is important to provide an effective method to guarantee their reliability and stability. Fault tolerance methods can be reactive or proactive. Reactive fault tolerance methods are, mainly, used to minimize the influence of faults on time and monetary costs. The most used reactive methods

are: replication and check pointing. However, proactive methods are probabilistic, and they are applied to predict to a possible extent the faults of virtual machines prior to their occurrence. The main goal of these methods is trying to avoid the occurrence of failures and then avoid recovery procedures of the reactive methods (Amoon, 2016).

In this paper, we propose a proactive fault-tolerance model that allows to make an appropriate decision in the event of a breakdown. The proposed model is based on supervised learning and multi-agent systems (MAS).

A multi-agent system (MAS) is formed by many agents connected to achieve the desired objectives specified by the design. Usually, in a multi-agent system, agents work on behalf of a user to achieve given goals (Arafat and Elbouraey, 2016). Thus, our system consists of three types of agents: Supervisor agent which is responsible for calculating the threshold through the reliability weight assigned to each processing node. The increase and decrease in reliability depend on the virtual machines to produce the results within the given time. The Supervisor Agent uses a metric to assess reliability. The metric evaluates the level of reliability of the node compared to the calculated threshold (Malik and Huet, 2011). The SA uses the decision tree, also called classification tree, to classify the nodes in two reliable and unreliable classes. It allows to determine the Trusted Nodes. The Replicator Agent (RA) is responsible for managing dynamic percentage for provisional replication of cloud service alternatives. Replication-based fault tolerance is about creating multiple copies of processes on different machines. When a service fails, it is replaced by one of its copies (replicas). Replication in a distributed system has several advantages. It provides better reliability by increasing the availability and reliability of data. Indeed, it allows avoiding the loss of cycle of execution in the event of failure by masking the faults (Ndiaye, 2013). The monitoring agent can detect changes in reliability of the different nodes which affects the percentage of replication.

The rest of the paper is organized as follows: Section 2 presents an illustration of some related work. Section 3 describes the problem, the proposed architecture and the functionality of the proposed model including the agent's behaviors. In Section 4, we introduce a case study to illustrate the functionality of our model. In section 5, we conclude the paper by a conclusion and some future work.

2 Related work

In a Cloud Computing environment, it is essential to ensure its reliability and robustness. To achieve these goals, faults should be assessed and handled effectively. Several fault detection methods, architectures and models have been proposed to increase fault tolerance competency of the cloud.

In (Rajesh and Kanniga Devi, 2014), the authors proposed a model to analyze how the system tolerates defects and decide based on the reliability of the processing nodes, i.e. the virtual machines. If the virtual machine can produce a correct result within the delay, its reliability increases, and if it fails to produce the result in time or the correct result, its reliability decreases. If the node continues to fail, it is deleted, and a new node is added. There is also a minimum level of reliability.

The model proposed in (Malik and Huet, 2011) is a Time Stamped Distributed fault tolerance model. This model incorporated the concept of time stamping with the outputs. All these models have been defined for real time systems based on standard computer architecture.

The authors in (Deepa and Ramachandran, 2015) delivered an intelligent data backup algorithm called a seed block algorithm. The purpose of the SBA proposal is to recover the files in case of cloud destruction or deletion of the cloud file. The major advantage of SBA is to take the minimum time for the recovery process.

In (Bashir et al., 2016), the authors provided an optimized approach to fault tolerance when a model is designed to tolerate faults based on the reliability of each compute node (virtual machine) and can be replaced if performance is not optimal. The preliminary test of the proposed algorithm indicated that the rate of the increase in the success rate exceeds the decrease of the failure rate and it also considered the forward and backward recovery using various software tools.

In (Lee et al., 2011), the authors proposed a fault tolerant and recovery system called FRAS system (Fault Tolerant and recovery Agent System) which uses multi-agent in distributed computing systems So, this is an agent-based system consisting of four types of agents.

Different from above approaches, the model proposed in this paper is based on both Supervised Learning techniques and MAS. So, we want to take advantages from these works such as MAS approach. This approach deals with the cloud services (SaaS, PaaS, IaaS) as external entities to create, call and manage them.

3 Proposed model

The cloud computing is the new promising technology for IT industry, it provides for any computer developer a set of services which allows to store and to perform big calculations. It is well known that Fault Tolerance constitutes a challenging and an open issue in Cloud computing. Fault Tolerance issue consists of fault detection, backup, and failure recovery. So, all Fault Tolerance approaches are based on the use of redundancy. Our research domain is a part of the fault tolerance in a cloud environment. Our goal is to propose a new method based supervised learning and MAS approaches.

3.1 overall architecture of our system

The general architecture of our system is illustrated in figure 1. Our architecture has different components, contains multiple agents' types and other necessary system components, they are:

1. ***Agent layer***: This layer contains all the agents that interact with each other, they can execute a service in a smarter way, where they transfer and obtain the messages. We have three types of agents:
   a. **The supervisor agent** which is responsible for calculating the reliability threshold of the nodes, using the "decision tree" technique. It can determine the reliable nodes.
   b. **The replicator agent** is responsible for managing the dynamic percentage of provisional replication of cloud service alternatives.
   c. **The monitoring agent** can detect changes in reliability of the different nodes which affects the percentage of replication.
2. ***Server layer***: it is composed of a set of nodes, which is the hardware part of the cloud.

3. ***Data Center layer***: This is the Information Sources Layer. It is a class that can receive queries, process them, and return the results. This layer groups all the multiple sources that provide the basic data of the system.

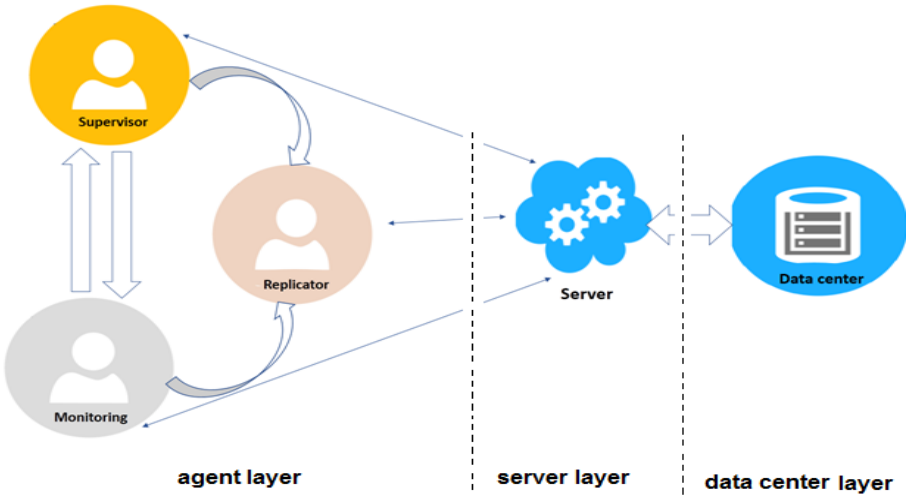


FIG. 1 – *General architecture.*

3.2 Functionality of the proposed model

As mentioned above, our system is composed of three types of agents: the supervisor which is responsible for the calculation of the reliability threshold of the nodes, using the "decision tree" it makes it possible to determine the reliable nodes. The replicator agent is responsible for managing the dynamic percentage of provisional replication of cloud service alternatives. And, the monitoring agent can detect changes in reliability of different nodes which affects the percentage of replication. The main steps of our model can be described as follows:

Step 1: The supervisor agent calculates the threshold.

Step 2: The supervisor agent decides the reliable and unreliable nodes (using the decision tree).

Step 3: The supervisor agent must associate to each class of nodes (reliable and unreliable) a monitoring agent.

Step 4: The supervisor agent sends the threshold to the monitoring agent.

Step 5: The supervisor Agent groups the services with the same features.

Step 6: The supervisor agent chooses a dynamic percentage of replication of the reliable nodes (according to the cost of transfer).

Step 7: The Supervisor Agent Sends the percentage and Groups of Services to the Replicator Agent

Step 8: The replicator agent replicates services based on the dynamic percentage.

Step 9: The monitoring agent checks the reliability (depending on the threshold).

Step 10: The monitoring agent sits eventual changes in the reliability of nodes.

Step 11: in the case where the reliability of the nodes changes, the supervisor agent must do an update of its classes.

Step 12: in the case where the reliability of the nodes changes, the replicator agent must eliminate all provisional replications.

Step 13: at the end of the request, the replicator agent must eliminate all provisional Replications.

We represent the functionality of our model by the sequence diagram illustrated below (see fig.2):

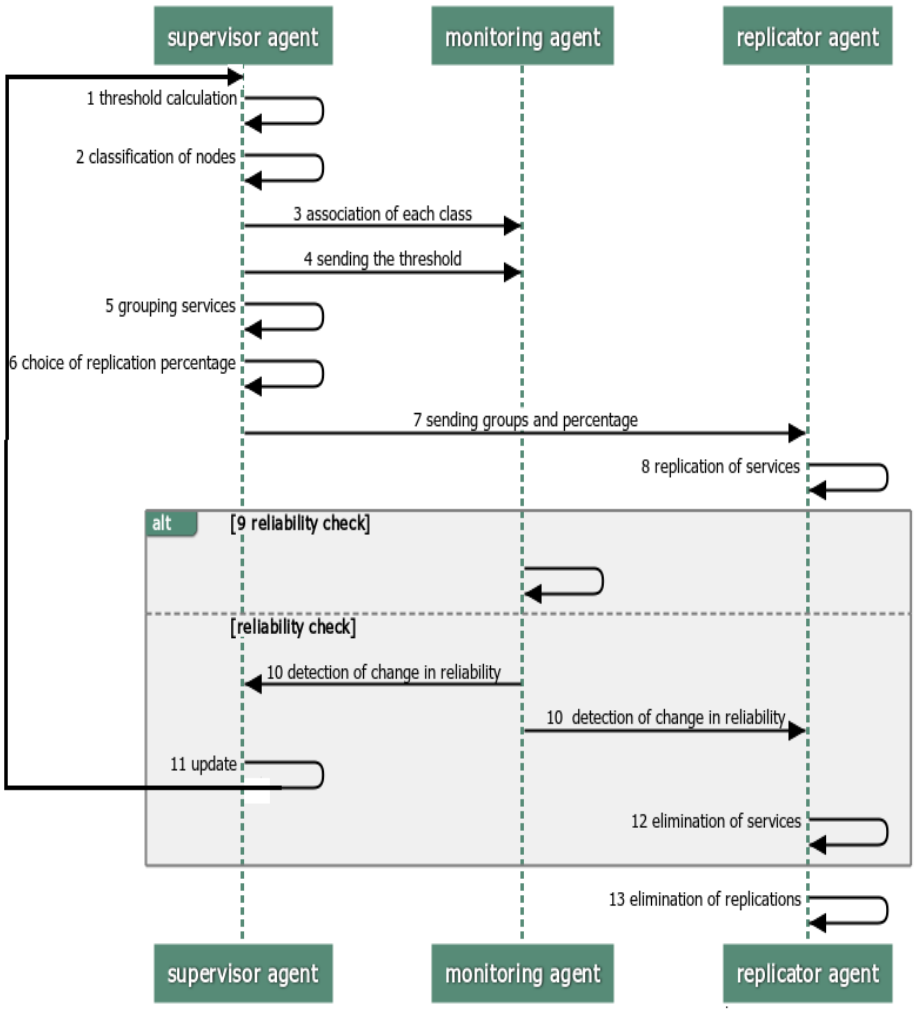


FIG. 2 – *Sequence diagram of the proposed model.*

3.3 Agents behavior

Our proposed model has three types of agents, as shown in fig 1. In this section, we present briefly the functionality of each type of agents.

3.3.1 Supervisor Agent

As discussed above, this agent is responsible of the threshold calculation. The threshold is calculated by the mean of the reliability weights of each node. It uses the decision tree to classify the nodes into two classes reliable and unreliable nodes. Also, it groups the services which have the same features. Then, it chooses a dynamic percentage of replication of the reliable nodes (according to the cost of transfer in term of time):

Cost of transfer= size of service* transfer time.

In the case where the reliability of the nodes changes, the supervisor agent must update its information. (see. Fig.3)

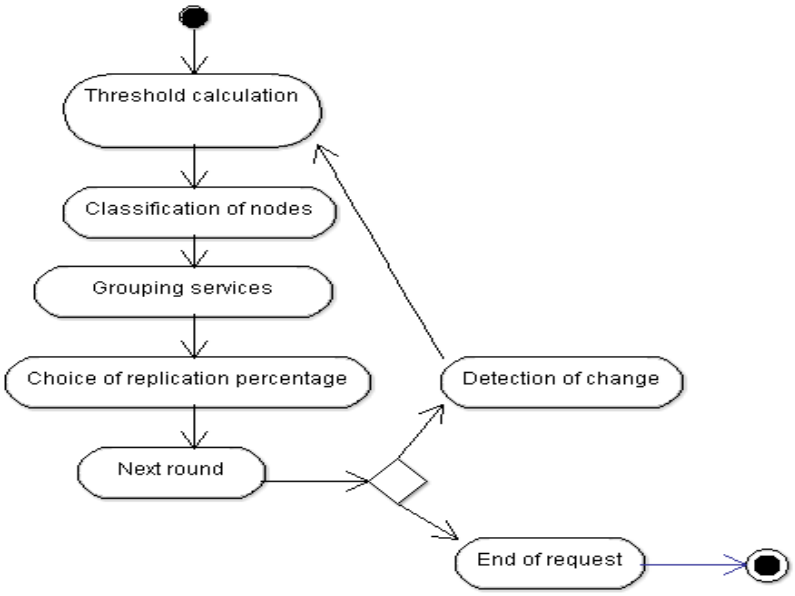


FIG. 3 – *The supervisor agent behavior*

In order to classify nodes, we use the below algorithm. This later has as input the different nodes which compose our system and a given threshold. As output, we will get a decision tree which allows to take an appropriate decision. So, we will get two classes of nodes: reliable and unreliable ones.

```
Generic learning algorithm:
 Input: threshold; the nodes
 begin
  Initialize to the empty tree; the root is the current node
  repeat
  Decide if the current node is reliable
  If the node is reliable then
   Assign it to a reliable class
  if not
   assign an unreliable class
  End if
 Move to the next node unexplored if there is one
 Until we get a decision tree
end
```

3.3.2 Replicator Agent behavior

When this agent receives the percentage and service groups from the supervisor agent, it replicates services based on the dynamically received percentage. In the case where the reliability of the nodes changes or at the end of the request, the replicator agent must eliminate all provisional replications made before. (see. Fig. 4)

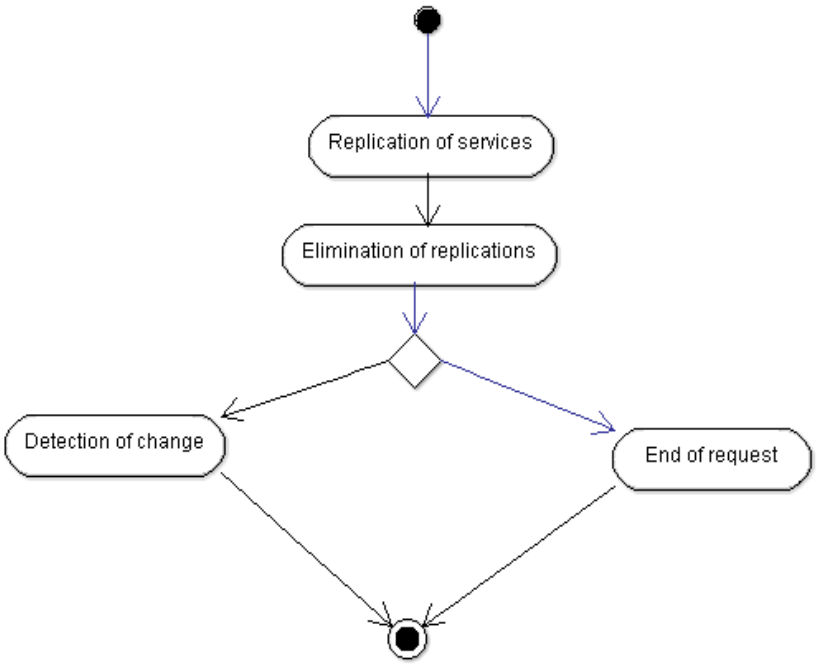


FIG. 4 – *The replicator agent behavior*

3.3.3 Monitoring Agent behavior

This agent checks reliability (depending on the threshold) and then inform the supervisor agent in order to take into consideration these changes. (see. Fig. 5).

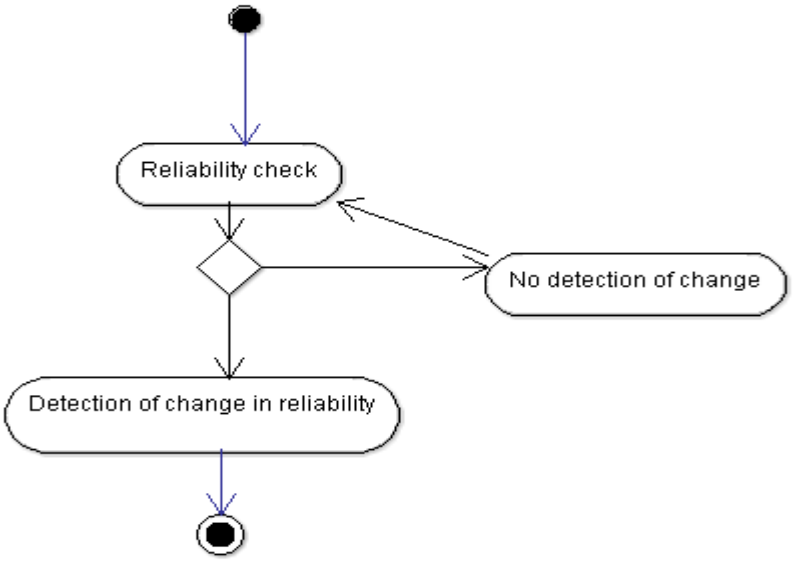


FIG. 5 – *The monitoring agent behavior*

4 Case study

We consider 10 alternatives of the requested service; these alternatives are distributed over 10 nodes; it is assumed that these alternatives are ordered (According to system requirements and user preferences) (see Table 1):

	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10
S1	X									
S2				X						
S3		X								
S4							X			
S5								X		
S6						X				
S7										X
S8									X	
S9					X					
S10			X							

TAB. 1 – *Distribution of services over nodes*

The supervisor agent calculates the threshold according to formula 2. It must first calculate the reliability weight of each node (see Table 2):

Nodes	Total request	Successful responses	Reliability
N1	7	5	71,42%
N2	7	6	85,71%
N3	10	7	70%
N4	2	2	100%
N5	8	2	25%
N6	5	1	20%
N7	4	3	75%
N8	9	8	88,88%
N9	3	1	33,33%
N10	6	3	50%

TAB. 2 – *Reliability of each node*

According to table 2:

- total request. total number of requests received in instant t1
- successful responses. number of successful requests
- reliability. (successful responses/total request) *100

  Example: reliability of N1 = (5/7) *100 = 71.42 %

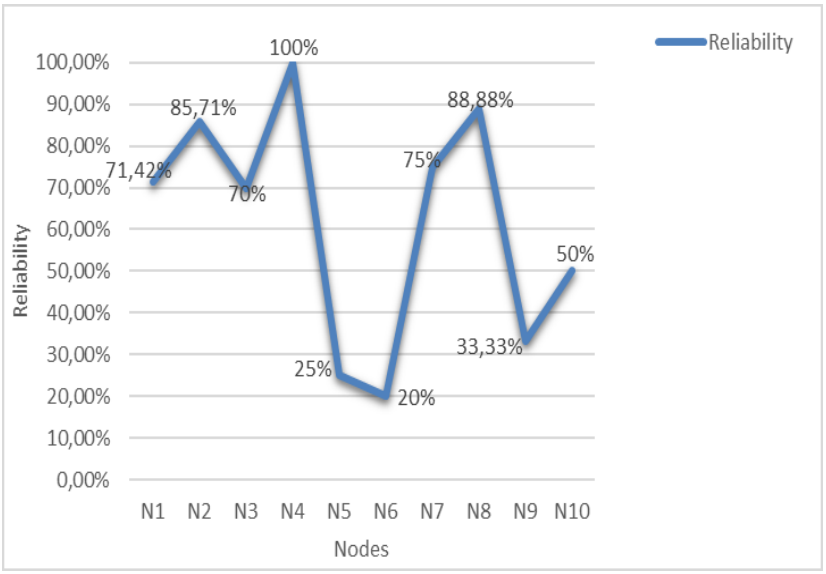


FIG. 6 – *Reliability of each node*

Threshold calculation

$$\text{threshold} = \sum \text{of reliability weights / number of nodes} \quad (1)$$

Threshold= (RN1+RN2+RN3+RN4+RN5+RN6+RN6+RN7+RN8+RN9+RN10)/10
Threshold= (71.42+85.71+70+100+25+20+75+88,88+33,33+50)/10
Threshold= 61.93 %
The Supervisor agent decides the reliable and unreliable nodes using the decision tree to classify the nodes into two classes: reliable and unreliable.

So, we obtain the following classes: **reliable class** is composed of {N1, N2, N3, N4, N7, N8} and **unreliable class** is composed of {N5, N6, N9, N10} (see Table 3).

```
Execution of algorithm for the first node N1:
Input: threshold= 61.93%; N1, N2, N3, N4, N5, N6, N7, N8, N9, N10;
 begin
Initialize to the empty tree; the root is N1
 repeat
Decide if N1 is reliable
If N1 is reliable then
  assign it to a reliable class
  if not
  assign an unreliable class
  End if
 Move to the next node unexplored if there is one
 Until we get a decision tree
end
```

Nodes	Reliable	Unreliable
N1	X	
N2	X	
N3	X	
N4	X	
N5		X
N6		X
N7	X	
N8	X	
N9		X
N10		X

TAB. 3 – *Classification of Nodes*

After that, the supervisor agent must associate to each class of reliable and unreliable nodes a monitoring agent, and it will send the threshold (61.93%) to each agent.

The supervisor agent chooses a dynamic percentage of replication of the reliable nodes. In the first iteration it chooses a percentage of 10%, then it sends the percentage and service

groups to the replicator agent. This later replicates Services Based on the Dynamic Percentage (10%) according to the formula 3.

$$\text{Services to replicate} = \text{number of services} * \text{percentage}; \quad (2)$$

So, we obtain: Services to replicate = (10 services) * 10%, it means Services to replicate is 1. So, the replicator agent will replicate the S1 service; The provisional replication of S1 is (S'1).

The monitoring agent verifies the reliability (depending on the threshold). After two iterations, the monitoring agent detects a change of reliability in the nodes N1 and N6. In N1, the total request is equal to 7 and the Successful responses equal to 4. So, the reliability of N1 is decreased to 57.14% (RN1 < threshold). In N6, the total request equal to 5 and the Successful responses is 4, so, the reliability of N6 is increased to 80% (RN6 > threshold).
After this change in the reliability of the nodes, the supervisor agent must do an update (recalculate the threshold). And, the replicator agent must eliminate all provisional replications such as (S'1) and so on.

5 Conclusion

Failure is one of the most important problem in Cloud Computing that we need to tackle, in order to achieve a high performance and insure the reliability. In this paper, we presented a new model for Fault Tolerance based on decision tree method, replication and multi-agent systems, to achieve the desired goals. In particular, we provided the architecture of our system. Furthermore, we explained the functionality of each component of this architecture, and of our model.

In future, we plan to more detail the functionalities of the different types of agents that make up our architecture. Since, the proposed classification is based on decision trees lacks more or less precision, and in order to make it more precise, we aim to introduce a fuzzyfication factor of 1 or 2 degrees (i.e. Fuzzy Type I or Type II) to have more certainty of decision-making. Also, we plan to introduce Circuit Breaker Pattern (Nygard, 2007) in our model because leveraging patterns like it can mitigate the corresponding loss to the lowest level (Balalaie et al., 2015). We aim, also, to evaluate our proposed model through some real systems case studies and repeat experiments with more queries (a large number of queries). Currently, we are working on the implementation of this model on a solid cloud computing and multi-agent platforms.

References

Amoon, M. (2016). Adaptive Framework for Reliable Cloud Computing Environment. IEEE Access, 4: 9469-9478.

Arafat, Y. and F., Elbouraey (2016). A Survey on Fault Tolerant Multi Agent System. International Journal of Information Technology and Computer Science, 9: 39-48.

Balalaie, A., A., Heydarnoori, and P., Jamshidi (2015). Migrating to Cloud-Native Architectures Using Microservices: An Experience Report. In: European Conference on Service-Oriented and Cloud Computing (ESOCC'15) : 201-215.

Bashir, M., M., Kiran, A., Irfan-Ullah. and M., Kabir (2016). Optimizing Fault Tolerance in Real-time Cloud Computing IaaS Environment. In: 4th International Conference on Future Internet of Things and Cloud, Vienna, Austria.

Deepa, S. and G., Ramachandran (2015). Disaster Recovery System Using Seed Block Algorithm in Cloud Computing Environment. International Journal of Advanced Research in Computer Science and Software Engineering, 5(2):309-314.

Gerardo, G. and F., de Assis López-Fuentes (2014). A storage service based on P2P cloud system. Research in Computing Science, 76: 89–96.

Lee, H., D., Park, H., Yu and G., Lee (2011). FRASystem: fault tolerant system using agents in distributed computing systems. Cluster Computing, 14 (1): 15-25.

Malik, S., and F., Huet (2011). Adaptive Fault Tolerance in Real Time Cloud Computing, IEEE World Congress on services, Washington, DC, USA.

Ndiaye, N.M. (2013). *Techniques de gestion des défaillances dans les grilles informatiques tolérantes aux fautes*. Thèse de doctorat, Université de Pierre et Marie Curie, Paris.

Nygard, M. (2007). *Release It!: Design and Deploy Production-Ready Software*. Pragmatic Bookshelf, Raleigh, North Carolina Dallas, Texas.

Rajesh. S. and R., Kanniga Devi (2014). Improving Fault Tolerance in Virtual Machine Based Cloud Infrastructure. International Journal of Innovative Research in Science, Engineering and Technology, 3:2163-2168.

Ying-Si, Z. and Z., Qing-An (2018). Secure and Efficient Product Information Retrieval in Cloud Computing. IEEE Access, 6: 14747-14754.

Résumé

Le Cloud Computing est un ensemble de services répartis sur l'internet. Ces derniers offrent un système en temps réel et un environnement virtuel avec d'énormes capacités de calcul. La plupart des systèmes en temps réel sont essentiels à la sécurité et nécessitent une grande fiabilité et un très haut niveau de tolérance aux pannes pour leur exécution. Dans ce travail, nous proposons un modèle de tolérance aux pannes qui permet de prendre une décision appropriée en cas de panne. Le modèle proposé est basé sur des techniques d'apprentissage supervisé et des systèmes multi-agents. Ainsi, notre système est composé de trois types d'agents: l'agent superviseur qui est responsable du calcul du seuil de fiabilité des nœuds, en utilisant l'arbre de décision. L'agent de réplication a pour rôle principal la gestion de la réplication provisoire des alternatives du service Cloud. L'agent de veille qui peut détecter les changements de fiabilité des différents nœuds.

Opinion and emotion analysis through the linked data lens

Leila Moudjari*, Karima Akli-Astouati**

*l.moudj11@gmail.com
**kakli@usthb.dz
RIIMA Laboratory, USTHB, Algiers, Algeria,

Abstract. The immense contribution of the social web has greatly motivated researchers. This led to the emergence of techniques that have proven their effectiveness in customised opinion and emotion modeling related to applications such as NLP, automatic learning etc.... On the other hand, when it comes to interoperability and a unique encoding of opinions and emotions, there are some weaknesses. This prompted a new research direction that combines opinion analysis works with those of "Linked Data". In this article, we will expose different solutions and projects by presenting some limitations. The reasons why we also believe that linked data is important and how we would like to conduct this research are also detailed in this article.

1 Introduction

Opinion and emotion analysis drew the researchers' attention from different fields since early 2000s and it became one of the Natural Language Processing (NLP) most active research areas. It was quickly sought after in the data mining field, web mining and information retrieval (Liu, 2012). The NLP community claims that emotions are briefer than sentiments and opinions are personal interpretations and not as emotionally charged as sentiments [1].
The social web, the source number one for subjective data. It interlinks people from different backgrounds and allows them to share their thoughts and express their feelings. They transmit information about the emotional state of the author, his or her judgment or assessment of a certain person or subject, or premeditated emotional communication (the emotional effect the sender wants to have on the receiver). Message receivers have the ability to sense senders' emotions through verbal cues such as emotional words and language markers (e. g., lexical or syntactic coding of emotions) as well as non-verbal paralinguistic cues (e. g., emoticons, feedback, reactions or sharing their messages).

The web 3.0 is a collection of interlinked datasets also called "Linked Data".
So what is linked data and why were researchers interested in applying it in this field?

The rest of the document is organized as follows, in section 2 we provide an overview of opinion and emotion analysis. Section 3 introduces linked data technologies and their use for opinion and emotion analysis. Section 4, presents projects related to the field and discusses their strengths and weaknesses. Conclusions and future work are outlined in section 5.

1. sewaproject.eu/files/fbbe2939-91c0-42bd-c87b-d2e98d45cbac.pdf

2 Opinion and emotion analysis so far

According to dictionaries sentiment, emotion and opinion are synonyms. The Cambridge online dictionary [2] defines an emotion as a strong feeling or sentiment, a sentiment as a thought, opinion, or idea based on a feeling about a situation, or a way of thinking about something, but an opinion is a thought or a belief about something or someone or a judgment that we make.
Sentiment analysis is a semantic analysis that aims to extract and analyze subjectivity out of user generated content on the social web, generally textual sources(Arti Buche, 2013).

It is an approach that relies mainly on text mining and semantic analysis to determine the "position" of the individuals studied with respect to a brand or event. It is based on other elements as well, such as the use of the "emoticons and emojis", likes, interactions, voice analysis or even facial expressions.

When based on textual searching, emotion analysis can include verbatim statements from social networks, blogs, microblogs, review sites, forums, etc. However, it is considered more vague and general since it is limited to labeling the document with one of the three classes (positive, neutral, negative) with respect to someone or something. Emotions analysis is similar to sentiment analysis. Nevertheless, it goes beyond a binary/ternary distinction. Since it provides the emotional class (neutral, anger, disgust, fear, joy, sadness, surprise). Some sources claim that they are the same, others claim that sentiment analysis is a part of the emotion analysis.
However, opinion analysis takes the process even further. It depends not only on the situation in which the opinion is expressed, but also on the person expressing the opinion. There are studies that consider all as the same, and are called opinion mining, sentiment analysis, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc. Other studies consider sentiments as the emotional part of opinions. Opinion mining provides an automatic analysis of feedbacks and a tool that answers questions such as "What do people think of the product?" (Arti Buche, 2013).

An opinion expressed could be subjective or objective (neutral). If it is subjective, it means the opinion carries an emotion. If we want to know its polarity we perform a sentiment analysis, if we want to go beyond this scope then we are talking about an emotion analysis. Therefore, we will accept the expression *"Opinion and emotion analysis"* to cover all meanings of the terms used, allowing us to detect the subjectivity and emotional state of a subject studied and understand the opinion expressed.
Several approaches were proposed in the literature, which facilitated several daily tasks.

2.1 Opinion and emotion use cases

Since the advent of the internet, data has become crucial. Opinion and emotion analysis has raised the bar even higher by simplifying multiple tasks. In this section, among several we mention;

— Huge volumes of opinion-rich data are published on social media at an unprecedented rate, opinion and emotion analysis facilitates the analysis of humanly impossible data sources to analyze (Cheng et al., 2017).
— Decision making, on a personal or business level.

2. https://dictionary.cambridge.org/dictionary/english/

— Rapid analysis of customers' feedbacks for branding or predicting future needs as well as trend analysis with forecasts for future trends1.
— Often used in monitoring social networks, tracking customer attitudes towards brands, politicians etc 1.
— Detection of groups and violent social movements (Zimbra and Chen, 2012).
— Prediction poll ratings.(Hu et al., 2013)
— Forecasting box office revenues of movies(Asur and Huberman, 2010).

2.2 Proposed works

Opinion and emotion analysis is an established field of research and a growing industry(Liu, 2012). Some recommended methods are automatic learning, NLP and data mining.
(Ravi and Ravi, 2015) presented a review of works from 2002 to 2015, (Soleymani et al., 2017), a survey on Multimodal Sentiment Analysis. Nevertheless, in this work we present some of the most used approaches and techniques and explain the interest in the use of linked data to solve the presented problem. As Ravi suggested, sentiment and opinion analysis approaches can be classified in three categories: machine learning based, lexicon based or hybrid approaches. Under lexicon based approaches, one can use either dictionary or corpus based approaches, where dictionaries are created with/without ontologies. Corpus approaches rely on probabilities, based on the occurrence of the overall sentiment in a document.
(Mataoui et al., 2016) proposed a Lexicon-Based Sentiment Analysis Approach for the Vernacular Algerian Arabic. A manually annotated dataset and three Algerian Arabic lexicons (negation words, intensification-words and a list of emoticons with their assigned polarities, and a dictionary of common phrases of the ALGD) were used for polarity computation.

Machine learning methods can be divided into supervised and unsupervised methods.
The supervised methods train a robust sentiment classifier generally from manually labeled training data. Some of the commonly applied classifiers for supervised learning are Decision Tree (DT), Support Vector Machines (SVM), Neural Network (NN), Naïve Bayes (NB), and Maximum Entropy(ME) ((Dixit et al., 2017), (Bilal et al., 2016), (Nasser et al., 2016)).
Sentiment Polarity Identification on Arabic, Algerian Newspaper Comments (SIAAC) a proposed approach by Rahab in (Rahab et al., 2017) to classify Arabic comments from Algerian Newspapers. SIAAC is a corpus dedicated to Arabic sentiment analysis. Using supervised learning classifiers SVM and NB, tests showed that both classifiers gave competitive results in term of precision, also the use of bigrams increased the results in the two models.

Unsupervised learning approaches are usually mixed with a supervised learning process, such as a Learning Word Vectors for Sentiment Analysis, a mix of unsupervised and supervised techniques to learn word vectors capturing both semantic and sentiment similarities between words in order to perform a document level polarity classification (Maas et al., 2011).
(Hu et al., 2013) presents an unsupervised sentiment analysis based on "emotional signals" such as Emoticons and product ratings, by providing a unified way to model two main categories of emotional signals: emotion indication and emotion correlation. (You et al., 2016) is another mix of supervised and unsupervised learning techniques that combines visual and textual Sentiment Analysis of Social Multimedia. A cross-modality consistent regression (CCR) model is proposed, which is able to utilize both the visual and textual sentiment analysis techniques, a fine-tuned convolutional neural network (CNN) for image sentiment analysis and a trained paragraph vector model for textual sentiment analysis. (Cheng et al., 2017) is a more

recent Unsupervised Sentiment Analysis with Signed Social Networks, that studies sentiment signals in textual terms and user interactions to determine the polarity of a text.
One of the biggest issues in the opinion and emotion analysis field is the fact that data is domain-based, having applications for specific domains can be beneficial. For a better exploitation, we need tools linking such applications to provide better understanding of data and a stronger support for domain and context to avoid ambiguities. For instance "Honey is not bad!", is it the food substance produced by bees? HONEY MOTORS the used car dealer in Dubai? or Honey Grove, the city in Texas? Let us take the work (Yerva et al., 2010), an effort to distinguish tweets about companies such as Apple, the company or the fruit.
(Abdulla et al., 2014) talked about the limitations of the lexicon-based approaches. He pointed out that the accuracy level is low, but it could be improved by "expanding the lexicon with a dialect-specific and domain-specific content, performing more thorough experimentations on bigger and diverse corpora, better handling of the very difficult cases such as sarcasm, etc".
We also believe that adoption of informal speech -used in everyday conversations on the social platforms- alongside the formal one can boost the comprehension of the analyzed opinions.
Xavier in (Glorot et al., 2011), talked about the problem of domain adaptation for sentiment classifiers. Using a deep learning system, the proposed protocol relies on Stacked Denoising Autoencoder (SDA) with rectifier units to extract features in a supervised manner. And a linear classifier SVM with squared hinge loss is trained on a labeled data of one domain and tested on other domains. The system proved its effectiveness compared to others.
(Shoukry and Rafea, 2012) presented a sentence-level Arabic sentiment classification (positive/negative) using a supervised learning process NB and SVM.
(Baly et al., 2017), proposed the first Arabic Sentiment Treebank (ArSenTB) that is morphologically and orthographically enriched. A key requirement for Recursive Neural Tensor Networks (RNTN: a deep learning network) is the availability of a sentiment Treebank; a collection of syntactic parse trees annotated for sentiment, currently only exists in English. The model achieved significant improvements at the sentence classification level better than SVM.
(Al-Smadi et al., 2017), a comparison between Recurrent Neural Network and SVM for aspect-based sentiment analysis of Arabic Hotels' reviews. SVM outperformed RNN in the identification, extraction and polarity calculation level. But RNN gave a better execution time.
A deep learning based framework for text sentiment classification in Arabic was presented in (Al Sallab et al., 2015). It consists of four architectures: three based on Deep Belief Networks and Deep Auto Encoders (DAE). The fourth, based on the Recursive Auto Encoder (RAE), tackles the lack of context handling in the others. DAE model gave a better representation of the input sparse vector. However, RAE was better, although it required no sentiment lexicon.
(Al-Radaideh and Al-Qudah, 2017) used the Rough Set theory for feature selection from Arabic tweets. Results indicate that the method can enhance the overall accuracy and reduce the number of used terms which will lead to a faster classification process, especially with a large dataset.

Opinion and emotion analysis field can benefit from an accuracy enhancement. Some rare works considered parameters such as gender to improve sentiment classification (Volkova et al., 2013) and cyber-bullying detection (Dadvar et al., 2012). Other parameters can be considered to further this enhancement.
There are more works in the literature. Nevertheless, opinion and emotion analysis is still considered a quite recent field and the door is open for new possibilities and suggestions. The next

section presents a new direction in this field, that opted for Linked Data technologies.

3 Linked data and the analysis of opinion and emotion

One of the factors that makes it difficult to exploit opinions is that emotions and subjectivity are highly sensitive to the context and highly dependent on the field. For example, the ice cream melted (negative) and the cheese melted (positive) in the food sector(Liu, 2012).
Due to the dependence of the analysis on the context and domain of the opinions and emotions expressed, a semantic network linking entities and features to their exact meaning in a phrase can help tremendously overcome this issue. It will also provide a uniform semantic interface for users and entities will have a uniform URI nomenclature according to the related data conventions.
The web of data is a collection of interrelated datasets on the web. To assist the use of such resources, the Semantic Web technologies help query databases, discover new relationships due to automatic procedures using the data itself or vocabularies (ontologies). Linked data designate a list of rules to respect so that the published data is standardized and structured in order to be interlinked and better exploited using the Semantic Web technologies [3].

In the following we cite some reasons that explain our interest in linked data for the analysis of emotions and opinions.

- The different formats for representing data in the opinion and emotion analysis are heterogeneous, generally they are modeled according to the need of the application despite the existence of several standards1.
- In social media, it is time and labor consuming to obtain sentiment labels, social media data is often unlabeled (Cheng et al., 2017), linked data can unveil and link already labeled published data to help advent the research.
- Although this area has fostered innovation in a range of SMEs (Small and medium-sized enterprises) with strong semantic and linguistic skills, the language resources developed remain limited to their clients. The main reason for this is the fear of losing competitiveness or missed returns on research / business investment [4].
- Easily accessible social platforms such as Twitter provide data from people of different origins, different languages and various topics discussed. To capture the meaning of tweets by considering all these factors is beyond the capabilities of existing formats for sentiment analysis, which hinders the emergence of applications that make deep sense of data. The use of linked data can help better exploit such information 1.
- Expressions such as: "it's cooooool" and "good n8 :)", are not standard, but are more acceptable and frequently used in social media (Hu et al., 2013), linked data can unveil the standard format of these expressions and their meaning therefore simplifying the opinion detection.
- The richness of data available on social networks beyond pure text can be exploited to offer a deeper understanding of user-generated content, and ultimately train a system to look for more targeted characteristics, process and categorize content. Apart from the actual content, it is also the context in which it was created that can serve as a rich

3. www.w3.org/standards/semanticweb/data
4. http://eurosentiment.eu/section/project/

source of information and be used to generate more powerful data analysis and make smarter business decisions1.

— There are commercial solutions, such as Lexalytics, Sentimetrix, Hootsuite, Klout and Tweetreach. Though they are quite generic, do not deal with the multilingual aspect and do not adapt to the enormous volumes of data to be processed. Nor to the integration of emotional analysis observations between data sources and/or modalities at a significant level and are not trained for domain-specific terminology, idioms and characteristics1.

— Domain-specific resources for multiple languages are potentially valuable, but not shared. Sometimes because of intellectual property and license considerations, but often because of technical reasons. Including interoperability, there are some initiatives such as Text Encoding. But there is not yet a widely accepted global solution for integrating and combining heterogeneous linguistic resources from different sources(Iglesias and al, 2017).

Using linked data is an interesting turning point for opinion and emotion analysis. Not only it can help standardize modeling of emotions and opinions, but also simplify the extraction of subjective data to ultimately reach a better understanding of opinions, while taking into consideration context, cultural background and other factors.
Some techniques have been implemented using a set of tools as presented in the following.

3.1 Tools

Here is a list of tools recommended by W3C, since it promotes web standards. However, there are others (Chinese Emotion Ontology(Hu et al., 2014), EMOTIME [5], Senti-TUT [6]).

— **WordNet Affect**, an affective lexicon providing affective polarity of words. It is based on WordNet, adding a new set of tags to a selection of synsets to annotate them with affective information (Strapparava et al., 2004).

— **SenticNet** a knowledge base that provides a set of semantics, sentics(emotion categorization values expressed in terms of four affective dimensions (Pleasantness, Attention, Sensitivity, and Aptitude)), and polarity (floating number between -1 and +1) associated with 100,000 natural language concepts [7].

— **EmoSenticNet** is a lexical resource that assigns six WordNet Affect emotion labels to SenticNet concepts. It is considered as an expansion of WordNet Affect emotion labels to a larger vocabulary [8].

— **Emotive Ontology** the work of Sykora et al. (Sykora et al., 2013) in 2013, proposed a mechanism to extract emotions from informal messages.

— **NLP Interchange Format (NIF)** an RDF/OWL-based format, it aims to achieve interoperability between NLP tools, language resources and annotations. NIF consists of specifications, ontologies and software. It defines a semantic format and an API for improving interoperability among NLP services. NIF can be extended via vocabularies modules(Hellmann et al., 2013). It uses Marl for sentiment annotations(Sánchez-Rada and Iglesias, 2016).

5. https://github.com/luca-m/emotime
6. http://www.di.unito.it/ tutreeb/sentiTUT.html
7. http://sentic.net/
8. https://www.gelbukh.com/emosenticnet

— **Provenance Ontology** provides information on the entities, activities and people involved in producing the data.(Sánchez-Rada and Iglesias, 2016)
— **Lemon** (lexicon model for ontologies) built on previous work on standards for lexical resources' representation, the Lexical Markup Framework (LMF2) but extends the underlying formal model and provides a native integration of lexica with domain ontologies. It was designed to model lexicon and machine-readable dictionaries and link them to the Linked Data cloud. Lemon-based ontology lexicalisation is the use of URIs for uniquely identifying all objects defined by the lemon model (lexicons, lexical entries, etc.). Its main purpose is representing language resources for opinion and emotion analysis in a Linked Data conform way (RDF-native form), leveraging existing Semantic Web technologies (SPARQL, OWL, RIF, etc)(Buitelaar et al., 2013).
— **Marl** a vocabulary for annotating and describing subjective opinions, it models feature level opinions. Marl follows the Linked Data principles as it is aligned with the Provenance Ontology, it represents lexical resources as linked data, and has been integrated with lemon.(Buitelaar et al., 2013)
— **Onyx** a semantic vocabulary of emotions with a focus on lexical resources and emotional analysis services. It follows a linguistic Linked Data approach.

These tools help represent opinions and emotions and provide their sentiment polarity. Onyx on the other hand tried to provide a tool that combines most of these tools. It has been aligned with vocabularies such as NIF and the Provenance Ontology. Also, integrated with lemon to represent lexical entries. Aligned with EmotionML and WordNet-Affect to work with different theories of emotion. The combination of subjective information from opinion and emotion analysis with facts already published as Linked Data would make it possible to offer a wide range of new services, which would require a widely accepted representation of emotionally related data. Onyx aims to fill this gap and allow interoperability of tools and resources (Sánchez-Rada and Iglesias, 2016).

4 Projects

In order to provide standardized metrics for opinion and emotion analysis, the World Wide Web Consortium (W3C), created the forum [9] "Linked Data Models for Emotion and Sentiment Analysis W3C Community Group" to promote sentiment analysis research, where the main topics discussed are Linked Data based vocabulary and models for emotion and sentiment analysis.
This section introduces projects and approaches that adopted the linked data philosophy. We also mention the advantages and disadvantages of each one:

— **Emotion Markup Language (EmotionML)**(Schröder et al., 2011), a markup language, recommended by the W3C community and widely used since it provides a common data representation for multiple opinion and emotion analysis applications.
— **ArsEmotica (AEO)** an ontology-driven approach, encoded in OWL and incorporates, in a unifying model, multiple ontologies describing different aspects of the connections between media objects (e. g. works of art), people and emotions. It helps identify emotions in art magazines, or excerpts from them. In addition, due to the need to model the

9. https://www.w3.org/community/sentiment/

link between words and the emotions they refer, AEO integrates lemon to provide the lexical model and WordNet-Affect synset to provide the affective information. With linkage to external LOD repositories (e. g. DBpedia). The ontology contains 87 emotional concepts to categorize emotion denoting words. Each class was populated by instances from English and Spanish words. Over 450 Italian emotional words were added using MultiWordNet and WordNet-Affect (Patti et al., 2015).

— **EuroSentiment** provides a pool of shared language resources for multilingual opinion and emotion analysis. The lexicon format is based on a combination of lemon (lexical concepts), Marl (opinion/sentiment) and Onyx (emotions). The corpus format uses NIF, Onyx and Marl for subjectivity. Each entry in the lexicon is described with part of speech, morphosyntactic information, links to DBpedia and WordNet and information on the identified opinion and/or emotion. The project provides a semantically rich lexical resources, a set of lexicons and corpus, conversion tools from inherited non semantic formats, an extension of the NIF format and an API for Web services in different programming languages to simplify the development of semantic services for the analysis. It contains resources for the electronics and hotel domains in six languages (Catalan, English, Spanish, French, Italian and Portuguese)(Sánchez Rada et al., 2014).

— **MixedEmotions** [10] continues the work of EuroSentiment, including other media (image and sound) in different languages. To develop new multi-dimensional multimodal data analysis applications to analyze a more complete emotional profile of user behavior using data from mixed input channels: multilingual text data sources, A/V signal input, social media and structured data. It implements an integrated Big Linked Data platform for emotion analysis based on heterogeneous data sources, languages and modalities, using existing tools, services and approaches to track the emotional aspects of user interaction. NIF, to assign URIs to language resources, to be annotated following linked data principles. Lemon to represent lexicons and linking lexical and semantic entities to lexical forms. Marl and Onyx for annotating sentiments and emotions.

— **SEWA** (European Sentiment Analysis in the Wild [11]) develops visualization, speech processing and automatic learning tools for the automatic understanding of human interactive behavior in naturalistic contexts for spatio-temporal analysis of sentiments.

— **OpeNER** Open Polarity Enhanced Named Entity Recognition provides a set of NLP tools. OpeNER uses the annotation format KAF, with ad-hoc elements to represent the characteristics of opinion and emotion. The results of the project include an annotated corpus of reviews and a Linked Data node that displays this information (Bosma et al., 2009). It offers a usable language analysis tool for six languages for applications such as Reputation Management and Information Access. The project focus was on a generic application domain than was adapted to the Tourism domain (Agerri et al., 2013).

— (Charfuelan and Steiner, 2013) presented a framework for synthesis of expressive speech based on MARY TTS (text-to-speech) using audiobook data and EmotionML. It creates expressive unit selection and HMM-based voices using audiobook data labeled according to voice styles. Data was split according to voice styles by principal component analysis (PCA) of acoustic features extracted from segmented sentences. EmotionML was used to represent and control expressivity in terms of discrete emotions/opinions.

10. https://mixedemotions-project.eu/linked-emotions-data/
11. https://www.sewaproject.eu/

— **Sentilo** an unsupervised, domain-independent system that performs sentiment analysis by hybridizing NLP and semantic Web technologies. It makes use of affective knowledge resources such as SenticNet and SentiWordNet. It recognizes the holder, detects the topics and subtopics of a sentence expressing an opinion, links them to relevant situations and events and evaluates the sentiment expressed, outputting an RDF graph. It uses FRED (semantic Web aware machine reader) that resolves the identity of entities involved in an opinion on resources like DBpedia and WordNet (Recupero et al., 2015).

Table 1 represents some limitations of the mentioned projects and some suggestions to overcome these limitations. Adding support for popular, cross border languages such as Arabic,

Project	Limitations	Suggestions
EmotionML	- Representation of suppressed emotions, simulated, masked by another emotion.	+ Add an attribute "master" to "emotion" tag to represent the explicit emotion. And use another emotion tag to represent the masked emotion($master = false$). + Use ontologies to define and link terms and map emotion vocabularies.
AEO	- Detect emotions in multi-word expressions, that do not explicitly convey emotions, but are related to concepts that do.	+ Incorporate more language resources to operate on a multilingual level. + Integrate support for informal speech.
Euro-Sentiment	- Restrained to six languages. - Limited set of resources and domains. - Handles only textual data.	+ Add other popular, cross border languages such as Arabic.
Mixed Emotions	- Restrained to six languages. - Lacks support for informal speech.	+ Needs more multilingual and multi-domain enrichment. + Integrate support for informal speech.
SEWA	- According to [11], it has some limitations regarding social signals and image processing.	+ Needs a multilingual enrichment. + Add support for informal speech.
OPENER	- Lacks multilingual support. - Lacks support for informal speech.	+ Adopt more domains and context support to avoid ambiguities.
MARY TTS	- Lacks multilingual support.	+ boost the expressivity of the generated audio with a domain-based extension.
Sentilo	- Lacks multilingual support. - Lacks support for informal speech.	+ Exploit WordNet for a multilingual support to enlarger the user base, therefore be a reference tool for topic/subtopic detection.

TAB. 1 – *Limitations and suggestions of opinion and emotion analysis projets*

since millions of Arab speaking individuals live in Europe [12]. Would enlarge the user base on an academic or industrial level. Not just formal languages, but informal ones as well. Integrating support for informal speech would enrich the understanding of emotion and opinions. The adoption of more domains and context support can help avoid ambiguities.
If we go back to parameters that can enhance the analysis like "gender" 2.2, linked data can facilitate the process by exploiting users' data already published and link profiles, which takes us to another point "revealing fake profiles on social media". Each person has at least three to four accounts on different social platforms, detecting fake profile can help diminish cyber crimes, but most importantly avoid suspected profiles linked to terrorism.

5 CONCLUSION

Our goal was to cover main techniques and approaches proposed in the literature for opinion and sentiment analysis and to extract their weaknesses on one hand. On the other hand, we aimed at exploring the linked data field and what it has to offer to enrich this field.
The mentioned approaches were mainly based on machine learning and NLP.
Based on the researches results SVM proved effective in detecting and classifying emotions and opinions. However, there are others, such as NB, DT, EM and deep learning techniques SDA, RNTN and RNN. The main issue these techniques suffer from is the lack of support for interoperability and a single agreed-on encoding for opinions and emotions. Since the modeling is custom-ade for each application, it is difficult to find trained models or labeled data.
The lexicon-based techniques on the other hand, lack semantics and are more vulnerable against ambiguities. That could be avoided if the correct meaning is located through an ontology, providing links to external LOD repositories, such as DBpedia. This holds true when dealing with data published by users from different origins, languages and a variety of topics discussed. Due to a restricted number of words in the dictionaries or because some parameters that can influence the opinion or emotion expressed in are not considered.
The use of linked data can help overcome some of these issues. Technologies like RDF, OWL and SPARQL provide standards for publishing structured data to be interlinked and queried.
The fact that projects such as MixedEmotions, EuroSentiment are using such technologies to deal with similar problems supports our view. To guarantee the success and adoption of such approach, we need common vocabularies. Onyx and Lemon are a first step in this direction.
Also we think that the enrichment of the data by meta-data like gender or age will improve the accuracy of an emotion classifier.
Another interesting branch is the analysis of informal speech, the most used on the social web. As mentioned there are some limitations in the multilingual support, especially for Arabic data. Several studies on the Arabic sentiment analysis have been carried out in recent years. They mainly focus on Modern Standard Arabic (MSA) among which few have investigated Arab dialects. The Algerian dialect is less normalized compared to MSA. It has been enriched by many languages which resulted in a complex linguistic situation (Mataoui et al., 2016).
Popular works in this area are generally limited to sentiment classification using supervised classifiers such as SVM. To the best of our knowledge, none used linked data. Thus, opportunities for continued research are large, not only text analysis, but other data formats as well.

12. https://apaelo.com/know/view_html.php?q=Arabs%20in%20Europe&sq=Arabs-in-Europe&language=en

In a future work, we will be concentrating our efforts on the use of linked data models for emotion and opinion analysis in Arabic texts, starting with the Algerian vernacular dialect.

References

Abdulla, N. A., N. A. Ahmed, M. A. Shehab, M. Al-Ayyoub, M. N. Al-Kabi, and S. Al-rifai (2014). Towards improving the lexicon-based approach for arabic sentiment analysis. *International Journal of Information Technology and Web Engineering (IJITWE) 9*(3), 55–71.

Agerri, R., M. Cuadros, S. Gaines, and G. Rigau (2013). Opener: Open polarity enhanced named entity recognition. *Procesamiento del Lenguaje Natural* (51).

Al-Radaideh, Q. A. and G. Y. Al-Qudah (2017). Application of rough set-based feature selection for arabic sentiment analysis. *Cognitive Computation 9*(4), 436–445.

Al Sallab, A., H. Hajj, G. Badaro, R. Baly, W. El Hajj, and K. B. Shaban (2015). Deep learning models for sentiment analysis in arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pp. 9–17.

Al-Smadi, M., O. Qawasmeh, M. Al-Ayyoub, Y. Jararweh, and B. Gupta (2017). Deep recurrent neural network vs. support vector machine for aspect-based sentiment analysis of arabic hotels' reviews. *Journal of Computational Science*.

Arti Buche, Dr. M. B. Chandak, A. Z. (June 2013). Opinion mining and analysis: A survey. *International Journal on Natural Language Computing (IJNLC) Vol. 2, No.3*.

Asur, S. and B. A. Huberman (2010). Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, Volume 1, pp. 492–499. IEEE.

Baly, R., H. Hajj, N. Habash, K. B. Shaban, and W. El-Hajj (2017). A sentiment treebank and morphologically enriched recursive deep models for effective sentiment analysis in arabic. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 16*(4), 23.

Bilal, M., H. Israr, M. Shahid, and A. Khan (2016). Sentiment classification of roman-urdu opinions using naïve bayesian, decision tree and knn classification techniques. *Journal of King Saud University-Computer and Information Sciences 28*(3), 330–344.

Bosma, W., P. Vossen, A. Soroa, G. Rigau, M. Tesconi, A. Marchetti, M. Monachini, and C. Aliprandi (2009). Kaf: a generic semantic annotation format. In *Proceedings of the GL2009 Workshop on Semantic Annotation*, pp. 1–8.

Buitelaar, P., M. Arcan, C. Iglesias, F. Sanchez-Rada, and C. Strapparava (2013). Linguistic linked data for sentiment analysis. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, pp. 1–8.

Charfuelan, M. and I. Steiner (2013). Expressive speech synthesis in mary tts using audiobook data and emotionml. In *INTERSPEECH*, pp. 1564–1568.

Cheng, K., J. Li, J. Tang, and H. Liu (2017). Unsupervised sentiment analysis with signed social networks. In *AAAI*, pp. 3429–3435.

Dadvar, M., d. F. Jong, R. Ordelman, and D. Trieschnigg (2012). Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*. University of Ghent.

Dixit, A., A. K. Pal, S. Temghare, and V. Mapari (2017). Emotion detection using decision tree. *Development 4*(2).

Glorot, X., A. Bordes, and Y. Bengio (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 513–520.

Hellmann, S., J. Lehmann, S. Auer, and M. Brümmer (2013). Integrating nlp using linked data. In *International semantic web conference*, pp. 98–113. Springer.

Hu, F., Z. Shao, and T. Ruan (2014). Self-supervised chinese ontology learning from online encyclopedias. *The scientific world journal 2014*.

Hu, X., J. Tang, H. Gao, and H. Liu (2013). Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 607–618. ACM.

Iglesias, S.-R. and al (2017). Linked data models for sentiment and emotion analysis in social networks. In *Sentiment Analysis in Social Networks*, pp. 49–69. Elsevier.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies 5*(1), 1–167.

Maas, A. L., R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 142–150. Association for Computational Linguistics.

Mataoui, M., O. Zelmati, and M. Boumechache (2016). A proposed lexicon-based sentiment analysis approach for the vernacular algerian arabic. *Research in Computing Science 110*, 55–70.

Nasser, A., K. Dinçer, and H. Sever (2016). Investigation of the feature selection problem for sentiment analysis in arabic language. *Research in Computing Science 110*, 41–54.

Patti, V., F. Bertola, and A. Lieto (2015). Arsemotica for arsmeteo. org: Emotion-driven exploration of online art collections. In *FLAIRS Conference*, pp. 288–293.

Rahab, H., A. Zitouni, and M. Djoudi (2017). Siaac: Sentiment polarity identification on arabic algerian newspaper comments. In *Proceedings of the Computational Methods in Systems and Software*, pp. 139–149. Springer.

Ravi, K. and V. Ravi (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems 89*, 14–46.

Recupero, D. R., V. Presutti, S. Consoli, A. Gangemi, and A. G. Nuzzolese (2015). Sentilo: frame-based sentiment analysis. *Cognitive Computation 7*(2), 211–225.

Sánchez-Rada, J. F. and C. A. Iglesias (2016). Onyx: A linked data approach to emotion representation. *Information Processing & Management 52*(1), 99–114.

Sánchez Rada, J. F., G. Vulcu, C. A. Iglesias Fernandez, and P. Buitelaar (2014). Eurosentiment: Linked data sentiment analysis.

Schröder, M., P. Baggia, F. Burkhardt, C. Pelachaud, C. Peter, and E. Zovato (2011). Emotionml–an upcoming standard for representing emotions and related states. In *International Conference on Affective Computing and Intelligent Interaction*, pp. 316–325. Springer.

Shoukry, A. and A. Rafea (2012). Sentence-level arabic sentiment analysis. In *Collaboration Technologies and Systems (CTS), 2012 International Conference on*, pp. 546–550. IEEE.

Soleymani, M., D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic (2017). A survey of multimodal sentiment analysis. *Image and Vision Computing 65*, 3–14.

Strapparava, C., A. Valitutti, et al. (2004). Wordnet affect: an affective extension of wordnet. In *Lrec*, Volume 4, pp. 1083–1086. Citeseer.

Sykora, M. D., T. Jackson, A. O'Brien, and S. Elayan (2013). Emotive ontology: Extracting fine-grained emotions from terse, informal messages. *IADIS Intelligent Systems and Agents Conference, Prague (Czech Republic)*.

Volkova, S., T. Wilson, and D. Yarowsky (2013). Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1815–1827.

Yerva, S. R., Z. Miklós, and K. Aberer (2010). It was easy, when apples and blackberries were only fruits. Technical report, WePS-3, colocated with CLEF.

You, Q., J. Luo, H. Jin, and J. Yang (2016). Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pp. 13–22. ACM.

Zimbra, D. and H. Chen (2012). Scalable sentiment classification across multiple dark web forums. In *Intelligence and Security Informatics (ISI), 2012 IEEE International Conference on*, pp. 78–83. IEEE.

Résumé

L'immense contribution du web social a grandement motivé les chercheurs. Cela a conduit à l'émergence de techniques qui ont prouvé leur efficacité pour la modélisation des opinions et des émotions liées à des applications comme le traitement du langage naturel et l'apprentissage automatique. Par contre, lorsqu'il s'agit d'interopérabilité et d'un encodage unique des opinions et des émotions, il y a des insuffisances. Cela a motivé une nouvelle orientation de recherche combinant les travaux sur l'analyse des opinions et ceux du "Web des données liées". Dans cet article, nous allons exposer les différentes solutions et projets en présentant certaines limitations. Les raisons pour lesquelles nous croyons également que les données liées sont importantes et la façon dont nous aimerions procéder à cette recherche sont également détaillées dans cet article.

Back Recovery Protocol Based Multi-Agent Planning for the Fault Tolerance of Composite Cloud Services

Abderazak Mimouni *, Amer Agoune *, Ouassila Hioual**

* Abbes Laghrour University of Khenchela, Algeria
{zaknio1, amer.agg09}@gmail.com
**Abbes Laghrour University of Khenchela, Algeria,
**LIRE Laboratory of Constantine, Algeria
ouassila.hioual@gmail.com

Abstract. Cloud services are considered as partial solutions that must be composed in order to provide a virtual service. Unfortunately, when running such a service some faults may occur. To remedy this problem, we propose a back recovery based multi-agent planning model. Our architecture is composed of two Agents: a Planner Agent (PA) and a Plan Controller Agent (PCA). The role of the PA is to create a set of plans as a graph where nodes are Cloud services and bows represent the composition order of these services. These bows are provided with the cost, the execution time and a probability of mistakes. This agent saves checkpoints in stable memory so that there are at least two possible paths. The PCA ensures that the running plan is working properly and in the case of a failure, it informs the PA to apply the back recovery technique and to select another sub-plan.

1 Introduction

Cloud computing is a way of computing where services are provided across the internet using models and levels of abstraction (Arockiam et al., 2011). It is built upon virtualization, distributed computing, utility computing, and more recently networking, web and software services. It also shares necessary software and on-demand tools for various IT Industries. Cloud services are considered as partial solutions that must be composed in order to provide to users a virtual service. This composite service should be automated and must be dynamism in order to respond quickly to the needs of users., but in such a case of performing certain services, a fault may occur and as a solution to this problem is to apply fault tolerance techniques which, refers to correct and continuous operation even in the presence of faulty components. It is the art and science of building computing systems that continue to operate satisfactorily in the presence of faults. A fault tolerant system may be able to tolerate one or more fault types including- transient, intermittent or permanent hardware faults, software and design errors, operator errors, or externally induced upsets or physical damage (Jhawaret al., 2012).

Various fault tolerance techniques can be used at either task level or workflow level to resolve the faults (Bala and Chana, 2012). Fault tolerance (FT) techniques can typically be categorized into two sets: reactive fault tolerance techniques and proactive fault tolerance techniques. Reactive fault tolerance techniques have the advantage of reducing the impact of failures on a system when failures have effectively occurred. There are several techniques based on this policy, among these techniques we can cite: checkpoint/Restart and retry, duplication, job migration and so on. However, it is also possible to anticipate failures and proactively take action before failures occur in order to minimize failure impact on the system and application execution (Vallee et al., 2008).

In this paper, we are interested in check pointing policy. This policy is used when doing task scheduling, the check-points are inserted to identify fault incidence. These techniques take less computation and less time as a result of the task is restarted at the previously checked point. There is no ought to restart the full task (Jhawar et al., 2013). The different checked points are selected through an oriented graph which represent the different plans. These plans represent the different alternatives of a composition solution. Several approaches have been proposed to solve the Fault tolerance problem in Cloud Computing, including multi-agent planning approach In this study, we propose a back recovery solution based on the agent paradigm, which has proved to be effective for distributed applications A Multi-agent system (MAS) is composed of multiple interacting intelligent agents, within a given environment. These agents co-operate to solve difficulties that are beyond the capability or knowledge of each agent. There are several key characteristics of agents, for which we chose to use this paradigm, such as adaptation, scalability, re-us ability, local view, autonomy, responsiveness and distribution. In order to achieve the necessary goals, agents are required to be able to communicate with many other agents in the environment (Byrski et al., 2015)

The remainder of the paper is structured as follows: In Section 2, we present some related work. In Section 3, we introduce the problem research and we present our proposed model. In Section 4 and through an illustrative scenario, we present some evaluations of our method. Finally, the Section 5 presents a conclusion and future work.

2 Related work

Several approaches and methods have been proposed to solve the fault tolerance problem including back recovery based checkpointing in Cloud Computing environments. There is, also, a wide set of works in intelligent cloud computing that tries to make clouds more intelligent by adopting intelligent agents to automate the interactions among clouds and between consumers and cloud (Sim, 2013). In this study, we propose a back recovery solution based on the agent paradigm, which has proved to be effective for distributed applications. A Multi-agent system (MAS) is composed of multiple interacting intelligent agents, within a given environment.

A reactive fault tolerance technique is proposed, using check-pointing, in (Kalanirnika, 2015). The proposed strategy called VM- μ Checkpoint framework protects VMs against transient errors. Also, Copy on Write – Presave in Cache algorithm is used and, in order to save the checkpoints of the tasks running in the VMs in advance, a cache memory is used. In

the proposed algorithm in-memory incremental checkpoints are taken so that restoration can be done in-place.

In (Wang et al., 2015), the authors presented an overview of workflow temporal checkpoint selection. The authors proposed a temporal checkpoint selection strategy to deal with business workflows. Several consistency models for business and scientific workflows such as Throughput based temporal consistency model, Response-time based temporal consistency model, probability based temporal consistency have been discussed.

In (Santosh and Ravichandran, 2013), earlier strategy included a non-preemptive scheduling with task migration algorithm. The proposed method had a major drawback of starting again the task in another virtual machine. The proposed solution included an algorithm that migrates the aborted task and starts the execution at a point where the latest checkpoint was saved. This leads to better performance and achieves QoS. The main defect of this method is that greatly increases the execution time of the migrated task.

The strategy proposed in (Di et al, 2013) implicates how to calculate the optimal number of checkpoints based on failure event distribution. Various parameters like check-pointing overload, time delay have an impact on the cloud system. For this, and in order to optimize the impact and for better performance an adaptive algorithm was designed. In (Di et al, 2013), the parameters that are considered for implementing the model are user request of multiple task, number of jobs, checkpointing cost, checkpointing position, probability of failure event, execution time and wall clock time. Experimental results showed the better adequacy of the system for large scale applications.

Most of these works take into account several QoS criteria to solve the fault tolerance problem, in a cloud computing environment. In our contribution, we want to take advantages from these works such as check pointing techniques because they provides an appropriate services to the system. In contrast to the above research works, we approach fault tolerance in a cloud environment by taking into consideration the problem when some faults may occur while running a composite cloud service. Our architecture is an agent-based one because agents are capable of solving problems independently and may collaborate with one another to achieve objectives. In our architecture, we assume that a Planner Agent (PA) has a complete knowledge of all services deployed in the Cloud. The Plan Controller Agent (PCA) ensures that the running plan is working properly and in the case of a failure, it informs the Planner Agent to apply the back recovery technique and to select another sub-plan. In the next section, we will explain our problem description, proposed model and some techniques and methods used to solve it.

3 Problem description and proposed model

Cloud services are provided either as computing services or storage services. A customer, and in order to be served, sends his request for service to the cloud provider with the necessary requirements to his request. Cloud services are normally partial solutions that must be composed to provide a single virtualized service to cloud consumers. This services composition should be done in an automated and dynamic way in order to promptly fulfil client requirements. It is expected that numerous failures will occur that will protract the time expected to carry out the customer requests and this will exhaust the cloud resources. For customers, they will not get their services in the time expected. For the cloud, failures will lead to loss of cloud resources and then money (Amoon, 2016). Our research problem is a

part of the fault tolerance in a cloud environment. Our goal is to propose a method based back recovery and multi-agent planning approaches. Thus, back recovery based check pointing consists in capturing enough data during fault-free distributed execution of tasks using a multi-agents planning system. In the literature, it is estimated that 60-90% of current computer errors are from software faults (Gray, 1991), so the faults that we treat in this work are software ones.

3.1 Overview of our system architecture

Multi-agent systems (MAS) provide a reliable and efficient approach for designing and constructing distributed and collaborative applications. A MAS can be seen as a set of autonomous entities, called agents, that interact and coordinate to solve a given problem. The MAS are becoming common solutions to address distributed applications problems such as Fault tolerance in Cloud computing.

The architecture of our system (see. Fig 1) is an agent-based one which is inspired from a previous work. The main agents of the architecture are as follows:

- Planner Agent (PA) which designs the different plans that represent the alternative plans necessary to have the appropriate composition of services responding to the user request. These plans constitute an oriented graph enabling it to define the ideal path (a sub-plan) and periodically it will save a check point in its stable memory provided that :
  - Availability of another sub-plan from the checkpoint site,
  - And, that the service is actually available at the time of the fault occurs.
- A Plan Controller Agent (PCA) whose role is to control and ensure the proper execution of the chosen plan. In the case of a fault appearance during the execution, it informs the PA in order to take the necessary measures.

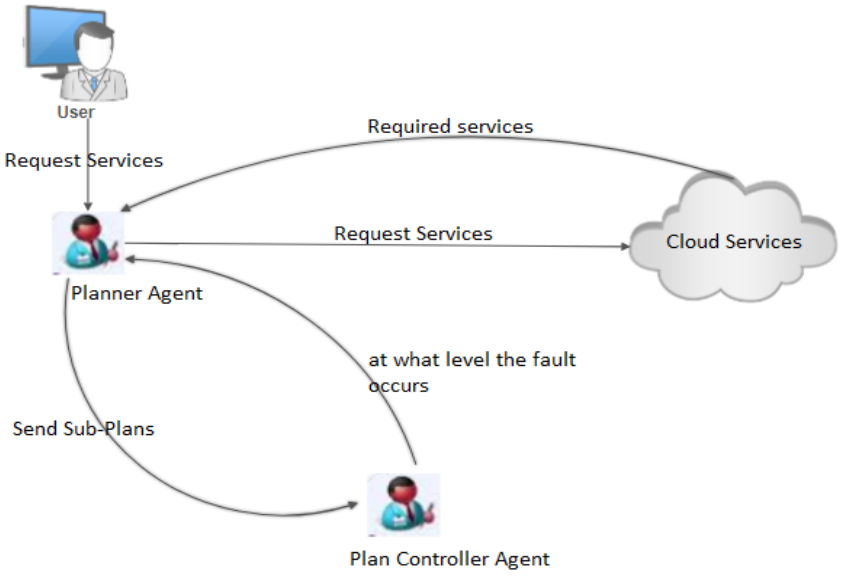


FIG. 1 – *The proposed architecture of our system.*

3.2 Functionality of the proposed method

In this section, we present a brief walk-through of our method. This is done through a sequence diagram as shown in Figure 2. As stated above, we use the back recovery based checkpointing technique and the planning based agents techniques.

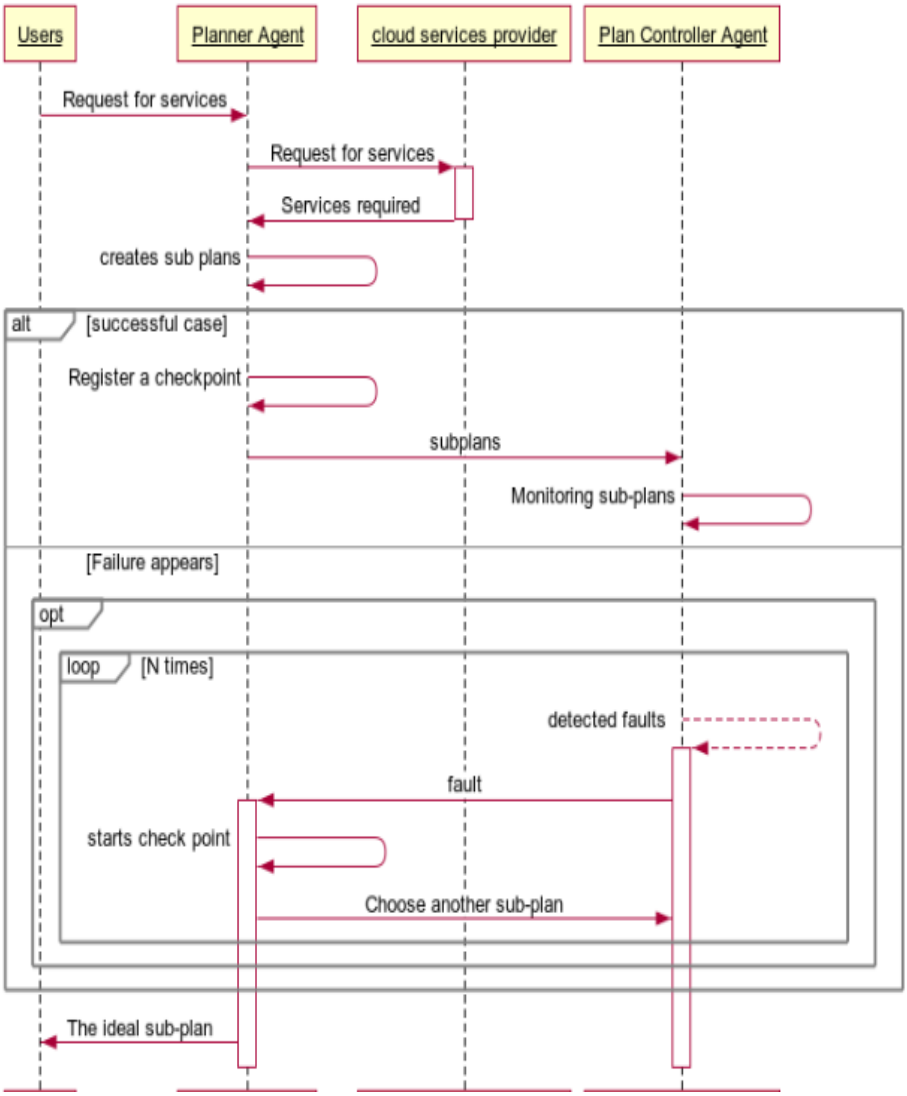


FIG. 2 – *Sequence diagram which represents the functionality of our proposed method*

According to the scenario shown in Figure 2, a planner agent after receiving a client request, it produces a set of plans (each plan represents an eventual solution for the composi-

tion of the requested composite service). Then, it sends to the PCA the initial plan to be executed. This later will control the proper functioning of the selected plan. Meanwhile, the PA saves periodically a checkpoint (see Fig 3). When a fault will be detected by the PCA, it informs the PA in order to choose an adequate sub-plan and so on.

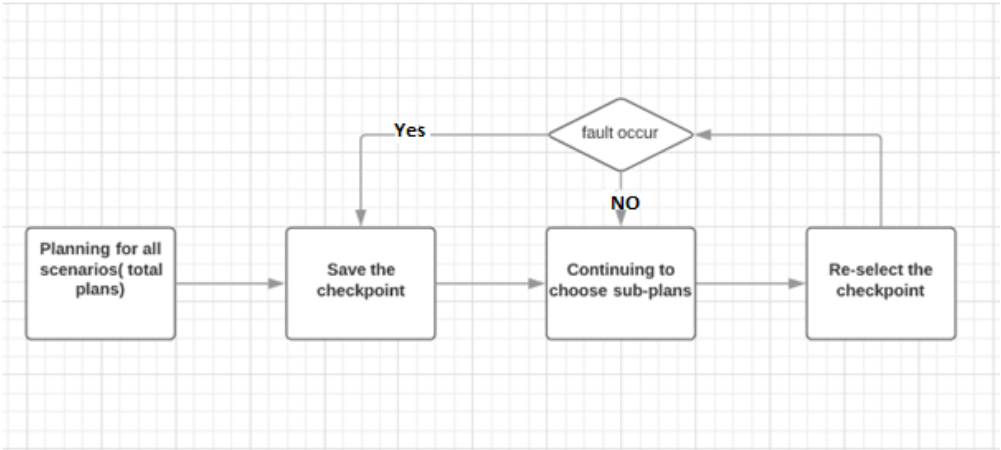


FIG. 3 – *How a PA saves a novel check point*

We assume that the different plans of a composite service constitute an oriented graph (see Fig 4). The root of this later is the initial cloud service that represents the start of composing services for all alternative plans of a given composite service. The nodes of this graph represent the component cloud services. An arc between two nodes is an oriented arc and represents the succession link, that is to say that an arc connecting the node a to the node b means that: the cloud service b will be executed after the good execution of the cloud service a. So, according to our method, a checkpoint is a node from which we can find at least two sub-plans (cf. Figure 4). The bows have values that represent weighted values between the cost of the cloud service, the estimated time of its execution and its reliability. These values are used by the PA in order to select the ideal path (initial plan or a sub-plan after a fault detection).

3.2.1 Planner Agent behavior

As mentioned above, the Planner Agent is responsible for creating cloud services composition plans in order to respond to a user request. In this work, we are not interested in how to create these plans. So, we suppose that the PA has a set of plans that it will represent them in the form of an oriented and evaluated graph.

The PA saves a checkpoint in a stable memory mechanism (Coghlan and Jones, 1992). Data to save are: *the service ID, the service length (it means the size of this service to be executed in a Cloud Resource), the serviceFileSize, the serviceOutputSize and the status of this service*.

This mechanism uses a set of parameters in function of services number and the exchanged messages between PA and PCA agents. We note SM the necessary stable memory for stocking different checkpoints. The functionality of this agent is illustrated by the activity diagram of Figure 5.

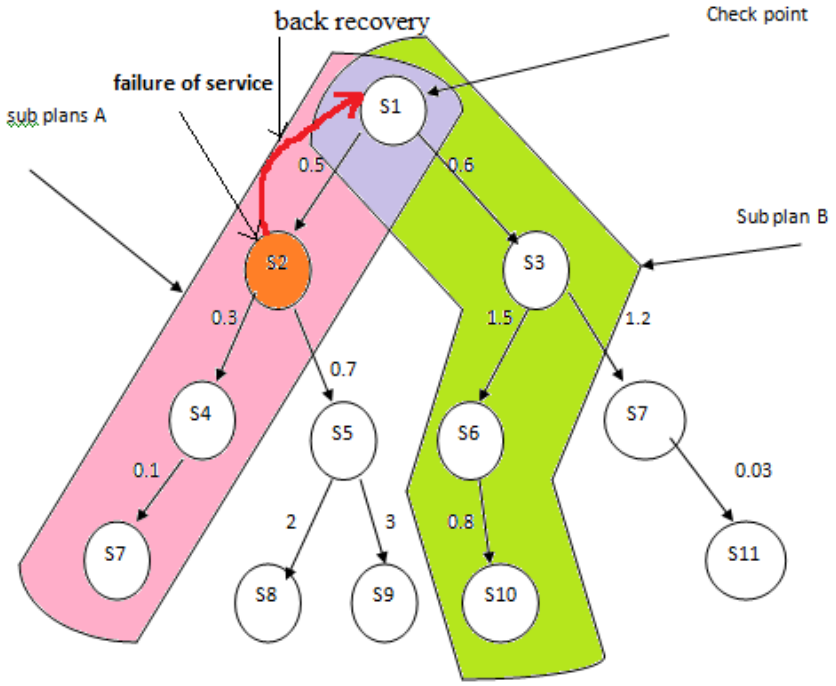


FIG. 4 – *Plans of a composite service in the case of failure*

3.2.2 Plan Controller Agent behavior

The PCA role is to control and ensure the proper execution of the chosen plan. When a fault occurs, it must react immediately by informing the PA, and that by sending it the useful information such as at which level (node of the current plan) the fault appeared. The overall functionality of this agent is illustrated in Figure 6.

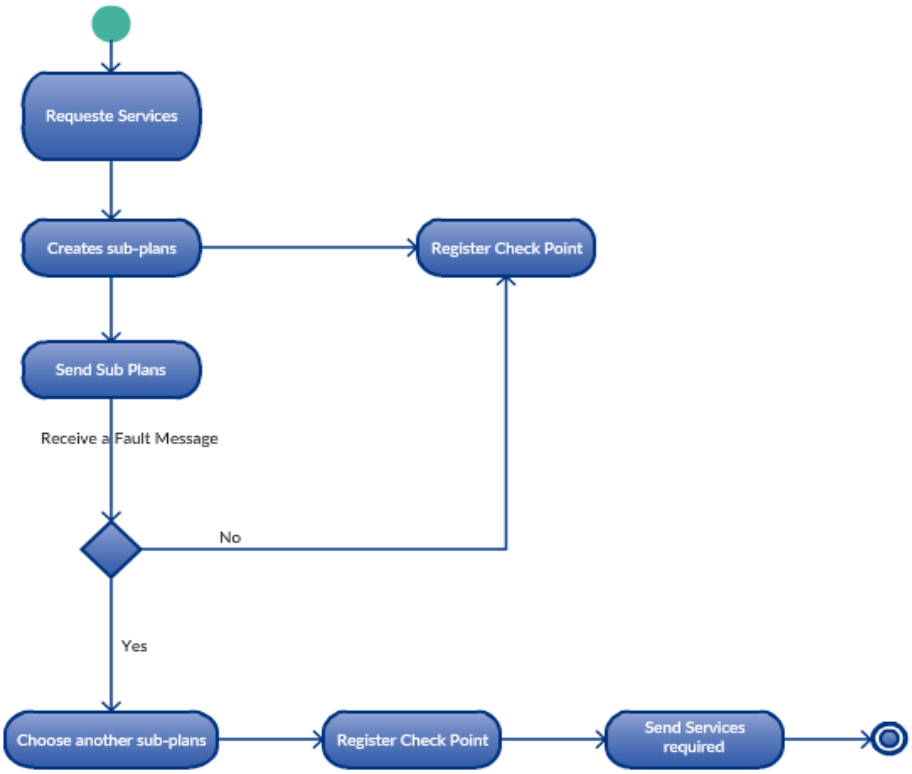


FIG. 5 – *Activity diagram of Planner Agent*

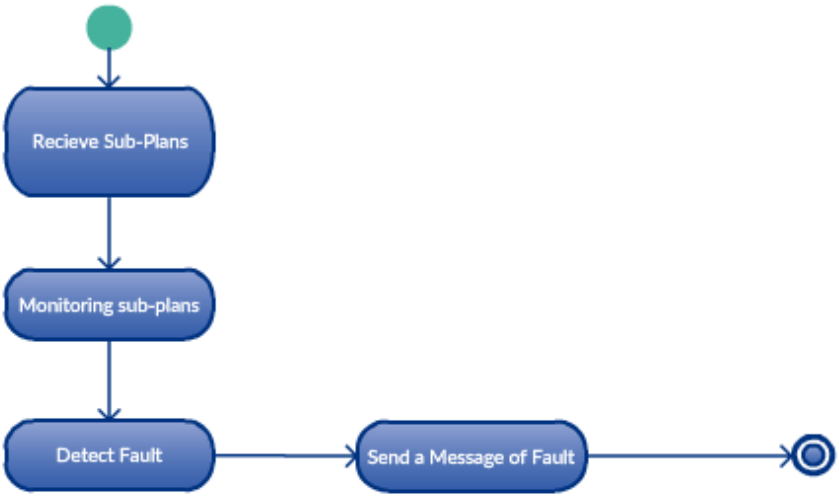


FIG. 6– *Activity diagram of Plan Controller Agent*

4 Illustrative Scenario

This section examines a scenario to illustrate, first, the steps or procedures that must be followed according to our method, and second, the interaction between the different agents of the proposed architecture. The scenario is as follow: we have chosen the full scheduled flight service and we assume that the Planner Agent may represent services as a scheme (an oriented graph: see Figure 7). Then, it applies Djikstra's algorithm (Djikstra, 1971) (Shortest distance calculation) to choose every time the path which has the less cost.

We suppose that we have the values that represent the estimated time to execute a service, price, and reliability of each service.

Time	**[5min –10min [**	**]10min—20min]**	**P>20min**
The cost	0.1	0.2	0.3

TAB. 1– *How to estimate the execution time of a service*

Service price($)	**[50 --100 [**	**]100--100]**	**P>100**
The cost	0.1	0.2	0.3

TAB. 2– *How to estimate the price of a service*

Number of users	**[10 K—100 K [**	**] 1K—10 K]**	**NU<1 K**
Reliability	0.1	0.2	0.3

TAB. 3– *How to estimate the reliability of a service*

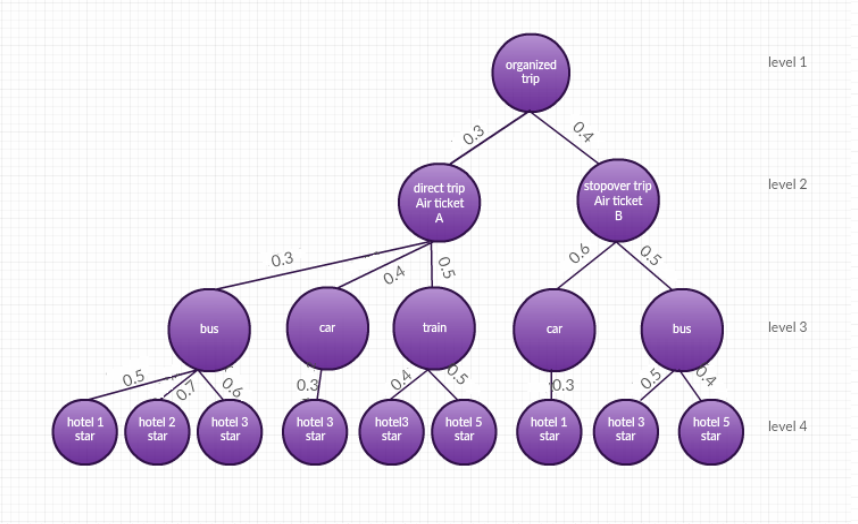


FIG. 7– *The oriented graph of the different plans of the requested service*

Firstly, the planner agent requests a set of organized tourism services from Cloud services provider, the services are organized according to their chronological order. Then (PA) will select the services from the graph using the Djkstra's algorithm. Every time that the PA sets out to visit a new node, it choose the node with the smallest known weight to visit first. At every step, it checks each of its neighboring nodes.

For each neighboring node, it calculates the weight for the neighboring nodes by adding the cost of the edges that lead to the node which had checked from the starting vertex.

Finally, if the distance (cost) to a node is less than a known distance, it will update the shortest distance that it has on file for that vertex.

The planner agent, after applying Djikstra algorithm, will get the following sub-schema: Level 1: organized trip➔ level 2: direct trip Air ticket A➔level 3: car ➔level 4: hotel 3 stars.

The planner agent will retain the first service (direct trip Air ticket A) as a checkpoint (see. Fig. 8) then the previous sub-schema will be sent to the Plan Controller Agent in order to ensure that the services do not contain any failure.

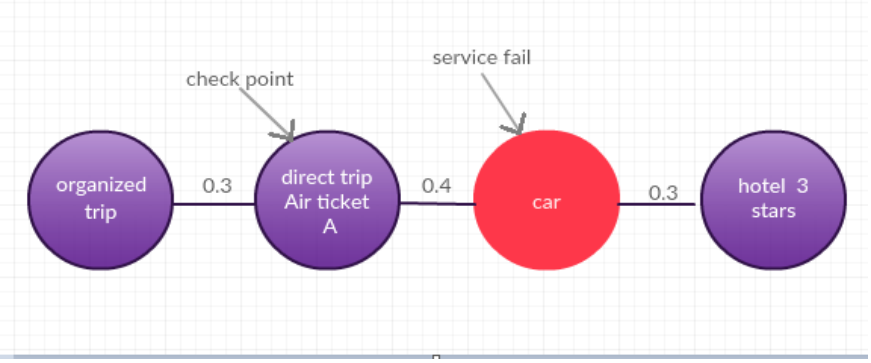


FIG. 8– *Report the failure after monitoring the sub-plans by The Plan Controller Agent*

The Plan Controller Agent Sends the fault report to the planner agent: Level 3: Service: means of transportation (car)

The planner agent apply Back Recovery to the check point Level 2: Service: (direct trip Air ticket A). It then reapply the djikstra algorithm to find a new sub-plan different from the latter where the failure occurred.

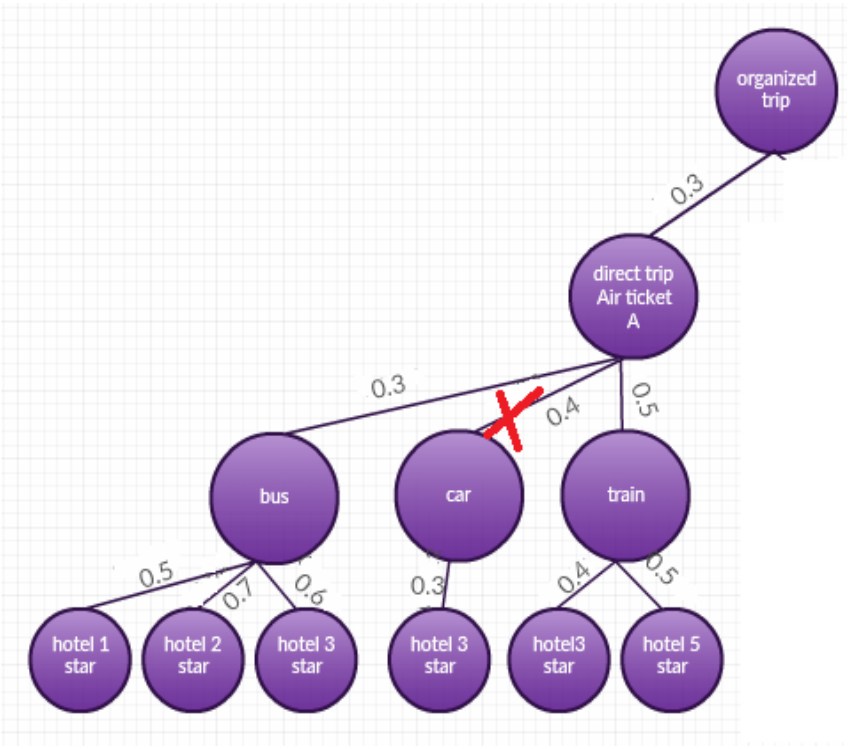


FIG. 9– *Possible sub-plans except the schema that contains the failed service*

So, after applying the Dijikstra algorithm, the following sub-plan will be sent to the Plan Controller Agent (see. Fig.10)

Level 1: organized trip➔ level 2: direct trip Air ticket A➔level 3: bus ➔level 4: hotel 1 stars

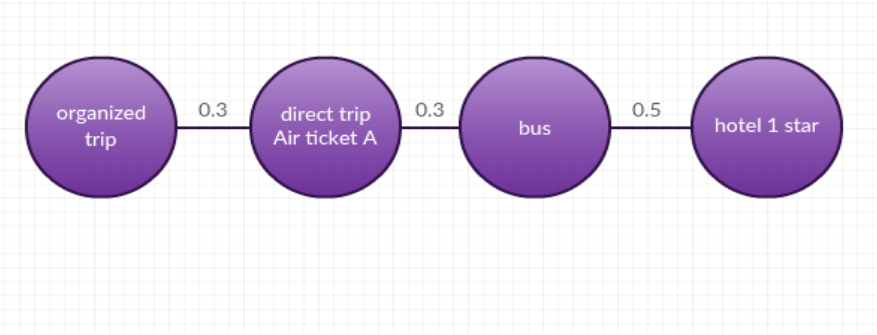


FIG. 10– *The new sub-plan to be taken into consideration*

5 Conclusion

In this paper, we have presented a method to enable fault tolerance using back recovery which relies on periodic check pointing. Our method takes profit of the specificities of multi-agent applications which is the planning. In the future, we aim to detail the functionalities of both the two types of agents that make up our architecture. We plan to treat the behavior of the system when a problem occurs when the checkpoints are not yet registered or the measurement data are not yet available (first execution for example). Also, we aim to evaluate our proposed method more extensively through some real systems case studies.

References

Amoon, M (2016). Adaptive Framework for Reliable Cloud Computing Environment. *IEEE Access Journal*, 4: 9469-9478.

Arockiam, L., S., Monikandan and G., Parthasarathy (2011). Cloud Computing: A Survey. *International Journal of Internet Computing*, 1(2): 26-33.

Bala, A. and I., Chana (2012). Fault Tolerance Challenges, Techniques and Implementation in Cloud Computing. *International Journal of Computer Science Issues*, 9(1): 288-293.

Byrski, A., D., Rafał, S., Leszek , and K-D., Marek (2015). Evolutionary multi-agent systems. *The Knowledge Engineering Review*, 30(2): 171-186.

Coghlan, B.A. and J.O., Jones (1992). *Memory Checkpointing*. Irish Patent Application 2784/92 and derivatives, Tolsys Ltd.

Di, S., Y., Robert, F., Vivien, D., Kondo, C-L, Wang and F., Cappello (2013). Optimization of cloud task processing with checkpoint-restart mechanism. In: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis. IEEE, Denver, CO, USA.

Djikstra, E-W (1971). A short introduction to the art of programming. *EWD316*, 1: 67-73.

Gray, J., and D. P., Siewiorek (1991). High-Availability Computer Systems. *IEEE Computer*, 24(9):39-48.

Jhawar R., V., Piuri and M., Santambrogio (2013). Fault tolerance management in cloud computing. A system-level perspective. *IEEE Systems Journal*, 7(2): 288–297.

Jhawar, R., V., Piuri and M., Santambrogio (2012). A Comprehensive Conceptual System-Level Approach to Fault Tolerance in Cloud Computing. In: 2012 IEEE International Systems Conference (SysCon), IEEE. Vancouver, BC, Canada.

Kalanirnika, G. and R.V.M.Sivagami (2015). Fault Tolerance in Cloud Using Reactive and Proactive Techniques. *International Journal of Computer Science and Engineering Communications*, 3(3): 1159- 1164.

Santosh, R. and T., Ravichandran (2013). Non-Premptive Real Time Scheduling using checkpointing Algorithm for cloud computing. *International Journal of computer Applications*, 80(9): 1-4.

Sim, K.M (2013). Cloud intelligence: agents within an InterCloud. *Awareness Magazine*. The Official Magazine for Future and Emerging Technologies Proactive Initiative [online] http://www.awareness-mag.eu/pdf/005153/005153.pdf (accessed 15 May 2015).

Vallee G., C., Engelmann, A., Tikotekar, T., Naughton, K., Charoenpomwattana, C., Leangsuksu and S., Scott (2008). A Framework for Proactive Fault Tolerance. In: Third Intemational Conference on Availability, Reliability and Security (ARES). IEEE, Barcelona, Spain.

Wang, F.T., X., Liu and Y., Yang (2015). Necessary and sufficient checkpoint selection for temporal verification of high-confidence cloud workflow systems. *Science China Information Sciences*, 58(5): 1-16.

Résumé

Les services Cloud sont considérés comme des solutions partielles qui doivent être composées pour fournir un service virtuel. Malheureusement, lors de l'exécution d'un tel service, certaines erreurs peuvent se produire. Pour remédier à ce problème, nous proposons un modèle de planification multi-agents basé sur le recouvrement arrière. Notre architecture est composée de deux agents: un Agent Planificateur (AP) et un Agent Contrôleur de Plans (ACP). Le rôle de l'AP est de créer un ensemble de plans sous la forme d'un graphe où les nœuds sont les services et les arcs représentent l'ordre de composition de ces services. Ces arcs comportent le coût, le temps d'exécution et une probabilité d'erreurs. Cet agent enregistre les points de contrôle dans la mémoire stable afin qu'il y ait au moins deux chemins possibles. L'ACP s'assure que le plan fonctionne correctement et en cas de défaillance, il informe l'AP afin de sélectionner un autre sous-plan.

Learning and Optimization for a Driving Assistance System

Manolo Dulva Hina*, Assia Soukane*
Amar Ramdane-Cherif**

* ECE Paris School of Engineering, 37 quai de Grenelle, 75015 Paris, France
manolo-dulva.hina@ece.fr, assia.soukane@ece.fr
http://www.ece.fr
**Université de Versailles St-Quentin-en-Yvelines, 78140 Vélizy, France
rca@lisv.uvsq.fr
http://www.lisv.uvsq.fr

Abstract. In this paper, an alternative driving assistance is presented. It is alternative in the since that it is not closed to proprietary constraints which is the case for those associated with car manufacturers. Machine learning is used to identify a driving event and optimization is done to identify the optimal actions that should be performed following the identified driving event. Ontology is used to represent knowledge and driving situations. For the learning part, a training set is created, storing driving situation patterns and from which an intelligent system can used to determine the driving situation. For the optimization part, algorithms are developed that maximizes scores for an action that corre-sponds to safe driving, green driving and comfortable driving. This work is a contribution to make driving assistance systems affordable and available to regular vehicles that common people possess.

1 Introduction

Advanced driving assistance system (ADAS) (Li et al. 2012) is intended to assist someone to drive safely and to some extent to consume less fuel. It also provides comfort for the driver. The embedded features of ADAS (Armand et al. 2014, Hina et al. 2018) are numerous and expensive to implement. These features are usually reserved for premium types of vehicles. Pending the deployment of these features on all vehicles, it is essential to propose an alternative, affordable and compatible solution for the existing and coming fleet. This is the main motivation for us proposing an alternative ADAS. Apart from this introductory content, this paper is structured as follows: Works related to ADAS are discussed in Section 2. The driving model is discussed in Section 3. Simulation and signal processing are discussed in Section 4. Machine learning and optimization details are presented in the sections 5 and 6. The paper is concluded in Section 7.

2 Related Work

We investigated various types of driver assistance provided by different vehicle manufacturers. Formerly reserved for top-of-the-range vehicles, a lot of assistance is now available on the vehicles in large series; others are optional and very costly. Advanced Driver Assistance Systems (ADAS) offer a means to enhance, among other things, active and integrated safety (Bengler et al. 2014). One of the first active assistance systems based on proprioceptive sensors was the anti-lock braking system (ABS) which started in 1978 (BOSCH 2016). The traction control system later improved the features of ABS system. With reference to the requirement on improving support for the driver, projects such as (CVIS , PREVENT , SAFESPOT), the number of safety-available functions on the vehicle is continuously increasing. This is in addition to already commercially available features such as: anti-block brake system, electronic stability control, advanced driver assistance system, collision avoidance, lane departure warning and blind-spot monitoring. Many types of ADAS systems exist but are very expensive and therefore available only to limited number of vehicles. An alternative ADAS is generic, not closed to proprietary constraints and can be utilized to all types of vehicles. This is where our work is situated.

3 The Driving Model

The driving model is the representation of driving situations and the rules and conduct of driving. It is important to show the driving model because it is through this that we will be able to analyze the driving event and the kind of driving assistance that is appropriate for such event. We will represent the driving model using ontology.

3.1 Knowledge Representation using Ontology

Ontology (Noy and McGuinness) is the whole structure of concepts and the relations representing the meaning of a given domain. It is applied in artificial intelligence and semantic web, and allows the representation of knowledge. There are various definitions of ontology and the most commonly used are those of (Neches et al. 1991), (Gruber 1993), and (Guarino 1995). Ontologies are formal because they are expressed as formalism with formal semantics. The use of ontologies in the modelling of accident black spots situations in transport (Maalel et al. 2011) and on assistance on search of data (Charest and Delisle 2006) take a significant boom because the contribution of semantic web to the realization of systems allows for the development of architectures with distributed components in the network. (Kannan et al. 2010) uses ontology in modelling an Intelligent Driver Assistance System (I-DAS) for Vehicle Safety. In order to design our ontology model, we use the software tool called Protégé (Protégé 2016). We also use VOWL (Visual Notation for OWL) (VOWL) as a plug-in for data visualization. All diagrams related to ontology that appear in this paper are data visualizations through VOWL. (Guarino 1995).

3.2 The Driving Context

The driving context is the fusion of parameters of the context of the driver, the vehicle and the environment. Ontology is used to represent the driving context. The "Environment" is an ontology class that describes all the entities that are on the road, including other vehicles, pedestrian, traffic signs, etc. The "Vehicle" is the class that represents our vehicle while the "Driver" is the class that describes the main actor of the driving context, the driver himself. See Fig. 1. The classes "Environment" and "Vehicle" are related through hasVehicle object property while hasDriver is the object property that links with its driver. The property is functional because only one driver can own a vehicle.

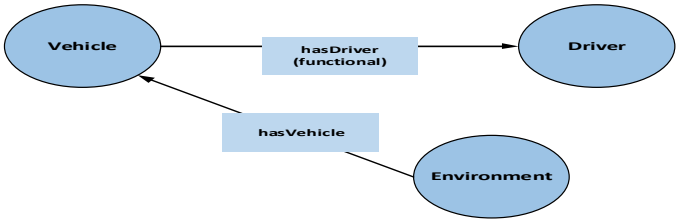


FIG. 1 – *Ontology for the driving context.*

3.3 The Context of the Driver

The ontological context of the driver is shown in Fig. 2. The "Driver" class is related to other classes, which describe the state of the driver: (i) MentalState: the current mental state of the driver. The class has the following subclasses: "Anger", "Stress", "Faint" and "Stress"; (ii) DriverProfile: this class contains pertinent attributes: Name, LicenseScore, DriverLicenseNumber, Age, etc; (iii) DriverViolations: this class contains the historical driving information. It has subclasses, such as "RedLightViolation", "OverSpeeding", etc; and (iv) FocusOnDriving: a class that contains many Boolean attributes, including "hasEyesOnTheRoad", "hasPhoneConversation", "hasPassengerOnBoard", "hasHandsOnTheSteeringWheel, etc.

3.4 The Context of the Vehicle

The ontological context of the vehicle is treated as follows: the class "Vehicle" is a subclass of "MovingObject" which is a class that describes all moving entities on the road, including pedestrians, cyclists, and other vehicles. A "Vehicle" has subclasses, namely "Truck/Bus", "Car", and "MotorBike". The "Vehicle" is described by various relationships with other classes: (i) TechnicalData: this class describes the technical data of our referenced vehicle. Its subclasses are "FuelType", "EmissionClass" and "TractionType"; (ii) Cockpit: a class that contains the status of all elements that are found in a vehicle's cockpit. For example, 'hasWindowsOpen' is a data property that has a Boolean value; (iii) ComponentsStatus: this contains as subclasses all components that we have to check to guarantee a good driving experience. Among these subclasses are "DirectionIndicator" (with values 'NoIndicator', 'RightIndicator', 'LeftIndicator', and 'DoubleIndicators'), "TyresPression", "LubricantTemperature", "EngineLubricantLevel" (with values 'LowLevel', 'HalfLevel' and 'FullLevel'), and "FuelQuantity". The class has also some Boolean properties to check if some components are active or not. Example is 'hasFogLightsOn'; (iv) hasPossibleCollision: this

is a property that associates our vehicle with the class "MovingObjects"; and (v) hasPhysics: this property links our vehicle with class "Physics" which describes our vehicle with attributes, such as speed and acceleration.

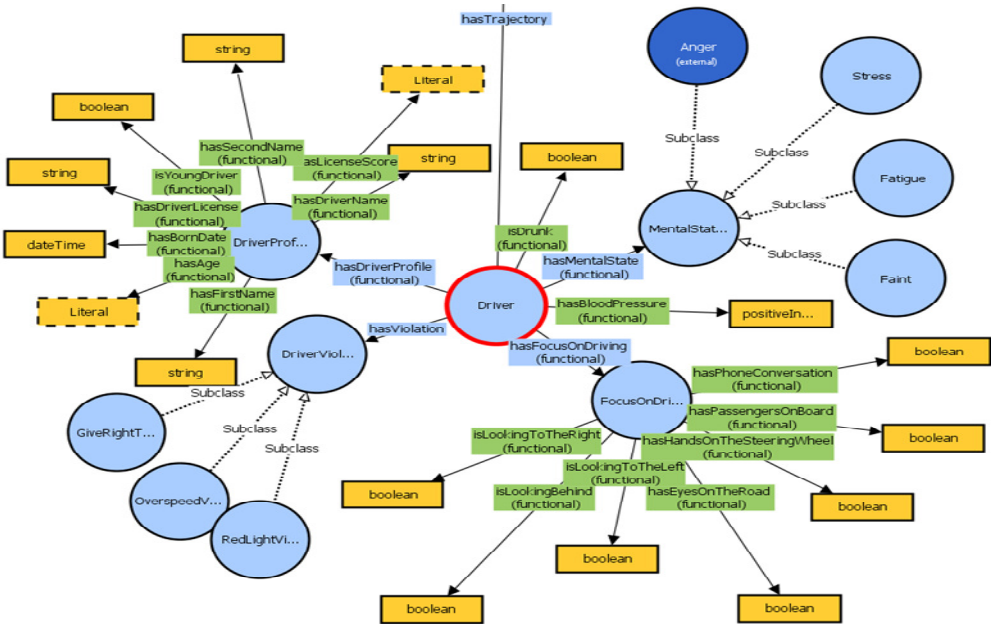


FIG. 2 – *The context of the driver.*

3.5 The Context of the Environment

The ontological context of the environment is treated as follows: the Environment is related to all the elements that belong to the scene where the driving event takes place. The "Environment" is an abstract class and general concept made up of cities where vehicles are present. The class Environment is related to other classes given as follows: (i) City: Here, an Environment is an area where we can find many cities. A city has two data properties, namely 'hasCityName' and 'hasLimitedTrafficZone' which is a Boolean value indicating if the city can be accessed only during some intervals of the day; (ii) DistrictArea: it contains as objects the different districts of a city, linked through 'hasDistrictArea' property. The position of the "Driver" is stored in the "PositionArea", a subclass of "Physics" and equivalent to "DistrictArea"; (iii) Road: a road has many data properties, such as 'hasMinSpeedLimit', 'hasMaxSpeedLimit', 'hasNumberOfLanes', 'hasContinuousLine' and 'hasLength'. A road is made up of three subclasses, as follows: (1) 'Urban', (2) 'ExtraUrban', and (3) 'Highway'. The Road is related to the class RoadSegment via property hasRoadSegment. A RoadSegment has three subclasses: Lane, Intersection, and RoundAbout. They describe the kind of road segments that we find in real life; and (iv) RoadProperty: it further describes the "Road". Its subclasses include "Visibility" (values are 'Low', 'Average' or 'High'), "Weather" (values are 'Fog', 'Sun', 'Rain' and 'Snow'), "AccidentHistory" (values are 'Unusual' or 'Frequent'), "TrafficCongestionHistory" (values are 'Low', 'Average' or 'Intense') and "CurrentTrafficCongestion" (values are 'Low', 'Average' or 'Intense').

4 Driving Simulation and Signal Processing

Our driving assistance concepts needs to be tested in the laboratory first before it is tested on the actual roads of Paris. As shown in Fig. 3(a) the driving scenario is a reflection of the realities on the road, showing the driver, the driver's vehicle, and the environment with other moving entities (i.e. vehicle, pedestrians). The driving experience mimics that of the vehicle cockpit. Indeed, using this set-up, we are able to drive in the lab as if we are driving a real vehicle. We also have the option of selecting the driving scenario to simulate as well as to monitor driving parameters throughout the duration of the driving exercise. The signal processing in this alternative ADAS system is shown in Fig. 3(b). The systematic step-by-step processes involved are described below:

- *Step 1*: A user drives the driving simulator. This yields true values for various parameters related to the context of the driver, the vehicle and the environment. For example, as soon as the user drives the virtual vehicle, the vehicle speed is noted. This value as well as other parameters' values are sent to the ontology templates.
- *Step 2a*: Actual simulated values are passed as input to the ontology template. This produces the instances of various classes, attributes and properties of our ontology.
- *Step 2b*: Given all the parameters obtained from the simulator, the next process involved is the identification of the current driving event. This involves machine-learning mechanism, specifically the classification of the event using supervised learning. In general, *Step 2* identifies the driving event.
- *Step 3a*: The output in the previous process becomes input in this process. Here, we are interested in classification of driving event.
- *Step 3b*: Using optimization algorithms and reinforcement learning, we identify the optimal action that is to be invoked for this event. By optimal, we mean the action that produces optimal score for safe driving, green driving and comfortable driving.
- *Step 4*: This process is about the identified optimal action for the driving event.
- *Step 5a*: The driving assistance mechanism intended for the driver is activated in this phase. This may mean sending an audio or written message to the driver.
- *Step 5b*: The driving assistance mechanism intended for the vehicle is activated by sending values to some vehicular signals, such as the fog light being "on".
- *Step 6*: The driving continues and the same processes repeat.

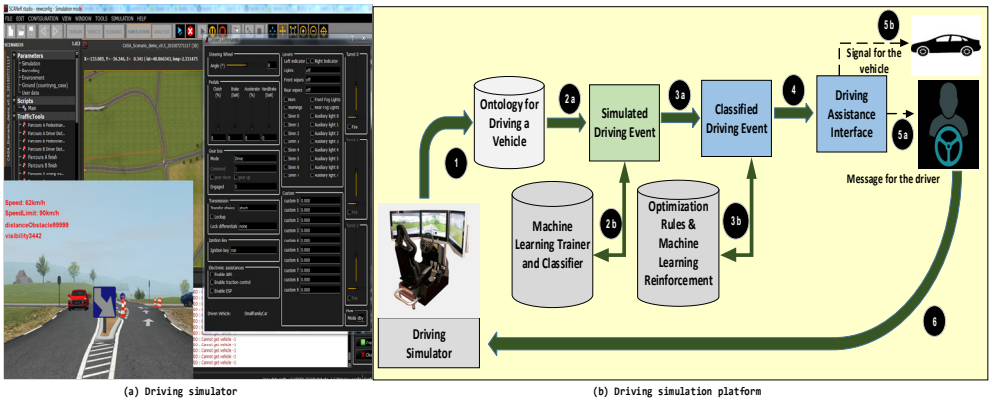


FIG. 3 – *(a) The driving simulator. (b) The driving simulation platform.*

5 Machine Learning for Driving Event Identification

Our driving assistance concepts needs to be tested. As shown in Fig. 3(a) the driving scenario is a reflection of the realities on the road, showing the driver, the vehicle, and the environment with other moving entities (e.g. vehicle, pedestrians). The driving experience mimics the vehicle cockpit. Using this set-up, we are able to drive in the lab. We also have the option of selecting the driving scenario to simulate, and monitor driving parameters during the driving exercise. The signal processing of this alternative ADAS is shown in Fig. 3(b).

5.1 Understanding Machine Learning

Machine learning is the discipline that allows machines to improve performances by learning from previous events (Vahidi and Eskandarian 2003). It can be divided into four learning types: (i) *supervised learning*, (ii) *unsupervised learning*, (iii) *reinforcement learning* and (iv) *deep learning*. Each type differs from others based on what and how the machine learns. Supervised and unsupervised learnings are the two main learning types (Tchankue et al. 2013), while reinforcement and deep learnings are special application of supervised and unsupervised learnings. Consider a normal *x*-*y* function. Given a set of input *x*, we define *y* as the output for a relation *f* between *x* and *y*. The differences between machine learning techniques may be explained using the basic notion of mathematics. See Tab. 1

- In supervised learning, *x* and *y* are known; the goal is to learn a model that approximates *f*.
- In unsupervised learning, only *x* is given, the goal is to find *f* for set *x*.

	Relation	Remarks
Supervised learning	y = f(x)	*x* and *y* are known ; the goal is to learn a model that approximates *f*
Unsupervised learning	f(x)	*x* is given and the goal is to find *f* for a given set of *x*.
Reinforcement learning	y = f(x); r	*r* is a reward that allows determination of *f* in order to obtain the optimal *y*

TAB. 1 – *Categories of machine learning algorithms. Red-colored notation means that such data are unknown while the blue-colored symbol means data are known.*

Supervised learning is used for model approximation and prediction while unsupervised is used for clustering and classification. Reinforcement learning is a particular case of supervised learning; it differs from the standard case not due to the absence of y but in the presence of delayed-reward *r* that allows it to determine *f* in order to get the right *y*. Deep learning is a supervised or unsupervised work based on learning data representation. It uses an architecture based on multiple-layer structure for data, using it for feature extraction and representation. Each successive layer uses as input the previous layer output.

5.2 Gathering Training Data Set

The identification of a driving event necessitates the collection of various signals coming from the driver, the vehicle and the environment. These signals are then interpreted accordingly. In our work, we collect the following data from our driving simulator: (i) `Image` (png file): the screenshot of the situation, used to label the data; (ii) `Look orientation` ("front", "left", "right", "behind"); (iii) `Steering angle`; (iv) `Throttle`; (v) `Brake`; (vi) `Speed`; (vii) `Orientation` ("North", "South", "East", "West"); (viii) `Previous Orientation`; (ix) `Going to` ("North", "South", "East", "West"); (x) `Speed` limit; (xi) `Blinker state`; (xii) `Number of lanes`; (xiii) `Continuous lane` (true, false); (xiv) `Current lane`; (xv) `Next lane`; (xvi) `Position on lane`; (xvii) `Relative vehicle rotation` (angle between -180° and 180°): Rotation of the vehicle relative to the road; (xviii) `Road type` ("straight", "T-cross", "curved"); (xix) `Coordinate on lane`; (xx) `RoadObject` (JSON): Road objects such as stop sign, pedestrian and other vehicle that our vehicle is aware of; and (xxi) `RoadMap` (JSON): Road map that contain all the road, lane, intersection, and the connection between them.

5.3 Driving Event Classification using Decision Tree

The driving event identification is a classification problem in nature. As far as machine learning is concerned, supervised learning is the solution for this case. Machine learning algorithms discover patterns and predict an output from a formatted input after training the algorithm on a sufficiently big set of training data. Decision tree for supervised learning uses a binary tree as a predictive model. See Fig. 4.

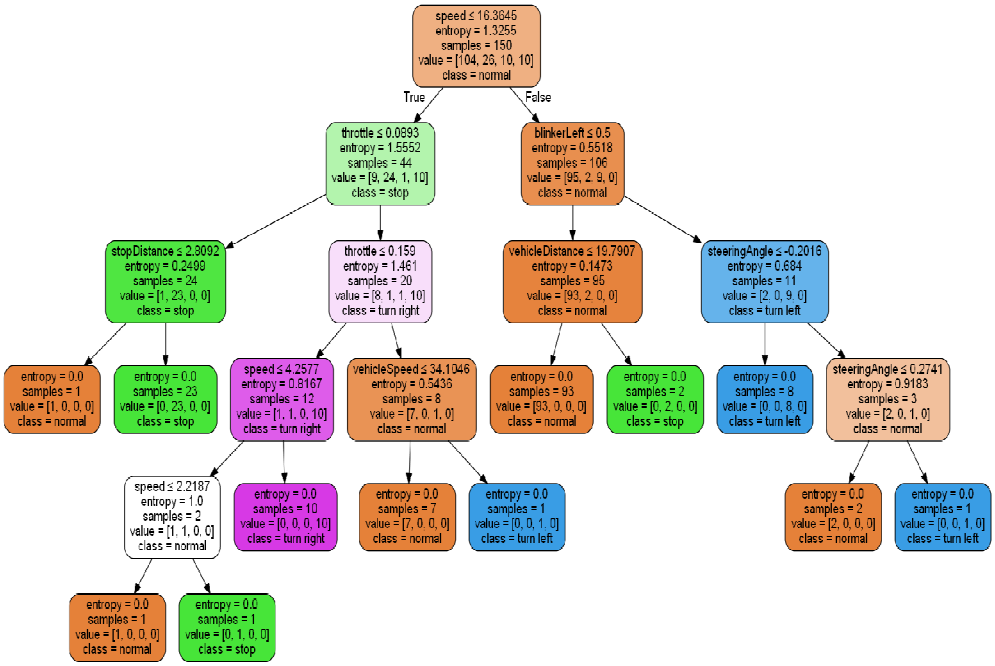


FIG. 4 – *The decison-tree learning algorithm's model for driving event classification.*

A decision tree is a flowchart-like structure in which each internal node represent a "test" on an attribute, each branch represent the outcome of the test and each leaf represent a class label for classification tree. A tree can be created by splitting the training set into subset based on an attribute value test and repeating the process until each leaf of the tree contains a single class label or we reach the desired maximum depth. Here, we can easily see the most important features, i.e. the ones that split the tree in meaningful way to separate efficiently all the classes. We used the machine learning algorithms from the scikit-learn library. We choose decision tree and k-nearest-neighbor algorithms because of their speed and simplicity, the possibility of data analysis after learning and because they are the ones that gave the best results. We obtained results of 96.18% of precision for the decision tree algorithm on our test set and 92.65% with the k-nearest neighbor. Tables Tab. 2 and Tab. 3 show the confusion matrix of each algorithms. The results indicate a good accuracy although the number of samples is low. For better results, we need more data with many different variables.

	Detected Normal	Detected Stop	Detected Turn Left	Detected Turn Right
Is Normal	226	5	1	2
Is Stop	1	52	0	0
Is Turn Left	2	0	25	1
Is Turn Right	1	0	0	24

TAB. 2 – *Confusion matrix of decision tree.*

	Detected Normal	Detected Stop	Detected Turn Left	Detected Turn Right
Is Normal	228	1	3	2
Is Stop	1	51	1	0
Is Turn Left	6	1	16	5
Is Turn Right	4	1	0	20

TAB. 3 – *Confusion matrix of K Nearest Neighbours (KNN).*

6 Optimization: Optimal Action for a Driving Event

It is necessary to learn the best driving behavior following a certain driving event, taking account of the safe driving and green driving goal. To test the concept, we choose a specimen event: "*turn right*". To perform this: we use simulation with steering wheel's direction automated and the vehicle driving on a predetermined path. The actions on the throttle, brake and direction indicators are controlled by the decision-making algorithm. Following turn right event, we simulate the reaction of the vehicle in each of these actions. Multiple actions can be combined. Given the 10 speed management actions and 3 blinker actions (a total of 30 possible combinations) given below, our goal is to find the most appropriate action:

- *Speed management action*: (1) stop accelerating, (2) accelerate low, (3) accelerate medium, (4) accelerate high, (5) stop braking, (6) brake low, (7) brake medium, (8) brake high, (9) maintain speed, and (10) no action
- *Direction signal action*: (1) toggle left blinker, (2) toggle right blinker,(3) no action.

6.1 Scoring Functions

A scoring system is used to measure optimality of an action. Three scores, for (i) *safe driving*, (ii) *green driving*, and (iii) *comfortable driving*, are calculated. Currently, the safe driving score is a function of vehicle speed, the blinker state and the driver state. The green driving score depends on the reduced consumption of gas via brake use while comfortable driving is still a work in progress (not used in this paper). For each of these scores, we add "*penalty action*" that make the score smaller when we use more actions. Every action costs (0,1) point. It allows us to differentiate between two action sets that may give the same results but with different number of actions. We then choose the best action by getting the sum of different scores to maximize the final score. Currently, the safety score is between -4 and 4, and the green driving score is between -1 and 1.

6.2 Safe Driving Score

The safe driving score represents the safety of the driver during a driving event. Such safety depends on the vehicle's speed, the driver's focus on driving and the appropriate use of the direction signals (blinkers). The safe driving score is the sum of all safety scores:

```
Safe driving score = speed score + direction score + driver focus score        (1)
```

The speed score is calculated such that going to safe speed yields a score of 1, going 10 km/h under or 5 km/h over yields a score of 0, and going 20 km/h under or 10 km/h over yields a minimal score of -1. See Fig. 5(a). The blinker score is 1 if in a situation, a blinker is needed and that it has been activated or if a blinker is not needed and both blinkers are not activated. It is -1 if a blinker is needed in the situation but no blinker is activated or the wrong one is put on, and if a blinker is not needed and yet one is activated. See Fig. 5(b). Next, we calculate the security distance between two vehicles. Based on the French law on security distance of vehicle (France 2001), such distance must be at least the distance traveled in two seconds. Given that distance `d = velocity v * time t`, and that `t = 2 seconds`, hence if `v` is in km/h then `d = 5/18 * v`. In meter/second, the security distance is 5/9 of the vehicle's speed. The theoretical braking distance can be found by determining the work required to dissipate the vehicle's kinetic energy [8]. The braking distance is :

$$d = v^2 / (2\mu g) \qquad (2)$$

where μ is the coefficient of friction between the road surface and the tires, g is the gravity of the Earth, and d is the distance traveled. The maximum speed for a braking distance d is then $v = \sqrt{(2\mu g d)}$. We use a coefficient of friction of 0.70 which is typically used for automobile (Jernigan and M. F. Kodaman), thus giving $2\mu g = 13.7$. See Fig. 5(a). The driver's focus score is 1 if the person's focus is on the road; this score decreases with distraction. For now, this value is limited to checking if the driver is focusing in front. See Fig. 5(d). For now, our representative driving event is "Turn Right" event. However, looking forward, some other events should be added in the calculation of safe driving score function.

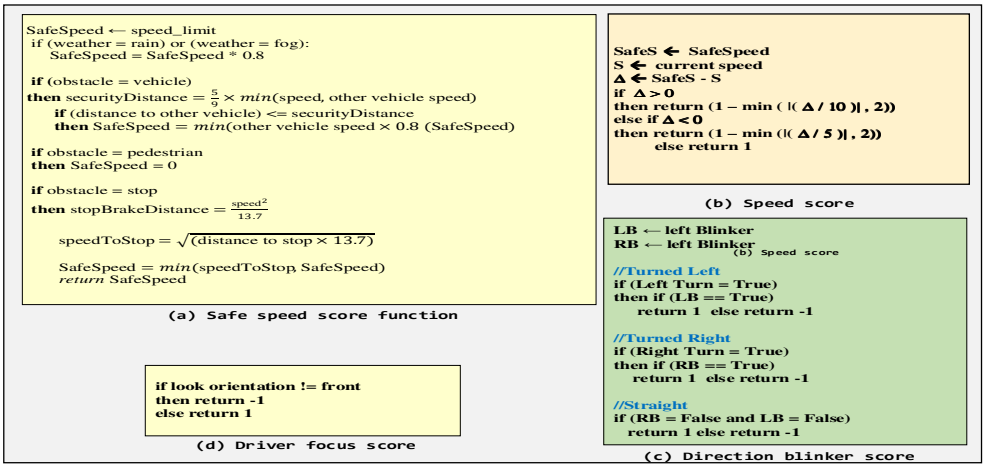


FIG. 5 – *The decison-tree learning algorithm's model for driving event classification.*

6.3 Green Driving Score and Exhaustive Action Search

The green driving score represents the ecological impact of a driver's action in a driving event. For now, it is limited to the unnecessary use of brake pedal. A vehicle consumes more petrol, hence more CO_2 emission, to slow down a vehicle moving in full speed. The more unnecessary brakes are done, the more CO_2 emissions are produced. Here, the green score starts at 1; as one brakes, the score goes down, up to -1 when breaking fully. The algorithm for green driving score function is given as: `brake ϵ [0,1]: return (1-brake² * 2)`. Given the finite set of actions, we search for an optimal action for the given driving event by starting from state `S` which is the start of the "right turn" event to final states `S'` after an action. We always end in state `S'` after using an action `A`. Thus, we can consider the reward of our action as function `f(A) = reward`. For this reason, an optimization algorithm would be the way to find the best action to take for a given situation. We only need to maximize the reward obtained from the function `f`. The method invoked is an exhaustive search due to the small number of possible actions and the use of a simulation. Fig. 6 shows a snapshot of the selection of actions for a right turn event, beginning from speed of 10 km/h without any blinkers, from the worst action to the best one. The action selected is "Blinker Right" without an action on speed which is a good action for the initial situation where our blinker was off.

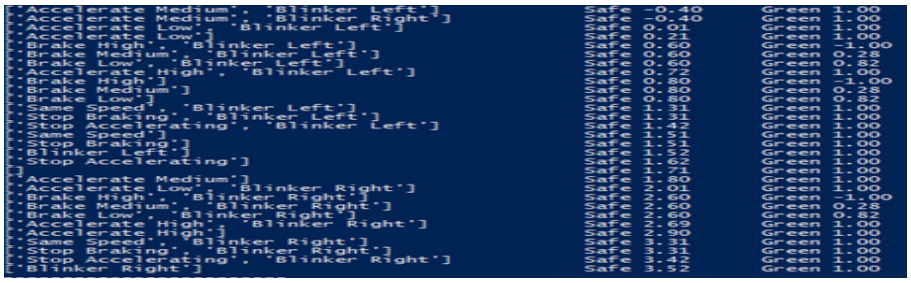


FIG. 6 – *Simulation: search for optimal action for a turn right driving event.*

7 Conclusions and Future Works

This paper presents the functionalities of an alternative ADAS. It is alternative because it is generic and can be adapted to any type of vehicle, contrary to the norm that ADAS is closed to proprietary constraints. The driving context is modeled using ontology and it is implemented in the driving simulator. The values obtained from the simulator are taken and passed to the ontological template to produce a real-time driving situation in which the actual values are those obtained from the simulator. Machine learning is invoked to identify the sampled driving event. Ours is a case of classification problem and we used supervised learning and decision tree learning in particular to solve the issue at hand. To decide what actions need to be invoked for the recognized driving event, an optimization solution is presented. The goal is to optimize driving features (safe, green and comfortable) of a proposed action to the driving event. Here, messages may be sent to the driver to assist him in driving. Future works include reinforcement learning to determine actions for the optimized action presented in this paper. We wish to add a cognitive user interface design (Peschl and Stary 1998) and the cognitive component (Kelly_III 2015) that reasons with purpose and interacts with humans naturally. Furthermore, this work forms part of a proposed European project on intelligent transportation to assist reducing road congestion during Paris 2024 Olympic Games.

References

Armand, A., et al (2014). Ontology-Based Context Awareness for Driving Assistance Systems. *IEEE Intelligent Vehicles Symposium*. Dearborn, Michigan, USA: 227–233.

Bengler, K., et al (2014). Three Decades of Driver Assistance Systems: Review and Future Perspectives. *IEEE Intelligent Transportation Systems*. 6: 6-32.

BOSCH. (2016). from http://www.bosch.com/assets/en/company/innovation/theme03.htm.

Charest, M. and S. Delisle (2006). Ontology-Guided Intelligent Data Mining Assistance: Combining Declarative and Procedural Knowledge. *IASTED International Conference*.

CVIS. from http://www.cvisproject.org/.

France (2001). Code de la route - Article R412-12.

Gruber, T. R. (1993). "A Translation Approach to Portable Ontology Specifications." *Knowledge Acquisition* 5(2): 199 - 220.

Guarino, N. (1995). "Formal ontology, conceptual analysis and knowledge representation." *Human-Computer Studies* 43(5-6): 625-640.

Hina, M. D., H. Guan, A. Soukane and A. Ramdane-Cherif (2018). "CASA: Fusion, Fission and Cognition of Driving Context for Driving Assistance System." *IEEE Transactions on Intelligent Transportation Systems* (submitted).

Jernigan, J. D. and M. F. Kodaman An investigation of the utility and accuracy of the table of speed and stopping distances, University of Michigan.

Kannan, S., et al (2010). "An Intelligent Driver Assistance System for Vehicle Safety Modeling Using Ontology Approach." *Intl Journal of Ubiquitous Computing* 1(3): 15 – 29.

Kelly_III, J. E. (2015). Computing, cognition and the future of knowing: How humans and machines are forging a new age of understanding: 1-11.

Li, L., D. Wen, N.-N. Zheng and L.-C. Shen (2012). "Cognitive Cars: A New Frontier for ADAS Research." *IEEE Transactions on Intelligent Transp. Systems* 13(1): 395 - 407.

Maalel, A., H. H. Mabrouk, et al (2011). "Development of an ontology to assist the modeling of accident scenario application on railroad transport." *Journal of Computing* 3(7).

Neches, R., R. Fikes, T. Finin, T. Gruber, R. Patil, T. Senator and W. R. Swartout (1991). Enabling Technology for Knowledge Sharing. *AI Magazine*: 36-56.

Noy, N. F. and D. L. McGuinness. "Ontology Development 101: A Guide to Creating Your First Ontology." Retrieved July 2017, from http://protege.stanford.edu/publications/

Peschl, M. F. and C. Stary (1998). "The Role of Cognitive Modeling for User Interface Design Representations." *Mind and Machines* 8(2): 203 - 236.

PREVENT. from http://www.prevent-ip.org/.

Protégé. (2016). "Protégé: Open-source ontology editor." from http://protege.stanford.edu.

SAFESPOT. from www.safespot-eu.org.

Tchankue, P., J. Wesson and D. Vogts (2013). Using Machine Learning to Predict Driving Context whilst Driving. *SAICSIT 2013, South African Institute for Computer Scientists and Information Technologists*. East London, South Africa: 47-55.

Vahidi, A. and A. Eskandarian (2003). "Research advances in intelligent collision avoidance and adaptive cruise control." *IEEE Transactions on Intelligent Trans.Systems* 4: 143–153.

VOWL. "Visual Notation for OWL Ontologies." from http://vowl.visualdataweb.org/v2/.

Résumé

Dans cet article, une aide à la conduite alternative est présentée. C'est une alternative dans la mesure où elle n'est pas fermée aux contraintes propriétaires, ce qui est le cas pour les constructeurs automobiles. L'apprentissage automatique est utilisé pour identifier un événement de conduite et une optimisation est effectuée pour identifier les actions optimales qui doivent être effectuées après l'événement de conduite identifié. L'ontologie est utilisée pour représenter les connaissances et les situations de conduite. Pour la partie d'apprentissage, un ensemble d'apprentissage est créé, stockant les modèles de situation de conduite et à partir duquel un système intelligent peut être utilisé pour déterminer la situation de conduite. Pour la partie optimisation, des algorithmes sont développés pour maximiser les scores d'une action qui correspond à une conduite sécuritaire, à une conduite écologique et à une conduite confortable. Ce travail est une contribution pour rendre les systèmes d'aide à la conduite abordables et disponibles pour les véhicules que les gens ordinaires possèdent.

La confidentialité des entrepôts de données dans le Cloud Computing à base de profil utilisateur

Amina EL Ouaazzani[1], Nouria Harbi[2], Hassan Badir[3]

[1 3]Laboratoire LabTIC, Ecole Nationale des Sciences Appliquées Tanger, Maroc
{a.elouazzani2000,hbadir}@gmail.com
[2] Laboratoire Eric, Université Lyon II, France
nouria.harbi@univ-lyon2.fr

Résumé.
Un entrepôt de données (ED) présente un facteur primordial de l'entreprise qui donne une vue clair sur ses activités et une source riche pour les décideurs. Il contient les données sensibles sur l'entreprisse et ses clients, et par conséquence elles ne doivent pas être accessibles sans contrôle d'accès.(**El ouazzani 2018**) La solution de l'hébergement de l'ED dans le CC (Cloud Computing) gagne progressivement plus de popularité dans les entreprises, car elle permet de surmonter l'expansion sans fin des données et bénéficier de sa capacité de traitement et le stockage de ces données. Cependant la confidentialité de ces EDs dans le CC a besoin de nombreuses améliorations et de la mise en place des normes précises, en raison de l'évolutivité et l'élasticité du paradigme CC, car il n'y a pas un protocole standard pour gérer la connectivité des utilisateurs du CC aux ressources hébergés en prenant compte la performance d'exécution des requêtes. L'objectif de nos travaux est de proposer un cadre garantissant la confidentialité des EDs hébergés dans le CC à base de profil utilisateur.

1 Introduction

Un ED représente est un facteur primordial pour la haute direction qui cherche à prendre les bonnes décisions stratégiques. Il présente une source des données cruciales de l'entreprise et la vie privée de ses clients telle que les données médicales, financières protégées par des lois, parmi ces lois, HIPAA[1] (Health Insurance Portability and Accountability Act HHS (1996)). Et par conséquence elles ne doivent pas être accessibles sans contrôle d'accès.

A un certain stade, et malgré l'excellent utilitaire, Le maintien de ces EDs de grande quantité de données au sein de l'entreprise implique un grand investissement matériels et ressources humaines afin de les gérer. Aujourd'hui, la solution de l'hébergement de l'ED dans le CC gagne progressivement plus de popularité dans les entreprises, car elle permet de surmonter l'expansion sans fin des données et bénéficier de sa capacité de traitement et le stockage de ces données.

Le contrôle d'accès est l'un des mécanismes de sécurité les plus importants des services de CC qui garantit la confidentialité des données, cependant le service CC ne peut pas appli-

[1]http://www.hhs.gov/hipaa/

quer le modèle de contrôle d'accès traditionnel en raison de son évolutivité et son élasticité, car il n'y a pas un protocole standard pour gérer la connectivité des utilisateurs du CC aux ressources hébergés en prenant compte la performance d'exécution des requêtes **(Blanco, et al., 2015)**.

Dans cet article, nous présentons nos travaux qui se composent de trois parties :

- La première partie décrit un mécanisme de contrôle d'accès qui aide le propriétaire à bien définir les permissions de chaque profil utilisateur selon son rôle dans l'entreprise, et de calculer le niveau de sensibilité de chaque élément de l'ED.
- Notre deuxième contribution permet de détecter des cas d'inférence d'un utilisateur qui occupe un ou plusieurs rôles en analysant la totalité des permissions en se basant sur cinq règles proposées.
- La troisième partie, traite la confidentialité des EDs hébergés dans le CC à base de profil usage d'un utilisateur, en augmentant le coefficient de confidentialité/performance afin de contrôler l'accès à l'ED contenant une grande quantité de données en maintenant l'évolution du système et l'optimisation de temps de réponse.

Après la présentation de la problématique dans la section 1, le reste de cet article est structuré comme suit. La section 2 présente une vue d'ensemble des travaux connexes. La Section 3 décrit l'architecture proposée dont l'accès est basé sur les profils des utilisateurs. La section 4 présente la mise en œuvre et le test de notre solution. Enfin, la section 5 présente nos conclusions et perspectives.

2 Etat de l'art et synthèse

2.1 Etude de l'existant

Nous avons organisé les travaux selon deux parties, la première est consacrée à la présentation des modèles de contrôle d'accès intégrant la confidentialité dans le processus de modélisation des EDs (niveau conception), Dans la deuxième partie nous traitons les modèles de contrôle d'accès pour un ED déjà mise en place (niveau exploitation) **(El ouazzani, et al.,2015).**

2.1.1 La confidentialité dans le processus de modélisation des EDs

Afin de garantir la confidentialité d'ED au niveau conception, certains auteurs **(Rosenthal, 2000), (Saltor, et al., 2002)** ont proposé l'utilisation des autorisations définit au niveau des sources de l'ED. Alors que d'autres auteurs **(Trujillo, et al., 2009) (Soler, et al., 2008)** ont considèrent cette proposition non performante puisque l'ED à ces propres caractéristiques. Dans des travaux recent, le langage UML (**U**nified **M**odeling **L**anguage) présente un standard afin de modéliser les regles de sécurité d'un ED. Dans ce sens, nous pouvons citer le travail **(Blanco, et al, 2015)** qui présente une architecture MDA (Model Driven Architecture) automatique pour sécuriser un ED, cette architecture est composé d'un modèle logique et ses transformations depuis le modèle conceptuel en utilisant l'extension de UML et le package CWM (**C**ommon **W**arehouse **M**etamodel). Ainsi que le travail **(Arora, et al., 2016)** qui modélise le contrôle d'accès à des données sensibles de l'ED dans la phase analyse

et conception, à l'aide des diagrammes UML et la programmation orientée objet est la dernière tendance dans l'industrie du logiciel en raison de ses différentes fonctionnalités.

A noter également que la gestion des inférences s'est inspirée et s'inspire encore aujourd'hui des travaux réalisés dans le domaine des ED ou les bases de données en général. On retrouve dans la littérature le travail de **(Triki, et al., 2013)** qui propose un modèle pour sécuriser les données multidimensionnelles contre les inférences précises et partielles. Cette approche consiste à identifier les éléments sensibles à protéger en interrogeant le concepteur de l'ED. Ensuite, le propriétaire construit un graphe permettant de détecter les combinaisons sensibles. Par contre le travail de **(Blanco, et al., 2010)** traite une approche basée sur le diagramme états-transactions pour détecter les inférences au niveau de la conception. Cette proposition se focalise sur les requêtes sensibles et ses évolutions. Ce travail indique que la combinaison de plusieurs permissions peut être plus sensible, ce qui est approuvé dans le travail de **(Sweeney, 2002).** Ce travail décrit un cas réel de l'inférence des données sensibles, par une démonstration d'identification du nom de l'ancien gouverneur « William Weld » et ses dossiers médicaux en se basant sur le croisement des données d'un groupe d'assurance, et une liste d'inscription des électeurs. Nous trouvons également, le travail de **(Accorsi, et al., 2013)** qui propose une approche dans laquelle les règles d'inférences sont connues par le moteur d'inférence sans mentionner comment les préciser. Le diagramme proposé montre un processus de détection des inférences qui se compose d'une politique composée dans lequel l'utilisateur compose la politique et les règles de confidentialité. Ensuite, le moteur d'inférence prend cette politique qui calcule à son tour toutes les fermetures d'inférence possibles de la politique entrée en se basant sur un algorithme. Et Le noyau teste pour chaque élément non noyau s'il est obtenu à partir d'un élément noyau.

2.1.2 La confidentialité des EDs au niveau exploitation

Le contrôle d'accès à un ED déjà mis en place prend en compte l'emplacement de l'ED qui peut être le site de l'entreprise ou chez un fournisseur CC :

- **La confidentialité d'un ED sur le site de l'entreprise :** la plupart des méthodes de contrôle d'accès à l'ED déjà proposées se basent sur le profil utilisateur. Ce dernier souffre toujours du problème de l'efficacité dans la gestion de l'intégrité. Dans ce sens les auteures **(Thangaraju, et al., 2016)** ont proposé un profil multi-utilisateurs orienté vers la gestion de l'intégrité basée sur la mesure des profondeurs d'accès en fonction du niveau des objets appelés et du niveau d'accès autorisé à l'utilisateur et du nombre d'objets auxquels l'utilisateur a accès. Alors que dans un autre travail **(Kechar, et al., 2015),** qui propose un système de contrôle d'accès basé sur les rôles en exploitant l'architecture de la norme XACML, afin de garder les performances des requêtes d'aide à la décision.
- **La confidentialité d'un ED hébergé dans le CC :** Malgré les avantages de la solution de l'hébergement d'un ED dans le CC, la confidentialité des données dans cet environnement reste un risque à traiter. Parmi les travaux qui traitent cette problématique on trouve (**Al-Aqrabi, et al., 2015)** qui se focalise sur la sécurité des systèmes décisionnels hébergés dans le CC et il décrit deux modèles de gestion des accès en tenant en compte le rapport temps de réponse/sécurité. Alors que d'autres travaux **(Bensaidi, et al., 2012) (Ray, et al., 2014)** se basent sur la notion de la confiance, on se focalisant sur la diminution du niveau de confiance affecté à l'utilisateur lors d'une tentative de violation des droits fixés. Nous trouvons également le travail **(Naushahi, 2016)** qui utilise le concept de liste de contrôle d'accès ACLs en intégrant la notion de Profile en définis-

sant des règles pour chaque profil afin d'accorder l'accès à un système et à des ressources hébergé dans le CC, Les résultats de la simulation montrent que cette solution offre un temps d'accès aux données réduit en diminuant les demandes d'authentification

2.2 Limitation des solutions existantes

La protection des ED contre les accès illégaux s'est fait sentir et traiter d'une manière incontestable dans plusieurs travaux **(Fernandez-Medina, et al., 2006) , (Soler, et al., 2008), (Trujillo, et al., 2009)**. Suite à l'étude des travaux existants, nous avons constaté les points suivants :

- La confidentialité des ED a été traditionnellement considérée dans la mise en œuvre définitive d'un ED **(Villarroel, et al., 2006) (Eavis, et al., 2012)**, par contre les travaux les plus récents **(Blanco, et al., 2015) (Rodriguez, et al., 2011)** considèrent son inclusion dans les stades de développement ce qui peut produire des solutions de qualité plus robustes, ainsi le système peut accueillir ces exigences de sécurité d'une façon plus naturelle.
- La majorité des travaux de recherche, surtout ceux intervenant dans la phase de modélisation conceptuelle, se sont appuyés sur le méta modèle CWM, dans le but de concevoir un ED sécurisé. Sachant que le modèle CWM est basé sur trois standards, à savoir UML, MOF et XMI. pour représenter correctement toutes les règles de sécurité et d'audit définies dans la modélisation conceptuelle des ED.
- Une architecture MDA pour une conception automatique, sécurisé d'un ED est appliquée dans **(Blanco, et al., 2015), (Inmon, 1991)**, mais les deux approches ont été incapables de comprendre des règles de sécurité qui sont complexes.
- La plupart des travaux modélisent le contrôle d'accès à base des politiques RBAC (**R**ole **b**ased **A**ccess **C**ontrol) et MAC (**M**andatory **A**ccess **C**ontrol), alors que le profil utilisateur est considéré comme une table isolée qui regroupe les données nécessaires pour l'accès d'un utilisateur d'une façon statique sans la prise en compte des priorités de l'utilisateur authentifié.
- Bien que les autorisations présentent l'axe principal pour garantir la confidentialité de l'accès à l'ED, cependant l'absence d'une norme qui gère la précision de ces autorisations peut provoquer des incohérences et des inférences comme conséquences. Dans ce sens, certains auteurs **(Rosenthal, 2000) , (Saltor, et al., 2002)** ont proposé de tirer le modèle de contrôle d'accès à l'ED, à partir des sources de données, tandis que d'autres auteurs **(Priebe, et al., 2001) (Fernández-Medina, et al., 2007)** ont considéré cette proposition difficile puisque les données sources proviennent de différents systèmes (avec des politiques différentes). Ainsi que les systèmes opérationnels utilisent le modèle relationnel alors que les systèmes OLAP utilisent le modèle multidimensionnel.
- A noter également, que la notion d'inférence a été citée dans plusieurs travaux en tant qu'élément essentiel pour garantir la confidentialité, et dont la maîtrise est cruciale. Néanmoins, malgré les risques élevés d'inférences, il n'est pas suffisamment pris en compte dans la phase conceptuelle.
- Aucun travail ne propose une méthode qui permet d'assurer la cohérence des permissions d'un utilisateur selon son profil.
- Aucun travail ne propose une méthode conviviale pour détecter les combinaisons sensibles qui peuvent provoquer des inférences.

Nous constatons que la plupart des travaux affecte la tâche de la classification des données selon leur niveau de sensibilité (Très sensible, sensible, confidentiel) au propriétaire de données. Sachant que selon le rôle de l'utilisateur, le propriétaire de données lui affecte un niveau de sensibilité des données pour accéder à des données possédant le même niveau de sensibilité ou inférieure. Le propriétaire de l'ED peut alors attribuer un niveau de sensibilité moins important à une donnée cruciale. Il en résulte cependant un problème de perte de confidentialité de l'information. De plus, les permissions définit au niveau des sources ne sont pas suffisamment exploitées pour aider le propriétaire à bien déterminer les permissions d'un utilisateur de l'ED.

A noter également que la solution de l'hébergement de l'ED dans le CC prend de plus en plus sa place dans les entreprises. Afin de bénéficier de ses avantages, des nouveaux défis concernant la sécurité des données hébergés ont été posés par la multi-location, l'élasticité et l'évolutivité de ce paradigme. Suite à l'étude des travaux existants dans ce sens, nous avons constaté aussi les points suivants :

- D'après le travail de **(Moussa, et al., 2013) et (Al-Aqrabi, et al., 2013)**, le mécanisme de contrôle d'accès ne doit pas influencer l'évolutivité et la performance de l'ED hébergé dans le CC en évaluant la charge des traitements sur l'échelle de temps, et en mesurant le nombre des requêtes traitées au cours d'un intervalle de temps. Un système évolutif, devrait maintenir le même nombre. Alors qu'ils n'ont pas proposé un mécanisme dans ce sens.
- les auteurs **(Naushahi, 2016)** ont utilisé le concept de liste de contrôle d'accès ACLs afin d'accorder l'accès à un système et à des ressources hébergées dans le CC, Les résultats de la simulation montrent que cette solution offre un temps d'accès aux données réduit. Par contre ils n'ont pas utilisé l'historique des accès qui peut minimiser le trafic et par conséquence réduire le temps de réponse.
- Plusieurs travaux **(Bensaidi, et al., 2012), (Ray, et al., 2014)** proposent d'utiliser la notion de confiance qui consiste à attribuer un niveau de confiance à chaque utilisateur, chaque tentative de violation provoque sa diminution, Après un nombre bien défini des tentatives malveillantes, le connecté perd tous ses privilèges au sein de l'entreprise. Ce qui peut dégrader la performance, augmenter la charge de traitement en recalculant le niveau de confiance à chaque tentative de violation, et retarder le travail d'un utilisateur en lui retirant ses autorisations initiales d'accès.

Donc, la migration des ED vers le CC devrait améliorer la satisfaction de l'utilisateur final et induire une plus grande productivité de l'entreprise. Ce qui nécessite une haute performance qui peut être garanti par la mise en œuvre de l'intra-parallélisme de requête qui consiste à décomposer une requête complexe en sous-requêtes, et les traiter sur plusieurs processeurs, et enfin effectuer le post-traitement pour présenter une réponse à la requête principale **(Moussa, et al., 2013)**. Alors que la mise en place d'un mécanisme de contrôle d'accès ne doit pas augmenter la charge des traitements dont le but est d'avoir un système évolutif et productif avec des données qui sont bien protégées contre l'accès aux données interdites puisque ces données seront confiées à un prestataire externe.

Ce mécanisme de contrôle d'accès ne doit pas influencer l'évolutivité de l'ED hébergé dans le CC en évaluant la charge des traitements sur l'échelle de temps, et en mesurant le nombre des requêtes traitées au cours d'un intervalle de temps **(Moussa, et al., 2013).** Dans la section suivante, nous présentons l'architecture de notre proposition qui pallie à ces limites.

3 Architecture globale proposée

D'après la synthèse des travaux réalisés qui traitent la confidentialité des ED, nous avons constaté que l'implémentation et l'administration des permissions en utilisant les modèles MAC, DAC et RBAC d'une façon manuelle est difficile et insuffisante, ce qui a motivé la création du modèle de contrôle d'accès dynamique à base de profil utilisateur. Ce modèle se compose de parties montrées dans notre architecture qui sont :

- **Interface Propriétaires de données** : l'interface administrateur permettant au propriétaire de données, d'accéder à la couche contrôle d'accès avec ses différents modules.
- **ED dans le CC** : présente la partie CC qui contient les données multidimensionnelles hébergées.
- **Couche contrôle d'accès à base de profils utilisateur** : Cette couche contient les trois modules proposés afin de contrôler l'accès à l'ED hébergés dans le CC, qui sont (Figure 1:
  - La classification dynamique des niveaux de sensibilité basée sur les profils utilisateur (1).
  - La détection des inférences par la combinaison de plusieurs profils (2).
  - La gestion des profils à partir des usages (3).

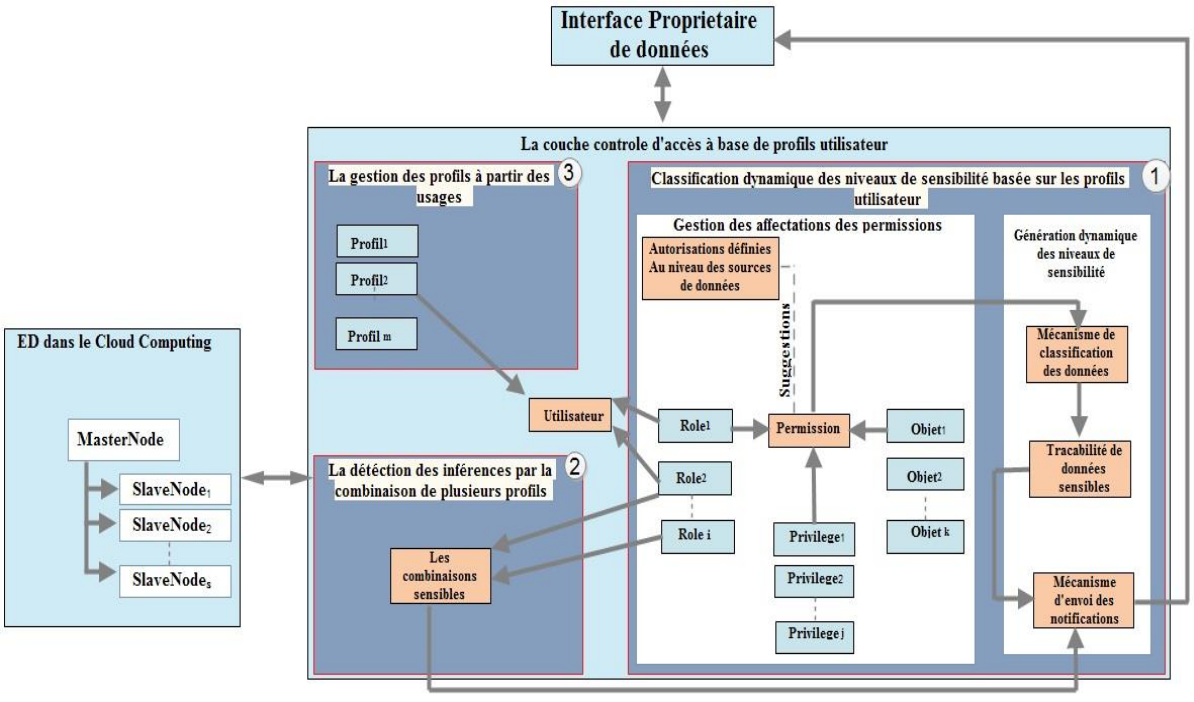


Figure 1 Architecture globale proposée

3.1 Description des modules de l'architecture proposée

Selon les travaux étudiés, et en se basant sur leur synthèse nous proposons deux contribution présentés dans les deux parties :

- La confidentialité des ED : consiste à définir les permissions ou les droits d'accès de chaque utilisateur selon son rôle, à générer le niveau de sensibilité des données, et à détecter les inférences en utilisant des combinaisons des données autorisées.
- L'implémentation de notre approche dans un environnement CC : consiste à gérer l'accès à l'DW hébergé dans le CC selon le profil usage de l'utilisateur.

3.1.1 La classification dynamique des niveaux de sensibilité basée sur les profils utilisateur

Ce module permet de générer automatiquement le niveau de sensibilité de chaque objet de l'ED à base des profils utilisateurs. Il se compose de deux phases (**El ouazzani, et al.** D**écembre 2016**) (**El Ouazzani.et al. 2018**) :

Phase 1 : Il s'agit de définir les permissions d'un rôle sur une donnée avec un privilège précisé, en prenant compte les autorisations définies au niveau des sources de données. Il s'agit de suggérer ces derniers à l'administrateur propriétaire des données de l'ED, lors de l'affectation des permissions.

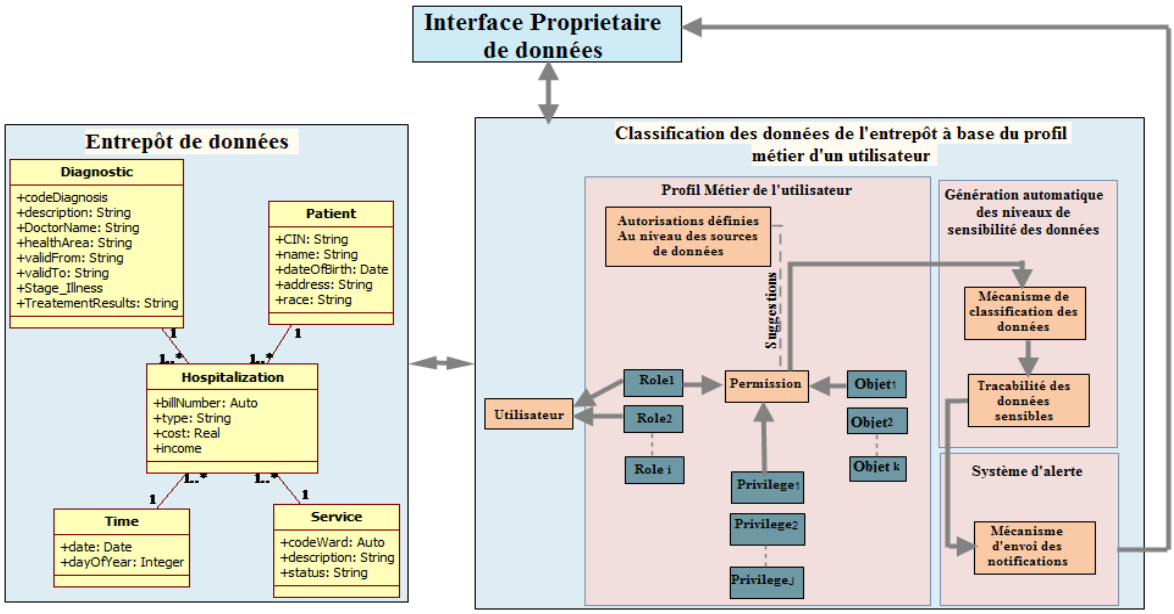


Figure 2 Classification des données de l'entrepôt à base de profil métier d'un utilisateur

Phase 2: génération automatiquement les niveaux de sensibilité des données de l'ED en se basant sur les permissions définies dans la phase 1. Ensuite le mécanisme **de traçabilité des données sensibles** permet de tracer les actions des utilisateurs sur les données qui ont un niveau de sensibilité élevé. Selon le niveau de sécurité détecté par le mécanisme de génération automatique des niveaux de sensibilité des données, notre système d'alerte permet d'envoyer des notifications à l'administrateur lors d'une tentative de violation d'une permission sur une donnée sensible.

3.1.2 La détection des inférences par la combinaison de plusieurs profils

Afin de permettre une extraction automatique des inférences à partir des permissions autorisées, nous proposons un modèle informatique visuel avec des règles à vérifier en se basant sur la présentation graphique des profils accordés à un utilisateur et les liens entre les données en utilisant le diagramme de classe source. Ce module est la suite du travail de (Triki, et al., 2013) qui propose une méthode de détection des inférences précises et partielles, cependant notre proposition consiste à détecter les combinaisons sensibles.

Sachant qu'un utilisateur peut avoir un ou plusieurs rôles au sein de l'entreprise, ce dernier accède à l'ED avec un ou plusieurs profils. Le but de notre système de détection des inférences est de détecter si l'utilisateur peut déduire indirectement des informations non autorisées en utilisant deux ou plusieurs permissions depuis un ou plusieurs profils différents.

Le module 2 de notre architecture globale que nous allons détailler dans cette partie, se focalise sur la détection visuel des inférences par la combinaison de plusieurs permissions par un même utilisateur. Afin d'envoyer les combinaisons sensibles détectées au propriétaire de données.

La figure 3 présente l'architecture détaillée de notre module de détection des inférences, qui permet d'analyser les permissions de chaque profil, afin de détecter les combinaisons sensibles, il se base sur deux entrées qui sont (**El ouazzani, et al. 2017**)

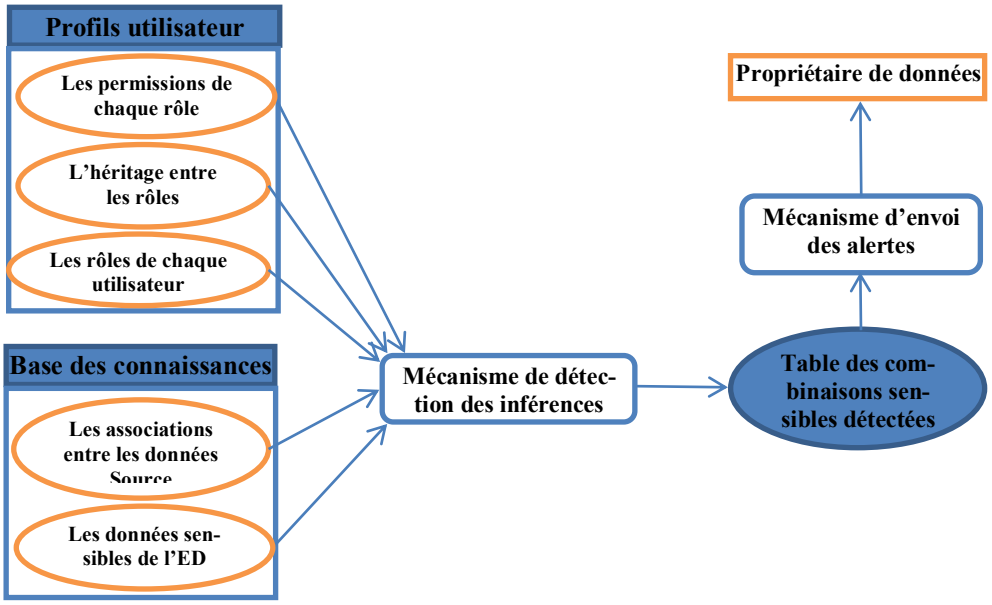


Figure 3 la détection des inférences par la combinaison de plusieurs profils

- **Les profils utilisateur :** présentent les utilisateurs qui endossent des rôles qui leurs sont attribués par le propriétaire de données. Ces rôles sont organisés en hiérarchie, et ils disposent des permissions qui ne peuvent pas être accordées directement aux utilisateurs.
- **La base des connaissances :** ce sont les données sensibles de l'ED à protéger contre les inférences, et les associations entre les données selon le diagramme de classe de la base de données source.

.

3.1.3 La gestion des profils à partir des usages

Notre but dans cette partie est de proposer une politique de contrôle d'accès performante à base des profils usage des utilisateurs PACUP (Performing Access Control based on Use Profiles) qui présente le module 3 de notre contribution. Cette politique économique combine entre le profil métier et le profil usage d'un utilisateur. Sachant que le profil métier était l'objet du module 1 (El Ouazzani, et al., 2016) de notre contribution qui définit le rôle et les permissions d'accès autorisées à un utilisateur et qui classifie les données de l'ED selon leur niveau de sensibilité. Donc PACUP permet de sécuriser l'accès avec le profil métier de l'utilisateur, et minimiser en même temps le trafic et l'échange de données entre le CC et l'organisation, en affectant un profil usage pour chaque utilisateur.

Nous cherchons à améliorer le premier scénario de (Al-Aqrabi, et al., 2015) qui se comporte d'une manière hautement sécurisé et qui garantit la facilité de gestion de l'application de la confidentialité. Nous proposons un nouveau modèle (figure 4) dont la politique de la gestion des profils utilisateurs PACUP est centralisé ce qui dit la facilité de la gestion, et en même temps cette politique optimise le trafic entre le CC et l'organisation. Cette solution minimise la charge de traitement et le temps de réponse par l'ajout d'un profil usage pour chaque utilisateur, qui gère l'utilisation et l'accès optimal à l'ED. Notre modèle comprend deux parties :

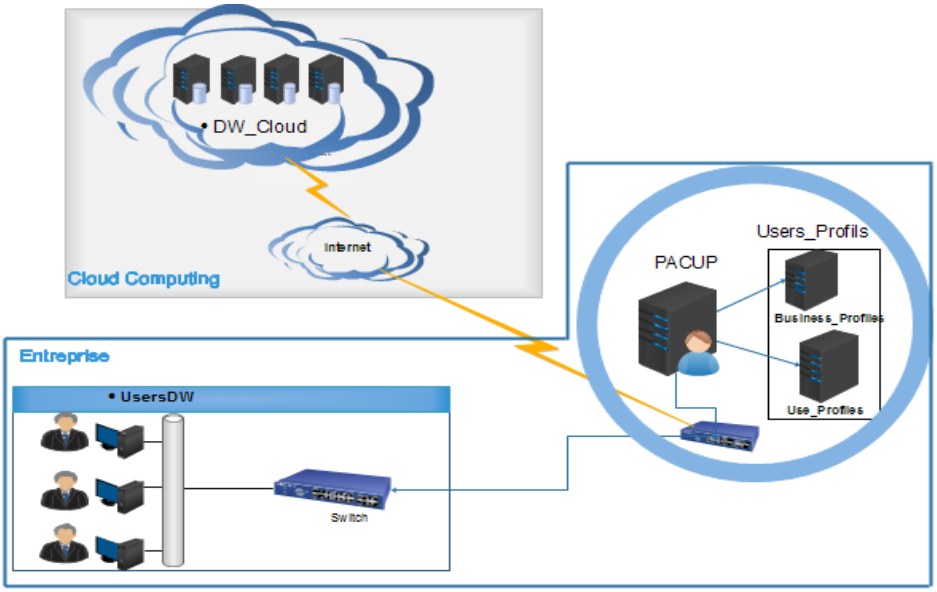


Figure 4 Architecture globale de PACUP

- **DW_Cloud :** présente les données multidimensionnelles hébergées dans le CC.
- **Entreprise :**
  - **usersDW** : c'est le réseau local de l'entreprise regroupant les utilisateurs.
  - **PACUP** : une politique qui permet de gérer l'accès à l'ED hébergé dans le CC en combinant entre le profil usage et les profils métiers d'un utilisateur. Elle permet également de gérer le cache. Donc les clients ont besoin de vérifier eux-mêmes au cours de l'accès aux services de CC en analysant les requêtes demandée selon :
    - **Profil Métier de l'utilisateur :** Il s'agit de définir les permissions d'un utilisateur selon son rôle sur une donnée avec un privilège précis, en prenant en compte les autorisations définies au niveau des sources de données (El Ouazzani, et al., 2016).
    - **Profil usage de l'utilisateur :** c'est un profil regroupant les droits d'usage de chaque utilisateur. Nous détaillons les composants de ce profil dans la partie suivante.

L'idée principale de l'approche est de créer pour chaque utilisateur un profil usage selon son rôle, et un cache partagé entre l'ensemble des utilisateurs. Ce cache se base sur les préférences et les usages précédents ou historiques des accès des profils. L'objectif est de faire face aux problèmes de performance d'accès à un ED dans le CC en réduisant le temps d'accès et en minimisant le trafic avec le CC.

Un ED de données hébergé dans le CC serve un grand nombre des utilisateurs. Le système de cache des requêtes dans les bases de données est un axe de recherche prometteur, en particulier dans les systèmes OLAP où l'utilisateur navigue interactivement dans un cube en lançant une séquence de requêtes. Les utilisateurs peuvent avoir des résultats qui ne les satisfont pas à cause du temps de réponse.

En outre, nous sommes devant la problématique de trouver les requêtes à garder en cache afin d'améliorer la performance de notre modèle de contrôle d'accès à l'ED hébergé dans le CC et optimiser le temps de réponse.

4 Implémentation

4.1 Outils et environnement de développement

Pour la mise en œuvre de notre contribution et la réalisation des expérimentations, nous avons utilisé une machine DELL PRECISION T1700 sous Windows 7 professionnel 64 bits. cette machine est dotée d'un processeur Intel Core i7-4770 de 3.40 Ghz et 8 Go de RAM. Nous avons utilisé les outils logiciels suivants :

- Le langage Java avec l'environnement de développement Eclipse. Ce choix a été motivé par les avantages qu'offre ce langage en termes de portabilité, de robustesse ainsi que la disponibilité de nombreuses bibliothèques ;
- Le SGBD Oracle version 12c pour concevoir l'entrepôt de données de tests. Ce choix a été principalement argumenté par la disponibilité de l'option OLAP offrant en particulier un moteur analytique OLAP, des espaces de travail et un gestionnaire d'espace de travail analytique (Analytic Workspace Manager-AWM) ;

- Le SGBD relationnelles MySql pour concevoir notre méta-modèle (Figure 7). Ce choix a été principalement argumenté par sa simplicité d'utilisation et ses interfaces pour effectuer diverses opérations.

4.2 Interface d'affectation des permissions

L'affectation des permissions est une tâche que seul l'administrateur s'en charge. Dans cette interface on choisit le rôle qu'on va affecter à un utilisateur. Le programme nous affiche les informations concernant cet utilisateur ainsi que ses objets autorisés au niveau des bases de données source, ce qu'on a déjà traité dans (El Ouazzani, et al., 2016), afin d'aider l'administrateur à bien définir les permissions. Cette interface permet de préciser le type d'objet (table/colonne/valeur), et les privilèges à autoriser.

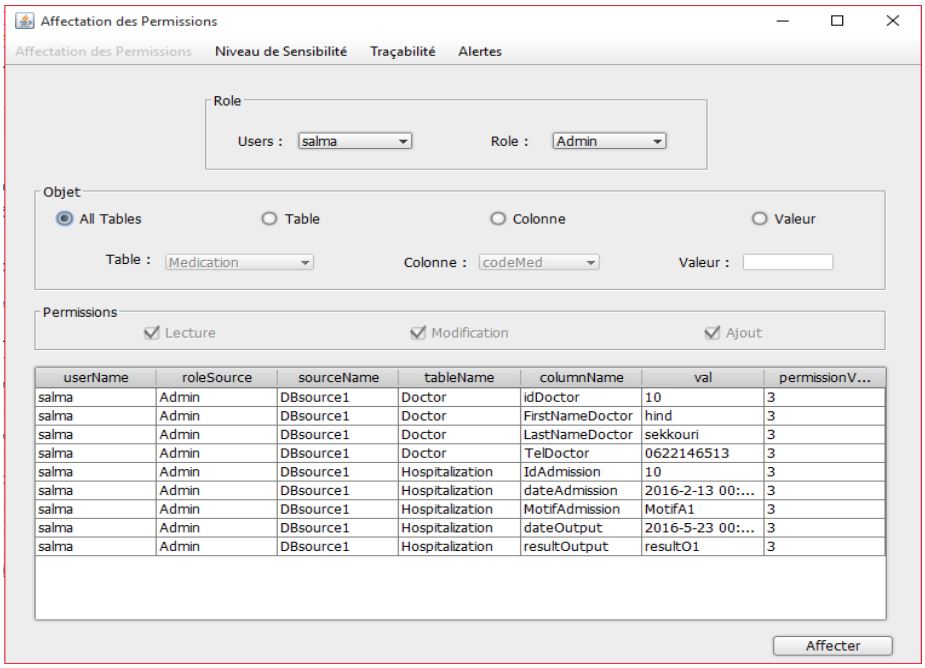


Figure 5 Interface d'affectation des permissions

4.3 Interface de détection des inférences

Cette interface permet d'afficher les profils affectés à l'utilisateur choisi. Ensuite l'administrateur peut vérifier l'existence d'une inférence selon les 5 règles proposées dans cet article. (**El Ouazzani.et al. 2018**)

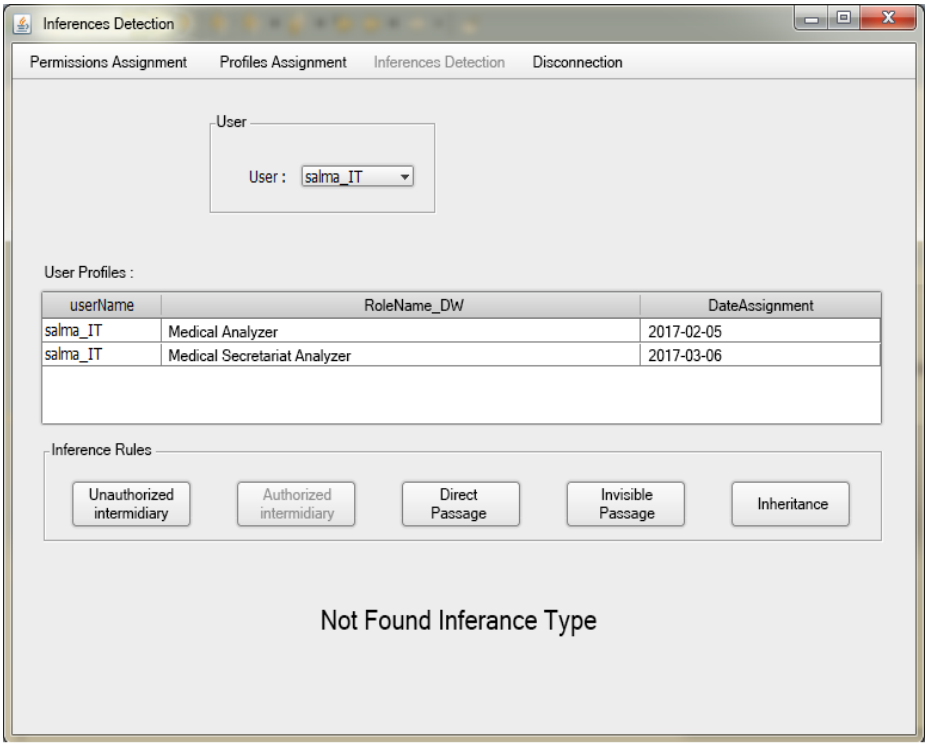


Figure 6 Interface de détection des inférences «Intermédiaire autorisé»

Dans cette capture d'écran (Figure 7), notre programme a détecté une inférence de type « Passage direct » entre les deux rôles « Analyseur Médical » et « Analyseur Secrétariat Médicale ». (**El Ouazzani.et al. 2017**)

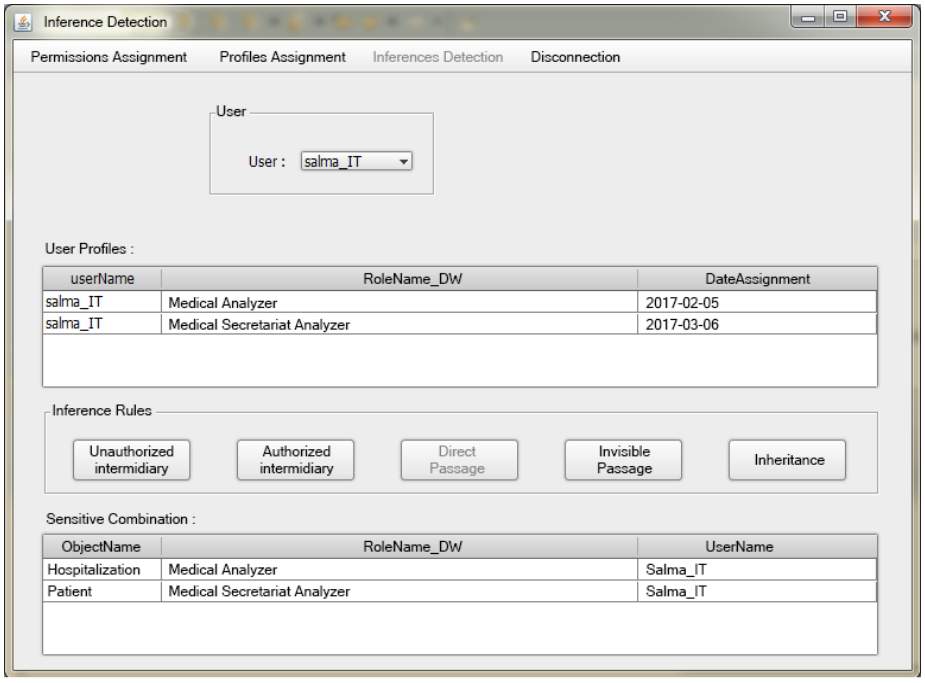


Figure 7 Interface de détection des inférences « Passage direct»

5 Conclusion

Nos travaux de recherche se situent dans le cadre de la confidentialité des ED hébergés dans le CC à base de profil utilisateur. Nous avons décomposé notre sujet en deux parties. La première partie consiste à garantir la confidentialité de l'ED contre les accès illégaux et les inférences, en se basant sur le profil métier de l'utilisateur. La deuxième partie consiste à adapter la première partie de notre contribution dont l'objectif est de contrôler l'accès d'une façon performante à un ED hébergé dans le CC. Nos propositions s'articulent autour de trois modules :

- Une méthode de classification des données de l'ED selon leur niveau de sensibilité.
- Une méthode de détection des inférences entres les profils utilisateurs.
- Une méthode de contrôle d'accès aux ED hébergés dans le CC à base des profils usage des utilisateurs.

Références

Accorsi, R. and Müller, G. 2013. Preventive inference control in data-centric business models. s.l. : Security and Privacy Workshops (SPW), 2013 IEEE (pp. 28-33). IEEE. , 2013

Al-Aqrabi, H., et al. 2013. Business intelligence security on the clouds: challenges, solutions and future directions. s.l. : Service Oriented System Engineering (SOSE) IEEE 7th International Symposium on (pp. 137-144). I, 2013.

Al-Aqrabi, H., et al. 2015. Business intelligence security on the clouds: challenges, solutions and future directions. . s.l. : Service Oriented System Engineering (SOSE), IEEE 7th International Symposium on (pp. 137-144). I, 2015.

Arora, D. and Kumar, U. 2016. Protecting Sensitive Warehouse Data through UML based Modeling. s.l. : Proceedings of the International Conference on Informatics and Analytics (p. 31). ACM., 2016.

Bensaidi, M., Aboulkalam, A. and Marzouk, A. 2012. Politique de contrôle d'accès au cloudcomputing: Recommandation à base de confiance. s.l. : Network Security and Systems (JNS2), National Days of (pp. 90-96). IEEE., 2012.

Blanco, C., et al. 2010. Towards the secure modelling of olap users behaviour. 2010.

Blanco, C., et al. 2015. An architecture for automatically developing secure OLAP applications from models. . s.l. : Information and Software Technology, 59, 1-16, 2015.

Eavis, T. and Althamimi, A. 2012. Olap authentication and authorization via queryre-writing. s.l. : The Fourth International Conference on Advances in Databases, Knowledge, and Data Applications, 130–139, 2012.

El Ouazzani, A., Harbi. N., Badir. H. 2015. Confidentialité des entrepôts de données dans le Cloud Computing: Etat de l'art et Perspectives.. 9ème édition de la conférence sur les Avancées des Systèmes Décisionnels. ASD, 2015.

El Ouazzani, A., Rhazlane, S., Harbi. N., Badir. H. 2016 . Dynamic management of data warehouse security levels based on user profiles. s.l. : Information Science and Technology (CiSt). 4th IEEE International Colloquium on (pp. 59-64). IEEE., 2016 .

El Ouazzani, A., Harbi. N., Badir. H. December 2016 . DYNAMIC CLASSIFICATION OF SENSITIVITY LEVELS OF DATA WAREHOUSE BASED ON USER PROFILES. s.l. : International Journal of Database Management Systems (IJDMS) on Volume 8, Number 6, 2016 .

El Ouazzani, A., Harbi. N., Badir. H. 2017. LA DETECTION DES INFERENCES PAR LA COMBINAISON DE PLUSIEURS PROFILS.. INTIS, 2017.

EL OUAZZANI, A., Harbi, N., & Badir, H. (2018). User Profile Management to protect sensitive data in Warehouses. *INTERNATIONAL JOURNAL OF NEXT-GENERATION COMPUTING*, *9*(1) 2018.

Fernandez-Medina, E., et al. 2006. Access control and audit model for the multidimensional modeling of dws. . s.l. : Decision Support Systems, 1270–1289., 2006.

Fernández-Medina, E., et al. 2007. Developing secure data warehouses with a UML extension. s.l. : Information Systems, 32(6), 826-856., 2007.

Inmon. 1991 . Building the data warehouse. 1991 .

Kechar, M. and &Bahloul, S. N. (2015, November). An Access Control System Architecture for XML Data Warehouse Using XACML. s.l. : Proceedings of the International Conference on Intelligent Information Processing, Security and Advanced Communication (p. 15). ACM., (2015, November).

Moussa, R and Badir, H. 2013. Data Warehouse Systems in the Cloud: Rise to the Benchmarking Challenge. s.l. : IJ Comput. Appl 20(4), 245-254, 2013. Vols. 20(4), 245-254.

Naushahi, U. M. A. 2016. Profile-Based Access Control in Cloud Computing Environments with applications in Health Care Systems. 2016.

Priebe, T. and Pernul, G. 2001. A pragmatic approach to conceptual modeling of olap security. . s.l. : Proceedings of the 20th International Conference on Conceptual Modeling (ER'01) 2224, 311–324., 2001.

Ray, I. and Ray, I. 2014. Trust-based access control for secure cloud computing. . s.l. : High PerformanceCloud Auditing and Applications (pp. 189-213). Springer New York, 2014.

Rodriguez, A., et al. 2011. Secure business process model specification through a uml 2.0 activity diagram profile. 2011.

Rosenthal, A. et S. Sciore. 2000. . View security as the basis for data warehouse security. . s.l. : In DMDW (p. 8)., 2000.

Saltor, F, et al. 2002. Building secure data warehouse schemasfrom federated information systems. 2002.

Soler, E., Stefanov, V. and Mazon, N.J. 2008. Towards Comprehensive Requirement Analysis for Data Warehouses: Considering Security Requirements, pp. 104–111. IEEE,. Los Alamitos : s.n., 2008.

Sweeney, L. 2002. k-anonymity: A model for protecting privacy. 2002.

Thangaraju, G. and Rani, X. A. K. 2016. Multi User Profile Orient Access Control based Integrity Management for Security Management in Data Warehouse. . s.l. : Indian Journal of Science and Technology, 9(22)., 2016.

Triki, S., et al. 2013. Sécurisation des entrepôts de données: de la conception à l'exploitation. Lumiere II Lyon : Rapport de thèse., 2013.

Trujillo, J., et al. 2009. A UML 2.0 profile to define security requirements for Data Warehouses. . s.l. : Computer Standards & Interfaces, 31(5), 969-983., 2009.

Villarroel, R., Fernandez-Medina, E. and Piattini, M. 2006. uml 2.0/ocl extension for designing secure data warehouses. . s.l. : Journal of Research and Practice in Information Technology 38, 31–43, 2006

Summary

A data warehouse (DW) is a critical business factor that gives a clear view of its operations and a rich source for decision-makers. It contains sensitive data about the company and its customers, and therefore they must not be accessible without access control. The hosting solution of ED in the CC (Cloud Computing) is gradually gaining more popularity in companies because it helps to overcome the endless expansion of data and benefit from its ability to process and store these data. However the confidentiality of these EDs in the CC needs many improvements and the setting up of precise standards, due to the scalability and elasticity of the CC paradigm, since there is not a standard protocol to handle connectivity of CC users to hosted resources by taking into account query execution performance. The objective of our work is to propose a framework guaranteeing the confidentiality of the EDs hosted in the CC based on user profile.

Big Data and Security Issues

Kassimi Dounya*, KAZAR Okba*
BOUSSAID Omar**, SAOULI Hamza*

*Laboratory LINFI, Department of Computer Science, University of Mohamed Khider, Biskra, Algeria
Dounya_kassimi@yahoo.fr, kazarokba@gmail.com, hamza_saouli@yahoo.fr.
**ERIC Laboratory - Warehouse, Knowledge Representation and Engineering Department of Psychology of Health, Education and Development (PSED), University Lumiere Lyon 2
omar.boussaid@univ-lyon2.fr

Summary. Big Data has become a famous area of research and innovation. Its attraction come frome the approaches and techniques used to treat huge volumes of information circulating on the Internet. However, it is not only very difficult to store big data and analyses them with traditional applications, but also it has challenging security and privacy problems. This paper relates to the security in big data, this security touch different levels: Application level, Network level, Classification level and Analytics level, including Data Classification, Authentication, Authorization, Crypto Methods, Logging and Supervising. We also discusses in this paper the ecosystem of big data and presents comparative view of big data privacy and security approaches in literature in terms of infrastructure, application, and data.

1 Introduction

According to the explosion of information volume and the evolution of information technologies as well as the variety and the complexity of the current data, all these factors push us to study this phenomena of Big data.

Big data defines large volume of data in general. The data can be both structured and unstructured and also widely used by both the individual users and businesses on a daily basis. When Big Data have emerged, it offers a set of technologies as Hadoop and MapReduce for processing and securing massive amounts of data which are measured by petabytes produced each day [33]. As a result of this technological revolution, the big data is becoming increasingly an important issue in the sciences, governments, and enterprises. Big Data is a data set, which is difficult to capture, store, filter, share, analyse and visualize on it with current technologies [2]. The complex computation environment, traditional security and privacy mechanisms are insufficient to analyse big data. This challenges in big data consist of computation in distributed and non-relational environments, cryptography algorithms, data provenance, validation and filtering, secure data storage, granular access control, and real time monitoring [54]. Identifying the sources of problems will result in more efficient use of big data and the use of big data in analysis would make the systems become safer.

The paper is organized as: Section 2 defines the Problem of security in Big Data; Section 3 defines the Level of security in Big, Section 4 Concludes with the Future work in the existing security research areas.

2 The Problem of Security in Big data

Malicious attacks on IT systems mainly target sources like email, content, and sites. These are highly complex malwares and new ones are constantly being developed by the attackers, since the fixes are also being developed simultaneously for existing malwares. Traditional solutions are insufficient when dealing with big data to ensure security and privacy. Encryption schemes, access permissions, firewalls, transport layer security can be broken; provenance of data can be unknown; even anonymized data can be re-identified. Big data security issues are widely discussed based on its role in a framework. The following section describes the framework in detail.

3 The Level of Security in Big Data

In this section we are going to present the framework that support users/organizations to manage Big Data Security. The Figure 1 [53] demonstrates how four levels define the Security Framework.

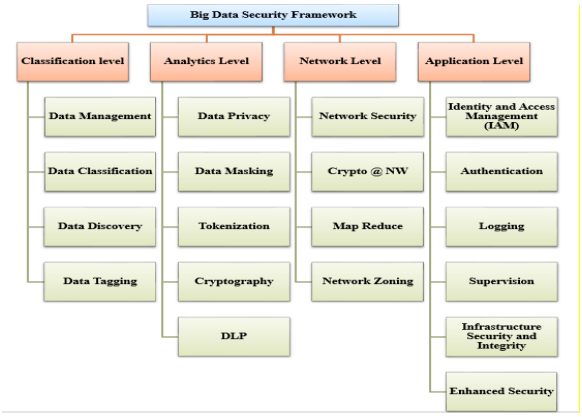


Fig 3.1.the zone of security in Big Data

3.1 Classification Level

It is not easy to use data that is unstructured. Data for analysis should be in a classified state. In simple terms, this can be a process to organize the data.

3.1.1 Data Management

Large data management is to ensure that data is of high quality and in a form that is easy for organizations whether it is structured or unstructured.

3.1.2 Data Classification

There is a lot of techniques of classification, among them Support Vector Machines (SVM) and Decision Trees (DT).Results [1] shown SVM performs better with the supervised classification technique, yielding better results. The other existing techniques used for classification are Supervised Classification Techniques, Geo-Metric Representation Learning Techniques with the Big Data Algorithms and Statistical Data Mining Methods with Remote Sensing through SVM, Naive Bayes and Image Classification. From these techniques listed above [1, 2] the articles explains the usage of logistic regression, tree ensembles, decision trees and random forests as best suited for classification and also their nature of being implementation-friendly make them the most viable choice with respect to security. The future research in this area are direction to incorporating deep learning [3, 4, 5] to guarantee security. The research extends further into Network Traffic Intrusion, with an aim to prevent intruder's activity in networks. The existing work [2] in this area uses 'Geo Metric representation' through Learning Algorithm to find the remote sensing population for animals in the Forest.

3.1.3 Data Discovery

Data Discovery is one of the substantial research areas in the analytics level due to immense storage of data. We can say that our best tool for data discovery is SAP Lumaria [6]. The techniques that are used in data discovery are non-convex objective, linear regression and supervised learning. In [8], the presented work also describes the data mining techniques like social graphs and connection discovery methods to refer to user and images for mining image data. BOFT technique was one of the best methods used for this issue. The recent works [5] uses visual representation of data for addressing the problem of data analytics. Linear Regression with Semi Supervised Learning Approach, Security Network Analysis and K-SVD (Single Value Decomposition), Sparse Coding are the various methods which can be used with the Data Dictionary [9, 8] to work on the data with S\images and text. The best method suggested is linear regression, being a statistical model, relationship between dependent and independent variable could either be linear or nonlinear [10].

3.1.4 Data Tagging

Tagging is one of the best methods for identification of the data which is used in the unstructured format. The research progresses with the grid and Data related research. Data grid deals with the data management which provides solution using most of the programming languages for distributed resources that contain large datasets. Peer 2 Peer network was the first extended research area in this regard. Most of the distributed systems like P2P provide one of the best support models for data tagging. Data tagging [6] has become more significant with smartphones, with respect to the Twitter messages etc. The smart phones are used to track the user location and tag them with user information to derive a holistic view of the

user. The major challenge here would be the number of tweets made by the users and the existing options to find the location. The latter cannot be exact with the existing bandwidth. The GPS (Global Positioning System) grid value accuracy is a vital input for this analysis. If the grid value fails the pattern will be missed during analysis.

3.2 Analytics Level

Analytics Level is a specially moulded area for the customers using Big Data. Analytics is a resultant study which defines the customer on their focus areas with the comparative study from their previous business outcomes. Security is high focus at this level, because customer's future investments remain at this level. So security remains a close watch in this area.

3.2.1 Data Privacy

Data privacy is one of the secure methods to protect the unstructured data. Data privacy gains momentum with data generation techniques. Data generation is of two types: Active and Passive. In active data generation the user generates the data and this is handed over to an organization they process the data. But with Passive data generation, the data is generated by the user. The user knows how and where the data flows through the application. As the data can be tracked, the risk associated with it is reduced and the passive generation is relatively safer than the former.

3.2.2 Data Masking

This is a process which will replace the critical data by the non-critical data. So the additional data will not impact the existing workflow in the applications. The research paper [11] explores removing the attributes from the sensitive data using a technique called, Format Preserving Encryption, also, execution of encryption data remains a serious issue. Hence, the latest work [4] explains about the need for the privacy with the drastic growth of the applications in the cloud using DED [Data Element Dictionary]. This work deals with making a selective encryption of data with various privacy methods. D2ES [Dynamic Data Encryption Strategy] algorithm results with advanced privacy protection on the data compared with Format Preserving Encryption, Computer Masked Data, Dynamic Data Encryption Strategy [4, 11, 12]. DED Algorithm was also developed to work in parallel to encrypt the data with different time constraints [2, 12,13]. Data encryption will be the future exploring areas for Data Masking.

3.2.3 Tokenization

Tokens are used for data security with third party vendors. This is a quiet expensive technique. Since, the device and entire required setup has to be procured with the vendor. Data Privacy explores its wings around various applications used in different layers including

SAAS, IAAS, PAAS. This Paper [14] also explains the security issues in several layers with the cloud which works on large scale data. It analyses the vulnerabilities associated with analytical tools that stores, process and keeps the active data. In [15], the works of the authors is along with Homomorphic Encryption. As a future work [13], the authors show the path to implement some of these protection techniques in an open source big data analytic tool. The effective methodologies used are Bdass, Bpaas, IAAS, SAAS and Homomorphic Encryption. The comparative results show Homomorphic Encryption [13, 15] remains a better technique. Various Optimization methods with the second generation implementation show the better results using the Homomorphic Encryption.

3.2.4 Crypto Methods

Crypto methods cover the encryption at network and data level. This paper [53] explains the security needed for the data that are stored, transferred and processed. With the huge volume of data transfer in the unsecure network the security is more than essential. Homomorphic Encryption, Verifiable Computation and Multiparty computation techniques are the various techniques that can support the advanced security on both transfer and processing. Intelligent selective encryption method [10] assists to extract the final output after the encryption with the multimedia data. Performance scales better with the selective scheme of real time applications. The various methodologies used are AES, DES and RSA, DW-AES Maps, Video Encryption and Data De duplication. In this the paper recommends AES, Data Duplication and Proxy - Re Encryption [11, 12, 15] as better methods for usage. These are the trusted secure methodologies [15] which can be implemented because of delegation and transitivity.

3.2.5 Data Loss Prevention: (DLP)

DLP explains about how the data can be transferred without being damaged by the intruders. The existing work refers to the administrators of the applications to insist users what data should be shared outside the secure network to protect confidential data. There are various tools to assist / warns the users on transferring the sensitive data. In the article [16], it speaks about the secured data transfer through the Internet. With the extensive growth of the data the article expands its view towards the transfer of the data which can be sensitive through the internet. In [17], the authors talked about the risk reduction during Data Mining is another area to be explored with the PPDM (Privacy Preserving Data Mining) trends. The techniques used with DLP are OLAP, Machine Learning and Privacy Preserving Data Mining [17, 18]. The recommended better method is Privacy Preserving Data Mining [17]. The recommendation is based on the generic solution [8, 19].

3.3 Network Level

Security breach is common in the data transfer between the networks. In [53], the authors illustrates the various Network Level security measures with their techniques.

3.3.1 Network Security

Network security is one of the critical areas that implement security within the Big Data Analytics zone. Network security works towards the security features used for the data transfer. The security ensures the data is transferred safe and secure.

3.3.2 Map Reduce

Map Reduce works with the large data sets processing, using a parallel, distributed algorithm in a cluster of nodes setup. Map Reduce is a combined model of Map procedure that works on any of the data integrity methods like filtering / sorting. Then it uses Reduce method that performs the final security operation with data privacy. The data [19] is derived and the required data to be processed is moved as a set and this is processed by a recommender system as defined by the authors. Map reduce framework [20] with Hadoop generates a similar result. The system uses Map-Reduce, exact reconstruction process. The algorithm developed the MDS [Multidimensional Scaling] with the various decode algorithm which can be used with the cloud environment. The security methods which can be used with Map reduce are Recommender System, Inter Image Cloud Platform, Classification Algorithm, and Machine Learning [12, 17, 19, 20] to work with large data sets. The better method suggested in this paper is Machine Learning and Classification Algorithm [21] because of Feature Learning, Parameter Optimization options available. Expanding with the Classification Algorithms the listing of Tree Ensembles, Support Vector Machines and Linear Regression makes the classification algorithms as best method for Map Reduce.

3.3.3 Network Zoning

The article [22] describes about two different types of Zoning, ‘storage and servers’. It also proposes a method for estimating actual speed for the link from very few sampling frequency. The taxi GPS can be traced without any software by the users. The method is based on a path inference process and is applied over a detailed road network in a large city region. The paper [23] describes how to use the software defined network architecture. This is also tested with the various switches. Also, this can be incorporated with various networks like Demoralized Zone. The methods used with Network Zoning are Map Matching, SDN (Software Defined Networking) – NeIF [23, 24]. The recommended method is SDN NeIF which uses centralized networking, easier management and are more secure and cheaper [4].

3.4 Application Level

Numerous customers explore the Big data for implementation. Banking sectors, Transportation, Geo Survey uses these emerging Techniques. So application level of security should be high to ensure the data is secured.

3.4.1 Identity and Access Management (IAM)

The Focus area with the Big Data Security relies on IAM. IAM Encryption is one way to prevent the open threats. The existing work in this technique analyses two different types of encryption: Hadoop clusters and disk level transparent encryption. There are various encryption techniques such as, Full-disk encryption (FDE), OS-native disk encryption (dm-crypt).

3.4.2 Authorization

Authorization is one of the basic levels of security which can be implemented with the Files. The paper [35] explains the data integrity issues with data files. BLS [Basic Life Support] Method, Intelligent Privacy Manager and Secure Remote Password Protocol are the various methods used with Authentication [25, 12]. Secure Remote Password Protocol is the best in the field as password threats are one of the challenging areas. The data is secured using the password. The passwords cannot be retrieved through any of the exiting brute force methods. So this is one of the best methodologies to protect the password [26].

3.4.3 Authentication

Authentication is also a major security area, which can be implemented with the user access level. The security can be implemented with login credentials. Authentication with users based on roles can be restricted. Unauthorized logins will be prevented. In general, in [27], the authors explain the authentication techniques for multiple level protocols. User certain information to be generated is based on these authentication methods, so the data can remain secure. Authors also explains about the various paradigm and model oriented paradigm technologies like date exchange methodology, graph analysis, mining and intelligent algorithm for querying data encryption and authentication protocols are the various techniques are used with the Authentication. Authentication Protocols and Querying Data Encryption are best recommended techniques to be used with authentication processes [24, 27].

3.4.4 Infrastructure Security and Integrity

Infrastructure plays a major role with any of the applications. Similarly, with Big Data it decides what applications / process will be efficient for a particular aim. Infrastructure Integrity that sounds for the data integrity to be used and the accuracy of the integrity required with the data.

3.4.5 Logging

Logging is a method to track the system, process and application actions as a record for further analysis. Our paper provides an approach to look upon the security of such an infrastructure using log information, inspired on data from a real telecommunications network. It presents an approach to identify malicious entities based on large log files from several devices without having to instruct the system about how entities misbehave. This is advantage with the logging as the system is not intruded during the process. The methods derived for logging are machine learning and learning classification, AWS cluster used Cloudera Hadoop distribution and Text Analytics. The better method suggested for use are Machine Learning [21]. Machine learning supports enhanced logging through feature learning and parameter optimization to monitor data effectively.

3.4.6 Supervision

Data Supervision is collecting, validating, and organizing data in an application. Process can monitor the data with different levels. Paper [28] defines the research on the accidents which can be caused by the elevators. The required Supervision is to be set with the elevator, so this remains an entry data for Supervision. So the accident can be prevented. As, the Supervision of the data is available for analysis the accidents can be avoided. The predictions remain easier. The data can be protected through Supervision through logs, login details and other credentials. The best and efficient methodology is Machine Learning with Tree Ensemble [4, 12]. As listed in Logging the features of machine learning has given a wide scope to improve the field of Supervision.

3.4.7 Enhenced Security

Big data have become a predominant technology that mainly uses Data. The simulation and experimental results in [22], show the advantages of the scheme in terms of high efficiency and low error rate for security situational awareness. Moreover, data integrity is also provided in the IBSC (Identity Based Signcryption) scheme. The paper also tells about ICT framework for the Grid. The techniques used are Security Situation Assessment and Association Analysis and identity based Security Schemes [29, 30, 31]. The better method recommended is IAM (Identity Access Management) Based Security method [29, 31] because the productive usage and users own device can be used along with providing fixes to the password probmens.

4 Categories of Big data Security and Privacy

Advanced technique and technologies to ensure security and privacy in Big data, that is because the traditional methods and solution are insufficient. The authors in [52] categorized the security and privacy issues for Big data under 5 titles as Figure 2 Show.

Fig 4.1. The category of Big Data Security & Privacy

4.1 Hadoop Security

Hadoop [31, 32] has become a popular platform but the problem is that this platform is not developed for security from the beginning. So it became a necessity to develop a Hadoop system that guarantees security and privacy of information on the cloud, for that two techniques were proposed to prevent a hacker who wants to get all data in cloud [33]. A trust mechanism has been implemented between user and name node which is component of HDFS and manages data nodes. In this step, SHA-256 which is one of the hashing techniques is used for authentication. Random encryption techniques such as RSA, Rijndael, AES and RC6 has been also used on data in order that a hacker does not gain an access whole data. MapReduce is executed encryption/decryption process in this approach. Another unit that cause the security weakness is Hadoop Distributed File System (HDFS). Three methods to increase HDFS security has been developed [33]. In order to achieve authentication issue, Kerberos mechanism based on Ticket Granting Ticket or Service Ticket have been used as first method. The second method is about monitoring all sensitive information in 360° by using Bull Eye algorithm. This algorithm has been used to make sure data security and manage relations between original data and replicated data. It is also allowed only authorized person to read or write critical data. To handle name node problems as final method, two name node has been proposed: one of them is master and the other is slave. If something happened to master node, administrator gives data from slave name node on condition that Name Node Security Enhance (NNSE) permission. Therefore latency and data availability problems succeeded in secure way.

4.2 Cloud Security

Data storage on clouds is one of the main problems nowadays. Therefore, some precautions must be taken by the service provider. Because of this, a secure way to handle and share big data on cloud platform has been presented [34, 35]. It includes many security methods like authentication, encryption, decryption, and compression etc. to store big data

securely. Authentication with email and password has been used for the authorized person. Data has been encrypted and compressed to prevent security issues. It also takes precautions in case of a natural disaster and uses three backup servers for this purpose. The classical encrypted technique is not enough for big data security on cloud. Consequently, new scheme to secure big data storage has been proposed [36]. This scheme uses cryptographic virtual mapping to create data path. According to the proposed scheme, big data has been separated into many parts and each part is located in different storage providers. As a security measure, if all data encryption are thought to be quite computational and useless, only storage path which shows critical information encryption seems enough, rather than all big data encrypts. The proposed scheme also supports some information encryption to increase the security level. To achieve availability, the scheme holds multiple copies of each part and their accessing index. Thus, if any data part is lost for some reason, information availability is successfully maintained.

4.3 Supervision and Auditing

Security Supervision is gathering and investigating network events to catch the intrusions. Security audit is a systematic measurable security policy to use different methods. To solve this problem, a security Supervision architecture has been developed via analysing DNS traffic, IP flow records, HTTP traffic and honeypot data [37]. Big data security event Supervision system model has been proposed which consists of four modules: data collection, integration, analysis, and interpretation [38]. Data collection includes security and network devices logs and event information. Data integration process is performed by data filtering and classifying. In data analysis module, correlations and association rules are determined to catch events. Finally, data interpretation provides visual and statistical outputs to knowledge database that makes decisions, predict network behavior and respond events.

The separation of non-suspicious and suspicious data behavior is one other issue of Supervision big data. Therefore, a self-assuring system which includes four modules has been suggested [39]. The first module contains keywords that are related to untrusted behavior and it is called library. The second module records identification information about event when a suspicious behavior occurs and this step is named as a low critical log. High critical log as the third module counts low critical logs' frequency and checks whether low critical logs reach the thresholds value. The last module is a selfassuring system and the user is prevented by the system if he/she has been detected as suspicious. While big data becomes a new phenomenon with 5V (Volume, Value, Veracity, Variety, Vilocity) features, new gaps are emerging for big data auditing such as data availability, consistency, integrity, identification, aggregation and confidentially. Hence, some precautions must be taken for all of these gaps in terms of big data. Data availability is satisfied with multiple replicas on big data environment [40]. Thanks to replica nodes, accessing information is quite easy and fast even though some data nodes may be damaged for any reason. These advantages sound good, but they lead to a few security problems like data integrity trouble. In [40], communication overhead

and public auditing and authentication problems have been solved with proposed scheme based on Multi-Replica Merkle Hash Tree.D.

4.4 Key Management

Key generating and sharing between servers and users is another big data security issue. However, using big data centers, quick and dynamic authentication protocols can be suggested. In [41], a layered model has been proposed for quantum cryptography for strong keys in less complexity and PairHand protocol for authentication in mobile or fixed data centers. This model has been not only increased efficiency but also reduced key search operations and passive attacks. The big data services consist of multiple groups that need group key transfer protocols for secure communications. For this reason, novel protocol without an online key generation center based on Diffie-Hellman key agreement and linear secret sharing scheme unlike existing protocols has been offered [42]. The protocol counter attacks via ensured key freshness, key authentication and key confidentiality reducing system overhead. In more complex systems, conditional proxy re-encryption (CPRE) is used for secure group data sharing. Accordingly, an outsourcing CPRE scheme has been proposed in cloud environment which reduces overhead without downloading all data from the cloud, encrypting them and uploading them to the cloud in a new condition unlike CPRE [44]. Security suit has been developed for data node consisting of different types of data and security services for each data type [45]. The proposed approach contains two stages, data analytics, and security suite. Firstly, filtering, clustering and classification based on data sensitivity level is done in data analytics phase. Then data node of databases is created and a scheduling algorithm selects the appropriate service according to section (identification, confidentiality, integrity, authentication, non-repudiation) and sensitivity level (sensitive, confidential, public) from security suite. For example, to provide privacy of sensitive text data, 3DES algorithm is selected.

4.5 Anonymization

Data harvesting for analytics causes big privacy concerns. Protecting personally identifiable information (PII) is increasingly difficult because the data are shared too quickly. To eliminate privacy concerns, the agreement between the company and the individual must be determined by policies. Personal data must be anonymized (de-identified) and transferred into secure channels [46]. However, the identity of the person can be uncovered depending on the algorithms and the artificial intelligence analysis of company. The predictions made by this analysis can lead to unethical issues. In [47], PII has been removed from Intel Circuit web portal usage logs to protect users' privacy. To provide privacy protection, an Adaptive Utility based Anonymization (AUA) has been proposed, which depends on association mining [48]. There are many classical methods to fulfil anonymization over data, but none of them is sufficient for big data because they suffer from scalability issues because of the vol-

ume of the data [49]. Consequently, a hybrid scheme has been proposed which combines two classical method such as Top-Down and Bottom-up for Sub-Tree Anonymization to raise scalability capabilities on big data using MapReduce. The suggested scheme has been tested and the results show that hybrid subtree approach has better per-formance than classical subtree anonymization. In another study, when compared with [50], a new scalable method for local recording scheme considering the proximity-aware privacy has been proposed in [51]. In this scheme, data sets have been generated at cell lev-el. To solve scalability problem, two steps have been planned and coded for MapReduce jobs. The first step is used to split dataset using tancestor clustering; the second step records data with the proximity-aware agglomerative algorithm.

5 Conclusion

This paper explains the various security frameworks used with the Big Data, and studies on big data security and privacy, comparatively. The aim of this paper is to explore the current research progress with the Big Data Security.

Big data privacy, safety and security are the biggest issues to be discussed more in the future, so new techniques, technologies and solutions need to be developed in terms of human-computer interactions or existing technologies should be improved for accurate results.

References

[1] X. Wu, X. Zhu, G.-Q. Wu and P. Ding: Data mining with big data. IEEE Trans. Knowl. Data Eng. 26(1), 2014, pp 97–107.

[2] M. Jhaveri and D. Jaheveri: Big data authentication and authorization using SRP protocol. Int. J. Comput. Appl. 130(1), 2015, pp 26–29.

[3] Triguero, I, Galar. M, Merino. D, Maillo. J, Bustince. H and F. Herrera: Evolutionary undersampling for extremely imbalanced big data classification under apache spark. In: 2016 IEEE Congress Evolutionary Computation (CEC), 24–29 July 2016

V. Gadepally, B. Hancock, B. Kaiser, J. Kepner, P. Michaleas and M. Varia: Computing on masked data: a high performance method for improving big data veracity. In: 2015 IEEE International Symposium Technologies for Homeland Security (HST), 14–16 April 2015

[4] Y. Gahi, M. Guennoun and H. T Mouftah: Big data analytics: security and privacy challenges. In: 2016 IEEE Symposium Computers and Communication (ISCC), 27 –30 June 2016

[5] G. Bordogna and A. Cuzzocrea: Clustering geo-tagged the paperets for advanced big data analytics. In: 2016 IEEE International Congress Big Data (BigData Congress), 27 June–2 July 2016

[6] K. Slavakis and G.B. Giannakis: Online dictionary learning from big data using accelerated stochastic approximation algorithms. In: 2014 IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP), 4–9 May 2014.

[7] D. Gonçalves, J. Bota2 and M. Correia1: Big data analytics for detecting host misbehavior in large logs, June 2016, pp. 25–27.

[8] M. Cheung and Z. Jie: Connection discovery using big data of user-shared images in social media. IEEE Trans. Multimedia 17(9), 1417–1428 (2015)

[9] A. Ouda: A framework for next generation user authentication. In: 2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC), 15–16 March 2016

[10] V.M. Bande and G.K. Pakle: CSRS: customized service recommendation system for big data analysis using map reduce. In: International Conference Inventive Computation Technologies (ICICT), 26–27 August 2016

[11] K. Sekar and M. Padmavathamma: Comparative study of encryption algorithm over big data in cloud systems. In: 2016 3rd International Conference Computing for Sustainable Global Development (INDIA-Com), 16 – 18 March 2016
[12] K. Gai1, M. Qiu, H. Zhao and J. Xiong: Privacy-aware adaptive data encryption strategy of big data in cloud computing. In: 2016 IEEE 3rd International Conference Cyber Security and Cloud Computing (CSCloud), 25–27 June 2016
[13] T. Kiblawi and A. Khalifeh: Disruptive innovations in cloud computing and their impact on business and technology. In: 2015 4th International Conference Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), 2–4 September 2015
[14] C. Xiao, L. Wang, Z. Jie1 and T. Chen: A multi-level intelligent selective encryption control model for multimedia big data security in sensing system with resource constraints. In: 2016 IEEE 3rd International Conference on Cyber Security and Cloud Computing.
[15] G. Geethakumari and A. Srivastava: Big data analysis for implementation of enterprise data security, Int. J. Comput. Sci. Inf. Technol. Secur. (IJCSITS) 2(4), 2012, pp.742–746.
[16] H. Raja and W.U. Bajwa: Cloud K-SVD: a collaborative dictionary learning algorithm for big, distributed data. IEEE Trans. Sig. Process. 64(1), 2016, pp.173–188.
[17] L. Xu, C. Jiang and Y. Ren: Information security in big data: privacy and data mining. IT Prof. 17(3), 2015, pp 1149–1176.
[18] S. Suthaharan: Big data classification: problems and challenges in network intrusion prediction with machine learning. ACM SIGMETRICS Perform. Eval. Rev. 41(4), 2014, pp.70–73.
[19] X. Qin, B. Kelley and M. Saedy: A fast map-reduce algorithm for burst errors in big data cloud storage. In: 2015 10th System of Systems Engineering Conference System of Systems Engineering Conference (SoSE), 17–20 May 2015.
[20] M. Hinkka, T. Lehto and K. Heljanko: Assessing big data SQL frameworks for analyzing event logs. In: 2016 24th Euromicro International Conference Parallel, Distributed, and NetworkBased Processing (PDP), 17–19 February 2016.
[21] B. Deng, S. Denman, V. Zachariadis and Y.J. Issue: Estimating traffic delays and network speeds from low - frequency GPS taxis traces for urban transport modelling. EJTIR 15(4), 2015, pp.639–661.
[22] U. Urkude: Big data analysis by classification algorithm using flight data set. IJIRT 2(10), 2016, pp.188–190.
[23] P.A. Prakashbhai and H.M. Pandey: Inference patterns from big data using aggregation, filtering and tagging a survey. In: 2014 5th International Conference Confluence The Next Generation Information Technology Summit (Confluence), 25–26 September 2014
[24] A. Abdullah, M. Othman, M.N. Sulaiman, H. Ibrahim and A. Othman: Data discovery algorithm for scientific data grid environment. J. Parallel Distrib. Comput. Spec. Issue Des. Perfor. Netw. Super Clust. Grid-Comput. Part II 65(11), 2005, pp.1429–1434.
[25] A. Ibrahim and A. Ouda: Innovative data authentication model. In: 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), October 2016, pp.13–15.
[26] K.S. Yim: Evaluation metrics of service-level reliability monitoring rules of a big data service. In: 2016 IEEE 27th International Symposium Software Reliability Engineering (ISSRE), 23–27 October 2016
[27] J. Wu, K. Ota, M. Dong, J. Li and M. Wang: Big data analysis-based security situational awareness for smart grid. IEEE Trans. Big Data PP (99) (2016).
[28] V.A. Ayma1, R.S. Ferreira1, P.N. Happ1, D.A.B. Oliveira1, G.A.O.P. Costa1, R.Q. Feitosa1, A. Plaza and P. Gamba: On the architecture of a big data classification. In: 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 26–31 July 2015.
[29] X. Wu, X. Zhu, G.-Q. Wu and P. Ding: Data mining with big data. IEEE Trans. Knowl. Data Eng. 26(1), 2014, pp.97–107.
[30] K. Dounya, K. Okba, S. Hamza and B. Omar: Design and Implementation of a New Approach using Multi-Agent System for Security in Big Data. International Journal of Software Engineering and its Applications 1, September 2017, pp.1-14.
[31] K. Dounya, K. Okba, S. Hamza, B. Omar, S. Safa and H. Iman: A new approach based mobile agent system for ensuring secure Big Data transmission and storage. 2017 International Conference on Mathematics and information Technology, Adrar, Algeria, December 4 - 5, 2017, pp. 196-200.
[32] S. Hamza, K. Okba and K. Dounya: Applications et enjeux des Big Data dans le contexte des défis mondiaux. Proceedings 10th of Les Avancées des Systèmes Décisionnels (ASD), Annaba, Algérie (2016) 14-16 May.
[33] P. Adluru, S.S. Datla and Z. Xiaowen: Hadoop eco system for big data security and privacy", Systems, Applications and Technology Conference (LISAT), Long Island, Farmingdale, NY, 2015, pp.1–6.

[34] B. Saraladevi, N. Pazhaniraja, P. Victer Paul, M.S. Saleem Basha and P. Dhavachelvan: Big Data and Hadoop-A Study in Security Perspective. Procedia Computer Science, vol. 50, 2015, pp.596–601.
[35] A.T.H. Ibrahim, Y. Ibrar, B.A. Nor, M. Salimah, G. Abdullah and U.K. Samee: The rise of "big data" on cloud computing: Review and open research issues. Information Systems, vol. 47, 2015, pp.98–115.
[36] A. Kumar, L. HoonJae and R.P. Singh: Efficient and secure Cloud storage for handling big data. Information Science and Service Science and Data Mining (ISSDM), Taipei, 2012, pp.162–166.
[37] H. Cheng, C. Rong, K. Hwang, W. Wang and Y. Li: Secure big data storage and sharing scheme for cloud tenants. Communications, China, vol. 12, issue: 6, 2015, pp.106–115.
[38] S. Marchal, J. Xiuyan, R. State and T. Engel: A Big Data Architecture for Large Scale Security Monitoring. Big Data (BigData Congress), Anchorage, AK, 2014, pp.56–63.
[39] L. Liu and J. Lin: Some Special Issues of Network Security Monitoring on Big Data Environments. Dependable, Autonomic and Secure Computing (DASC), Chengdu, 2013, pp.10–15.
[40] A. Gupta, A. Verma, P. Kalra and L. Kumar: Big Data: A security compliance model. IT in Business, Industry and Government (CSIBIG), Indore, 2014, pp.1–5.
[41] L. Chang Liu, R. Ranjan, Y. Chi, Z. Xuyun, W. Lizhe and C. Jinjun: MuRDPA: Top-Down Levelled Multi-Replica Merkle Hash Tree Based Secure Public Auditing for Dynamic Big Data Storage on Cloud. Computers, vol. 64, issue 9, 2015, pp. 2609–2622.
[42] T. Vijey and A. Aiiad: Big Data Security Issues Based on Quantum Cryptography and Privacy with Authentication for Mobile Data Center. Procedia Computer Science, vol. 50, 2015, pp. 149–156.
[43] H. Chingfang, Z. Bing and Z. Maoyuan: A novel group key transfer for big data security. Applied Mathematics and Computation, vol. 249, 2014, pp. 436–443.
[44] S. Junggab, K. DongHyun, R. Hussain and O. Heekuck: Conditional proxy re-encryption for secure big data group sharing in cloud environment. Computer Communications Workshops (INFOCOM WKSHPS), Toronto, ON, 2014, pp. 541–546.
[45] M.R. Islam, M.E. Islam: An approach to provide security to unstructured Big Data. Software, Knowledge, Information Management and Applications (SKIMA), Dhaka, 2014, pp. 1–5.
[46] T. Omer, P. Jules: Big Data for All (2013). Privacy and User Control in the Age of Analytics. Northwestern Journal of Technology and Intellectual Property, article 1, vol. 11, issue 5.
[47] J. Sedayao, R. Bhardwaj and N. Gorade (2014). *Making Big Data, Privacy, and Anonymization Work Together in the Enterprise: Experiences and Issues*. Big Data (BigData Congress), Anchorage, AK, 601–607.
[48] J.P. Jisha, S.P. Anitha (2013). *Adaptive Utility-based Anonymization Model: Performance Evaluation on Big Data Sets*. Procedia Computer Science, vol. 50: 347–352.
[49] Z. Xuyun, L. Chang, S. Nepal, Y. Chi, Wanchun Dou; Jinjun Chen (2013). *Combining Top-Down and Bottom-Up: Scalable Sub-tree Anonymization over Big Data Using MapReduce on Cloud*. Trust, Security and Privacy in Computing and Communications (TrustCom), Melbourne, VIC, 501–508.
[50] Z. Xuyun; D. Wanchun, P. Jian, S. Nepal, Y. Chi, L. Chang, C. Jinjun (2015). *Proximity-Aware Local-Recoding Anonymization with MapReduce for Scalable Big Data Privacy Preservation in Cloud*. Computers, vol. 64, issue 8: 2293–2307.
[51] M. Thangaraj and S. Balamurugan (2017*). Survey on Big Data Security Framework*, International Conference on Knowledge Management in Organizations, 470-481.
[52] D. S., R. Terzi and S. Sagiroglu (2015). *A Survey on Security and Privacy Issues in Big Data*, the 10th International Conference for Internet Technology and Secured Transactions, 202-207.
[53] Cloud Security Alliance Big Data Working Group. *Expanded Top Ten Big Data Security and Privacy Challenges*, April 2013.

Résumé

Le Big Data est devenu un domaine de recherche et d'innovation reconnu. Il est non seulement très difficile de stocker des données volumineuses et de les analyser avec des applications traditionnelles, mais cela pose aussi des problèmes de sécurité et de confidentialité. Ce document porte sur la sécurité des données dans Big Data, cette sécurité touche différents niveaux: Application, réseau, classification et niveau analytique, y compris classification des données, authentification, autorisation, méthodes cryptographiques, journalisation et supervision.

Multi-agent parallel implementation to solve nonlinear equality constrained multiobjective optimization problem

Adil Jaafar*, Mohammed Mestari*

*Laboratoire SSDIA, ENSET, University Hassan II Casablanca,
Bd. HassanII, Mohammedia, 28820 Morocco
jaafar.adil@gmail.com
mestari@enset-media.ac.ma

Abstract. This paper investigates a decomposition-coordination method using Multi-agent systems implementation for solving the nonlinear equality constrained multiobjective optimization problem (NECMOP), where several nonlinear objective functions must be optimized in a conflicting situation. The Java Agent Development Framework (JADE) was chosen for the implementation because it adheres to FIPA communication standards, widely used and open source. The NECMOP is converted to an equivalent scalar optimization problem (SOP). The SOP is then decomposed into several-separable subproblems. These subproblems are independent from each other, which makes them processable in parallel and allows nonlinearity to be treated at a local level by an agent which lives in a container. With ACL (Agent Communication Language), the agents communicate the results to the master agent who coordinates the intermediate solutions using Lagrange multipliers and requests a new loop until it finds an optimal solution that satisfies all the constraints.

1 Introduction

In the relentless quest for more performance in computing different simulations, mathematicians, computer builders, and programmers are continually searching for new methods and techniques for more accurate results while reducing calculations' time.

This performance challenge was shared between mathematicians who develop algorithms and programmers who implement them in one hand, and the computer microprocessors manufacturers, on the other hand.

However, expectations were often directed to manufacturers who "promised" to double the power of microprocessors every 18 months, in accordance with Moore's Law (Schaller, 1997), by doubling the number of transistors per circuit of the same size, and with the same cost. A boon for programmers who see their algorithms gain more performance without significant effort on their part.

But this method has reached its limits: The frequency increase of a processor requires the increase of the electric power supplied, and therefore of the thermal energy generated, which must be dissipated (Thompson and Parthasarathy, 2006). It is to get around this limit that

manufacturers have turned to the fragmentation of chips and the construction of several cores in the same processor, hence the birth of multi-core. This new deal has pushed mathematicians to rethink their algorithms and programmers to review their source code, to fully exploit the characteristics of these processors.

In this context, this paper intends then to use the decomposition-coordination method, which distributes the nonlinearity to many local levels and reaches coordination at the upper level thanks to Lagrange multipliers, in order to solve the nonlinear NECMOP system. The principle of locating nonlinearities is in fact a key element of this work's direction. This approach consists in converting the nonlinear system into a scalar optimization problem (SOP) with a single cost objective function (Ouarrak *et al.*, 2016, 2017). The system is solved thanks to the communication between an independent multiagent (Ferber, 1999) and a master agent which provides coordination and ensures convergence.

This paper is structured as follows: Section 2 describes the NECMOP and the conversion to an equivalent scalar optimization problem (SOP) with a single objective (cost) function. Section 3 introduces the Java Agent Development Framework (JADE) and the implemented distributed algorithm. Section 4 presents an example of a practical application of the method. Finally, the findings of the study as well as the major inferences will be discussed in the conclusion (section 5).

2 Statment of the problem

Consider the nonlinear discrete-time systems presented as follows:

$$\begin{cases} \min\limits_{\{w_k^* \mid 0\le k\le N-1\}} \{J_1(x,w), J_2(x,w), \dots, J_p(x,w)\} \\ x_{k+1} = f(x_k, w_k), \qquad x_0 = x(0) \; and \; x_N = x_d \; given \end{cases} \tag{1}$$

$$x = [x_0^T, x_1^T, \dots, x_N^T]^T \tag{2}$$

$$w = [w_0^T, w_1^T, \dots, w_{N-1}^T]^T \tag{3}$$

Where $x_k \in \mathbb{R}^n$ and $w_k \in \mathbb{R}^m$ are respectively the data and the control inputs of the system at time k. Our aim is to determine the control inputs which lead to the desired output x_d at time N. The objective functions $J_i(x,w)$ are conflicting, and in general, there is no complete optimal solution v^* which satisfies: $J_1(v^*) \le J_1(v), J_2(v^*) \le J_2(v), \dots, J_p(v^*) \le J_p(v)$. Therefore, solving this problem requires determining a set of feasible points of the decision space, representing a compromise according to the different objectives of the problem. Using the Minimax method, which gives the smallest value of the maximum values of all the objective functions J_i, we define ω_i as the weight of the k component with $\sum_1^p \omega_i = 1$, and the objective functions $(J_1, J_2, \dots, J_p)$ such as:

$$E(x,w) = \max_{1\le i\le p} \{\omega_i J_i(x,w)\} \tag{4}$$

We then, obtain the associated optimization problem:

$$\begin{cases} \min\limits_{\{w_k^* \mid 0 \le k \le N-1\}} E(x,w) \\ s.t. \quad x_{k+1} = f(x_k, w_k) \quad with\, x_0 \, and \, x_N = x_d \, given \end{cases} \tag{5}$$

$$x = [x_0^T, x_1^T, \dots, x_N^T]^T$$
$$w = [w_0^T, w_1^T, \dots, w_{N-1}^T]^T$$

The solution of problem (5) represented in (fig.1) is difficult because it might sometimes require countless computations.

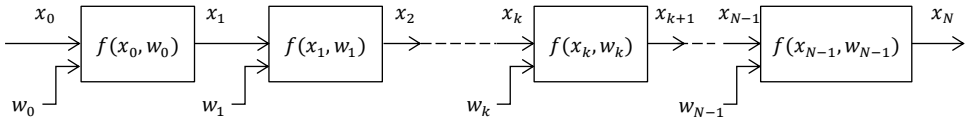


FIG. *1 – Original system made up of N stages*

To solve the problem (5), we decompose the system into a group of N interconnected sub-systems by introducing intermediate outputs y_k of subsystem k (fig.2)

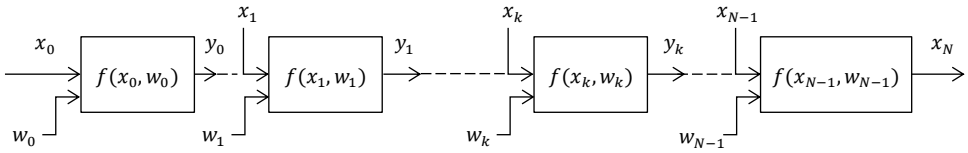


FIG. *2 – Overall system made up by N interconnected subsystems*

Where

$$y_k = f(x_k, w_k), \quad k = 0, 1, \dots, N-1 \tag{6}$$

$$x_k = y_{k-1}, \quad k = 1, 2, \dots, N-1 \tag{7}$$

Therefore, the problem (5) can be written as follows:

$$\begin{cases} \min\limits_{\{w_k^* \mid 0 \le k \le N-1\}} E(x,w) \\ s.t. \quad y_k = f(x_k, w_k), \qquad k = 0, 1, \dots, N-2 \\ and \; x_k = y_{k-1}, \qquad k = 1, 2, \dots, N-1 \; with \; x_0 = x(0) \end{cases} \tag{8}$$

The Lagrange function for the constraints (6)-(7), where μ_k (n components) and β_k (n components) are the Lagrange multiplier, is written as:

$$L = E(x,w) + \sum_{k=0}^{N-1} L_k \tag{9}$$

Where

$$L_0 = \mu_0^T (f(x_0,w_0) - y_0) \tag{10}$$

$$L_k = \mu_k^T (f(x_k,w_k) - y_k) + \beta_k^T (x_k - y_{k-1}), \qquad for\ 1 \le k \le N-2 \tag{11}$$

$$L_{N-1} = \mu_{N-1}^T (f(x_{N-1},w_{N-1}) - x_d) + \beta_{N-1}^T (x_{N-1} - y_{N-2}) \tag{12}$$

The derivations of the ordinary Lagrange function (9) enable us to transform the equality constrained minimization problem (8) into a set of differential equations. A stationary point $(x_k^*, w_k^*, \mu_k^*, y_k^*, \beta_k^*)$, satisfies the following:

$$\nabla_{y_k} L = -\mu_k^* - \beta_{k+1}^* = 0, \qquad for\ 0 \le k \le N-2 \tag{13}$$

$$\nabla_{\beta_k} L = x_k^* - y_{k-1}^* = 0, \qquad for\ 1 \le k \le N-1 \tag{14}$$

$$\nabla_{x_k} L = \frac{\partial E}{\partial x_k} + \frac{\partial f^T}{\partial x_k}\mu_k^* + \beta_k^* = 0, \qquad for\ 1 \le k \le N-1 \tag{15}$$

$$\nabla_{w_k} L = \frac{\partial E}{\partial w_k} + \frac{\partial f^T}{\partial w_k}\mu_k^* = 0, \qquad for\ 0 \le k \le N-1 \tag{16}$$

$$\nabla_{\mu_k} L = f(x_k^*, w_k^*) - y_k^* = 0, \qquad for\ 0 \le k \le N-1 \tag{17}$$

This set of differential equations is decomposed into two levels : upper (13)-(14) and lower (15)-(17)

using the forward Euler rule. The system of differential equations (15)-(17) can be converted into the system of different equations and written in the following scalar form:

$$x_{kq}^{(l+1)} = x_{kq}^{(l)} - \lambda_x \left(\frac{\partial E^{(l)}}{\partial x_{kq}} + \sum_{i=1}^{n} \frac{\partial f_i^{(l)}}{\partial x_{kq}} \mu_{ki}^{(l)} + \beta_{kq}^{(j)} \right), \qquad k = 1, \dots, N-1\ and\ q = 1, \dots, n \tag{18}$$

$$w_{kq}^{(l+1)} = w_{kq}^{(l)} - \lambda_w \left(\frac{\partial E^{(l)}}{\partial w_{kq}} + \sum_{i=1}^{n} \frac{\partial f_i^{(l)}}{\partial w_{kq}} \mu_{ki}^{(l)} \right), \qquad k = 1, \dots, N-1\ and\ q = 1, \dots, m \tag{19}$$

$$\mu_{kq}^{(l+1)} = \mu_{kq}^{(l)} - \lambda_\mu \left(f_q\left(x_k^{(l)}, w_k^{(l)}\right) - y_{kq}^{(l)} \right), \qquad k = 0, \dots, N-1\ and\ q = 1, \dots, n \tag{20}$$

where $\lambda_x > 0, \lambda_w > 0$ and $\lambda_\mu > 0$.

The upper level is responsible of continuously improving y_k and β_k through the j-iterations making them coming closer and closer to the satisfaction of equations (21) and (22). Thus, the

coordination parameters at iteration j + 1 are improvements of the coordination parameters at iteration j, that are then proposed the lower level that handles equations (18)-(20).

$$y_{kq}^{(j+1)} = y_{kq}^{(j)} - \lambda_y \left(-\mu_{kq}^{*}\left(y_k^{(j)}, \beta_k^{(j)}\right) - \beta_{k+1,q}^{(j)}\right), \qquad k = 0, \dots, N-2 \; and \; q = 1, \dots, n \tag{21}$$

$$\beta_{kq}^{(j+1)} = \beta_{kq}^{(j)} - \lambda_\beta \left(x_{kq}^{*}\left(y_k^{(j)}, \beta_k^{(j)}\right) - z_{k-1,q}^{(j)}\right), \qquad k = 1, \dots, N-1 \; and \; q = 1, \dots, n \tag{22}$$

where $\lambda_y > 0$ and $\lambda_\beta > 0$

the convergence of the system is proved (Mestari *et al.*, 2015) by introducing an adaptative coefficient $\lambda = \lambda_y = \lambda_\beta$ for the coordinative level. λ must be chosen as:

$$0 < \lambda < \left|\frac{B(j)}{A(j)}\right| \tag{23}$$

Where

$$A(j) = \sum_{k=0}^{N-1} \Delta e_{y_k}^{(j)^T} \Delta e_{y_k}^{(j)} + \Delta e_{\beta_k}^{(j)^T} \Delta e_{\beta_k}^{(j)} \geq 0 \tag{24}$$

$$B(j) = \sum_{k=0}^{N-1} e_{y_{k+1}}^{(j)^T} \Delta e_{y_k}^{(j)} + e_{\beta_k}^{(j)^T} \Delta e_{\beta_k}^{(j)} \tag{25}$$

In this case: $e_{x_k} \to 0,\ e_{w_k} \to 0,\ e_{\mu_k} \to 0$ if $e_{y_k}^{(j)} \to 0$ and $e_{\beta_k}^{(j)} \to 0$

3 Algorithm Details

JADE is a FIPA-compliant (Fipa, 2002) agent framework that facilitates the development of multi-agent systems. According to (Anandampilai, 2007): "It includes

- A runtime environment where JADE agents can "live" and that must be active on a given host before one or more agents can be executed on that host.
- A library of classes that programmers can use (directly or by specializing them) to develop their agents.
- A suite of graphical tools that allows administrating and monitoring the activity of the running agents."

The coordination agent and local agents will live in an instance of the JADE runtime environment called a Container (fig.3). A single special Main container has the AMS (Agent Management System) that provides a unique name for each agent and the DF (Directory Facilitator) that enables an agent to find other agents providing the services he requires to achieve his goals.

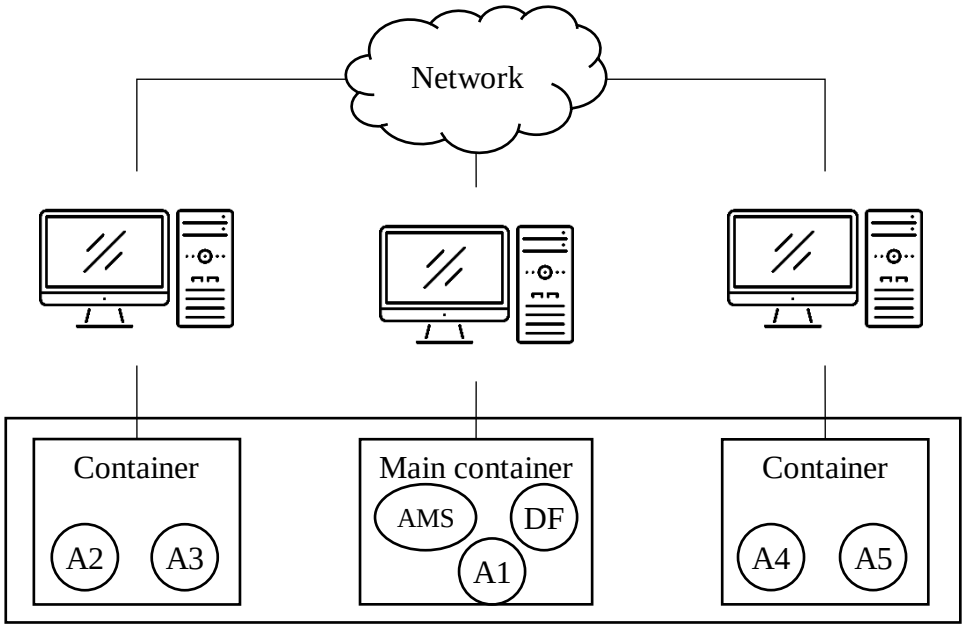


FIG. 3 – *JADE containers and Platforme*

The algorithm organigramme is illustrated in fig.4

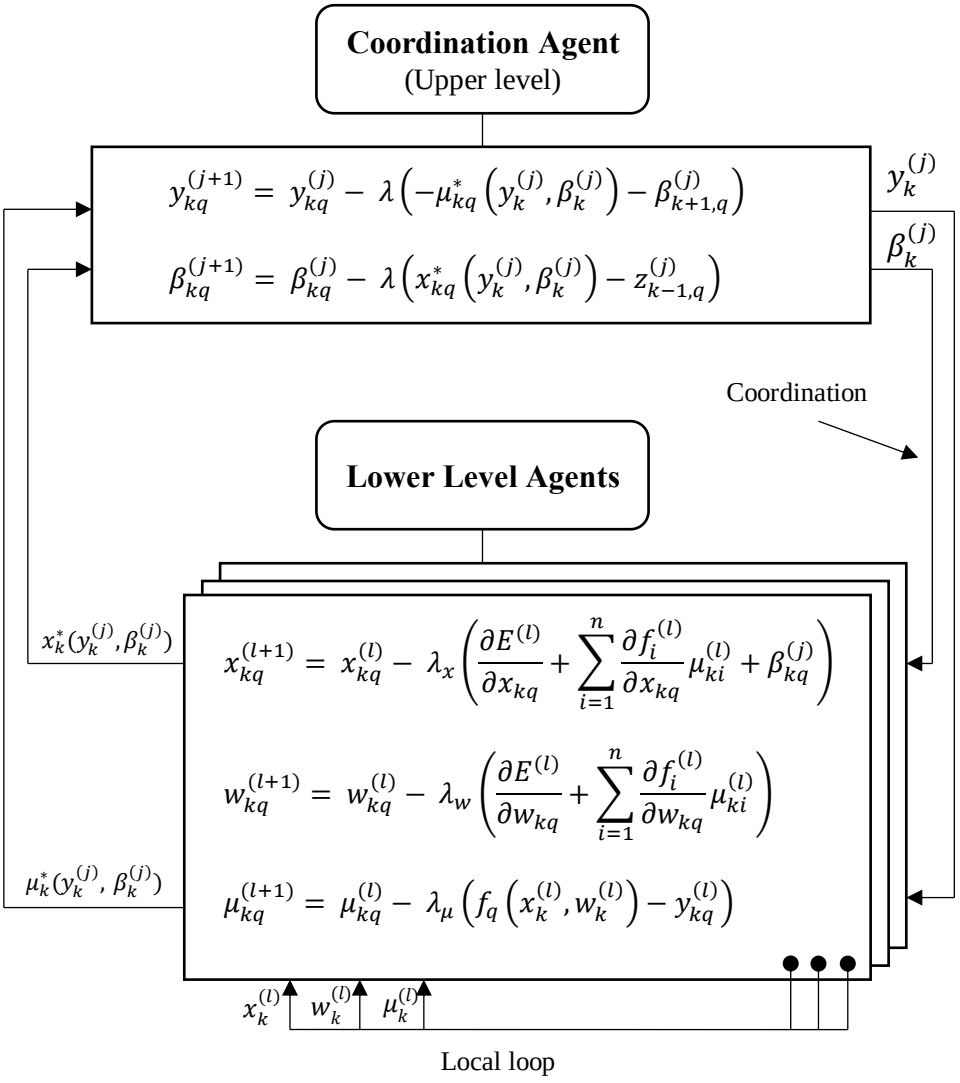


FIG. *4* – *Multiagent algorithm organigramme*

4 Practical application example

Consider the system described by the nonlinear discrete-time equation:

$$\begin{cases} x_{k+1} = f(x_k, w_k) = x_k^2 + w_k \\ x_0 \ given \end{cases} \tag{26}$$

In this system, associate the following optimization problem:

$$\begin{cases} min\dfrac{1}{2}(x_0^2 + w_0^2 + x_1^2 + w_1^2) \\ s.t. \quad y_k = f(x_k, w_k), \qquad k = 0, 1, \dots, N-2 \\ and\ x_k = y_{k-1}, \qquad k = 1, 2, \dots, N-1\ with\ x_0 = x(0)\ and\ x_N = x_d \end{cases} \tag{27}$$

In this example we take $N = 4$, the graph (fig.5) and (fig.6) present evolution of the errors calculated in iteration j.

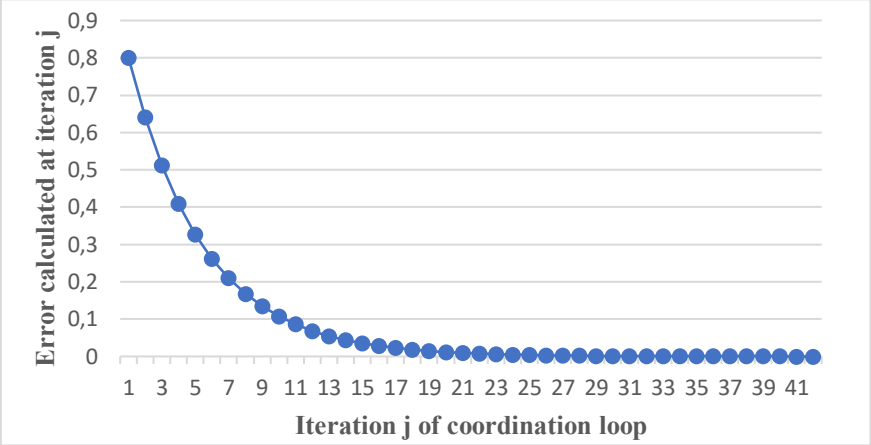


FIG. 5 – *Evolution of the error e_y, on the variable y*

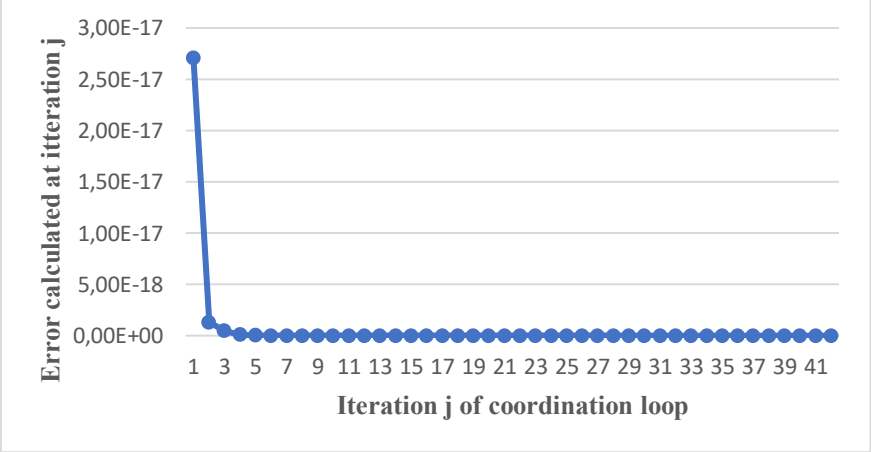


FIG. 6 – *Evolution of the error e_β, on the variable β*

5 Conclusion

The given numerical application demonstrates that the decomposition method used here may be profitably applied to SOP and to solving problems that require complex calculations. One of the best characteristics of the Decomposition Coordination Method is that it is easily adaptable to distributed computation, meaning that it could possibly be implemented in an analog neural network (ANN). However, Since the coordination phase is based on a single central agent, there is a risk of creating proforma problems when the number of iterations is very important. In this case, a potential solution could be decentralizing the coordination task by using a distributed model (Tamura, 1975). Future research will be conducted on the enhancement of the method by introducing parallel computations in the lower level agents.

References

Anandampilai, B. (2007) 'JADE Tutorial - Programming for Beginners', *International Journal of Soft Computing*, 2(3), pp. 422–425.

Ferber, J. (1999) *Multi-agent systems: an introduction to distributed artificial intelligence*. Addison-Wesley Reading.

Fipa, A. C. L. (2002) 'Fipa acl message structure specification', *Foundation for Intelligent Physical Agents, http://www. fipa. org/specs/fipa00061/SC00061G. html (30.6. 2004)*.

Mestari, M. *et al.* (2015) 'Solving Nonlinear Equality Constrained Multiobjective Optimization Problems Using Neural Networks', *IEEE Transactions on Neural Networks and Learning Systems*, 26(10), pp. 2500–2520. doi: 10.1109/TNNLS.2015.2388511.

Ouarrak, H. E. *et al.* (2017) 'A reactive path planning approach for a four-wheel robot by the decomposition coordination method', in *2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, pp. 1–6. doi: 10.1109/EECSI.2017.8239164.

Ouarrak, H. El *et al.* (2016) 'Trajectory planning for a four-wheel robot using decomposition-coordination principle', *Proceedings of 2015 IEEE World Conference on Complex Systems, WCCS 2015*, pp. 0–5. doi: 10.1109/ICoCS.2015.7483265.

Schaller, R. R. (1997) 'Moore's law: past, present and future', *IEEE Spectrum*, 34(6), pp. 52–59. doi: 10.1109/6.591665.

Tamura, H. (1975) 'Decentralized optimization for distributed-lag models of discrete systems', *Automatica*. Pergamon, 11(6), pp. 593–602. doi: 10.1016/0005-1098(75)90073-4.

Thompson, S. E. and Parthasarathy, S. (2006) 'Moore's law: the future of Si microelectronics', *Materials Today*, 9(6), pp. 20–25. doi: 10.1016/S1369-7021(06)71539-5.

Multi-agent parallel implementation to solve NECMOP problem

Résumé

Cet article étudie une méthode de décomposition-coordination utilisant l'implémentation de systèmes multi-agents pour résoudre le problème d'optimisation multiobjectif à contrainte d'égalité non linéaire (NECMOP), où plusieurs fonctions objectives non-linéaires doivent être optimisées dans une situation conflictuelle. Le JADE (Java Agent Development Framework) a été choisi pour la mise en œuvre car il adhère aux standards de communication FIPA, largement utilisés et open source. Le NECMOP est converti en un problème d'optimisation scalaire équivalent (SOP). Le SOP est ensuite décomposé en plusieurs sous-problèmes séparables. Ces sous-problèmes sont indépendants les uns des autres, ce qui les rend aptes à être traités en parallèle et permet à la non-linéarité d'être traitée au niveau local par un agent qui réside dans un conteneur. Avec ACL (Agent Communication Langage), les agents communiquent les résultats à l'agent maître qui coordonne les solutions intermédiaires en utilisant les multiplicateurs de Lagrange et demande une nouvelle boucle jusqu'à ce qu'il trouve une solution optimale qui satisfait toutes les contraintes.

Scalable Solution for Profiling Potential Cyber-criminals in Twitter

Soufiane Maguerra*
Azedine Boulmakoul*
Lamia Karim**
Hassan Badir***

*LIM/IOS, FSTM, Hassan II University of Casablanca, Mohammedia, Morocco
{maguerra.soufiane,azedine.boulmakoul}@gmail.com,
**Higher School of Technology EST Berrechid, Hassan 1st University, Morocco
lkarim.lkarim@gmail.com
***National School of Applied Sciences Tangier, Abdelmalek Essaâdi University, Morocco
hbadir@gmail.com

Abstract. The harm caused by cyber-criminals is intractable, and the number of its victims raises in an exponential manner. Social media play a major role in this expansion by spreading the criminals' knowledge. Consequently, the concern of the detection of these cyber-criminals has become primal. Despite the effort, this issue has to the best of our knowledge no efficient and scalable solution yet. This is mainly due to the large real-time streams flowing through social media. Thus, the necessity of a real-time solution based on Big Data technologies. In this paper, we propose a distributed real-time architecture to answer this issue in Twitter's Social Network. Regarding collection and tweets analytics ecosystems; our solution reuses same foundations as our previous work. This solution is based upon the same architectural design principles and involves the Apache Spark and Kafka Ecosystems. Yet it surpasses it by involving the semantics of an ontology and stores the assertions in a Neo4J Knowledge Graph. This conception assures the consistency of the assertions and the inference of further knowledge.

1 Introduction

In late years, there has been an exponential rise in the number of victims fallen to cyber-criminality. This increase can be partially explained by the mal-intended exploit of social media. Crackers leverage the functionalities of social networks to achieve their desired aims. Consequently, phishing messages and links to potentially harmful websites have invaded social walls. Profiles are been sniffed for information that can be used to gain the trust of others. Then, attain their secrets or steal their accounts, and oblige them to pay a ransom to keep their secrecy or get their belongings back. Still, black hats are not contended by that; they also share their knowledge in social media. The identification of these harmful users' profiles can secure

others from further loss, and prevent the rise of other black hats. However, this task can not be taken in a light manner.

On the one hand, social media contain billions of users; each user has a dynamic state, tends to have thousands of posts, and several relationships. These factors make the identification problem impossible to resolve, unless we project it in the context of Big Data technologies as a respond to the real-time massive nature of social media's data. On the other hand, users tend to have a fuzzy nature which makes their classification by conventional approaches unprecise. Thus, the need for fuzzy classification algorithms to efficiently identify these cyber-criminals. In our past work Maguerra et al. (2017), we used a predefined corpus of cyber-crime related words to create streams of tweets. The users of this tweets have been analysed, and clustered using a fuzzy relational clustering algorithm to gain knowledge about the liveliness of cyber-crime. However, we could not assure the efficacy of the obtained results because of the single word nature of the corpus. In addition, we did not consider the transitivity of the users' similarity relation which has leaded us to a large number of clusters.

In this paper, we answer the problem of the identification of cyber-criminals in Twitter's social network while using both the technologies of Big Data to enable scalability, and the max-min transitivity (Yang and Shih, 2001) over the fuzzy subjective similarity relation to gain efficiency. In contrast of our past work, we overcome the supra drawbacks by considering multiple keywords as entries of our corpus, e.g., follow link to free bitcoins, learn to hack any password. Our architecture filters the tweets and stores them at a production level in the *Apache Kafka Ecosystem* [1]. Differently then our past model, we consume these messages and store the attributes' values in a Neo4j [2] Twitter's users database. This upgrade enables us to fully exploit the statistical expressivity of the graph modelisation. The modelisation of the graph is based upon our own Twitter Cyber-security Ontology which has been constructed by following the Methontology method (Fernández-López, 1999). The ontology respects the OWL2-DL profile, and it is edited by Protégé [3]. Our ontology can be considered as a first step towards consistency and inference of new knowledge from the graph. This time, the clustering process is applied periodically for a past specific time interval, and it is achieved by constructing a similarity relation computed by the Gower's similarity index (Gower, 1971). The mixed data contains the tweets, their location, the location of the users, and the cited malicious links. On the one hand, the similarity between the different places is measured in the graph by a modified version of the Wu and Palmer Similarity index (Wu and Palmer, 1994). On the other, the users' tweets are grouped after some preprocessing to construct the documents. Each document is modeled in a vector space of terms where we compute the Term Frequency-Inverse Document Frequency (TF-IDF). Then, the cosine similarity (Huang) is applied to define the similarity between the different documents. Concerning the malicious links we simply apply the Jaccard Index (Anderberg, 2014). After obtaining the proximity relation, depending on a specific α-level we compute the max-min transitive closure to obtain the fuzzy similarity relation which is modelized as a graph. This graph is manipulated using the Spark GraphX library to obtain the connected elements. These elements form our clusters. The information related to the obtained clusters is stored in the graph foreach chosen time periode to enable further analysis over the dynamic of the users.

1. https://kafka.apache.org/
2. https://neo4j.com/
3. https://protege.stanford.edu/

This paper contains the following sections: Section II serves as a remainder for some preliminary notions needed to understand our clustering process, Section III gives an overview over the existing literature related to the subject of our research, Section IV details our novel proposed architecture, Section V discusses the results and the ecountred issues, and Section VI concludes our paper while stating some possible future works.

2 Preliminaries

Definition 1 Let the Universe of Discourse $X = \{x_1, x_2..., x_n\}$. The fuzzy set characterizing a lexical attribute, and involving these elements to a certain membership degree has the form $A = \{\langle x, \mu_A(x)\rangle \mid x \in X, \mu_A \in [0,1]\}$. Hence, a lexical variable can have several lexical attributes as the answers, i.e., the elements can be included in several fuzzy sets over a certain membership value (Zadeh, 1965).

Definition 2 A binary fuzzy relation $R = \{\langle (x,y), \mu_R(x,y)\rangle \mid x \in X, y \in Y\}$ relates the elements of two crisp sets X and Y. The strength of their relationship is marked by the function $\mu_R : (X,Y) \rightarrow [0,1]$. Each binary fuzzy relation has the resolution form $R = \bigcup_{\alpha} \alpha R_\alpha$ where α is a threshold between $[0,1]$, and R_α is a crisp relation with

$$\mu_{R_\alpha}(x,y) = \begin{cases} 1\text{ , }if\ \mu_R(x,y) \geq \alpha \\ 0\text{ , otherwise} \end{cases}$$

A binary fuzzy relation is denoted a proximity relation if it respects the following properties :

— (Reflexivity): $\forall x \mid X, \mu_R(x,x) = 1$;

— (Symmetry): $\forall x, y \mid X \times Y, \mu_R(x,y) = \mu_R(y,x)$;

This proximity relation becomes a similarity relation if it respects transitivity :

$$\mu_R(x,z) \geq max(T(\mu_R(x,y), \mu_R(y,z)))$$

The transitivity depends on the chosen T-norm. Several t-norms exist; still, in our study we consider only the minimum $T_{min}(x,y) = min(x,y)$. Because the obtained similarity relation characterizes each crisp relation in the resolution form as an equivalence relation. These different equivalence relations can be exploited to obtain a partition tree for the different α-levels (Yang and Shih, 2001).

Definition 3 A similarity relation is extracted from a proximity relation over a series of compositions. The composition is defined as

$$R \circ R = \max_y (\underset{x,z}{T}(R(x,y), R(y,z))$$

in our case the T-norm refers to the minimum.

Definition 4 Let $d_i = \{t_1, ..., t_n\}$ be a document vector of terms and D the set of documents. The Term Frequency $TF(t_j, d_i)$ represents the number of occurences of the term t_j in the i-th document. Since the high frequency of a term in a single document is not sufficient to give

insight of its importance in the whole corpus, we define the importance of each term in the corpus using the Inverse Document Frequency with $IDF(t_j, D) = \log \frac{|D|+1}{DF(t_j,d)+1}$ where DF represents the number of documents containing the j-th term (Huang). If a term is present in a separated group of documents, then the IDF measure increases (otherwise it decreases). The TF-IDF measure is the product of these two measures, and it is a very efficient weighting measure because it takes the importance of terms in consideration. Hence, a term has the highest weight when it is highly present in a document and in a minimal number of other documents.

3 State of the Art

Several studies have been conducted over analysing social media for cyber-security related subjects, especially in Twitter. However, there are relative few works who answered the problem of identifying cyber-criminals in social networks while respecting both efficiency and scalability. Yang et al. (2012) were the first to deeply analyse social spammers' groups to identify their interaction with outsiders and help identify new relationships. Lau et al. (2014) classified in an automated manner social media's discussions into transactional or collaborative categories while detecting the different users' interactions; however, the discussions have been manually extracted and the classification process is purely supervised. Klavans (2015) has indicated the necessity of analysing the users' vocabulary towards less cyber-criminality loss. The Linconly Laboratory Campbell Jr et al. (2015) leveraged the linear supervised SVM and Logistic regression techniques to classify english posts extracted from Twitter, Stack Exchange and Reddit. Still, the issue with the black hats is that they do not tend to follow a linear pattern which makes these algorithms not always a best choice. In addition, they presented the idea of a Twitter's Users Meta-graph stored in Neo4J which contains information of users and their different relationships. Semi-Supervised classifcation techniques based on collective inference are applied over this graph with the aim to infer other cyber-criminals. Altough their work proves to be efficient, the graph contains several types of users which can lead to intractable computations. Mittal et al. (2016) presented a CyberTwitter Framework that mines the twitter's discussions to extract and infer knowledge over vulnerabilities and their solutions. The system is based on a cyber-security ontology including the Unified Cyber-security Ontology, and an intelligence ontology wich modelizes the temporal aspects. Tweets are checked for different concepts using the Security Vulnerability Concept Extractor (SVCE), these concepts are then linked to the open knowledge graph. Then, SWRL rules are applied to infer new knowledge and push alerts to users depending on their specified profile while keeping them in track with the latest threats of interest. In the context of Big Data, Romsaiyud et al. (2017) identified cyber-bullying discussions in social media while using K-means as a pre-clustering step. Then, using Naive-Bayes to classify the disccusions in an Apache Hadoop Yarn environment. The Naive-Bayes is implemented using the Apache Mahout library, and the training set is extracted from the Perverted Justice Foundation. Nevertheless, the pre-clustering step is not distributed, and K-means is not really an adequat choice while dealing with categorical data. They indicated their aim to use Spark as a mean to enhance the scalability. Decidedly Ramalingam and Chinnaiah (2017), indicated that there is a lack for a scalable and efficient solution resolving our problem.

In our past work Maguerra et al. (2017), we used the Spark Streaming functionalities to achieve the desires of Romsaiyud et al. (2017), and we fully distributed the production and consumption process. Additionally in this work, we were inspired by Campbell Jr et al. (2015) to form a Neo4j Twitter's Users Graph. Still, the graph does contain only the users of the tweets respecting our cybercrime corpora. The construction of the graph is achieved in a distributed manner while preserving the data's uniqueness. Like Mittal et al. (2016), we constructed our graph while respecting our proper Twitter Cyber-security Ontology. The construction has been achieved while keeping in mind the integration of other ontologies, as a mean to enlarge our knowledge graph. Also, the ontology is indespensable to analyse the tweets' semantics for concepts and relations. All while keeping the consistency of our graph, and infer new knowledge by respecting some description logic axioms. In the clustering process, we include differently than the supra works four types of similarities. The transitive closure of the proximity relation is computed in an iterative distributed manner; then, we extract the partitions depending on a specific α-level.

4 The Architecture

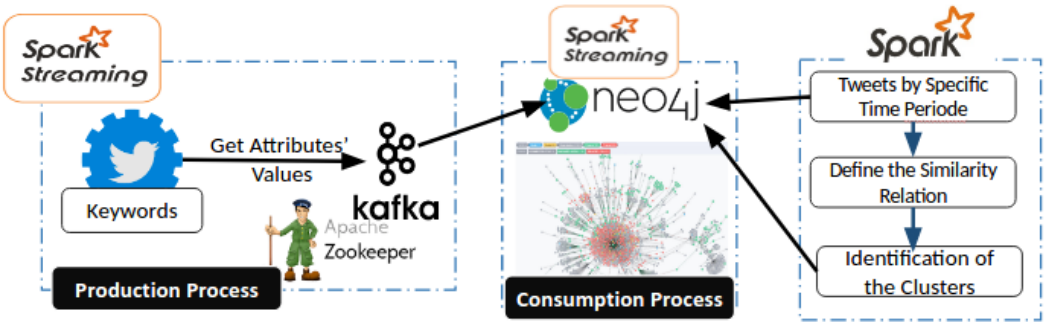


FIG. 1: The Architecture.

Our novel architecture can be seen in the figure 1. The architecture is implemented using the Scala language known for its scalability, and it exploits majorly the functionalities of Spark. This exploit is due to its highly fast distributed processing nature, the different libraries enabling the manipulation of different datatypes, and the flexible integration with other Big Data technologies. The architecture like the old one is composed of three major parts: the production of the tweets, the consumption, and the analysis of the tweets. Each of these parts are explained respectively in the next subsections.

4.1 The Production of the Tweets

In this part of the architecture, we use the functionalities of the Spark Streaming library to create batches of tweets. This is achieved via the *TwitterUtils* class, each batch is treated in parallel while considering only the tweets respecting the cyber-criminality keywords. Tweets

are mined for their identifier, text, creation date, place, user's identifier, name, screen name, description, location and creation date. The tweet's place as well as the user's description and place are not always available; but when they are we store the tweet's country code, country name, full place name and user's desription and location. Then, these attributes are serialized in json format using the json4s library [4]. As an example: (Note that the different dates are considered as timestamps)

```
{
          "id":1235155866,
          "text":"follow this link for free bitcoins, http:\\....",
          "creation_date":1519038734,
          "country_code":"uk",
          "country":"united kingdom",
          "full_place_name":"dartmoor, uk",
          "user_id":12345221563,
          "user_name":"hack p",
          "user_screen_name":"hack p",
          "user_description":None,
          "user_location":"london, united kingdom",
          "user_creation_date":1508021000
}
```

This json value is stored in Kafka as a message with the tweet's identifier as the key. Each slave in the cluster produces messages for each of its partitions depending on the batch load. This way we achieve high scalability. The choice of kafka is supported by its highly scalable and fault-tolerant nature. In contrast of the past work, this time we store all the tweets in a single topic which is partitioned over several brokers. This part of the architecture known for its temporal persistence is fundamental to evade trivial computational processes, since not all attributes are available and it does not infer with the consumption process.

4.2 The Consumption of the Tweets

The consumption is the part responsible for feeding the Neo4J graph. The graph modelization is considered for its statistical relational expressivity which enables us to efficiently store the knowledge, and have an idea about different statistical relational properties. Also, it reduces the complexity while being integrated with an ontology for the consistency validation process. The choice of Neo4J is backed by its fault-tolerance, linear scalability for any depth traversals, and handling over billions of nodes all while being the leader graph database management system. The Neo4J graph is based upon the ontology visualized in the figure 2 which has been edited in Protégé, reasoned over by the Pellet Reasoner [5] for the OWL2-DL profile, and visualized by OntoGraph [6]. The ontology has been constructed while respecting the Methontology method which has been chosen because of its middle-out strategy for the concepts' identification, adaptation to several domains, and its life cycle charcterized by evolving prototypes (Fernández-López, 1999). The ontology also contains some universal restrictions such as:

4. https://github.com/json4s/json4s
5. https://github.com/stardog-union/pellet
6. https://github.com/protegeproject/ontograf

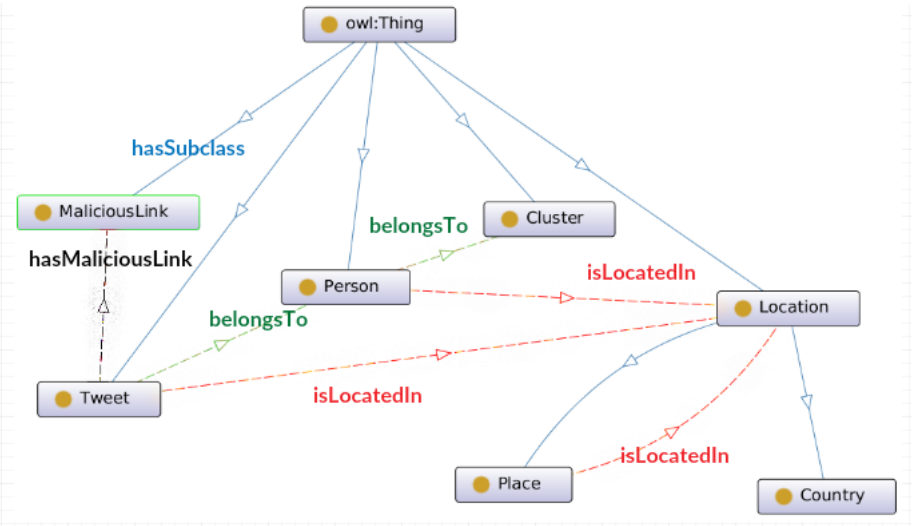


FIG. 2: The Twitter Ontology.

```
Tweet: belongsTo only Person
Tweet: isLocated only Location
Tweet: hasMaliciousLink only MaliciousLink
Person: isLocated only Location
Person: belongsTo only Cluster
Place: isLocated only Location
```

Also, some qualified cardinality restrictions which involve both datatype and object properties, as well as a minimum cardinality restrction concerning the Cluster's data type property:

```
Person: hasCreationDate exactly 1 xsd:long
Person: hasDescription exactly 1 xsd:string
Person: hasName exactly 1 xsd:string
Person: hasScreenName exactly 1 xsd:string
Tweet: hasCreationDate exactly 1 xsd:long
Tweet: hasText exactly 1 xsd:string
Tweet: belongsTo exactly 1 Person
Cluster: hasCreationDate min 1 xsd:long
```

These axioms can be taken as a first step to serve for checking the consistency of the graph. The graph is extended by another stream scheduled for a specific batch interval. First, the messages are consumed in parallel from the brokers. Each slave starts its Neo4j driver session, and transactions are created for each message while using the Neo4J Java Driver library [7]. The Neo4J Cypher transaction queries are :

— Creation of the tweet node if it does not already exist in the graph:

```
MERGE (:Tweet{id:id, text:text, date:created_at})
```

7. https://github.com/neo4j/neo4j-java-driver

— Creation of the malicious links while linking them to their tweet: (the links are extracted by using the linkedIn URL Detector [8])

```
MERGE (:MaliciousLink{text:url})
MATCH (l:MaliciousLink{text:url}), (t:Tweet{id:tweet.id})
  MERGE (t)-[:hasMaliciousLink]->(l)
```

— Creation of the person while checking the uniqueness of the identifier, if the person does not already exist we create it and we attribute its corresponding information: (if the user's description also exist we attribute it to the person node)

```
MERGE (p:Person{id:id}) ON CREATE SET p+={name:name,
  screenName:screen_name, date:created_at}
```

— Creation of the unique link between the tweet and person indicating that it belongs to him:

```
MATCH (t:Tweet{id:tweet.id}), (p:Person{id:user.id})
```

— For the user's location, its value consists of two locations. The first location is considered as an instance of the Place class (having the Place label), and the second as a standard location (if it exists). The following queries create these locations, the link indicating the person's place, and the link indicating that this place is included in the specified location :

```
MERGE (pl:Place{name:place}); MERGE (:Location{name:location})
MATCH (p:Person{id:user.id}),(pl:Place{name:place})
       MERGE (p)-[:isLocatedIn]->(pl)
MATCH  (l:Location{name:location}), (pl:Place{name:place})
       MERGE (l)<-[:isLocatedIn]-(pl)
```

The same above pattern is applied for the tweet's place where the places is located in the location, and the location is located in the country (when there is no location then the place is located in the country). After commiting those transactions for each slave, we check the consistency of the nodes and links related to the location information in the master node by commiting the following transaction:

— Match the places and locations with the same name as to extract the outgoing and ingoing links from the places and pass them to the locations. Then, we suppress the places after the extractions is achieved: (same process is applied for the locations and places containing the same name as the countries)

```
MATCH (pl:Place),(l:Location) WHERE pl.name=l.name WITH pl, l
   MATCH (n)-[:isLocatedIn]->(pl)  WITH n,pl,l
     MERGE (n)-[:isLocatedIn]->(l)
MATCH (pl:Place),(l:Location) WHERE pl.name=l.name WITH pl, l
   MATCH (n)<-[:isLocatedIn]-(pl)  WITH n,pl,l
     MERGE (n)<-[:isLocatedIn]-(l)
MATCH (pl:Place),(l:Location) WHERE pl.name=l.name
   DETACH DELETE pl
```

— Then, we check if any loops exists and if a country is located in another node to inverse the relation:

8. https://github.com/linkedin/URL-Detector

```
MATCH (n)-[r:isLocatedIn]->(n) DELETE r
MATCH (c:Country)-[r:isLocatedIn]->(n)
   MERGE (n)-[:isLocatedIn]->(c) DELETE r
```

4.3 Analysing the Graph

At last, we analyse the given graph for a specific time period. First, we acquire the data from the graph using the following cypher query:

```
MATCH (t:Tweet) WHERE t.date>=past AND t.date<=now WITH t
  MATCH (t)-[:belongsTo]->(p:Person) WITH p,t
   OPTIONAL MATCH (t)-[:hasMaliciousLink]->(l:MaliciousLink)
   WITH p ,t, collect(l.text) as links
     OPTIONAL MATCH (p)-[:isLocatedIn]->(ppl),
     (t)-[:isLocatedIn]->(tpl) RETURN properties(p), ppl.name,
     head(labels(ppl)), collect({tweet:properties(t),
     place:tpl.name, place_label:head(labels(tpl)), links:links})
```

This query matches the tweets respecting a specific time interval and the persons associated with them. An optional match is applied over the results to get the places of the tweets and persons if they exist, otherwise we get nullable values. Then, we obtain rows containing the properties of the person, the person's place name; a collection containing the properties of the tweets, their places, and their malicious links. The query is executed via the Spark Neo4J Connector library [9] to obtain the data in a Resilient Distributed Dataset (RDD) format. After obtaining the data, we define four similarity coordinate matrices. A distributed datatype proper to the Spark ML library. The first matrix concerns the tweets which have been filtred from emojis using the Emoji-Java library [10]. A dataframe is formed where each row represents a document involving the tweets of a person. This dataframe is handled via the Spark MLlib transfromers. In particular, we respectively use the *RegexTokenizer* to split each document into tokens, the *StopWordsRemover* to remove stopwords, the *Stemmer* for tokens stemming. These are the preprocessing steps; next, we use the *HashingTF* transformer to compute the term frequencies and we train the *IDF* estimator over the obtained frequencies to get the model. This model is the transformer used to computed the TF-IDF frequencies. After computing these frequencies, we construct our document vector space coordinate matrix M. To compute the similarities, we conduct the inner join of M with M^t. This process is optimized by conducting the join of the M columns with the M^t rows. The cosine similarity index (Huang) is used foreach record with $S(d_i, d_j) = \frac{d_i . d_j}{\|d_i\| \|d_j\|}$ (Note that since the documents' weights are positive, the similarity is between $[0, 1]$).

After defining the first subjective similarity relation, we compute the similarities between the users' places using a modified version of the Wu and Palmer index. This index has been chosen because we wanted to take consideration of the different heights in the trees, i.e, a couple of places has the highest similarity if it is the closest to the tree's root. Let N be the distance between the Lowest Common Ancestor (LCA) and the root of the tree, N_1 the distance between the i-th place and the root, and N_2 the distance between the j-th place and the root.

9. https://github.com/neo4j-contrib/neo4j-spark-connector
10. https://github.com/vdurmont/emoji-java

The similarity is defined as :

$$S(place_i, place_j) = \begin{cases} 1, \text{if the places are not defined} \\ 0, \text{if the places are not connected in the graph} \\ \frac{2N+1}{N1+N2+2}, \text{otherwise} \end{cases}$$

we choose the value 1 as the similarity when the places are not defined because the users of interest share an unkown behaviour and to normalize the final subjective similarity relation. Also, we incremented each distance by one to assure that the measure remains a similarity and to distinguish between two places when the root equals the LCA (otherwise we will obtain zero). The same index is used for defining the similarity between the tweets' places. Each user is characterized with a pattern of places, we conduct our analysis over the highly frequent places by considering the median of the places' frequencies. Each couple of users is characterized by a similarity matrix, we sum up the similarities and divide them by the length of the longest pattern to obtain the final similarity

$$S(p_i, p_j) = \frac{\prod_{k,h} S(place_{ik}, place_{jh})}{max(|P_i|, |P_j|)}$$

with p_i as the i-th person and P_i his places. To evade redundant computations, we first define the similarity between the users' places; then, we perform some outer joins in a distributed manner to define the remaining tweet's places where we the similarity is still undefined. To define the similarity between two places in Neo4J we use the infra cypher queries for respectively identifying the LCA, N, $N1$, and $N2$. The queries have been executed in a distributed manner where each slave has its partitions, the connection is opened once for each partition and each transaction is performed over a slave of the Neo4J cluster. Note that in the queries, we specify the places labels because it can happen that a user's link or twitter's link gets altered towards a location or country.

```
OPTIONAL MATCH path=
(n:li{name:pi)-[:isLocatedIn*]->(e)<-[:isLocatedIn*]-(m:lj{name:pj})
RETURN count(path), e.name, head(labels(e)) ;

MATCH path=(e:label_LCA{name:LCA)-[:isLocatedIn*]->() WITH
  max(length(relationships(path))) AS path
  RETURN CASE path  WHEN null THEN 1 ELSE path+1 END;

MATCH path=(p:li{name:pi})-[:isLocatedIn*]->(e:label_LCA{name:LCA})
WITH min(length(relationships(path))) AS path RETURN path+1;

MATCH path=(p:lj{name:pj})-[:isLocatedIn*]->(e:label_LCA{name:LCA})
WITH min(length(relationships(path))) AS path RETURN path+1
```

Concerning the malicious links, we characterize each user with a set of malicious links. Let L_i be the set of distinct malicious links, the similarity between a couple of users is determined by the Jaccard Similarity index (Anderberg, 2014) as $S(p_i, p_j) = \frac{|L_i \cap L_j|}{|L_i \cup L_j|}$. All the above computations have been conducted in a distributed manner, and the final subjective similarity

is obtained by three full outer joins over the different entries. The final similarity is a form of the Gower similarity index (Gower, 1971) where $S(i,j) = \frac{\sum_{k=1}^{4} S_k}{4}$.

At last, the transitive closure is computed in an iterative ditributed manner after filtering the similarity by a specific α-level. In the literature, several algorithms exist for achieving this task. The most known ones are the Seminaive and Smart Algorithm. The first takes fewer iterations than the second; however, at the expense of more expensive joins (Kabler et al., 1992). In our work, we consider the Seminaive algorithm because it has been known for the majority of cases to outperform the Smart in a distributed environment (Gribkoff). In our implementation, we consider a recursivity of a maximum n-calls with n^2 representing the size of the relation (R^0 is the initial proximity relation). The computation is achieved via the algorithm 1 which calls the algorithm 2 to compute the composition between the current and initial relation (Note that each relation is represented as a coordinate matrix). The obtained similarity relation is mined for the vertices and their edges to construct a GraphX graph which is analysed for the connected components that form the temporal clusters.

Algorithm 1: Computing the Transitive closure

Data: R^k: current relation, R: initial relation
Result: Subjective Similarity Relation
begin
  if R^k*.entries.subtract(*R*.entries).isEmpty* **then** (R^k)
  else composition(R^k)

Algorithm 2: Relation Composition

Data: R^k
Result: Subjective Similarity Relation
begin

```
val R^k_entries=R^k.entries.map({case MatrixEntry(i,j,s)=>(j,(i,s))})
val entries=R^k_entries.join(R_entries).map({
    case (_,(((i,wi)),((j,wj))))=>((i,j),(min(wi,wj)))
}).reduceByKey(max).map({
    case ((i,j),w)=>MatrixEntry(i,j,w)
})
new CoordinateMatrix(entries)
```

5 Experimental Results and Encountred Issues

The infra figures give a limited insight over the graph. The figure 3 visualizes the tweets belonging to the persons. The locations where the tweets have been posted can be seen in the figure 4 (for further infomation about the location class object relations please refer to the ontology 2). The malicious links included in the tweets are visualized in the figure 5. After achieving the clustering process, we obtain the clusters visualized in the figure 6. These figures only give a brief overview of the complexity of the graph. In our experiment, we used a cluster containing four slaves where a set of 4000 tweets have been analysed for approximately 2 minutes. Each slave has four cores and in Spark's configuration we specified 2 GB as the

executors memory. In the Neo4J High Availability Cluster cluster, we specified a maximum memory heap size of 2GB over a three slave cluster. Unfortunately, the Neo4J clusters does not support the partitioning of the graph. Hence, each machine needs to have approximately the same profile to be able to store the whole data. Other issues are related to the Twitter's Streaming API only a 1% percent of the data is publicly available,low production especially when the keywords filter is used, and there is a limit of 400 keywords by filter. To overcome these issues there is a need for different application accounts and a single production local machine related to each application account.

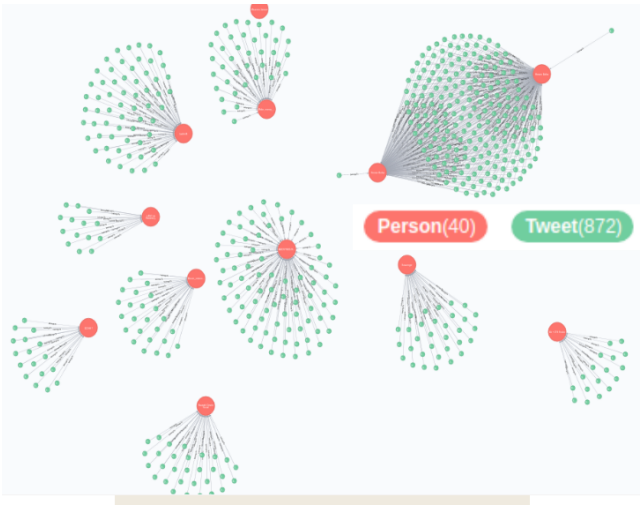


FIG. 3: The Persons' Tweets.

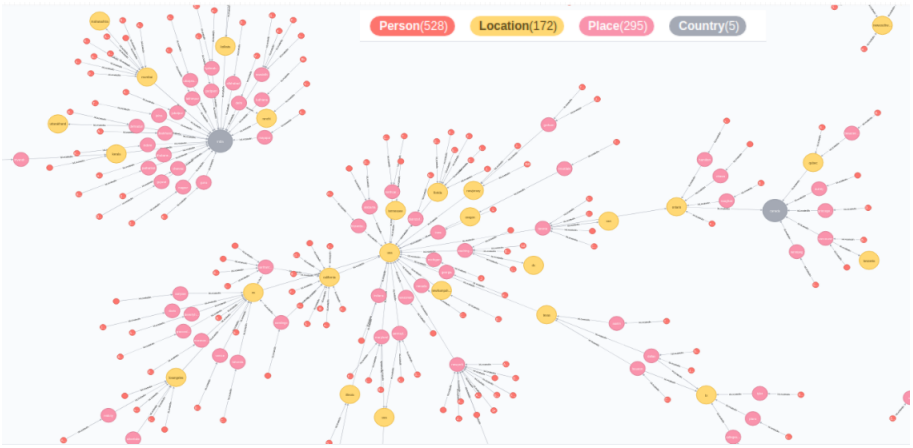


FIG. 4: The Tweets' Different Locations.

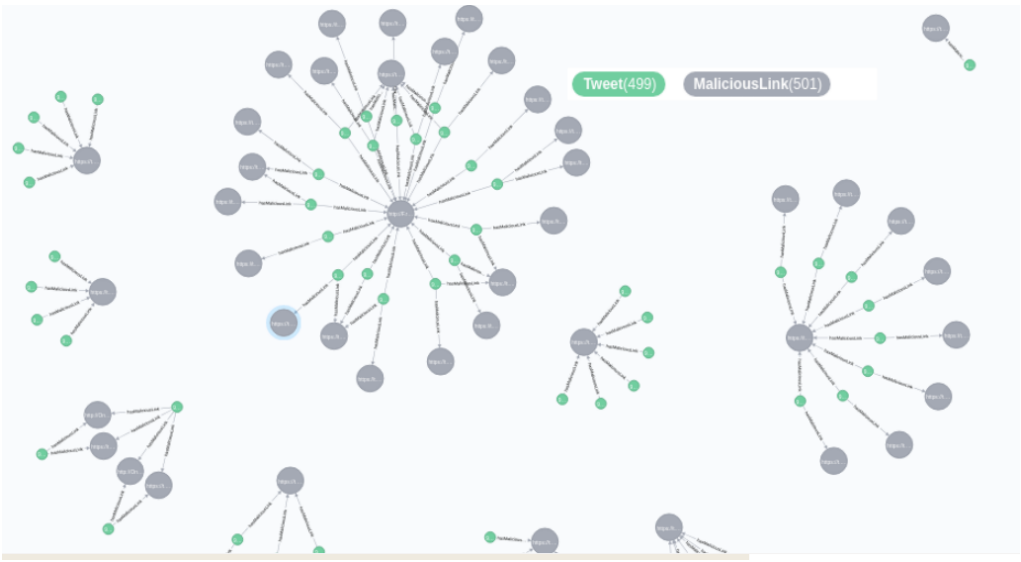


FIG. 5: The Tweets' Malicious Links.

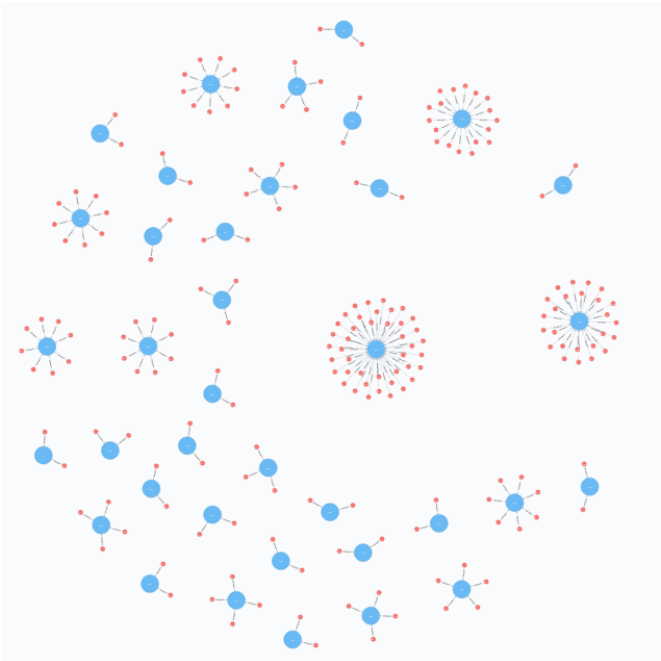


FIG. 6: The Tweets Belonging to the Different Clusters.

6 Conclusion and Future Works

The architecture described in this paper can be used to identify potential cyber-criminals in Twitter. The Neo4J graph has enabled us to perfom some complex queries especially for the measure of similarities. There has been no redundant data in the graph, only the users that have a related cyber-crime profile have been detected. The usage of the different type of similarities has enabled us to exploit the different information. In the future, we would like to exploit more the data and study the effect of time. All while using the Apache Zeppelin Ecosystem to construct a real-time dashboard, enabling us to extract knowledge of the temporal liveliness of cyber-crime. In addition, we would like to consider the max-delta transitivity while computing the transitive closure. Then, apply a fuzzy relational clustering algorithm to identify the partitions. There is also a need to study the effect of the different α-levels to choose the most compact and separated clustering scheme. The ontology needs also to be integrated in the analysis process, it has to be expanded to enable the detection of more knowledge while using the tweets' semantics. The semantics can lead us to infer more knowledge from the graph. These knowledge can be used to efficiently detect cyber-criminals and their clusters. The dynamic of the clusters has to be also analysed. There is also a necessity for a detailed runtime evaluation of the analysis process over different configurations.

References

Anderberg, M. (2014). *Cluster Analysis for Applications: Probability and Mathematical Statistics: A Series of Monographs and Textbooks*. Probability and mathematical statistics. Elsevier Science.

Campbell Jr, J., A. C. Mensch, G. Zeno, W. M. Campbell, R. P. Lippmann, and D. J. Weller-Fahy (2015). Finding malicious cyber discussions in social media. Technical report, MIT Lincoln Laboratory Lexington United States.

Fernández-López, M. (1999). Overview of methodologies for building ontologies.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 857–871.

Gribkoff, E. Distributed algorithms for the transitive closure.

Huang, A. Similarity measures for text document clustering.

Kabler, R., Y. E. Ioannidis, and M. J. Carey (1992). Performance evaluation of algorithms for transitive closure. *Information systems 17*(5), 415–441.

Klavans, J. L. (2015). Cybersecurity-what's language got to do with it? Technical report.

Lau, R. Y., Y. Xia, and Y. Ye (2014). A probabilistic generative model for mining cybercriminal networks from online social media. *IEEE computational intelligence magazine 9*(1), 31–43.

Maguerra, S., A. Boulmakoul, L. Karim, and B. Hassan (2017). Real-time distributed architecture for detection, profiling and monitoring of cyber-criminals in social networks.

Mittal, S., P. K. Das, V. Mulwad, A. Joshi, and T. Finin (2016). Cybertwitter: Using twitter to generate alerts for cybersecurity threats and vulnerabilities. In *Advances in Social Networks*

Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on, pp. 860–867. IEEE.

Ramalingam, D. and V. Chinnaiah (2017). Fake profile detection techniques in large-scale online social networks: A comprehensive review. *Computers & Electrical Engineering*.

Romsaiyud, W., K. na Nakornphanom, P. Prasertsilp, P. Nurarak, and P. Konglerd (2017). Automated cyberbullying detection using clustering appearance patterns. In *Knowledge and Smart Technology (KST), 2017 9th International Conference on*, pp. 242–247. IEEE.

Wu, Z. and M. Palmer (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pp. 133–138. Association for Computational Linguistics.

Yang, C., R. Harkreader, J. Zhang, S. Shin, and G. Gu (2012). Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. In *Proceedings of the 21st international conference on World Wide Web*, pp. 71–80. ACM.

Yang, M.-S. and H.-M. Shih (2001). Cluster analysis based on fuzzy relations. *Fuzzy Sets and Systems 120*(2), 197–212.

Zadeh, L. A. (1965). Fuzzy sets. *Information and control 8*(3), 338–353.

Résumé

Les dégâts causés par les cyber-criminels sont intractables, et le nombre des victimes s'accroît exponentiellement. Les médias sociaux ont un rôle majeur dans cette croissance en partageant les connaissances de ces criminels. Par conséquant, l'intérêt d'identifier ces cyber-criminels est devenu primaire. Malgré tous les efforts, à notre connaissance ce problème n'a pas encore de solution scalable et efficiente. Cela est dû aux larges real-time streams qui coûlent à travers les médias sociaux. Ainsi, la nécessité d'une solution basée sur les technologies du Big Data. Dans ce papier, nous proposons une architecture scalable et efficiente pour répondre à ce problème dans Twitter. Similaire à notre travail passé cette solution est basée sur le même principe et elle inclut les mêmes Ecosystèmes d'Apache Spark et Kafka. Cependant, ses points forts se manifestent sur les faits qu'elle inclut la sémantique introduit par une ontologie et stocke les assertions sous forme d'un graphe de connaissance dans Neo4J. Cette conception assure la consistance des assertions et l'inférence d'autres connaissances.

A reinforcement learning technique for web service composition using new multi-layer agent coalition architecture

Asma Bendahmane*, Hamza Saouli*
Okba Kazar*, Khaled Rezeg*, Imane Sriti*

*LINFI Laboratory, Departement of Computer Science
University Mohamed Khider Biskra, 07000, Biskra, Algeria

assoum_bd@yahoo.fr
hamza_saouli@yahoo.fr
okbakazar@yahoo.fr
rezeg_khaled@yahoo.fr
imenesriti@gmail.com

Abstract. Selection, composition and ranking are the basic operation that should be performed by any web services search engine to satisfy the client needs and requirements. The complexity and the dynamic character of composition lead us to propose a new multi-layer based agent architecture that uses coalition to choose the best services among hundreds of thousands deployed web services in the Net. Agents are autonomous and auto-adaptive, combined under coalition technique, can select and find the optimal services without visiting all population, which reduces response time and composition complexity. Finally, the composition results are encouraging due to the combination between Quality of services parameters and a new coalition algorithm.

Keywords: Multi-Agents System, Coalition, Cooperation, Composition, Learning, Quality of Services, Web Services.

1 Introduction

Web services are the best tools to ensure portability and operability between data and application in Internet Gowri et Lavanya (2013), which lead companies to use them as a composed chains in order to satisfy the users' queries Abdullah et Li (2016).

The goal of composition operation is to satisfy all users' sub-queries and reach a certain degree of service quality (QoS).

The growth number of web services deployed on Internet and the dynamicity character of the QoS Sun et al. (2013) push us to use Agent technology in order to represent web services composition as a cooperation and coalition process, because these two collaboration methods are analogue to the composition problematic Cetnarowicz et al. (2011).

A Web service knows only about itself, but not about the others or its users but agents are self-aware at meta level, and through learning and model building, gain awareness of other agents. Their interactions capabilities offer them the ability to cooperate negotiate and coor-

dinate over distributed environments. Agents are inherently communicative, whereas Web services are passive until invoked. Agents can provide alerts and updates when new information becomes available.

In this paper we propose a multi-layer agent based architecture for web services composition through a coalition scenario; moreover, a learning process based reinforcement is used to increase the composition accuracy Bendahmane et Kazar (2011).

The rest of the paper is organized as follows: the second section describes the proposed approach, the third section shows the implementation of the proposed approach; the fourth section presents a summary of the main related work with a comparison study between them and the last section is the conclusion of this paper.

2 Proposed architecture

In this section we present the global architecture of the proposed web services based composition system. The architecture is mainly based on the cooperation and coordination between a multi-agent system where each agent represents a web service so that each web service can benefit from agents' characteristics namely: autonomy and auto-adaptability. Figure 1 shows the overall architecture of the proposed system:

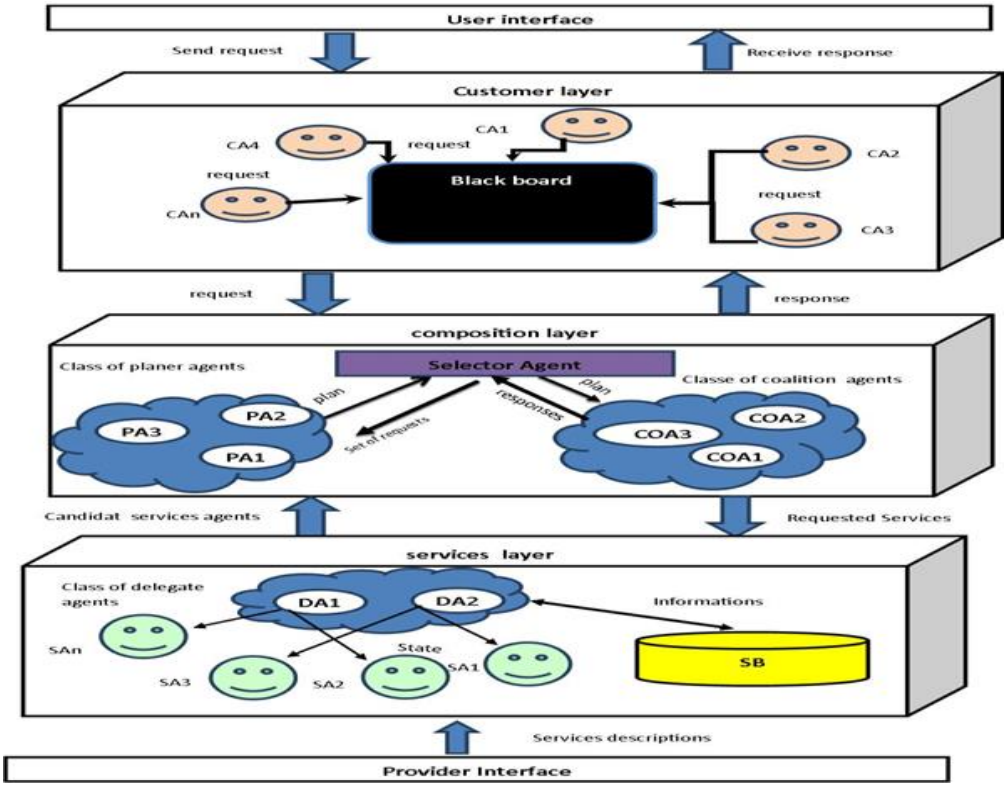


FIG. 1 – *The proposed global architecture.*

As shown in the Figure 1 above, the system architecture is composed of five layers:

- **User interface**: The user interface allows the clients to express their functional needs in a textual format, and their quality requirement as a set of numerical values.
- **Customer layer**: this layer is responsible of receiving the client query through a client agent that transmits this latter to the compositions layer. It is also responsible of treating and displaying the client query on a black table so that all planer agents can access it.
- **Composition layer**: By using three kinds of agents (planning, selector and coalition agent) this layer ensures the process of web services composition.
- **Services layer**: By representing each web service with an agent, this layer offers the necessary knowledge to coalition agents in order to increase and benefit from the cooperation and autonomous characteristics of agent technology. Moreover, this layer uses a delegated agent to gather the services that belongs to the same domain in order to increase the composition accuracy and reduce the response time.
- **Providers' layer:** it allows web services providers to publish their web services description parameters such as: Name, URL, textual description, Quality and so on.

2.1 Agents

2.1.1 Service agent

It is responsible of updating the web services dataset and seeking the best delegated agent (that belongs to the same category). Moreover, if a new web services category appears the service agent creates a new delegated agent in order to manage it. The figure 2 illustrates the internet architecture of the Service agent:

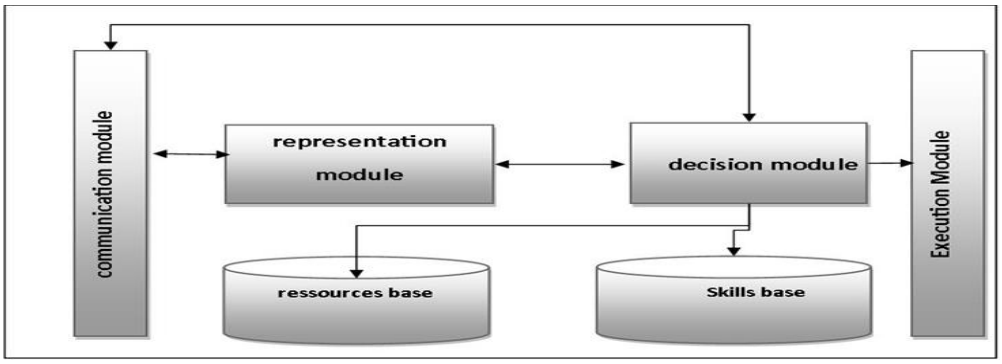


FIG. 2 – *Service agent components.*

2.1.2 Selector agent

It is responsible of : (1) grouping the client queries in unified categories, (2) differencing the simple queries from the complexes ones (3) creating a planner and coalition agents for

each similar complex queries and (4) finding the best web services, through the delegated agent, in case of simple queries. The figure 3 illustrates the Selector agent components:

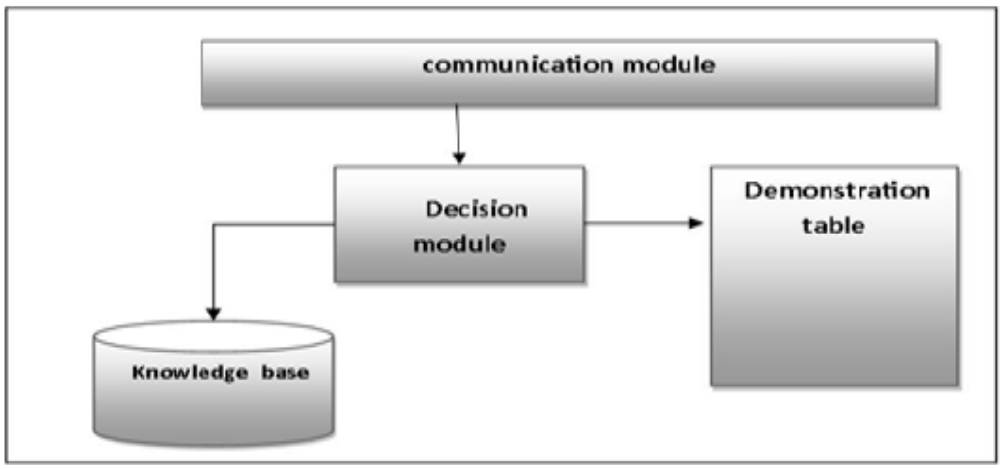


FIG. 3 – *Selector agent components.*

2.1.3 Delegated agent

It is situated between composition and service layers. The roles of the delegated agent are: (1) controlling all service agents that belong to the same category, (2) verifying the availability of service agents, (3) updating the web services index and (4) checking the actual agents' states. The Figure 4 illustrates the Delegated agent components:

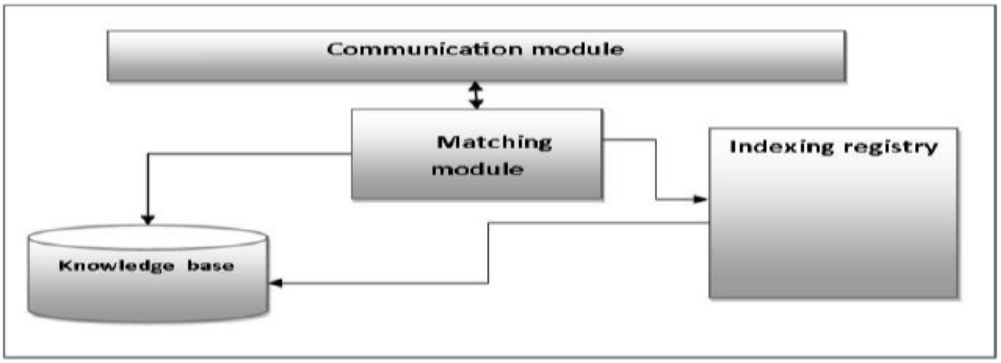


FIG. 4 – *Delegated agent components.*

2.1.4 Planning Agent

It is responsible of creating a composition plan that answer to a complex client query. the answer is a set of tasks executed sequentially or in parallel, the following source code represents an example of a plan model: Nod:<id, input, output>, StopPoint :<Nod, Nod, enchain-

ment>, Plan :<input, output, root>, Root :<0,Nod>. The figure 5 illustrates the planning agent components:

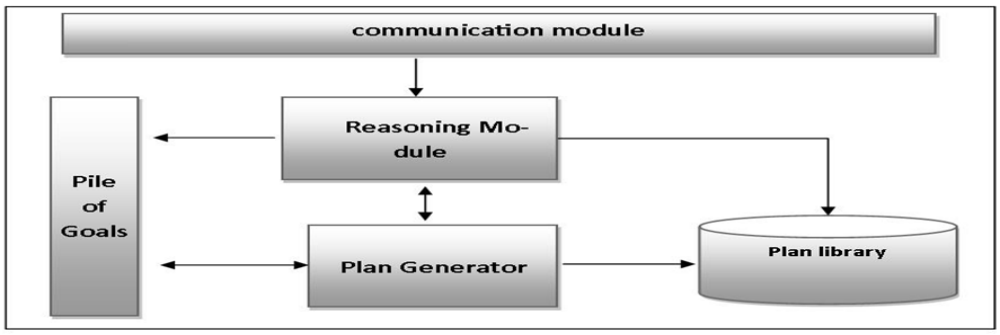


FIG. 5 – *Planning agent components.*

2.1.5 Coalition Agent:

It is responsible of web services composition operation by benefiting from the representation of each service by its own agent. The coalition algorithm below illustrates the agent behavior. Let's $Ag=\{a1, a2,\ldots\ldots an\}$ denotes a set of agents, any sub set $C \subset Ag$ may form coalition. $C = Ac, Uc, Lc$ where $Ac\ Ag$, Uc is the payoff expected after execution, received by coalition. $Uc= ai \in Uai$, Uai is the payoff of each agent ai participating in C. Lc is the goal of coalition.

2.2 Coalition sub-system

The coalition sub-system represents the core of the proposed web services composition system:

2.2.1 Definition of Coalition Value (CV):

Let's consider coalition as two vectors $QW= \{i=1kQwi. /k\ is\ the\ number\ of\ services\ included\ in\ the\ composition\}$ *and* $a=\{i=1kUai/k\ is\ the\ number\ of\ agents\ included\ in\ the\ coalition\ \}; V =1D(Ua, QW\ )$ *where* $DUa, QW = i=1k(Qwi- Uai\ )2$

2.2.2 Coalition algorithm

The coalition algorithm is based on three main steps: initialization, web services selection then evaluation:

Input : Q is a required query, $EC=\{\tilde{C}\ \}$ is the set of coalitions output of learning algorithm.
Variables:
TS: threshold of integer, Lc: is a goal of coalition of Boolean ,$\tilde{O}$: set of composition ,Timer is an array of $|EC|$ integer, Reply is an array of $|\tilde{C}\ |$ state ,$\tilde{C}$: potential coalition

1. Initialization

$\tilde{C} \leftarrow$ ø; Lc $\leftarrow$ false; i $\leftarrow$1; Reply[] $\leftarrow$ waiting; Count $\leftarrow$0; z $\leftarrow$1;

while (Lc = false) **do**

For ($\tilde{C}_z \subset EC$) **do**

activate (Timer[z]);

j $\leftarrow$1;

while ((Timer[z]) **and** ($\exists$ reply[j] == unknown) **do**

MemberShipRequest($\tilde{C}_z[a_j]$) */* return the response of the agent based on the value of μ_c if it is low than the threshold negative response is given else the negotiation is accepted and an OK is send */*

Wait();

j $\leftarrow$j +1;

switch (|reply[z] == OK|)

case (=$\tilde{C}_z$) : **goto** evaluation

case (> $|\tilde{C}_z|/2$) : **goto** selection

else z $\leftarrow$z +1

end;

2. Selection

For all ($a_k \in EC$) **do**

If ($a_k \uparrow w_k \subset Q \uparrow I^R$) **then**

Update($\tilde{O}$);

If ($efficient(\tilde{O})$= **true) then**

$\tilde{C} \leftarrow \tilde{C} \cup \{ a_k \}$;

3. Evaluation

for all ($a_k \in \tilde{C}_z$) **do**

if ($Tr_k > \tau$) and ($\alpha_c >>$) **then** */* τ is threshold required */*

if (*VC* >ρ) **then** */* ρ is threshold required */*

MemberShipOffer (a_k);

else

goto (selection);

else

send_reject(a_k);

Lc$\leftarrow$ true;

for all ($a_k \in \tilde{C}_z$) **do**

update(α_c, Tr_k);

2.2.3 Coalition components

Mainly, the coalition agent is composed of two modules:

- **Coalition Module**: it is responsible of extracting the best possible web services alliance which corresponds to the client query. In case where there is no adequate alliance, this module sends a request to the learning module in order to obtain other alliances as new web services collections that meet the client needs. Finally, this module evaluates every alliance through its functional and non functional parameters in order to choose the best one.
- **Learning Module**: it is responsible of improving the services agents' alliances by using genetic algorithms and reinforcement learning to choose the best services using theirs QoS parameters.

Figure 6 illustrates the coalition agent components:

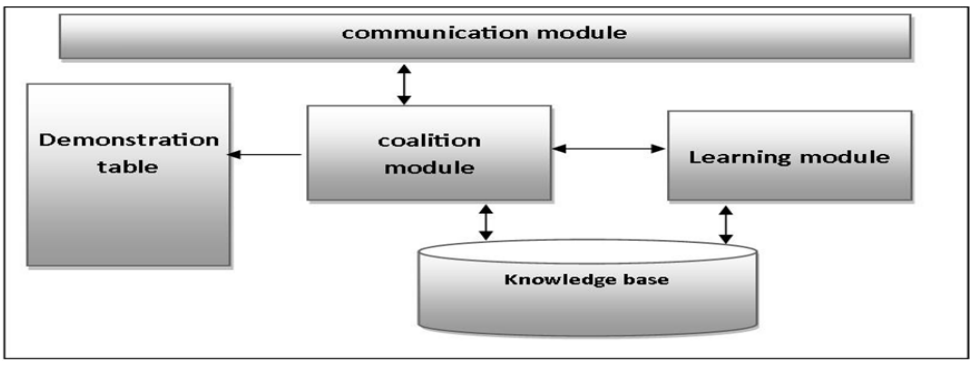


FIG. 6 – *Coalition agent components.*

2.3 Learning sub-system

Using learning techniques and algorithms give to any information system the capability to auto-adapt its self to uncertainty and unexpected cases. The two learning methods adopted in this work are the reinforcement learning combined with the genetic algorithm. This choice is based on the effective results of the study described in Bendahmene et Kazar, (2011).

2.3.1 Definition of Reinforcement learning

The Reinforcement learning measures the performance of the agent in time line by including its past experiments Bendahmene et Kazar, (2011). A benefit *r(t)* on a made action is equal to the variation of coalition's performance between $t-1$ and t.

The parameter of reinforcement G is the assignment of credit which is affected by the environment (reward or penalty). A rate of the positive accumulations of the assignments of credit G^+ is the percentage of positive rewards won further to the actions executed in the past by report the total number of the assignments and obtained credits. A negative rate of

the accumulations assignments credit G^- is the percentage of negative rewards won further to the actions executed in past by report the total number of the assignments obtained credits. The measure of performance *P* of an agent (process of learning) at the moment *t* is defines by two values, the rate of the positive and negative assignments credit, G^+ and G^- successively accumulated until moment *t*.

2.3.2 Protocol of reinforcement learning

Each period *t*, the agent follows the following protocol to estimate its performance:

1. The agent is in a state *x(t)* receives the new web service data from the WS index.
2. It updates its data and generates its *r(t)* and its assignment corresponding credit G
3. Record *r(t)* and G.
4. Calculate G^+ and G^-.
5. Generate the measure of performance *P (t).*
6. If *P (t)* < 0 then send a signal of change for the process of learning and decrease the reliability value α_c of coalition member's. Otherwise nothing is changed about the learning process but α_c is increased.

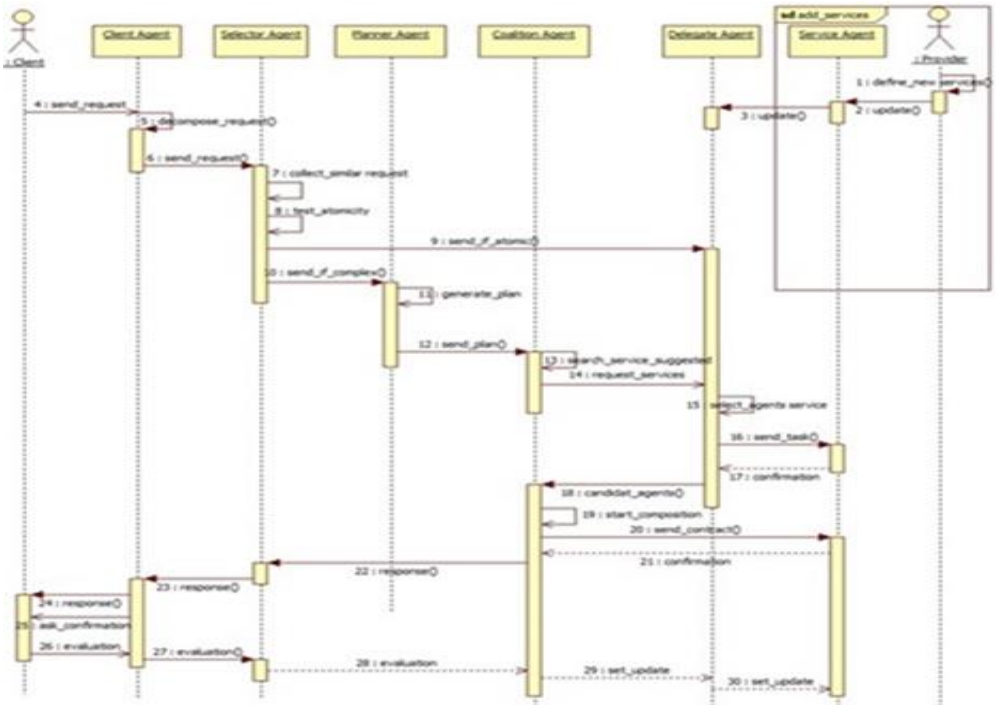


FIG. 7 – *Interaction and communication between system's components.*

2.4 Communication and interaction

As shown in the Figure 7 above, the agents' interaction follows nine main steps:

1. The client agent takes the user query and QoS parameters.
2. The client agent decomposes the query on a set of subqueries and put them in the black table.
3. A signal of query announcement is sent towards the composition layer.
4. As soon as the selector agent receives the signal, it starts to analyze the query complexity level to decide whether it is simple or compound.
5. If the query is a compound one, the selector agent contacts the service layer to coordinate with the service agents.
6. The services agent launch the composition process
7. The planning agent generate the composition plan and send a signal to the coalition agent to start with coalition process
8. The delegated agent supervises the negotiation between service agents taking into account their web services categories.
9. Once the coalition reaches a satisfactory solution the composite services are displayed to the user.

3 Implementation and results

To validate the proposed approach we have implemented a prototype based agent with JADE library. We have also used a bank QoS data of 150 web services to test the performances of the proposed system. The experiments results use the following parameters: availability, price, reputation, Response time, Penalty contract, Trust, and Credibility. We have also use three kind of web services to compose theme: Hotel reservation, Flight reservation and cars agencies. Before launching the creation of the composite web services the client has to determine: the departure and arrival flight information, the number of persons to book a room hotel, and the dates of taking and returning the booked car using the interface illustrated in Figures 8.

FIG. 8 – *Booking interface.*

After that the client has the possibility to define the Quality parameters that have the priority to him using the interface of the Figure 9:

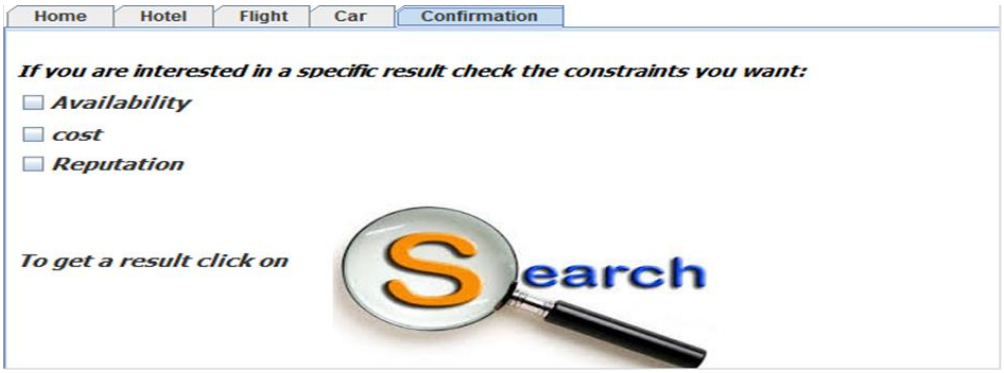


FIG. 9 – *Quality constraints interface.*

As shown in Figures 10, the coalition value rate is reduces, each time we add new web services to the dataset, which indicated that the Euclidian distance between the desired QoS (which represented the learning base) and the obtained QoS reduces, due mainly to the coalition algorithm and the used reinforcement learning technique.

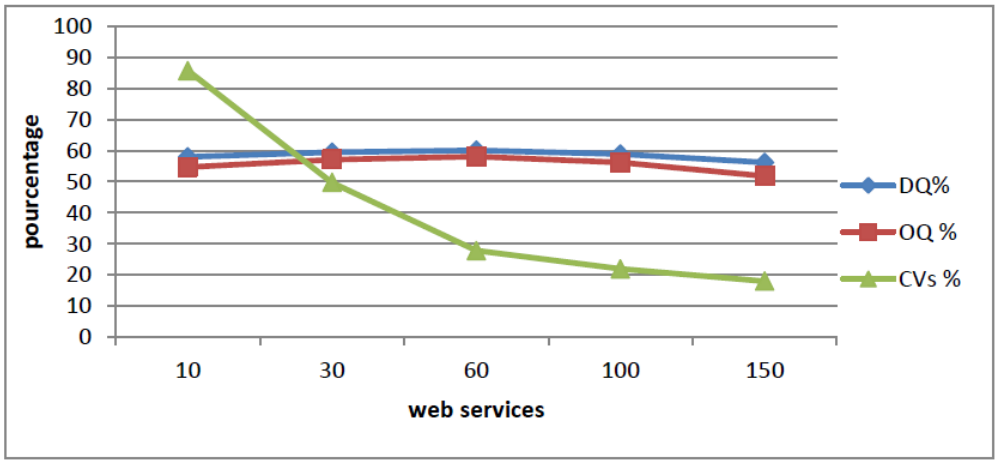


FIG. 10 – *Comparison between desired quality (DQ), Obtained quality (OQ) and the coalition values (CV).*

Finally, As shown in Figures 11, the obtained composite web services QoS parameters are highly best than the atomic web services QoS parameters, due to the use of dele-gated and selectors agents that choose the best composite web services chains using the genetic algorithm and the reinforcement learning processes:

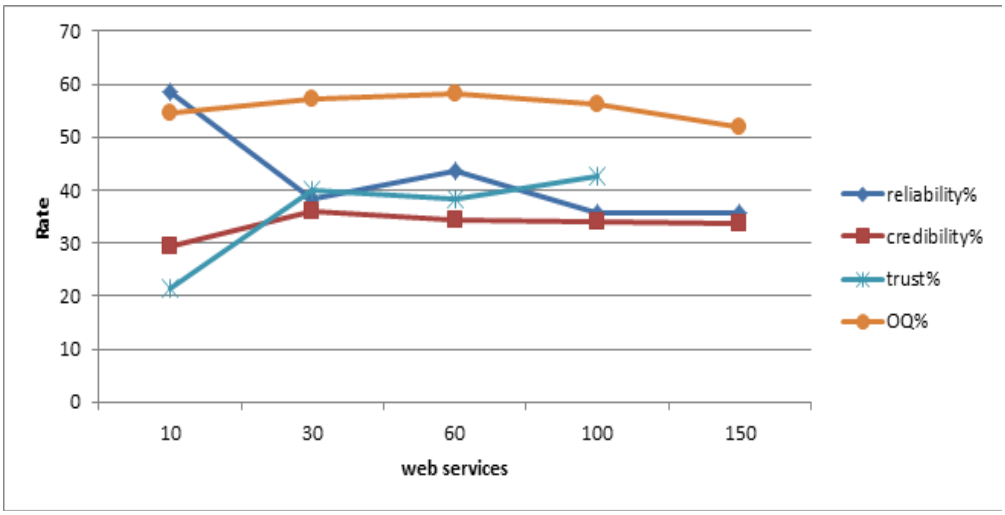


FIG. 11 – *Comparison between Obtained quality(OQ) and some QoS parameters.*

4 Related works and comparison

Ton et al. (2011) proposed an algorithm distributed of composition based planning, with the proposition of a model agent service. A mechanism of exception management for the composite service is missing and doesn't guarantee any adaptation of the system. Besides the lack of a dynamic inter-reasoning process within the service agent, it also doesn't include any optimization on the qualities of services used in the composition.

In Zheng et al. (2008) establish a framework for the needs necessary to the formation and to the negotiation of the coalition's inter-agent. They proposed a function of FO ontology to guarantee the terminological consistency between the agents' services and the agents' applicants. The FO is used to study the stability of a coalition. Therefore, a language of communication MSL bFO is developed. In this work, the authors didn't meet on the possibility of collaboration failure because they used a reduced negotia-tion mechanism.

Lomuscio et al. (2008) developed an approach of specification and verification of conform-ity of the agents through contracts to be respected, with positive and negative evaluations of the members included in the contract of which they proposed syntax of a formal language. One of the inconveniences of this work is that the complication of the proposed language is not automatic.

Maya (2012) proposed a model of negotiation that offers to the suppliers of ser-vices the possibility to participate in the process of composition. The considered crite-ria are at the same time linked to the suppliers as well as their supplied services. The composition is insured by means of coalition inter-agents formation. This approach lacks to treat other criteria of the QoS and it doesn't have a faculty to react in case of blockage or breakdown.

Bourdon et al. (2007) proposed to solve the problem of composition by using the distributed planning multi agents. The objective consists in transforming the plan generated in an OWL-S description of composite service. This latter receives a calculation of QoS according to composed services. The procedure to follow focuses on a demon-strator implemented in

CORE. The detected inconveniences are that the version of the CORE available to this time is not even more developed, it lack of control mechanism on the respect of constraints of the QoS and therefore absence of an assessment pro cedure after the execution. On the other hand, the basis of knowledge of the system is built manually.

Kumar et Mishra (2008) proposed an approach describes services composition based agent by using the seman-tics of the web services. The authors propose two models which differ in terms of using or not an agent coordinator to control the process of composition and to synthe-size the efficiency of each. The obtained results showed that there exist several points to be reviewed like the conditions of negotiation, the validation of entries of a request as well as to reinforce the approach of selection of the service agent's suppliers.

Finally, as illustrated in Table 1 below, the proposed method take into consideration the main QoS parameters in comparison with web services composition approaches which increase its efficiency regarding clients queries and sub-queries. However, this efficiency is not only due to QoS parameters but on the way which they are combined and planned Mehdi et Zarour (2016) by using a reinforcement learning algorithm that added an important dimension to coalition operation in comparison with other approaches that do not give to learning its importance in the creation of composite web services chains. Although the coordination is one of the major asset of the proposed architecture but the lack of ontological presentation and internal interaction protocol to carry out atomic web services linking remains a challenge to improve the proposed architecture.

	Response time	**Reliability**	**Trust**	**Reputation**	**Adaptation**	**Learning**	**protocols**	**Ontology**	**Coordination**
Bourdon et al. (2007)	✓	×	×	×	×	×	×	×	✓
Kumar et Mishra (2008)	✓	×	×	×	×	×	×	×	×
Lomuscio et al. (2008)	×	×	×	×	×	×	✓	×	×
Zheng et al. (2008)	×	×	×	×	×	×	✓	×	×
Tong et al. (2011)	×	×	×	×	×	×	✓	×	×
Maya (2012)	✓	×	×	×	✓	×	✓	×	×
Proposed architecture	✓	✓	✓	✓	✓	✓	×	×	✓

TAB. 1 – *A comparison study between web services composition approaches.*

5 Conclusion

In this paper we proposed a Multi-layer agent architecture based coalition for web services composition. Moreover, the proposed system uses a learning process to facilitate discovery and selection. The system benefits from the analogy between composition and coalition process in order to establish an inter-relationship between the composite services and the atomic ones. Service failures are handled by replacement of any service agent with its equivalent to ensure the high composite services availability.

In the future, we plan to propose a logic representation Xie et al. (2015) to model web services in-side the representing agents. Moreover, deploying the proposed system in a Cloud platform Merizig et al 2018 will certainly enhance the response time and allow us to take into consideration more QoS parameters and increase the system accuracy.

References

Abdullah, A., & Li, X. (2016,). Agent-based model to web service composition. *IEEE International Conference on Services Computing (SCC),* pp. 523-530.

Bendahmene, A & Kazar. O (2011). An Approach Based on Genetic Algorithm for the Learning of an Agent. In the *4th International Conference on Information Systems and Econimic Intelligence*, pp 308-314.

Bourdon, J., Beaune, P., & Fiorino, H. (2007). Architecture multi-agents pour la composition automatique de web services. *Actes de l'atelier Intelligence Artificielle et Web Intelligence, Plateforme AFIA.*

Cetnarowicz, K., Kozlak, J., & Zabinska, M. (2011). Multi-agent approach for com-position and execution of scenarios based on web services. In *International Conference on IEEE Complex, Intelligent and Software Intensive Systems (CISIS),* pp. 478-483.

Kumar, S., & Mishra, R. B. (2008). Multi-agent based semantic web service composition models. *INFOCOMP*, *7*(3), pp. 42-51.

Maya, S. B. (2012). A coalition formation based model for Web service com-position. In *the IEEE Second Inter-national Workshop on Advanced Information Systems for Enterprises (IWAISE),* pp. 28-33.

Mehdi, S., & Zarour, N. E. (2016).Composition of web services using multi agent based planning with high availability of web services. *2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP),* pp. 10-15.

Merizig, A. Kazar, O., & Lopez-Sanchez , M. (2018) A Dynamic and Adaptable Service Composition Architecture in the Cloud Based on a Multi-Agent System. *International Journal of Information Technology and Web Engineering (IJITWE),* 13(1), pp. 50-68.

Gowri, R., & Lavanya, R. (2013). A novel classification of web service com-position and optimization approach using skyline algorithm integrated with agents. In *IEEE International Conference on Computational Intelligence and Computing Research (ICCIC),* pp. 1-8.

Lomuscio, A., Qu, H., & Solanki, M. (2008). Towards verifying compliance in agent-based web service compositions. In *Proceedings of the 7th international joint conference on Autonomous agents and multi agent systems,* pp. 265-272.

Sun, W., Zhang, X., Yuan, Y., & Han, T. (2013). Context-aware Web service composition framework based on Agent. In *International Conference on IEEE Information Technology and Applications (ITA)*, pp. 30-34.

Tong, H., Cao, J., Zhang, S., & Li, M. (2011). A distributed algorithm for web service composition based on service agent model. *IEEE Transactions on Parallel and Distributed Systems, 22*(12), pp. 2008-2021.

Xie, P. Song, Y. Wang, Y. Luo Y. & Zhang, Y. (2015). A solution for web service composition based on logic-interface orchestration, *IEEE 19th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, Calabria, pp. 555-560.

Zheng, L., Tang, J., & Jin, Z. (2008). Requirement driven service agent coali-tion formation and negotiation. In *The IEEE 9th International Conference for Young Computer Scientists*, pp. 322-329.

Résumé

La sélection, la composition et le classement sont les opérations de base qui doivent être effectuées par n'importe quel moteur de recherche de services Web pour satisfaire les besoins et les exigences du client. La complexité et le caractère dynamique de ces tâches nous amènent à proposer une nouvelle architecture multi couches basée agents qui utilise la coalition pour choisir les meilleurs services parmi des centaines de milliers de services web déployés sur le Net. Les agents sont autonomes et auto-adaptatifs, combinés sous forme de coalition, peuvent sélectionner et trouver les meilleur services sans visiter toute la population, ce qui réduit le temps de réponse et la complexité de la composition. Enfin, les résultats de la composition sont encourageants à cause de l'utilisation des paramètres de Qualité des services et d'un nouvel algorithme de coalisation.

Build intelligent in Distributed Embedded System: Wireless Sensor Network as a case study

Amjad Ratrrout*, Abeer Z'aroor**, and Azhar Hamdan***

*The Arab American University, Faculty of Engineering and Information Technology, Computer Science Department, Jenin, Palestine
Amjad.ratrrout@aauj.edu
**The Arab American University, Faculty of Engineering and Information Technology, Computer Science Department, Jenin, Palestine
Abeer.Zaroor@aauj.edu
*** The Arab American University, Faculty of Engineering and Information Technology, Computer Science Department, Jenin, Palestine
Azhar.Hamdan@aauj.edu

Abstract. Wireless sensor networks (WSNs) are networks of distributed autonomous devices that can sense or monitor physical or environmental conditions cooperatively. Managing and routing hundreds or thousands of sensor nodes distributed in a large area consider a very complex task and a huge challenge, with the limited resources and energy. In this paper we will discuss WSNs as a distributed system, Complex adaptive system (CAS) and how to build intelligent using Multi-agent system (MAS). Also we will clarify how MAS reduce the complexity, and how they make intelligent decision in data aggregation and routing to reduce time, cost and energy consumption. In addition, we conduct a simulation using NetLogo to explain how agents communicate with each other to update their belief using filtering algorithm.

Keywords *:Wireless Sensor Network (WSN), Distributed embedded systems, Energy-consumption, complex adaptive system (CAS), Multi-agent system (MAS).*

1 Introduction

Smart environments represent the next evolutionary development step in many fields. For instance, in building, industrial, home, and transportation systems automation. Furthermore, the smart environment relies first and foremost on sensory data i.e. (embedded system) from the real

world (N. Pushpalatha, 2012). To understand what do we mean by sensory data, first we should define the embedded system. Consequently, the embedded system is a special purpose computer system designed to perform one or few dedicated function often with real time computing constraints. Thus, embedded system controls an environment by collecting and receiving data, processing them, and finally returning them quickly to affect the environment at that time. In this paper, we will focus on the distributed embedded system which is a collection of embedded systems that are distributed along a large area and work as a single coherent system. In addition, we take the wireless sensor network as a special case of the embedded distributed system. In this context, we can define the Wireless Sensor Networks (WSN) as a large number of small, inexpensive, disposable and autonomous sensor nodes that are generally deployed in an ad hoc manner in vast geographical areas for remote operations (D. J. Dechene, 2007). These sensor nodes are grouped in clusters, and each cluster has a node that acts as the cluster head.

In this paper we can summary our contributions as the following:

- We will discuss how the WSN is a distributed system, in addition Complex adaptive system (CAS).
- We will discuss how to build intelligent decision sensor networks using Multi-agent system (MAS).
- Finally, we build a simulation using NetLogo to clarify the filtering algorithm. And highlight the benefit of this algorithm which is used to update the belief of an agent by communicates with other agent to know their observation.

The remainder of this paper is organized as follows. In section 2, we explain how the WSNs are distributed system. In section 3, we explain how the WSN is a complex adaptive system and explain all the characteristic of complex system and how it will be adaptive to the environment. Then to solve the complexity problem, we talk in section 4 about Multi-agents in Wireless Sensor Network and explain the important role they play in data aggregation and how they used in the reduction of power consumption. In section 5, we will provide a simulation for the filtration algorithm using Netlogo simulator. Finally, we discuss the conclusion.

2 WSN as a Distributed System

As we said in the previous section, Wireless Sensor Networks (WSNs) are networks of distributed autonomous devices that can sense or monitor physical or environmental conditions cooperatively. The WSNs are having the ability to add large quantities of nodes (sensors) without affecting the whole system or the performance or the effectiveness of the system so the WSN

network is a scale system. Another important characteristic in the WSN that it has the ability to add or delete new nodes at any time without affecting the characteristic of the system so we can say the WSN is a distributed embedded system. In the following sections we will explain all these characteristics of the distributed embedded system and specially in the WSN.

2.1 Transparency in WSN

The main goal in the distributed system is to hide the fact that the component and devices that form the distributed system are allocated widely in separate logical area. Thus, it will appear to the users as a single coherent system. Generally, the transparency concept can be applied to several aspects of a distributed system. In this paper, we will discuss how transparency can be achieved in the embedded distributed system as a case study in the wireless sensor network. The location information is one of the most important issues in the WSN (Raghavendra V. Kulkarni, 2011) because it helps the system to detect the events and to rout packets between nodes to share its information location. For example, across a large building or a large area as a forest, the main task of this network is to send the data to a sensor which is a destinations sensor so the position of this sensor or at least the position of relative sensor node should be detected in the network. To carry out this task, we have two ways to find the positions of these nodes. The first way is using GPS (global position system), but this solution is a costly solution, because in WSN the size of the sensor node should be as small as possible and the energy conservation has been considered as a important issue and cost so we can't overcome these constraint by using the GPS so because all of these constraints of the GPS other researchers used another alternative solution which is the Multipath routing algorithm (Raghavendra V. Kulkarni, 2011)which is a cost and energy conservable solution, it's requiring the sensor nodes to be able to locate itself in various environment by the localization which determines the position of the node by using a routing protocol that use the power of received beacon signal of three anchor node which is a sensor node has prior knowledge of its location coordinates when it is deployed in the network environment since it is equipped with GPS, then the sensor node position will be determined and all theses information about the locations of these sensor should be hidden from the users to achieve location ***transparency***. Sensor nodes are grouped in clusters, and each cluster has a node that acts as the cluster head. All nodes forward their sensor data to the cluster head, which in turn routes it to a specialized node called sink node (or base station) and all information about how these agents can access theses base station or reach the resources should be hidden form the users to achieve the access transparency. Replication transparency in WSN can be describe as simple as that agents may make a copy of themselves in each data collection (replicated child) and placed in the same node the parent located in and give the child half the energy it has and then send that

child that have the same policy behavior and type (gene) to the base station to report data (P. Boonma, 2008).

Fault tolerance in WSN A distributed system has components of the system spread over a wide area and is communicating by cabling. There is no guarantee that communication with a particular node will always be available so the system must continue to operate if a node fails to communicate. It also must automatically reconnect The fault tolerant systems are close related to the description of the dependable systems, which are enclose following requirements availability, reliability, safety, and maintainability (M.Vinyals, 2010).

2.2 Scalability in WSN

The scalability means that the system is still scale and keeps its performance even if we add more nodes in large quantities to the network. Therefore, the scalability in the WSN allow hundreds of sensors within a building to be added in such a way that allows the system to still be able to achieve its tasks without affecting the its performance. Since WSN are usually composed of thousands of nodes making infeasible approaches where the computational cost is exponential to the number of sensors. Therefore, the WSN need for scalability which gives the network the ability to add a large number of sensors without limiting the effectiveness or the performance of the network (Gershteyn, 1996).

2.3 Openness in WSN

First, the distributed system means that a system consists of independent systems or components that work together as a single coherent system. So the openness concept in distributed system means that the ability of the system to add new components or delete existing one without affecting the whole system and if we want to explain this idea by an example of a distributed system we can talk about the WSN. The openness in the WSN distributed system is clear in that we have the ability to add new sensor every day and or delete new sensor without affecting the performance of the WSN and the network will still have the same characteristics even if we add more nodes (sensors). Here we can ask ourselves to which degree we will still able to add more nodes system in an open system??

The answer is that the system has a **border** (edge, threshold) this border shows to which degree the system can understand more nodes by the openness characteristics and when we add more and more nodes to the system this may leads to the harmony which is not acceptable, because we have the edge which deiced to which degree the open distributed system can accept more nodes.

3 WSNs from a complex adaptive systems perspective

3.1 Complexity in Wireless Sensor Networks

WSNs contain of hundreds or even thousands of sensor that collect information or detect even and exchange data between them, so to send the data from node to node to the base station, transmitting data will create a level of complexity: time complexity, message complexity and energy cost complexity for some tasks, such as collecting raw data from all nodes to a sink and data aggregation. The flow of information will be many-to-one since all sensors will send the data to the base station. Data-centric mechanisms that perform in-network aggregation of data are needed in this setting for energy-efficient information flow (B Krishnamachari, 2002). So there is no easy way to manually design a sensor network that acts properly in all possible environmental and network changes i.e. the complexity will be how to design sensor network adapt to changes in structure, and have minimal communication cost (Localization) and less power consumption (power management).

3.2 Adaptation, co-evolution and dynamics

Sensor networks are dynamic systems because of the effect of its internal changes or the effects of external forces change over time. For example, sensors can appear/ disappear over time in an unpredictable way. Hence, a sensor network operation should be able to retrieve and adapt to the current network states or changing network conditions including changes in network topology, node energy level, and the coverage and exposure bounds of WSNs. They also co-evolve to ensure survival in the new environment. Co-evolution is very important to increase the fitness with the environment by reproduction of agent to evolve to adopt the environment needs and changes.

The Co-evolution happened once agents arrive to base station then base station evaluate these agents according to their objective and select the best performing agents and propagate them to individual nodes. Agent running on each node performs reproduction with one of the propagated agents and the **reproduced agent** inherits a behavior policy (gene) from parents via crossover and mutation (new child are mutated through small, random genetic changes **in order to increase diversity** the process is repeated generation after generation until either a fit-enough solution is found or a previously set computational limit is reached) (RV Kulkarni, 2011)and then the new child replace existing agent. This behavior of the agent aims to evolve and improve agents to make them fit better to the environment by make agent whose fitness to the current network condition is very high (i.e. agents have effective behavior policy such moving to base station in short latency)

and eliminates agents that have fitness low for example consuming too much power. The selection of the mating parent based on agent selects one of the elite agents that have the most similar gene.

Co-evolution in which initially "dumb" individuals evolve through cooperation or competition and become fit enough to survive. Evolutionary algorithms model the natural evolution, which is the process of adaptation with the aim of improving survival capabilities through processes such as natural selection, survival-of-the-fittest, reproduction, mutation, competition and symbiosis (RV Kulkarni, 2011)

3.3 WSNs from a complex adaptive systems perspective (John Holland)

As CASs are formed of agents interacting with each other, adapting and co-evolving (M Rupert, 2008). They can be used to model phenomena where global behavior emerges from the local behavior of system entities and components. According to John Holland, he identified seven basic elements of a CAS (J. Holland, 1995). we follow his module and apply it on the WSNs:

- Aggregation: is the property by which agents aggregate sensed data and send them to sink node. Sensor networks contain too much data for an end-user to process. Therefore, automated methods of combining or aggregating the data into a small set of meaningful information is required by agent and this process very important in reducing consumption of power because the power needed to transform these data to base station will be more than the energy needed in aggregating data.
- Tagging: is the mechanism that used for agent identification. In WSNs attribute or tags contain information about the agent for example agent type is data collection agent or event detection agent behavior policy, sensor data to be reported to a base station, and the ID of a node where the data is collected (P. Boonma, 2008).
- Non-linearity is the property in which the emergent behavior of the system resulting from the interactions between aggregate agents is more complicated than a simple summation or average of the simple agents.
- Flows: The flow of information in sensor network will be many-to-one because all sensors will send the data to the base station.
- Diversity: The diversity of skills, experiments, strategies and rules of different agents ensures the dynamic adaptive behavior of a CAS.
- Internal models or schemas are the functions or rules agents use to interact with each other and with their environment. For example agents use Foundation for Intelligent Physical Agents (FIPA) specifying how agents themselves should communicate and interoperate in a standard way (Poslad, 2007).

- Building blocks: constructing the system from pre-define component instead of build it from scratch this approach called building blocks. In WSNs use middleware packages that provide basic building blocks. SNACK (CL Fok, 2005)is a middleware that provides a high-level language and a library of application-level services. Build over TinyOS it helps to reorganize the program to maximum efficiency and the compiler has great flexibility in rearranging components for higher efficiency, it provides a richer syntax for specifying parameter's values for example instead of using constraint value it can use in between, at least, at most allowing the complier to rearrange the control flow.

3.4 Ant Colony Model and Wireless Sensor Networks

Success rate of data transmissions from individual nodes to base stations is an important objective because higher success rate ensures that base stations have more data to make better informed decisions. At the same time, the **latency** of data transmissions from individual nodes to base stations is another important objective. Lower latency ensures that base stations can collect sensor data for more quickly and make more timely decisions, but success rate and latency conflict with each other, in success rate it's apply hop-by-hop and this can degrade latency. For improving latency, nodes may transmit data to base stations with the shortest paths; however, success rate can degrade because of traffic congestion on the paths. Ant colony autonomously satisfies conflicting objectives for example search for food, maintaining temperature inside a nest and minimizing the number of dead drones. If ants focus only on searching for food, they fail to satisfy their well-being. And this lead us in WSNs both latency and success rate can be achieved. Agents (ant) have: **attribute** which contain information about the agent for example agent type, behavior policy (gene), sensor data, **body** and **behaviors** what behave according to the sensed conditions will the agent make. We can classify the Agent behaviors into 7 stages as follows:

- **Food gathering and consumption**: agent periodically reads sensor data to gain energy and consumes a constant amount of energy for living.
- **Pheromone emission**: Agents emit different type of pheromone like, migration pheromone and alert pheromone. Migration pheromone emit on their local node when they migrate to neighbor node to show the destination node an agent migrate to. Alert pheromone emits when agent fail migration within a time period and this prevent link/node failure.
- **Replication**: agents make a copy of themselves in each data collection (replicated child) and placed in the same node the parent located in and give the child half the energy it has and then send that child that have the same policy behavior and type (gene) to the base station to report data.
- **Migration**: Agent may move from one node to another to transmit agent to base station. The migrate agent decide the path according to 3 factors: alert pheromone, migration pheromone

and base station. In base station emit pheromone to the individual node to help the agent to know where the base station locate and this pheromone decrease from hop to hop the agent sense the pheromone and move toward the station and this path will be the shortest path. The agent may not go to this path if the pheromone density very high and this emphasis that there is a heavy load in this path and many agent in this path so it select another path and this prevent latency. In alert pheromone agent avoid moving to a node referenced by an alert pheromone bypassing link/node failures.

- **Swarming**: agent may **merge** with others on their way to base station multiple agent become one. Resulting agent aggregate sensor data contained in other agent and use behavior policy of the best agent in term of power consumption and latency and this reduce power consumption because processing data take less energy than transmission.
- **Reproduction**: Agent running on each node performs reproduction with one of the propagated agents and the reproduced agent inherits a behavior policy (gene) from parents via crossover and then the new child replace existing agent.
- **Death**: Agent periodically consume energy for living and for invoke their behaviors, and due to they can't balance the gain and expenditure, agent die because the lack of energy. Then local platform removes the agent (ineffective behavior policy) and release all resources allocated to the agent (P. Boonma, 2008).

4 Multi-agents in Wireless Sensor Network

When we design WSN we consider some parameter to make it more reliable. Using low memory space with low performance CPU will not reduce the productivity. However, this will increase it. low performance resources in WSN will help it to work for long time with less power consumption which lead to reliable network of nodes each one of them has play important role in collection of the data. WSN system is categorized as distributed embedded system, and the essence of embedded systems is each component does a single function with low performance resources, so poor resources are suitable to this environment. In some environments, they use rich resources because it used to do more than one task but this will put the burden on the consumption of energy. The data that collected from nodes by agents should not transmit as it is, because it will consume energy and time. some WSNs should perform considerable processing tasks to reduce the costs. According to Franklin and Graesser (1996) define agent as system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future.

WSN are mostly consisting of base-station and one or more of sensor nodes. The role of base-station is to send some commands to nodes to control them or to change their tasks or in case that we have problem in some nodes and we need to recover the problem. The data that collected by agents can be transmitted directly to the base-station or through other nodes (Ad-hoc manner), and

this depend on the schema or technique that the network built on (R Tynan, 2005). Using agents will improve data collection and aggregation.

4.1 Multi-agent based routing in wireless sensor network

WSNs have some limitations of several network resources such as limited energy supply, limited computing power, and limited bandwidth of the wireless links connecting sensor nodes. In this section we will discuss some of the routing challenges and design issues that affect routing process in WSNs. The challenges are explained as the following:

- **Scalability of WSN:** Since the WSN is scalable and consist of large number of hundred and thousand nodes, it should provide with a routing protocols scalable to different network sizes**.**
- **fault tolerance**: if we have a failure in sensor nodes that should not affect the overall task of the sensor network. The sensor node should have self-recovering and in many nodes that fail the routing protocol should find new links to route data to base station.
- **Quality of Service**: since we have different application in the sensor network that means we have different quality of services requirement. In some applications, data should be delivered within a certain period of time from the time it sends until it reaches its destination. If the data late from this specific time data will be useless. So the latency in data delivery is an important condition in time-constrained applications. Another important condition in WSN is conservation of energy, which is directly related to network lifetime. So the network protocol design should consider the quality of service requirements for a specific application (Kumar, 2010)
- **Production Costs (low node cost)**: the cost of a single node should be low because it is very important to justify the overall cost of the whole networks because the network consist of large number of these nodes. (Rajashree.V.Biradar, 2010)

5 Simulation

The multi-agent filtering problem is to efficiently represent and update the agents' beliefs through time as the agent's act in the world (L Zettlemoyer, 2009). In our simulation we used NetLogo simulator to do filtering algorithm. Filtering algorithm use to update the belief of an agent by communicates with other agent to know their observation.

In our simulation we have num-nodes (agents) connected to each other. Each of these agent will do the filter algorithm to make sure that every agent connected with other agent does not have the same solution (solutions are: red, green or blue). For instance, as shown in Figure 1, we can see that agents have the same solutions before applying the filtering algorithm.

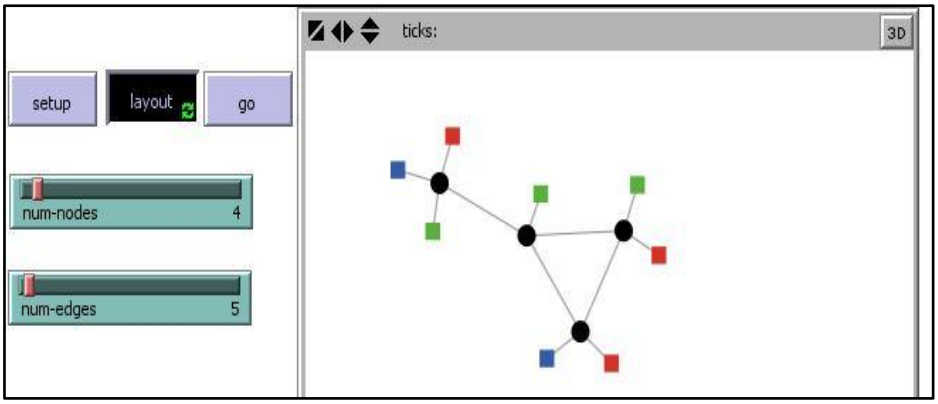


FIG. 1 – *Agents before applying the filtering algorithm have the same solutions*

Each agent communicates its domain to its neighbor and then removes values that cannot satisfy constraints from its domain. Agent x performs the following procedure revise for each neighbor.

```
to revise [other-node]
  let my-domain item who domain
  let his-domain item ([who] of other-node) domain
  if (length my-domain = 0) [stop]
  if (length his-domain = 0) [stop]
   if (length his-domain > 1) [stop]
   let his-color first his-domain
   if( member? his-color my-domain) [
     let my-new-domain (remove his-color my-domain)

     set domain replace-item who domain my-new-domain

     ask edge-neighbors [
        set domain replace-item ([who] of myself) domain my-new-domain
      ]
   ]
```

Check if (his-color) is in my-domain then it will remove his-color from my-domain. If some value of the domain is removed by performing the procedure revise, Agent x will send the new domain to its neighboring. If a new domain is received from a neighbor, call procedure revise again and do the same process.

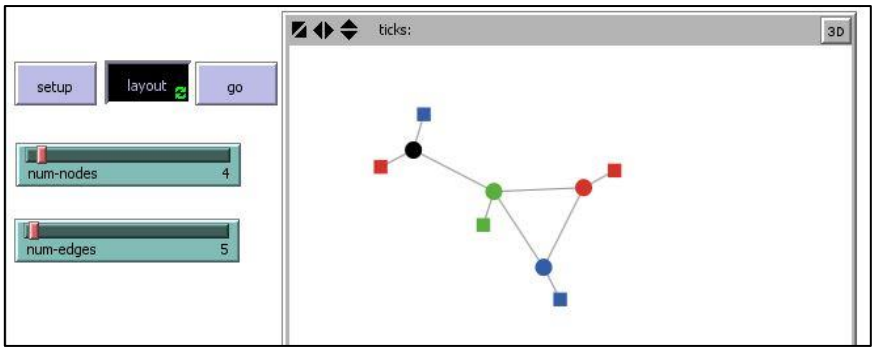


FIG. 2 – *Agents after applying the filtering algorithm have the different solutions*

As we see in Figure 2, after applying the filtering for the graph each agent does not have the same solution with its neighbor agent.

6 Conclusion

In this paper we discussed WSN as a distributed system, explain it from complex adaptive system perspective. In addition, we show the complexity in WSN, and we discuss how we can maximize the life time of the network and minimize the power consumption by some proposed techniques. Finally, we provided a simulation for agent's behavior, and how they can communicate to avoid conflict by applying a filtering algorithm.

References

B Krishnamachari, D. E. (2002). The impact of data aggregation in wireless sensor networks.

CL Fok, G. R. (2005). Software Support for Application Development in Wireless Sensor Networks.

D. J. Dechene, A. E. (2007). *Clustering Algorithms for Wireless Sensor Networks.*

Gershteyn, Y. (1996). Fault Tolerance in Distributed Systems.

J. Holland, H. O. (1995). How Adaptation Builds Complexity.

Kumar, S. (2010). Routing Protocols in Wireless Sensor Networks – A Survey.

L Zettlemoyer, B. M. (2009). Multi-Agent Filtering with Infinitely Nested Beliefs.

M Rupert, A. R. (2008). The web from a complex adaptive systems perspective. *Journal of Computer and System Sciences 74*, 133-145.

M.Vinyals, J. R.-A. (2010). A Survey on Sensor Networks from a Multi-Agent perspective.

N. Pushpalatha, D. (2012). *Shortest Path Position Estimation between Source and Destination nodes in Wireless Sensor Networks with Low Cost.*

P. Boonma, J. S. (2008). MONSOON: A Co-evolutionary Multi-objective Adaptation Framework for Dynamic Wireless Sensor Networks.

Poslad, S. (2007). Specifying Protocols for Multi-agent System Interaction.

R Tynan, G. O. (2005). Multi-agent System Architectures for Wireless Sensor Networks.

Raghavendra V. Kulkarni, A. F. (2011). *A Survey Computational Intelligence in Wireless Sensor Networks.* IEEE.

Rajashree.V.Biradar, V. .. (2010). Classification and comparison of routing protocols in wireless sensor networks.

RV Kulkarni, A. F. (2011). Computational Intelligence in Wireless Sensor Networks: A Survey.

Probabilistic failure prediction technique using neural networks in cloud computing

Youcef Bezza* , Ouided Hioual*
* Abbes Laghrour University of Khenchela, Algeria
{youcefbezza044,ouided.hioual}@gmail.com

Abstract. In this paper we present a probabilistic failure prediction technique using supervised probabilistic neural networks in cloud computing. we interested in the prediction at the hardware layer of the cloud. the proactive technique used predicts system failures in the future and replaces suspicious components, Assuming that this layer contains a set of different machine processors that works with the cloud, the prediction is done with respect to the run time and the failure probability, where each neuron represents a processor with these selected characteristics and the output layer the optimal execution time. Supervised learning is a phase of neural network during which the behavior of the network is modified until the desired behavior is obtained, we used for this modification the load balancing algorithm to calculate the load of each processor and the optimal execution time, The obtained results , show that this method gives us good results.

1 Introduction

Cloud Computing is rapidly becomes one of the most technologies in the world of engineering and computer science. This technology proposes combination of soft-ware and resources which shows dynamic scalability in nature (Sutari et al, 2017). It guaranties real costs effective and agility to organizations.

Cloud computing serves the demands of a number of individuals and organizations. Demands may be advanced end services, data or any other computing resources (Mahalkari and. Tondon , 2014).Cloud computing delivers IAAS (Infrastructure As A Service), PAAS (Platform As A Service) and SAAS (Software As A Service) (Yang and Ma , 2008).Google, Amazon, windows Azure etc, are the famous popular cloud service providers. Each service provider provides different services built on the demand of the users.

For example Amazon provide IaaS service Google Provides all services like SaaS, PaaS and IaaS, (Rajesh and KannigaDevi, 2014).

ANN is essentially familiarized from the topic of biology where neural network plays an essential and main role in human body. Neural network is responsible of the work of human body (Vidushi et al , 2012).

The Proactive fault tolerance strategy is to maintain a strategic distance from recovery from fault, and failure by predicting them and replace the suspected component means detect the issue before it actually come (Muijnck-Hughes and Hons,2011)

Cloud computing face many problems to detect and predict failure of its services, especially in hardware layer , for this fault tolerance is the biggest challenge cloud computing

face, because fault tolerance is an important point to ensure the availability and reliability of serious services and the execution of the application.

As fault tolerance is very important we propose a probabilistic failure prediction technique using supervised probabilistic neural networks in cloud computing which is a virtual machine composed of a set of unstructured resources that are classified into three layers. In our work, we interested in the prediction at the level of the hardware layer of the cloud. The prediction technique used is the proactive technique. The neurons of the networks represent the processors of the cloud. the neuron values of the input layers are the execution time and the failure probability , the output layers the optimal execution time, Using Supervised learning method.

The remainder of this paper is structured as follows: In section 2, we introduce some related work In section 3, we present the proposed architecture and the functionality of the proposed model, In Section 4, we introduce a case study to illustrate the functionality of our mode Finally the section 5, present a conclusion and future work.

2 Related work

There is a lot of work has been done about fault tolerance in cloud computing, we examine this:

In (Karahroudy, 2011), The authors proposed a fault tolerance middleware which implement a synchronized server replication plan, where a failed server is maintain with a consistent state.

In (Labaf ,2007), The authors proposed a model name AFTRC(Adaptive Fault Tolerance in Real-time Cloud computing)This scheme tolerates the faults on the basis of reliability of each computing node. The proposed scheme is a good option to be used as a fault tolerance mechanism for real time computing on cloud infrastructure.

The authors in(Rajasekaran and VijayalakshmiPai, 2011), proposed a system autonomic fault tolerance The experimental results demonstrate that the proposed system can deal with various software faults for server applications in a cloud virtualized environment.

In (Zhao et Al , 2010), authors, proposed an adaptive mechanism for replica distribution for effective fault tolerance in cloud computing, which can effectively used to achieve higher level data availability. The proposed replica distribution instrument recognizes the machine to take the backup or do the retrieval in the situation of the cloud data neglects to load on end clients machines.

The authors in(Malik and Huet ,2011), proposed a methode which is combination between EIPR and SBA algorithm for task scheduling and replication process in the cloud with efficient and effective performance

In (Singh et al ,2013), authors proposed technique can provide better performance in terms of accuracy and detection speed, which is critical for the cloud system.

The authors in (Arunkumar and Kesavamoorthi , 2016), proposed an optimized fault tolerance approach where a model is designed to tolerate faults based on the reliability of each compute node (virtual machine) and can be replaced if the performance is not optimal. Obtained results are suggests a good performance of our model compared to current existing approaches.

3 Proposed model

One of the principal canons of Cloud Computing is the 'as-a-Service' paradigm in which certain service is offered by a Service Provider to a User for use. This service can also be classified according to the application domain of its deployment (Muijnck-Hughes and Hons, 2011), Cloud computing has three major layer : Software as a Service(SaaS),Platform as a Service(PaaS),Infrastructure as a Service(IaaS).

Infrastructure as a Service (IaaS) the user is delivered with the capability to processing, storage and any software which they need to run and the operating system which they select on the cloud infrastructure. The user does not control the cloud infrastructure but networking components like host firewall, storage, operating systems and deployed applications are controlled by the consumer.(Yang and Ma, 2008),IaaS is mentioned to as hardware as a service. It is a delivery model in which an organization outsources the equipment used to support operations, including storage, hardware, servers, and networking components (Kara-hroudy , 2011).Our fault tolerance prediction technique is applied in this layer (IaaS) . Fault tolerance is the action of looking for faults and weakening in a system.

If a fault take place or there is a hardware failure or software failure, then also the system should function properly. Failures should be manipulated in a dynamic way for good Cloud Computing. we are interested in our work on fault tolerance at the level of the material layer of the cloud using probabilistic neuron networks according to the proactive technique, This method is to avoid extra effort for recovering the failed tasks, nodes, by predicting the fault in before and replace them with other working parts,(Tchana et al , 2012),

3.1 Architecture of our system

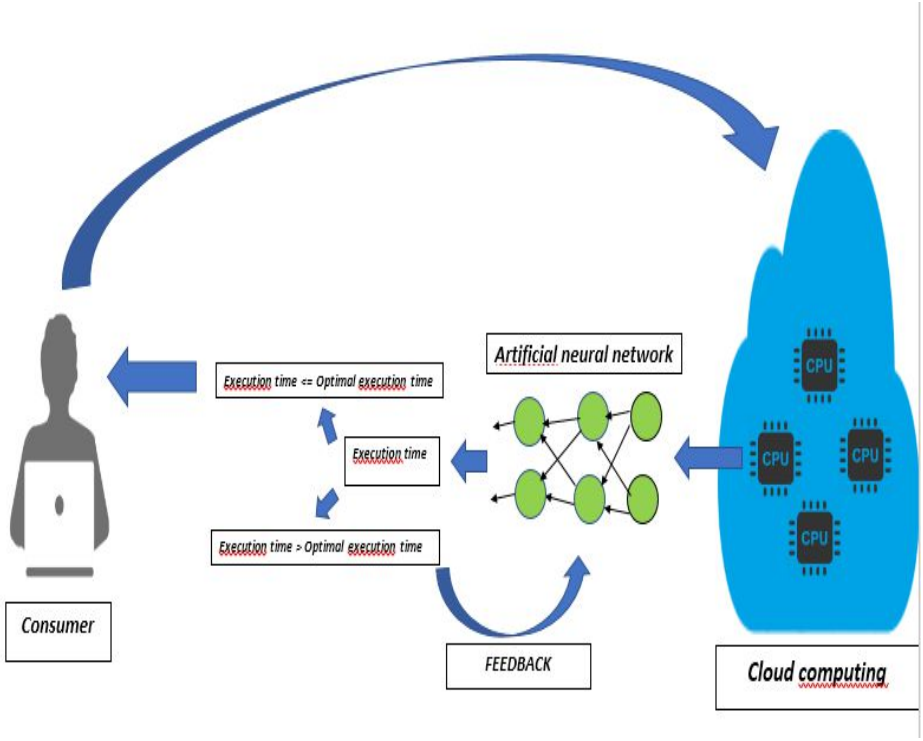


FIG. 1 – *.Architecture of our model*

Figure 1 illustrates the general architecture of our system which consists of cloud components as well as the neural network used

- Cloud Computing is a platform that makes processing and storage available to end users as services. In our work we interested to the hardware layer of the cloud, assuming that this layer contains a set of different machine processors that works with cloud.

- Artificial Neural Network (ANN)
  ANN is a machine learning prediction algorithm that is inspired by animal nervous system,(Kakoulli et al ,2012), A neural network is a composite structure which contain a collection of interconnected neurons which delivers a very exciting alternatives for difficult complex problem answering and other application which can play important role in computer science field (Kumar1 and Sharma, 2014), Network in Neural Network mean the interconnection between neurons. Present in various layers of a system. Every system based on input and output layer and one or more hidden layer. The input layer has input neurons which transfer data via synapses to

the hidden layer, and similarly the hidden layer transfers this data to the output layer via more synapses. The synapses stores values called weights which helps them to manipulate the input and output to various layers ,(MehtaniRoll, 2011).

The different classes of neural networks are: Single Layer Feedforward Network, Multi-layer Feedforward Network, Recurrent Networks.Learning methods, in NNs :

- Supervised Learning: every input pattern that is used to train the network is associated with a target or the desired output pattern.
- Unsupervised Learning: In this learning method, the target output is not presented to the network.

3.2 Functionality of the proposed model

In our work we are interested of the prediction at the hardware layer level of the cloud, assuming that this layer contains a set of different machine processors that works with last, the prediction is done according to the execution time and failure probability.

Initially, when the user requests a service (in our case it assumes that the tasks are independent and unitary), the cloud dispatches these tasks to the different machines.

So if a processor has longer run time and a lot of tasks its failure probability. is highest, and likewise if a processor has smaller run time and lots of tasks so its failure probability is highest also.

To raise this problem, and predict failures at the hardware cloud level? We have use a probabilistic neural network with supervised learning, the input layer is the characteristic of the processor (execution time, failure probability) the output layer is the execution time .

The input layer transfers the data to the hidden layer, the last applicate a probabilistic functions in these data and transfer the result to output layer, the desired behavior is the optimal execution time, and the output layer is the execution time of the processor, if the optimal execution time if great then the execution time of the processor no failure will be in the future, but if the optimal execution time is less then the execution time of the processor we do a back propagation .

Supervised learning is a phase in the development of a neural network which during the behavior of the network is modified until the desired behavior is obtained meaning modify the weight of the connections between the neurons and for modify the weights we use an algorithm that calculates the load of each processor.

Since it is assumed that the tasks are independent and unitary, we will use a static load balancing algorithm that is based on an optimal algorithm that already exists.

the Optimal reliable allocation for independent unitary task algorithm used , (Legrand and Robert , 2005)., uses the result of an optimal algorithm that already exists.

The optimal algorithm is based on the principle of distributing a set of independent tasks on a set of processors while respecting the execution time of each one, which aims to assign to the processor having the fastest processing speed a maximum number of tasks to be executed in order to obtain an optimum execution time and consequently obtain a better load balancing ,(Dongarra et al ,2007).

The load balancing manager distributes the tasks without performing a request to know the current load on the system processors. The algorithm used here must therefore be able to calculate the loads in advance, in order to allow an effective load balancing. So the load

produced by a task must be provided by the task itself or by other sources. Also, the different capacities of the processors or network nodes must be known.

Algorithm 1 : Optimal Allocation (Legrand and Robert ,2005):

Distribution ((t_1,….,t_p),M)
{Initialization: calculation of values ni as $n_i * t_i \approx$constant and $n_1 +$ + $n_m \leq N$}
1-For i=1 à m :

2- $n_i =$ $\left\lfloor \frac{\frac{1}{t_i}}{\sum_{i=1}^{m} \frac{1}{t_i}} \times N \right\rfloor$

Iteratively increment the ni which minimize the execution time while

$\sum_{i=1}^{m} ni < N$

3- while

$\sum_{i=1}^{m} ni < N$

4- find kϵ{1,…..,N} where $t_k * (n_k+1) = \min \{t_i*(n_i+1)\}$
5- $n_k = n_k+1$
6- return ($n_1+n_2+\ldots\ldots+n_K$)

Where :
M : The number of processors.
N : The number of tasks.
n_i : The load of the ith processor.
t_i : The execution time of the ith processor.

Algorithm 2 (Dongarra et al ,2007):

```
Distribution ((t1,….,tp),M)
Distribution ((b1,….,bp),M)
Input:  qϵ [0, 1[
Compute Top1= q *Top  using algorithm1
Sort the processor by increasing t1[i]
Sort the processor by decreasing b[i]
Sort the processor by increasing t1[i]* b[i]
 X → 0
for i=1:m
    if ( X < N)
        n[i] →  min( N-X,  int (Top/t[i])
    else
        n[i]=0
X=X+n[i]
```

Where :
n[i] : The load of the ith processor.
t[i] : The execution time of the ith processor.
b[i] : The failure probability.
Top: Optimal execution time.

4 Case study

Cloud Computing is a virtual machine composed of a set of unstructured resources, as we have already said. we are interested in this work in the prediction at the level of the hardware layer of the cloud, supposing that this layer contains a set of different machine processors , the prediction is done according to the execution time and failure probability.

Assuming that, we have 10 processors

This code calculate the run time of a task, the input is the task and the output is the time, we put the task between two Instructions. The two Instructions are long start-time:=system.nanoTime() and long endtime=system.nanoTime().The run time is the difference between the end time and start time.

Execution time algorithm:

```
begin
long starttime:=system.nanoTime()
long endtime=system.nanoTime()
double seconds = (endtime - starttime)/1e9
write("the time of operation is :" +seconds+"seconds")
end
```

Then, we attribute for each processors a failure probability based on the principle that the processor that has the smallest execution time its probability is high.

When the user requests a service , the cloud dispatches these tasks to the different machines.(for example for 200 tasks)

The table below shows for each processor: its execution time, failure probability and its load **.**

N° of processor	Execution Time	Failure probability	The load of the processor
1	0.2531038099098497	0.8165435232144264	4
2	0.8074930112528398	0.9992409774751406	44
3	0.7305904208846935	0.305024712213016702	11
4	0.25919653042495416	0.9779898941576612	5
5	0.7392568580963457	0.49558396931055504	11
6	0.7806177027951934	0.29648073832828514	5
7	0.8384233501055515	0.2563282751909879	10

8	0.13334908826862668	0.9935366037199109	4
9	0.9519213329858853	0.27063924366699776	3
10	0.7725732696203601	0.038143588509292226	3

TAB. 1 – *Distribution of load for processor*

if we takes for example the case of the processor number 2 in table 1, we note that his execution time , failure probability is high and it has the biggest load, therefore this processor probably will have failed, and the response time to meet the demand of the user is very high.

So, to solve this problem we used a probabilistic neural network which is build up of three layer.

The first layer contains 10 neurons where each neuron represents a processor with these selected carectistics and the output layer the optimal execution time

Concerning the modification of the weights of the neuron networks, we used a load balancing algorithm mentioned in section 3. and each time we compare the execution time of each processor after allocation of the load with the optimal execution time of the output layer , if the execution time of the neuron after the allocation of the spots is superior to the optimal execution, we will go backwards thanks to the proactive prediction technique used and we reallocate the loads until the best load, in this way so no processor breaks down. this process is called learning in the neural network.

N° of processor	**Execution Time**	**Failure probability**	**The load of the processor**
1	0.2531038099098497	0.8165435232144264	27
2	0.8074930112528398	0.9992409774751406	4
3	0.7305904208846935	0.305024712213016702	3
4	0.25919653042495416	0.9779898941576612	24
5	0.7392568580963457	0.49558396931055504	5
6	0.7806177027951934	0.29648073832828514	3
7	0.8384233501055515	0.2563282751909879	2
8	0.13334908826862668	0.9935366037199109	29
9	0.9519213329858853	0.27063924366699776	1
10	0.7725732696203601	0.038143588509292226	2

TAB. 2 – *The best Distribution of load for processor*

5 Conclusion

Fault prediction is used to provide system availability and robustness when system have hardware or software fault. The work presented in this paper concerns a probabilistic failure prediction technique using supervised probabilistic neural networks in cloud computing and proactive technique, we made the prediction at the level of hardware cloud, we have used the

load balancing algorithm at the intermediate layer of the neuron network in order to have a better balancing , We would like to evaluate our proposed model through some real systems case studies.

References

Arunkumar, B., and M. Kesavamoorthi (2016).Task Scheduling and Seedblock Based Fault Tolerance in Cloud, Volume 11, Number 6 .

Dongarra, J. et al (2007) . Bi-objective Scheduling Algorithms for Optimizing Makespan and Reliability on Heterogeneous Systems, SPAA'07, June 9–11.

Kakoulli, E., V.Soteriou, and T.Theocharides (2012). Intelligent Hotspot Prediction for Network-on-Chip-Based Multicore Systems, in Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on , vol.31, no.3, pp.418-431.

Karahroudy, A. (2011), Security Analysis and Framework of Cloud Computing with Parity-Based Partially Distributed File System.

Kumar1, E., and E.Sharma (2014). Artificial Neural Networks-A Study, Volume 2, Issue 2, PP 143-148.

Labaf, M. (2007), Esfahan University, Rule Extraction From Artificial Neural Networks Under Background Knowledge.

Legrand, A., and Y. Robert (2005). Algorithlmique Parallèle. Dunod.

Mahalkari, A., and P.RituTondon (2014) . A Replica Distribution Based Fault Tolerance Management ForCloud Computing ,Vol. 5 (5).

Malik, S., and F. Huet (2011). Adaptive Fault Tolerance in Real Time Cloud Computing , IEEE World Congress on Services.

MehtaniRoll, P. (2011). Pattern Classification using Artificial Neural Networks.

Muijnck-Hughes, J., and B.Hons (2011). Data Protection in the Cloud.

Rajasekaran, S., and G.A.VijayalakshmiPai (2011). Neural Network, Fuzzy Logic and Genetic Algorithm, Prentice Hall of India, pg-13-20.

Rajesh, S., and R.KannigaDevi (2014). Improving Fault Tolerance in Virtual Machine Based Cloud Infrastructre, Volume 3, Special Issue 3.

Singh, T., G. TarakaramaRaviTeja, and P.SrinivasaPappala (2013). Fault Tolerance- Challenges, Techniques and Implementation in Cloud Computing, International Journal of Scientific and Research Publications, Volume 3, Issue 6.

Sutari, V., M.Bhavsar, and V.K Prasad (2017). Fault Prediction and Mitigation in Cloud Computing ,Volume 8, No 3 .

Tchana, A., L. Broto, and D. Hagimont (2012). Approaches to cloud computing fault tolerance, pp. 1–6.

Vidushi, S., S.Rai, and A.Dev (2012). A Comprehensive Study of Artificial Neural Networks, Volume 2, Issue 10.

Webbing, Z. et Al (,2010). Fault Tolerance Middleware for cloud computing .Third International Conference on Cloud Computing.

Yang, K., and J. Ma (2008). Implementation of IEEE802.1x in OPNET, in Proc. 7th Asia Simulation Conference on System Simulation and Scientific Computing (ICSC), pp.1390-1394.

Résumé

Une technique de prédiction des pannes dans le cloud computing en utilisant des réseaux de neurones probabilistes supervisés est presenté . nous nous sommes intéressés à la prédiction dans la couche matérielle du cloud. la technique proactive utilisée prédit les défaillances du système à l'avenir et remplace les composants suspects. supposant que cette couche contient un ensemble de processeurs différents, la prédiction est faite en fonction du temps d'exécution et de la probabilité de tombé en panne. Chaque neurone représente un processeur avec ces caractéristiques sélectionnés et la couche de sortie le temps d'exécution optimal. L'apprentissage supervisé est une phase du réseau neuronal au cours de laquelle le comportement du réseau est modifié jusqu'à l'obtention du comportement désiré, nous avons utilisé pour cette modification l'algorithme d'équilibrage de charge pour calculer la charge du processeur et le temps optimale, Les résultats obtenus montrent que cette méthode nous donne de bons résultats.

Modélisation et répartition d'un Big Data Warehouse

Mourad Ghorbel*, Karima Tekaya**
Abdelaziz Abdellatif***

*Université de Tunis El Manar, Faculté des Sciences de Tunis,
Département informatique, LIMTIC, El manar 2092, Tunis, Tunisie.
ghorbel.fst@gmail.com,
**Université de Tunis, Ecole Supérieure des Sciences Economiques
et Commerciales de Tunis, Montfleury 1089, Tunis, Tunisie.
karima.tekaya@gmail.com
***Université de Tunis El Manar, Faculté des Sciences de Tunis,
Département informatique, LIPAH, El manar 2092, Tunis, Tunisie.
abdelaziz.abdellatif@fst.rnu.tn

Résumé. Poussées par la croissance continue des données, les approches des entrepôts de données doivent être adaptées. Généralement, les modèles en étoile, en flocon de neige ou en constellation sont utilisés comme des modèles logiques. Tous ces modèles sont inadéquats lorsqu'il s'agit de données massives qui ont besoin de systèmes évolutifs et flexibles.
Nous proposons dans cet article une modélisation d'un Big Data Warehouse. Cette modélisation sera par la suite utilisée pour une solution de répartition d'un Big Data warehouse issue du Benchmark TPC-DS. Cette solution a été implémentée en Java.

1 Introduction

Face à la mondialisation et à la concurrence grandissante, la prise de décision est devenue cruciale pour les dirigeants d'entreprises (au sens large du terme, entreprises privées, publiques, institutions, organisations...). L'efficacité de cette prise de décision repose sur la mise à disposition d'informations pertinentes et d'outils d'analyse adaptés. L'objectif des entreprises est de pouvoir exploiter efficacement d'importants volumes d'informations, provenant soit de leurs systèmes opérationnels, soit de leur environnement extérieur, pour l'aide à la décision.
L'informatique décisionnelle a connu et connaît aujourd'hui encore un essor important. Elle permet l'exploitation des données d'une organisation dans le but de faciliter la prise de décision. La diversité des solutions proposées pour la fragmentation des ED montre son importance. Cependant, l'efficacité de ces solutions doit évoluer au niveau des Big Data Warehouse. Les charges de travail volumineuses peuvent dégrader les performances du système de gestion des bases de données (SGBD), et ainsi, ralentir les applications et augmenter ainsi le temps de réponse au client, souvent exigeant dans les délais, en particulier lorsqu'il s'agit d'un décideur. Malgré l'intérêt accordé aux techniques de fragmentation des données et la diversité des solutions proposées, nous avons constaté que ce problème n'a pas eu l'attention qu'il mérite en

dépit de son importance.
Dans ce qui suit, nous présentons un exemple d'état de l'art qui se focalise sur la modélisation des Big Data et la fragmentation des ED. Ensuite, nous allons proposer une solution pour la modélisation et la répartition d'un Big Data Warehouse. Enfin, nous allons appliquer cette solution sur le Benchmark TPC-DS. Cette solution a été implémentée en Java.

2 Etat de l'art

Nous commençons ci-dessous par un ensemble de travaux sur la fragmentation des ED.

- Tekaya (2011) propose une approche de fragmentation du modèle logique global d'un ED qui garantit, d'une part, la conformité des modèles obtenus avec les principes de la modélisation multidimensionnelle et, d'autre part, leur adéquation aux besoins spécifiques par site. Cette approche de répartition des fragments sur différents sites tient compte des fréquences d'utilisation des différents fragments, de la contrainte de communication et de la contrainte de chargement. L'intégration d'une méthode de contrôle permet de calculer et d'évaluer les performances d'une répartition donnée selon des seuils de performance.

- Mahboubi (2008) s'intéresse à la FH des entrepôts de données XML afin de les répartir sur plusieurs sites. Pour ce faire, il se base sur l'adaptation des techniques de fragmentation existantes sur les entrepôts XML comme la FH primaire basée sur les prédicats et celle basée sur les affinités de prédicats. Ensuite, il présente une nouvelle méthode qui se base sur les concepts de fouilles de données à savoir la fragmentation basée sur la classification des prédicats.

- Boukhalfa (2009) propose un ensemble d'approches permettant d'optimiser les entrepôts de données. Ses approches d'optimisation reposent sur l'utilisation de trois techniques d'optimisation : la FH primaire, dérivée et les index de jointure binaires

- Darmont (2006) propose le banc d'essais ocb (object clustering benchmark) et le modèle de simulation voodb (virtual object-oriented database), qui ont pour objectif de remédier à des problèmes de performence en se positionnant comme des outils génériques, paramétrables et adaptés à l'étude du regroupement d'objets.

- Arres et al. (2014) proposent un algorithme basé sur MapReduce qui permet, à partir des données publiques du Ministère Français de la Communication et de la Culture, de définir un classement des galeries et musées nationaux selon leurs degrés d'accessibilité aux personnes handicapées.

Nous présentons ensuite un ensemble de travaux sur les Big Data.

- Yangui et al. (2016) proposent de nouvelles règles pour transformer un modèle conceptuel multidimensionnel en deux modèles NoSQL : orienté colonne et documentaire des modèles. Pour chaque modèle, ils distinguent deux types de transformation : simple et hiérarchique. Pour valider leur transformation des règles, ils ont mis en IJuvre quatre entrepôts de données en utilisant Cassandra comme un système NoSQL axé sur les colonnes et MongoDB comme système NoSQL orienté document. Ces systèmes ont été implémentés en utilisant des routines Java dans l'outil Talend Data Integration et évalué en termes de "Write Request Latency" et de "Read request Latency" en utilisant le test de performance TPC-DS.

- Atzeni et al. (2016) présentent les notions traditionnelles liées à la modélisation de données qui peuvent également être utiles dans ce contexte. Plus précisément, ils proposent NoAM

(NoSQL Abstract Model), un nouveau modèle de données abstrait pour les bases de données NoSQL, qui exploite les points communs des différents systèmes NoSQL. Ils proposent également une méthodologie de conception de base de données pour les systèmes NoSQL basés sur NoAM, avec des activités initiales indépendantes du système cible spécifique. NoAM est utilisé pour spécifier une représentation indépendante des systèmes des données d'application et, ensuite, cette représentation intermédiaire peut être implémentée dans des bases de données NoSQL cibles, en tenant compte de leurs caractéristiques spécifiques. Dans l'ensemble, la méthodologie vise à soutenir l'évolutivité, la performance et la cohérence, selon les besoins des applications Web de la prochaine génération.

- Dehdouh (2016) ont présenté la construction des cubes OLAP à partir de grands entrepôts de données implémentés en utilisant le modèle NoSQL en colonnes. L'utilisation de modèles NoSQL est motivée par l'incapacité du modèle relationnel, généralement utilisé pour implémenter l'entreposage de données, à permettre facilement l'évolutivité des données. En effet, le modèle NoSQL en colonnes convient pour stocker et gérer des données massives, en particulier pour les requêtes décisionnelles. Cependant, les SGBD NoSQL axés sur les colonnes n'offrent pas d'opérateurs d'analyse en ligne (OLAP). Leur contribution principale est de définir un nouvel opérateur de cube appelé MC-CUBE (MapReduce Columnar CUBE), qui permet de construire des cubes NoSQL en prenant en compte les aspects non relationnels et distribués lorsque les entrepôts de données sont stockés.

- Chevalier et al. (2015b) définissent un ensemble de règles pour mapper des schémas en étoile dans deux modèles NoSQL : axés sur les colonnes et sur les documents. La partie expérimentale est réalisée en utilisant le référentiel de référence TPC. Leurs expériences montrent que les règles peuvent effectivement instancier de tels systèmes (schéma en étoile et réseau). Ils analysent également les différences entre les deux systèmes NoSQL considérés. Dans leurs expériences, HBase (orienté colonne) est plus rapide que MongoDB (orienté document) en termes de temps de chargement.

- Data Warehousing et OLAP sur Big Data deviennent un des émergents défis pour la recherche des prochaines générations, avec un accent particulier sur le cloud data-intensive infrastructures. En conséquence, plusieurs études focalisent l'attention sur ce problème, et divers problèmes ouverts se posent. Cette preuve a inspiré l'étude de Cuzzocrea (2015), qui fournit un aperçu complet sur les problèmes de recherche ouverts réels dans le contexte de l'entreposage de données et OLAP sur Big Data, avec une discussion critique profonde sur les futures orientations de recherche à être pris sous cette route si difficile.

- Cuzzocrea et al. (2013a) mettent en évidence les problèmes ouverts et les tendances de recherche actuelles dans le domaine de Data Warehousing et OLAP sur Big Data, un terme émergent dans l'entreposage de données et la recherche OLAP. Ils ont aussi dérivé plusieurs pistes de recherche novatrices dans ce domaine et ont mis l'accent sur les contributions possibles à réaliser par de futures recherches.

- Chevalier et al. (2015a) ont étudié la mise en place d'entrepôts de données multidimensionnels Systèmes NoSQL. Ils définissent des règles de mappage qui transforment le conceptuel modèle de données multidimensionnel aux modèles axés sur les colonnes logiques. Ils considèrent trois modèles logiques différents et les utilisent pour instancier des entrepôts de données. Ils se concentreront sur le chargement de données, la conversion de modèle en modèle et le cuboïde OLAP calcul.

- Scabora et al. (2016) étudient trois conceptions physiques d'entrepôt de données pour

adapter le Benchmark Star Schema pour son utilisation dans les bases de données NoSQL. En particulier, leur enquête principale fait référence au traitement des requêtes OLAP sur des bases de données orientées colonnes utilisant le framework MapReduce. Ils analysent l'impact de la distribution des attributs parmi les familles de colonnes dans HBase sur les performances de la requête OLAP. Leurs expériences ont montré comment le temps de traitement des requêtes OLAP a été impacté par une conception physique de l'entrepôt de données en ce qui concerne le nombre de dimensions accédées et le volume de données. Ils concluent que l'utilisation de distributions distinctes d'attributs parmi les familles de colonnes peuvent améliorer les performances des requêtes OLAP dans HBase et, par conséquent, constituer la référence plus adapté à OLAP sur les bases de données NoSQL.

- Cuzzocrea et al. (2013b) étudient les solutions basées sur le partitionnement des données pour la construction parallèle de cubes de données OLAP, adaptés à nouveaux environnements Big Data, et ils proposent le framework OLAP, avec le benchmark associé TPC-Hd, une transformation appropriée du référentiel d'entrepôt de données bien connu TPC-H. Ils démontrent grâce à des mesures de performance, l'efficacité de la proposition framework, développé au dessus du serveur ROLAP Mondrian.

Les solutions existantes de la fragmentation et de la répartition des ED ont été étudiées récemment dans Ghorbel et al. (2016). La plupart de ces solutions se limite sur les ED centralisés. Elles prennent en considération la diminution du temps d'exécution des différentes requêtes spécifiques et du nombre de prédicats ainsi qu'au nombre de fragments, mais elles se bloquent par une grande complexité face au nombre de prédicats et par le non contrôle du nombre de fragments générés qui peut dans certains cas s'accroître.
Les travaux présentés dans cet état de l'art se basent sur la modélisation d'un Big Data selon les modèles des ED. Ils se focalisent à rendre toutes les données cohérentes. Mais, ils se limitent au niveau fiabilité des données et temps d'exécution des requêtes. Nous avons voulu travailler sur des données cohérentes en essayant d'assurer leur sécurité et minimiser le temps d'exécution des requêtes. Nous avons voulu présenter le problème d'un autre axe. C'est pour cela, la répartition des Big Data Warehouse devient une nécessité vu le nombre des données et l'augmentation des besoins des utilisateurs.
Dans ce qui suit, nous allons présenter une modélisation et une répartition d'un Big Data Warehouse.

3 Démarche préconisée

Nous présentons dans cette section, une approche de modélisation d'un Big Data Warehouse et une approche de répartition de celui-ci.
Dans ce qui suit, nous allons travailler sur un Big Data Warehouse en couvrant deux dimensions d'un Big Data qui sont le Volume et la Vitesse. Les données dans notre exemple sont structurées. On touchera la Variété dans nos prochains travaux.

3.1 Modélisation d'un Big Data Warehouse

En raison de l'énorme quantité de données, l'intégration de ces données externes avec les données internes de l'entreprise dans un entrepôt de données sont une approche promet-

teuse. Cette dernière aborde une importance primordiale et attire l'attention de nombreuses recherches.
Cependant, les méthodologies actuelles d'entreposage avec des bases de données relationnelles ne peuvent pas être appliquées avec succès pour gérer la complexité croissante et le volume de données généré à partir de l'entrepôt de données. Les règles conçues pour les données relationnelles ne peuvent pas être appliquées aux données générées par les services des Big Data.
C'est pour cela, nous proposons une approche hybride qui permet de transformer les entrepôts de données en Big Data Warehouse.
Cette architecture permet la fusion des avantages Big Data avec les avantages des entrepôts de données.
Avec l'arrivée du Big Data : le besoin de l'évolutivité dans l'architecture des ED est devenu crucial avec l'approche traditionnelle.
Le Big Data couvre trois dimensions : volume, vitesse et variété.
Volume : les entreprises sont submergées de volumes de données croissants de tous types, qui se comptent en téraoctets, voir en pétaoctets.
Vitesse : pour les processus chrono sensibles tels que la détection de fraudes, le Big Data doit être utilisé au fil de l'eau, à mesure que les données sont collectées par votre entreprise afin d'en tirer le maximum de valeur.
Variété : le Big Data se présente sous la forme de données structurées ou non structurées (texte, données de capteurs, son, vidéo, données sur le parcours, fichiers journaux, etc.). De nouvelles connaissances sont issues de l'analyse collective de ces données.
Le Big Data Warehouse est un gros entrepôts de données qui couvre que le volume et la vitesse du Big Data.
Les avantages des entrepôts de données sont : la fiabilité et d'assurez qu'aucune donnée n'est perdue.
L'avantage des Big Data est : le flux de données.
L'avantage du Big Data Warehouse est d'assurer la sécurité d'un gros entrepôts de données. D'où nous gardons les avantages acquises en gagnant au niveau sécurité.
Ci-dessous un exemple de modélisation d'un Big Data Warehouse.

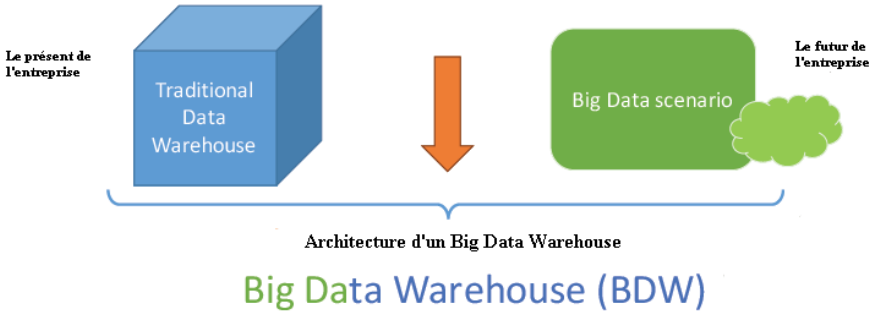


FIG. 1 – *Modélisation d'un Big Data Warehouse*

3.2 Répartition d'un Big Data Warehouse

Nous considérons que la méthode de classification par l'algorithme Bond Energy Algorithm (BEA) comme étant une méthode originale pour la fragmentation. Elle permet de contrôler à l'avance le nombre de fragments et d'intégrer les caractéristiques de la base et de l'utilisation des données sous un format quantitatif dans des matrices simples à utiliser. Nous nous sommes inspirés du travail réalisé par Mahboubi (2008) pour la conversion des prédicats et du travail réalisé par Tekaya (2011) au niveau des fréquences d'utilisation des prédicats.
Nous proposons dans cette partie une solution pour la répartition d'un Big Data Warehouse. Nous commençons par générer une liste de prédicats de l'algorithme. Ensuite, nous détaillons les différentes étapes de la démarche de répartition au niveau logique (fragmentation des tables) du Big Data Warehouse.
Dans notre travail, nous nous sommes concentrés sur la liste initiale des requêtes. Cette liste constitue un point d'entrée commun aux solutions de répartition les plus connues dans l'état de l'art. Nous considérons l'augmentation des données comme étant un problème important qu'il faut traiter par répartition. En effet, les prédicats sont extraits à partir des requêtes les plus fréquentes. Dans ce qui suit, nous proposons une solution pour la répartition d'un Big Data Warehouse suivant l'algorithme BEA. Son but est de minimiser le temps de réponse des requêtes utilisées et le coût de chargement des données. La solution proposée se déroule en cinq phases :
- phase de sélection des prédicats ;
- phase d'utilisation des prédicats ;
- phase de codification ;
- phase de classification ;
- phase de répartition ;

Nous avons opté pour une solution dirigée par la classification des prédicats car elle permet de contrôler à l'avance le nombre de fragments utilisant comme entrée une liste de prédicats déjà minimale et complète.
La solution proposée nous permet de répartir notre Big Data Warehouse selon les besoins des utilisateurs. Nous réduisons le nombre de fragments générés et nous minimisons le temps d'exécution des requêtes ainsi que le coût de chargement des données. Dans la section qui suit, nous présentons un exemple détaillé de notre solution, ainsi que son application sur un Big Data Warehouse réel issu du banc d'essai TPC-DS. Dans ce qui suit, nous allons valider notre solution par le benchmark TPC-DS avec une implémentation en Java de cette solution.

4 Etude expérimentale

Pour valider notre travail, nous avons utilisé un Big Data Warehouse réel issu du benchmark TPC-DS PilHo (2014). Ce benchmark demeure le plus utilisé par les travaux abordant le problème de la fragmentation des Big Data Warehouse que nous considérons similaires à notre contexte de travail. Le choix de TPC-DS nous permettra de comparer nos résultats à ces travaux aussi bien au niveau technique que pratique de la solution. Nous avons utilisé le TPC-DS pour valider la méthode avec un SF=1TB de données.
Sur ce Big Data Warehouse, nous avons exécuté un ensemble de 4 requêtes réparties sur trois sites géographiquement distants.

4.1 Exemple d'application sur le benchmark TPC-DS

Nous commençons par la 1ère phase :
La requête R1 contient les prédicats P1 et P3.
La requête R2 contient les prédicats P2 et P3.
La requête R3 contient les prédicats P2 et P4.
La requête R4 contient les prédicats P3 et P4.
Nous avons utilisé 4 requêtes et 4 prédicats.
Nous passons à la 2ème phase : Comme indique la figure 2 de la MUP que chaque cellule montre qu'un prédicat donné appartient à la requête correspondante ou non.
La requête R1 contient les prédicats P1 et P3.
La requête R2 contient les prédicats P2 et P3.
La requête R3 contient les prédicats P2 et P4.
La requête R4 contient les prédicats P3 et P4.
Ensuite, nous poursuivons par la phase 3. La figure 3 est un exemple de MFU. Chaque cellule

	P_1	P_2	P_3	P_4
R_1	1	0	1	0
R_2	0	1	1	0
R_3	0	1	0	1
R_4	0	0	1	1

FIG. 2 – *Matrice d'utilisation des prédicats (MUP)*

contient la fréquence d'utilisation d'une requête par les utilisateurs sur un site donné.
Au finale, la requête R1 est appelée 45 fois.
La requête R2 est appelée 5 fois.
La requête R3 est appelée 75 fois.
La requête R4 est appelée 3 fois.

Puis, nous passons à la 4ème phase. On a comme données pour le moment :
R1 : P1 P3 45
R2 : P2 P3 5
R3 : P2 P4 75
R4 : P3 P4 3
Nous aurons donc notre MA suivante pour la première étape :

Après le remplissage de la Ma (Figure 4), nous appliquons l'algorithme BEA sur cette matrice comme deuxième étape. On fixe au début les deux premières colonnes comme données (Figure 5) puis on calcule le BOND de chaque colonne pour avoir le maximum d'affinité des

	S_1	S_2	S_3
R_1	15	20	10
R_2	5	0	0
R_3	25	25	25
R_4	3	0	0

FIG. 3 – *Matrice des fréquences d'utilisation (MFU)*

	P_1	P_2	P_3	P_4
P_1	45	0	45	0
P_2	0	80	5	75
P_3	45	5	53	3
P_4	0	75	3	78

FIG. 4 – *Matrice d'affinité (MA)*

prédicats. D'où, la 3ème colonne sera insérée soit au début, soit au milieu ou soit à la fin des 2 premières colonnes.

	P_1	P_2	P_3	P_4
P_1	45	0	45	0
P_2	0	80	5	75
P_3	45	5	53	3
P_4	0	75	3	78

	P_1	P_2	P_3	P_4
P_1	45	0		
P_2	0	80		
P_3	45	5		
P_4	0	75		

FIG. 5 – *Initialisation de la Matrice de classification des affinités (MCA)*

On calcule maintenant l'ensemble de contribution :
Cont(A0,A3,A1) = 8820
Cont(A1,A3,A2) = 10150
Cont(A2,A3,A4) = 1780
Voici le calcul de Cont (A0,A3,A1) :
-Cont(A0,A3,A1)=2(bond(A0,A3)+bond(A3,A1)-bond(A0,A1))
-bond(A0,A3)=aff(A1,A0)*aff(A1,A3)+aff(A2,A0)*aff(A2,A3)+
aff(A3,A0)*aff(A3,A3)+aff(A4,A0)*aff(A4,A3)= 0 + 0 + 0+ 0=0
-bond(A3,A1)=aff(A1,A3)*aff(A1,A1)+aff(A2,A3)*aff(A2,A1)+
aff(A3,A3)*aff(A3,A1)+aff(A4,A3)*aff(A4,A1)=45*45+5*0+53*45+3*0=4410
-bond(A0,A1)=0
-Cont(A0,A3,A1)=2*4410=8820
On trouve que la meilleure composition est (A1,A3,A2) qui donne le maximum d'affinité des prédicats. Nous permutons au début la 2ème et la 3ème colonne (Figure 6).

	P_1	P_2	P_3	P_4
P_1	45	0	45	0
P_2	0	80	5	75
P_3	45	5	53	3
P_4	0	75	3	78

	P_1	P_3	P_2	P_4
P_1	45	45	0	
P_2	0	5	80	
P_3	45	53	5	
P_4	0	3	75	

FIG. 6 – *Permutation de la MCA (1ère étape)*

De même pour la 4ème colonne, la contribution (A3,A2,A4) est la meilleure composition qui donne le maximum d'affinité des prédicats. (Figure 7)

	P_1	P_2	P_3	P_4
P_1	45	0	45	0
P_2	0	80	5	75
P_3	45	5	53	3
P_4	0	75	3	78

	P_1	P_3	P_2	P_4
P_1	45	45	0	0
P_2	0	5	80	75
P_3	45	53	5	3
P_4	0	3	75	78

FIG. 7 – *Permutation de la MCA (2ème étape)*

De même que la permutation des colonnes, on permute les lignes. Dans notre exemple, nous avons permuté les colonnes 2 et 3, donc on termine par permuter les lignes 2 et 3. La

figure 8 est la matrice résultante.

	P_1	P_2	P_3	P_4
P_1	45	0	45	0
P_2	0	80	5	75
P_3	45	5	53	3
P_4	0	75	3	78

	P_1	P_3	P_2	P_4
P_1	45	45	0	0
P_3	45	53	5	3
P_2	0	5	80	75
P_4	0	3	75	78

FIG. 8 – *Permutation de la MCA (3ème étape)*

Nous abordons la dernière phase de classification des prédicats. Nous aurons comme résultat deux classes : C1=P1 et P3 C2= P2 et P4. Nous fragmentons notre Big Data Warehouse selon ces deux classes (Figure 9).

	P_1	P_3	P_2	P_4
P_1	45	45	0	0
P_3	45	53	5	3
P_2	0	5	50	75
P_4	0	3	75	78

FIG. 9 – *MCA*

A partir de l'ensemble des requêtes présenté en annexe, nous avons collecté 4 prédicats de sélection sur les tables de dimension. Sur ces prédicats, nous avons appliqué les 5 phases de notre solution. Nous aurons au finale deux classes de prédicats donc deux fragments.

4.2 Présentation et implémentation de la solution

Pour la répartition du Big Data Warehouse sur ces 2 machines, nous avons utilisé un réseau privé virtuel Pham (2002). Un VPN repose sur un protocole, appelé "protocole de tunnélisation" qui permet aux données passant d'une extrémité à l'autre du VPN d'être sécurisées par des algorithmes de cryptographie. Pour la génération des tables du banc d'essai TPC-DS, nous avons installé Oracle 10g sur la machine 1 et Oracle 11g sur la machine 2 que nous avons

trouvé les plus adaptées aux capacités des machines. Nous allons présenter ci-dessous des captures d'écran de la solution implémentée en Java (Figures 10, 11 et 12) : La première partie

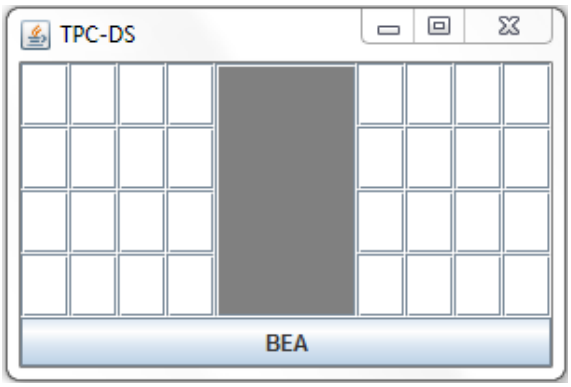


FIG. 10 – *Interface graphique (1/3)*

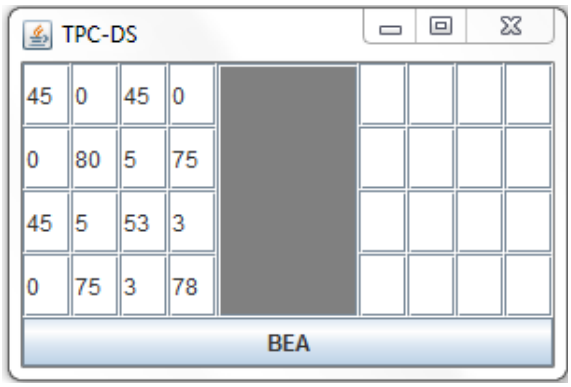


FIG. 11 – *Interface graphique (2/3)*

est le remplissage de la matrice. Ensuite, avec un simple clic sur le bouton BEA, la matrice devient classer. Nous trouvons les classes colorées en jaune.

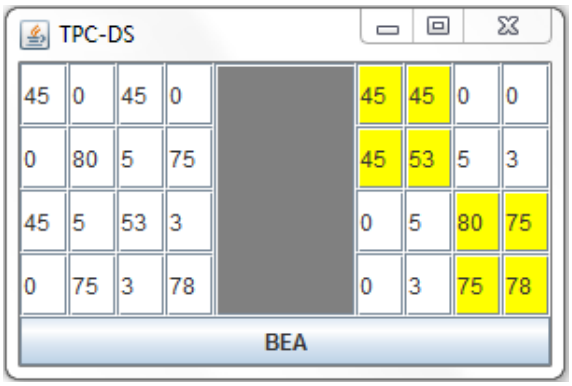


FIG. 12 – *Interface graphique (3/3)*

4.3 Interprétation des résultats obtenus

Dans notre exemple, la dernière phase de classification a engendré 2 classes de prédicats. Nous avons utilisé la conjonction de ces prédicats par classe pour le partitionnement des tables sur deux fragments. Les fragments engendrés ont été par la suite, alloués aléatoirement sur les deux machines distantes. Pour la validation de notre solution, nous avons commencé par mesurer les temps d'exécution des requêtes dans le cas d'un Big Data Warehouse centralisé, ensuite, réparti en appliquant l'algorithme BEA.
Pour les requêtes numéro 1, 2, 3 et 4, les temps d'exécution ont diminué, ce qui constitue pour nous un gain non négligeable. Le temps d'exécution global des requêtes dans un contexte réparti a diminué de 70% par rapport au contexte centralisé (Figure 13). Pour la 1ère et la 2ème

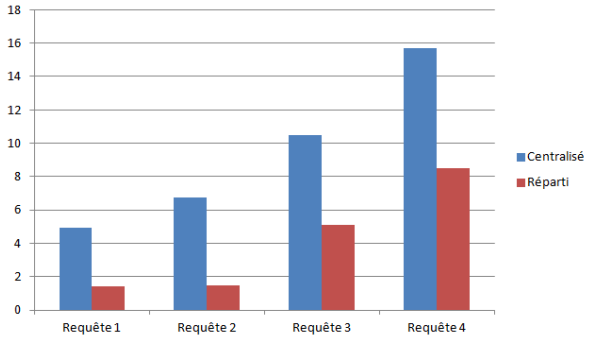


FIG. 13 – *Temps d'exécution de chaque requête en minutes*

requête, la diminution du temps d'exécution dépasse le 2/3 jusqu'au 3/4 du temps du centralisé vers le réparti. Par contre, pour la 3ème et la 4ème requête, le temps d'exécution diminue mais ne dépasse pas la moitié du temps d'exécution centralisé. Cela est dû par l'augmentation du nombre de jointures pour ces deux requêtes.

Nous avons donc diminué le temps d'exécution des requêtes pour un Big data Warehouse réparti avec une minimisation des fragments engendrés et minimisation du coût de chargements des fragments. Nous avons passé de trois sites à deux. La répartition nous minimise le temps d'exécution des requêtes, le coût de chargements des données, mais cela ne veut pas dire que l'augmentation des fragments minimise encore le temps d'exécution des requêtes. En effet, l'augmentation des jointures entre les fragments avec quelques fois des problèmes de réseaux peuvent impliquer des inconvénients sur la répartition en temps d'exécution des requêtes.

5 Conclusion et perspectives

Dans cet article, nous nous sommes intéressés à la répartition d'un Big Data Warehouse selon l'algorithme BEA. Nous avons proposé une solution à ce fait. Elle repose sur cinq phases : une phase de sélection des prédicats, une phase d'utilisation des prédicats, une phase de codification, une phase de classification et une phase de répartition. La première phase se base sur la collecte des prédicats des requêtes utilisées. La deuxième phase se base sur la MUP pour générer la matrice d'utilisation des prédicats. La phase de codification consiste à produire une représentation de la MFU selon l'utilisation des requêtes dans les différents sites. La quatrième phase utilise l'algorithme BEA pour obtenir la MCA. La phase de répartition est la phase de la classification engendrée par BEA pour fragmenter notre Big Data Warehouse. Pour l'évaluation de la solution proposée, nous l'avons appliquée sur un Big Data Warehouse réel issu du banc d'essai TPC-DS que nous avons réparti selon notre démarche de fragmentation en utilisant les différentes requêtes proposées comme exemple d'application. Les résultats obtenus sont motivants et garantissent une utilisation plus adéquate et plus souple des données au sein de l'entreprise. A l'issue de ce travail, nous estimons que quelques axes de recherches restent à étudier et à approfondir. Le premier est relatif à l'allocation (ou répartition) des données du Big Data Warehouse en un ensemble de Magasins de données (MD). Le problème de l'allocation des données dans un contexte de Big data Warehouse doit tenir compte des contraintes de répartition notamment la contrainte d'accès à un MD à partir des sites distants, le stockage des données, le délai de réponse et la fréquence d'utilisation de chaque MD par les différents sites de l'entreprise. Il faut aussi tenir compte de la contrainte de chargement d'un MD dans tous les sites.

Dans la plupart des cas, la répartition d'un Big Data Warehouse est fondée sur des critères de fragmentation (attributs, prédicats de sélection, affinité, etc.) et/ou des critères de réplication (fréquence d'utilisation, coûts d'accès, coûts de stockage, etc.). Ces critères évoluent selon les besoins des utilisateurs. Pour faire face aux changements, une mise à jour périodique du schéma de répartition est nécessaire.

Références

Arres, B., N. Kabachi, F. Bentayeb, et O. Boussaid (2014). Application du paradigme mapreduce aux données ouvertes cas : Accessibilit é des personnes à mobilité réduite aux musées nationaux. *EDA*.

Atzeni, P., F. Bugiotti, L. Cabibbo, et R. Torlone (2016). Data modeling in nosql world. *Computer, Standards and Interfaces*.

Boukhalfa, K. (2009). *De la conception physique aux outils d'administration et de tuning des entrepôts de données*. Thèse de doctorat, Université de Poitiers.

Chevalier, M., M. E. Malki, A. Kopliku, O. Teste, et R. Tournier (2015a). Implementation of multidimensional databases in column-oriented nosql systems. *ADBIS*.

Chevalier, M., M. E. Malki, A. Kopliku, O. Teste, et R. Tournier (2015b). Implementing multidimensional data warehouses into nosql. *ICEIS_CEKTT*.

Cuzzocrea, A. (2015). Data warehousing and olap over big data : a survey of the state-of-the-art, open problems and future challenges. *Int. J. Business Process Integration and Management*.

Cuzzocrea, A., L. Bellatreche, et I. Song (2013a). Data warehousing and olap over big data : Current challenges and future research directions. *ACM*.

Cuzzocrea, A., R. Moussa, et G. Xu (2013b). Olap : Effectively and efficiently supporting parallel olap over big data. *MEDI*.

Darmont, J. (2006). *Optimisation et évaluation de performance pour l'aide à la conception et à l'administration des entrepôts de données complexes*. Thèse de doctorat, Université Lumière Lyon 2.

Dehdouh, K. (2016). Building olap cubes from columnar nosql data warehouses. *Model and Data Engineering*.

Ghorbel, M., K. Tekaya, et A. Abdellatif (2016). Réduction du nombre des prédicats pour les approches de répartition des entrepôts de données. *ISI*.

Mahboubi, H. (2008). *Optimisation de la performance des entrepôts de données xml par fragmentation et répartition*. Thèse de doctorat, Université Lumière Lyon 2.

Pham, C. (2002). Vpn et solutions pour l'entreprise. *SaaS*.

PilHo, K. (2014). Transaction processing performance council (tpc). *Guide d'installation*.

Scabora, L., J. Brito, R. Ciferi, et C. Ciferi (2016). Physical data warehouse design on nosql databases olap query processing over hbase. *International Conference on Enterprise Information Systems*.

Tekaya, K. (2011). *Fragmentation et allocation dynamiques des entrepôts de données*. Thèse de doctorat, Faculté des sciences de Tunis.

Yangui, R., A. Nabli, et F. Gargouri (2016). Automatic transformation of data warehouse schema to nosql data base : Comparative study. *KES*.

Summary

Driven by the continued growth of data, data warehouse approaches need to be adapted. Typically, star, snowflake or constellation models are used as logical models. All of these models are inadequate when it comes massive data that require scalable and flexible systems. We propose in this article a modeling of a Big Data Warehouse. This modeling will then be used for a distribution solution of a Big Data warehouse from the TPC-DS Benchmark. This solution has been implemented in Java.

Chapter

3

Internet of Things & Banking

ASD'2018

Content

A roadmap to lead risk management in the digital era

Fadoua Khanboubi and Azedine Boulmakoul
LIM/Innovative Open Systems, FSTM, Hassan II University of Casablanca, B.P. 146 Mohammedia, Morocco,
khanboubi.fadoua@gmail.com, azedine.boulmakoul@gmail.com

Summary. Most banks have already integrated digital in recent years. The banking sector has even pioneered online services. Only, since the digital revolution, new entrants have intruded into financial era and give to traditional banks a hefty shove. Also, the consumption patterns of customers have evolved and their service requirements are in constant development. All these elements, united together, make the role of banks, their products and the very profession of banker destabilized and challenged. This work locates challenges and raises fundamental issues in risk management for Banks' digital transformation.

1 Introduction

I do not think it's too much to say that we're at a turning point in terms of innovation in financial services. Some expect that new technologies will completely disrupt traditional financial institutions, allowing entrepreneurs to access banking business. These innovations in financial services, known as financial technologies (fintech) generate a great deal of enthusiasm. The number of searches for the term "fintech" in Google has increased more than 30 times over the last 6 years (Wilkins C., 2016). Thus, they have the potential to transform a wide range of services within the financial system. And that's a good thing, because there are big gaps in efficiency that can be fixed. Also, while some technologies may seem revolutionary, their overall effect on the financial system is considered “evolutionary”… All those challenges give the financial institutions that adapt a chance to survive, because more and more new service providers will integrate into the financial ecosystem.

Furthermore, we find that large, well-capitalized companies outside the financial space, such as Apple and Google, are beginning to offer financial services. One of the main advantages of traditional institutions is the trusting relationship they have built with their clients. The technology giants also have a large customer base and loyalty to their brand, which could facilitate the adoption of the services they offer. These companies use the information they have about their customers and their existing platforms to offer attractive services at competitive prices and thereby attract established customers to the most profitable business sectors of financial institutions. Their range of services could expand over time. Moreover, mobile use will be, in coming years, the focal point of digital banking. Specially due to improving connection technology and the huge number of smartphones and tablets sold today. Technology trends such as Internet of Things and the full penetration of smartphones and tablets give rise to customer expectations.

However, with the growing spread of digitalization in financial area, most banks have to make extend data and security protection against cyberattacks. Cybersecurity will gain notoriety as banks store more data about their customers and the exposure to cyberattacks will increase in number.

Finally, the time has come for financial institutions, new participants and policy makers to work together. This is the best way to create an environment conducive to the modernization of the financial sector and for a careful management of rising risks. The remainder of the text is organized as follows: Details of the 4th industrial revolution are presented in Section 2. Section 3 develops the impact of digitalization on risk management. Section 4 shows a roadmap for successful banks digital transformation. Section 5 gives the basic principles for dealing with the new dimensions of digital risks. Finally, this work is concluded and future works are highlighted in Section 6.

2 The 4th industrial revolution

2.1 Digitalization in the manufacturing industry

Digitalization is “the conversion of text, pictures or sound into a digital form that can be processed by a computer” (Stevenson A., 2010). Nowadays, the digitalization is leading industry to a "Fourth Industrial Revolution" with offering to businesses considerable opportunities of development, especially, with the increasing number of Internet of Things devices (21 billion connected devices by 2020 according to Gartner, 2017). Thus, **“**industry 4.0” is one of the most important concepts that lead each enterprise to stay competitive becoming a real digital actress in the industrial era (PwC, 2016).

Industry 4.0 is defined as the use of automation, big data, cyber-physical systems CPS, Internet of things and cloud, to construct an intelligent factory that allows the perfect harmony between people, new technologies and innovation (I-Scoop, 2017). It began, with the Internet of connected objects and cloud computing, to produce products through intelligent systems, such as simulation systems and sensors (for example). The 4th industrial revolution represents the highest level of digitalization and defines a new organization of factories, also named smart factories. Their objectives are to better serve their customers through increased flexibility of production.

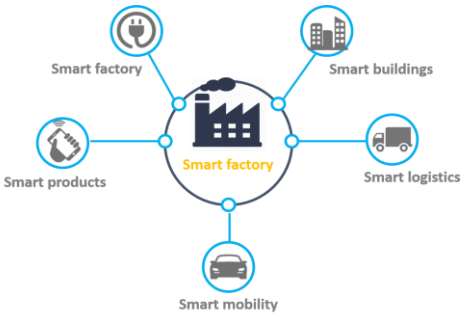


FIG. 1 – *Industry 4.0 of connected and smart world* (*Bauernhansl, ten Hompel, Vogel-Heuser, 2014)*

2.2 Zoom on digitalization in the banking model

Banking digitalization designates the use of all available digital technologies, in order to improve business performance, and contribute to an overall rise in the standard of living. Today, the digital does not influence only the customer-bank relationship but the banking model as a whole.

Thanks to digitalization, **the customer experience** becomes richer… more interactivity, simplicity and development are the key principles that govern this relationship. Also, it will be necessary to identify the processes most affected by digitalization in order to subsequently provide the roadmap for transforming and modifying **banking processes**. Thus, we could see that the overall **organization of the bank** may change by using new technologies and communication utilities. Like any new principle… it is very likely that a **new regulation** will ensure and help understand and follow this digital transformation. As bellow, the main changes that affect the banking model through digitalization:

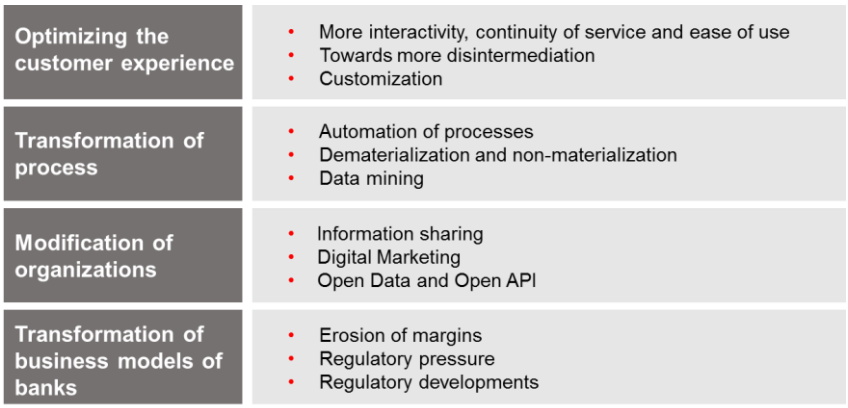


FIG. 2 – *The main changes of banking model due to digitalization*

2.3 Why do banks have to digitalize?

2.3.1 The arrival of new entrants

In the digital age, banks must struggle to preserve their value and relevance in a rapidly evolving industry. Indeed, many digital technological developments have opened the field to new players, who sometimes do not even come from the banking sector. The strong point of these new entrants is their power on cutting costs, improving the service quality and offering a new and attractive decor to financial landscape (The Economist, 2016).

Fintech, made up of start-ups in finance or other fields, is a serious competitor to banks and is gaining more and more market share. Average 95% of banks believe that their business is at risk due to the growing investment of fintechs that hit 8.4 billion USD in Q2 2017 (KPMG,2017). The most challenging point come from the fact that those start-ups encrust to a multiple financial activities: (Lee I., Jae Shin Y., 2017) identify in their work the six fintech business models that have indeed, a direct impact on financial activities: payment, wealth management, crowdfunding, lending, capital market, and insurance services.

2.3.2 New consumption modes and customer expectations

Banks face cost-cutting pressure and increased competition from new players as customer behaviors evolve. Indeed, customers, especially the Y and Z generations, expect banks to increase their online service offering (Vivekanandan L. and Jayasena V., 2011). Although the majority of people are rather traditional in their relationship with their banks, the agency channel takes the first position during the need for follow-up in the key stages, exe-

cution of complex operations, complaining or giving sharp advices (CIO, 2017). Thus, customers expect their banks to digitalize to be in line with their new way of life.

As reported by (Accenture, 2015), 20% of bank customers are digital-only user. To ensure their longevity, banks will have to succeed in their digital transformation, rethink and of course reposition their agencies because now “every bank customer is a digital customer”.

2.3.3 Digital banking agencies and advisors 4.0

Customers have adopted the concept of digital banking 4.0 but do not want a dehumanized bank. If each bank undertakes the digital transformation of its agencies differently, all have put the valuation and strengthening of the customer relationship at the center of their objectives. To bring together the right ingredients, the bank agencies must evolve towards a model of digital agencies, integrate solutions, digital software but also rely on consultants "4.0" (Cisco, 2014). Indeed, the banking agencies of the future have to introduce more digital media (such as tablets, interactive kiosks, automated self-service…), and complement them by the presence of advisers with roles more versatile and transverse than in the past.

3 The impact of digitalization on risk management

3.1 The historical evolution of risk management

3.1.1 Introduction

Risk management is considered relatively as a new function in financial sector. To understand its development, it is crucial to highlight some historical benchmarks. Since the early 1970s, the concept of financial risk management has evolved greatly…

International risk regulation also began in the 1990s, and financial firms developed internal risk management models to protect their business against unanticipated risks and to reduce regulatory capital (Friedman J. and Kraus W., 2011). It was also during these years that the governance of risk management became essential. The table 1 bellow, shows the major dates of the evolution of risk management (Dionne G., 2013):

TAB. 1 – *Historical dates for the development of risk management*

Dates	Events
1730	First future contracts on the price of rice in Japan.
1864	First futures contracts on agricultural products at the Chicago Board of Trade.
1900	Thesis of Louis Bachelier "Theory of Speculation"; Brownian movement.
1932	First issue of the Journal of Risk and Insurance.
1946	First issue of the Journal of Finance.
1952	Publication of the article "Portfolio Selection" by Markowitz.
1961-1966	Treynor, Sharpe, Lintner and Mossin develop the CAPM model.
1972	Future currency contracts by the Chicago Mercantile Exchange.
1973	Black and Scholes and Merton Option Valuation Formulas.
1974	Merton default risk model.
1980-1990	Exotic options, swaptions and equity derivatives.
1980-1982	First OTC contracts in the form of swaps: currency and interest rate swap.
1985	Creation of the Swap Dealers Association (OTC exchange standards).
1987	First risk management department in a bank (Merrill Lynch).
1988	Basel I
Late 1980s	Value at Risk (VaR) and the calculation of the optimal capital.
1992	Article by Heath, Jarrow and Morton on the term structure in Econometrica.
1992	Integrated Risk Management.
1992-1997	Risk Metrics & Credit Metrics (J.P. Morgan).
1994-1995	First bankruptcies associated with misuse (or speculation) of derivatives
1997-1998	Asian, Russian and LTCM default.
2001	Bankruptcy of Enron.
2002	New Sarbanes-Oxley and NYSE Governance Rules.
2004	Basel II
2007	Financial crisis.
2009	Solvency II
2010	Basel III
2013	Definition of new capital requirements

3.1.2 Towards Basel III

Due to the inadequacy of Basel I and the considerable evolution of banking activity, the Basel Committee has proposed a set of recommendations to measure credit risk in a more relevant way, taking into account the quality of the borrower. In addition, he introduced in his device market risk and operational calculations. The purpose of the scheme is to properly assess bank risks through prudential supervision and transparency. Basel II standards should replace standards set up by Basel I in 1988 and aim to set up the ratio McDonough (the new solvency ratio) to replace the Cooke ratio. The Basel II recommendations are based on three pillars: the minimum capital requirements, the process of prudential supervision and financial communication/market discipline. After the financial crisis of 2007, the Basel III reform was implemented to strengthen the financial system. The aim of this revision is to restore credibility in the calculation of RWA by (BIS, 2017):

- Improving standardized approaches for credit and operational risk.
- Forcing the use of internally modelled approaches
- Complementing the risk-weighted capital ratio with a robust capital floor

3.1.3 The different risks according to Basel's vision

- **Credit risk**: Risk of loss resulting from the inability of customers, issuers or other counterparties to meet their financial obligations. Credit risk may be aggravated by

concentration risk, resulting from a large exposure to a given risk or one or more counterparties, or to one or more groups of similar counterparties.

- **Market risk**: Risk of losses related to changes in the prices of financial products, volatility and correlations between these risks. These variations may relate in particular to interest rate fluctuations, as well as prices of securities and commodities, derivatives and other assets, such as real estate assets.
- **Operational risk**: Risk of loss or sanctions due to failures of internal procedures and systems, human errors or external events.
- **Liquidity risk**: refers to the lack of available liquidity to meet the receivables. To mitigate this risk, the Basel Committee incorporates into its regulatory framework the implementation of two liquidity ratios: a short-term liquidity ratio (or LCR for liquidity Coverage Requirement) and a long-term liquidity ratio (or NSFR for Net Stable Funding Ratio).

3.2 Digitalization as a tool management in industry

The work of (Schauppa E., Abele E. and Metternich J., 2017) proposed a method for using digitalization as a tool management in industry. The implementation process of digitalization is completed in 3 steps: The first is the definition of company's objectives, the most common one is: high tool availability, a low tool inventory and a high product quality. The second step consists on the definition of digitalization levers and their impact on the three company's targets mentioned before. The study of this impact shows three main levers: the development of employee's competencies, the integration of databases and the usage of track and trace technologies. Finally, the third step presents a readiness model including these levers to bridge the gap between the actual company's process and the target one.

3.3 Digital trends altering the current risk management model

Digitalization brings several changes that impact the banking model of risk management. Today, the data analytics, in the bank as in other fields, begins to have more importance; thanks to the use of new models and technics to extract, store and analyze data ... Also, the development and the change in customer habits drive any organization to evolve and revisit its whole management model.

Among the trends that can alter also banking model is the arrival of new regulatory standards and new companies that do not necessarily belong to the banking sphere. Finally, openness usually rhymes with risk ... the bank is today and more than ever exposed to cyber risks and attacks that can damage its system...

We can resume five trends that affect the banks' current business model: Importance of effective data analytics, evolving customer's behaviors, tough regulatory control, new entrants in the banking sphere and security risks in the digital transformation age.

Trends altering banking management	Impact on risk management (examples)
• Importance of effective data analytics	• Risk-based pricing, targeted segmentation with machine learning • Developing detection techniques to identify losses and risks exposures
• Evolving customers behaviors	• Peremptory customer request for online and mobile experience (mobile payments are expected to grow four times by 2020...)
• Tough regulatory control	• Proposing new regulations
• New entrants in the banking sphere	• Enabling banks to compete and/or collaborate with fintech companies on products and customer experience
• Security risks in the digital transformation age	• Changing security perimeters and cyber risks demand a holistic security approach for digital business

FIG. 3 – *Digital trends altering the current risk management model*

4 The way to digital transformation

The road to digital transformation is winding and requires a good preparation. To reach their goals, financial institutions have to achieve four principles components:

4.1 Data management

There is nothing more personal than his banking data… especially when this data evolves throughout the customer's life. Today's clients expect their agents to receive specific support that takes into account their financial, family and patrimonial situation; as a result, a personal response adapted to their own needs. To offer this quality of service, the sector has no choice but to equip itself with big data solutions enabling them to collect process and analyze all the data from all points of contact and best support their customers. Furthermore, (Cerchiello P., Giudici P., and Nicola G., 2016) proposed a framework that can estimate systemic risk using big data tools and proved also that such a model can predict the default probability of a bank.

To face these new challenges, banks will nevertheless have to rethink their IT architecture. The digitalization project then relies on the actor's ability to think of a system that does not complicate the customer journey but on the contrary simplifies it. Also, he has to think this system with a real Omni channel logic, a seamless system where all points of contact are shared in real time and accessible by all stakeholders.

4.2 Automating of processes

The emergence of new technologies such as process automation or RPA (Robotic Process Automation), cognitive computing and the Internet of Things (IoT) will play a key role in the digitalization of financial sector. Process automation technologies will accelerate market evolution and reduce costs for financial institutions. Those who succeed will be those who will adopt these robotic technologies to achieve their goals.

As the pressure for digital adoption intensifies, financial institutions must make technological advances available to their resources. RPA is at the heart of the human-machine interface and provides financial services with a virtual workforce governed by rules and connect-

ed to corporate systems such as users. With robotics, it is possible to automate and build an automation platform for front office, back office and support functions.

The benefits of robotics for financial services

Financial institutions operate in a highly regulated industry and face significant audit trail, security, data quality... Process automation allows the most innovative institutions to meet these requirements and achieve satisfactory operational efficiency (Accenture, 2017):

- cost savings: up to 80% cost reduction
- increase service quality: improving quality by reducing the risk of human error
- time savings: up to 80%-90% reduction in the execution time of tasks

Which processes can benefit from process automation technologies?

Multiple processes of financial services could be automated using robotics, these are some examples: Entering a new account on multiple systems, reporting on different systems, VAT declaration, Support and validation of the Audit, Loan Not Paying Notifications: sending emails to customers…

The impact of automation process on banking risks

Studies show that we can cut more than 15% of costs of different risks using automating (up to 60% for credit risk). A great majority sees benefits from increased precision, believe automation will improve compliance with regulation and expect an increase in customer and employee satisfaction.

4.3 External ecosystem

The work of (Kosmidou K., Kousenidis D., Ladas A. and Christos Negkakis C., 2017) notices the crucial role of a bank in the financial sector and it's clustering with other banks for predicting risks of contagion and protecting the financial sphere balance against crashes. To preserve this perfect harmony, we must identify bank business lines that market-leading digital (fintechs…) may snatch:

4.3.1 Impact on banking business lines

Leaders of financial institutions and infrastructure operators make critical strategic decisions about which areas of their organization they want to protect and grow, and which they want to reduce. More, the digital impacts various businesses of banks (seelings…) but requires that all employees evolve on certain fundamentals: communication, tools, risks…

a. Selling power

Account Managers (individuals and professionals) are significantly impacted

- Development of new forms of interactions with customers in all channels
- Strengthening of the advice posture towards the most connected (and therefore knowledgeable) clients and tutor with less technophile customers
- Strengthening the mobility of sales teams

Management Advisors have taken a lead and are more moderately impacted

The activities of receptionist and customer service could gradually disappear with the digitalization of their activity

Agency Directors are the most heavily impacted

- Managerial model switching from a control approach to a commercial "coach approach" using digital tools to animate the team
- Vigilance to develop on digital risks

The managers of business units are confronted to the same issues

b. Processing jobs

Digital reinforces the role of computer scientists and project managers

- Necessity to be continually aware of developments (infra and soft) due to the development of the innovation cycle
- Less competency internally and on the market and more specializing profiles
- New development methods (Agile, Open API, Open Data…)

The strengthening of treatment automation should increase operational efficiency in the back and middle office and could reduce the number of those resources

- Increased interactivity with internal customers to meet the requirement of instant
- Reinforcing the level of control of the management teams
- Conduct of change to be acquired at management team level

4.4 Skills, qualifications and risk culture

The digital transformation is now a reality whether in the retail banking sector or in the corporate finance and investment bank. The major developments in digital banking are undoubtedly the change in behavior and customer practices and a profound restructuring of the operating model of these banks. In contrast to the image of innovation reflected today by banks, Human Resources jobs in the banking sector are not always perceived as functions at the forefront of new technologies.

In the digital age where all HR functions are affected by digitalization (recruitment, skills management, training), HR needs to revisit their model and their own needs. Indeed, they have a key role to play in the digital transformation: the digitization-HR job is more united than ever.

Digitalization and adaptation of business lines

HR functions have undoubtedly benefited from digital innovations and largely transformed their business processes. The use of social networks in the recruitment process is commonplace. With digital, a new conception of the definition and evaluation of competences appears. It is necessary to imagine the jobs of tomorrow and the skills expected in connection with the digital strategy of the company.

Digitization is actually creating new business lines or even transforming existing businesses with the digital component. For example: In marketing, data scientists are appearing on the frontier between the exploitation of data, a domain previously reserved for CIOs, and marketing.

5 Dealing with the new dimensions of digital risks

5.1 Cybercrime: the priority number one of banks

For more than ten years, cybercriminals have been paying close attention to the banking and financial sectors. They target both institutions and their clients for (obviously and) eminently economic reasons. During these years, their techniques of attack or scam evolved: to a hardening of the security measures and to new norms of the sector, they responded by an increasing sophistication of their methods.

Threats to banking institutions and their clients are ubiquitous in time and geography. Better knowledge and constant monitoring of them certainly provides better protection. This is not only the fight against bank cybercrime, but also and above all the confidence of customers, guarantee growth in the short, medium and long terms.

5.2 Data security

No need to remember that data is a major issue for companies. At the beginning of the summer an IBM study showed that 73% of CEOs were convinced that data would play a major role in their business in the coming years, with an expected ROI of 15% on their initiatives in cognitive computing (ARMONK, N.Y., 2017). However, if everyone is passionate about what data can bring to business it seems that awareness about security / privacy is still not enough.

It all starts with a Cap Gemini Consulting study on the state of the art of data security in the banking sector. 83% of consumers trust banks in terms of data security, a score certainly flattering compared to other sectors of activity and certainly the image of rigor that he likes to project. A score all the more flattering that the reality seems less glowing.

On the other hand, only 21% of bank security managers have confidence in their ability to detect a cybersecurity gap and the study shows that only 29% of banks have both strong practices in terms of data privacy and strong security strategies.

6 Conclusion

When the web appeared in the late 90s, it was hard to imagine how much the internet would dramatically change the business ecosystem and consumer habits. In just a few years, banks - like other industries - have seen their business model strongly impacted. Being "jostled" has one essential virtue that of being forced out of one's comfort zone. Thus, the changes that were put in place (such as the electronic signature of documents) were unimaginable just a few years ago. The banks have made a lot of effort despite a difficult economic environment that has persisted since the fall of Lehman Brothers in 2007. They have managed to build new models, new offers to adapt to new expectations. But they have to go further. Banks must not live a simple evolution, but a revolution.

Every bank now has the obligation to position itself as a lifelong partner (family events, studies, projects, company creation, etc.). In this relationship, you have to think "Customer" and not "Product". The goal is to build customer loyalty by being the key interlocutor at each stage of its existence. The counselor then resumes his place, with an accented role of listening and advice. This risk is part of the banks' DNA. Prudent, they are so by nature. Thus, internally, some IT Departments are still cautious about the challenges related to digital, because these technologies are new and not all employees are well trained. However, it is

absolutely necessary for them to dare to open up to the digital world in order to be uberized by new actors.

Banks will need new ideas to help them transform their model, because the Fintech force the historical players of the banking and financial sector to quickly achieve their digital change. Their investments in these new generation start-ups are one of the keys that will allow them to continue their growth, these young shoots giving them the agility they need. Keeping banks in the running will also involve setting up new acquisition channels, improving customer relations, optimizing risk management through Big Data, or setting up new innovative services. Banks will also have to train their staff and guide their HR policy, in order to attract the talents that will help them in the construction of this new model. The bank of the future is on the move. Financial institutions will fully succeed in this challenge if they know how to draw lessons from the past, concretize the present and anticipate the future. As such, we will conceive in our future work a framework for measurement and holistic evaluation of the risks of digital transformation in the banking sector.

References

Accenture (2015). *Banking Customer 2020, Rising Expectations Point to the Everyday Bank* https://www.accenture.com/t20150710T130243__w__/us-en/_acnmedia/Accenture/Conversion-Assets/DotCom/Documents/Global/PDF/Dualpub_17/Accenture-Banking-Consumer-Pulse.pdf

Accenture (2017). *ROBOTIC PROCESS AUTOMATION, Future technology in financial services.* https://www.accenture.com/us-en/insight-financial-services-robotic-process-automation

ARMONK, N.Y. (2017) *IBM Report: Half of Surveyed Chief Executive Officers Plan to Adopt Cognitive Computing by 2019.* http://www-03.ibm.com/press/us/en/pressrelease/52753.wss

Bauernhansl T, ten Hompel M, Vogel-Heuser B (2014). *Industrie 4.0 in Produktion, Automatisierung und Logistik*. Springer Vieweg.

BIS: Bank for international settlements (2017). *Basel III: Finalising post-crisis reforms.* https://www.bis.org/bcbs/publ/d424.htm

Capgemini (2017). *The Currency of Trust: Why Banks & Insurers Must Make Customer Data Safer & More Secure.* https://www.capgemini.com/consulting/resources/data-privacy-and-cybersecurity-in-banking-and-insurance/

Cerchiello P., Giudici P., and Nicola G. (2016). *Twitter data models for bank risk contagion.* Neurocomputing, doi: 10.1016/j.neucom.2016.10.101 NEUCOM 18605

CIO (2017). *Decoding banks digital customers' expectations.* https://www.cio.com/article/3188478/leadership-management/decoding-banks-digital-customers-expectations.html

Cisco (2014). *Reimagining the Digital Bank. How U.S. Banks Can Transform Customer Interactions To Increase Profitability.* https://www.cisco.com/c/dam/en/us/solutions/collateral/executive-perspectives/Internet-of-Everything-executive-summary.pdf

Dionne G. (2013). *Gestion des risques: histoire, définition et critique*. CIRRELT.

Friedman J. and Kraus W. (2011) *Engineering the financial crisis.* University of Pennsylvania press, Philadelphia. 19104-4112

Gartner (2017). *Gartner Says 8.4 Billion Connected "Things" Will Be in Use in 2017, Up 31 Percent From 2016*. Egham, UK, https://www.gartner.com/newsroom/id/3598917

I-Scoop (2017). *Industry 4.0: the fourth industrial revolution – guide to Industrie 4.0.* https://www.i-scoop.eu/industry-4-0/

Kosmidou K., Kousenidis D., Ladas A. and Christos Negkakis C. (2017). *Determinants of Risk in the Banking Sector during the European Financial Crisis.* Journal of Financial Stability

KPMG (2017). *The Pulse of Fintech Q2 2017.* Global analysis of investment in fintech. https://assets.kpmg.com/content/dam/kpmg/xx/pdf/2017/07/pulse-of-fintech-q2-2017.pdf.

Lee I., Jae Shin Y. (2017). *Fintech: Ecosystem, business models, investment decisions, and challenges.* School of Computer Sciences, Western Illinois University, Macomb, IL 61455-1390, U.S.A and Hankyong National University, Anseong 17579, South Korea

PwC (2016, March). *Blurred lines: How FinTech is shaping financial services.* Global *Fintech* Report. https://www.pwc.de/de/newsletter/finanzdienstleistung/assets/insurance-inside-ausgabe-4-maerz-2016.pdf

PwC (2016). *Industry 4.0: Building the digital enterprise.*2016 Global Industry 4.0 Survey. *https*://www.pwc.com/gx/en/industries/industries-4.0/landing-page/industry-4.0-building-your-digital-enterprise-april-2016.pdf

Schauppa E., Abele E. and Metternich J. (2017). *Potentials of digitalization in tool management.* Institute of Production Management, Technology and Machine Tools (PTW), Otto-Berndt-Str. 2, 64287 Darmstadt, Germany

Stevenson A. (2010) Oxford Dictionary of English. 3rd ed. New York, London: OUP Oxford

The Economist, 415(8937), 13 (2015). *The FinTech revolution. A wave of startups is changing finance—for the better*. https://www.economist.com/news/leaders/21650546-wave-startups-changing-inancefor-better-fintech-revolution

Vivekanandan L. and Jayasena V. (2011). *Link between the Expectations of Retail Banking Customer and Electronic Banking Solutions*. Department of Computing, Informatics Institute of Technology Colombo 6, Sri Lanka and Dept of Computer Science & Engineering University of Moratuwa, Moratuwa, Sri Lanka

Wilkins C. (2016). *Les technologies financières et l'écosystème financier : évolution ou révolution?* Bank of Canada, Calgary (Alberta)

Résumé

Durant ces dernières années, la plupart des banques ont déjà intégré le numérique… le secteur bancaire a même été le pionnier des services en ligne. Néanmoins, plusieurs nouveaux entrants ont incrusté la place financière et ont bouleversé les banques traditionnelles. De plus, les consommateurs sont de plus en plus exigeants et leurs habitudes de consommation n'ont cessé d'évoluer. Tous ces éléments, réunis ensemble, rendent le rôle des banques, leurs produits et l'ensemble de leurs activités déstabilisés et controversés. Ce travail souligne les défis et soulève les questions fondamentales pour assurer une gestion maîtrisée des risques dans le cadre de la transformation digitale du secteur financier.

Algorithms and soft computing for credit scoring: State of the art

Yasser Zairi and Azedine Boulmakoul

LIM/Innovative Open Systems, FSTM, Hassan II University of Casablanca, B.P. 146 Mohammedia, Morocco,
zairiedu@gmail.com, *azedine.boulmakoul@gmail.com*

Résumé. This article intends to describe the most used techniques for Credit scoring as a way to assess the creditworthiness of clients. It focuses also on the pros and cons of each model so the reader can know when and why to use each one, and what would be the tradeoff in term of decision making.
Moreover, it illustrate the use of several models on the same dataset to show the performance of each one.

1 Introduction

Decision making has become a suitable ground for Data mining researchers and professionals to prove and to enhance the power of forecasting and prediction algorithms. Whether it be for trend identification, customers profiling, creditworthiness or healthcare applications.

One thing that all the prediction modeling community agreed upon is that there is nothing such as a good model for all applications. The choice depends on many parameters like field of application, purpose of the prediction, nature of the data and so on. Therefore, the challenge faced by data mining professionals is how to make sure that the model used is fairly adapted to the purpose of the application.

Many related works explores the development, application, and evaluation of predictive models in the credit business Crook, et al.(2007); Kumar & Ravi (2007). These models evaluate credit worthiness using a set of explanatory variables. Data employed comes from different sources; Data from financial ratios, balance sheets, or macro-economic KPIs for Corporate risk models, while data from customer demographics, application forms, and transactional data from the client history is used for retail models Thomas (2010). As each modeling type uses different types of variables, some challenges arise in Corporate as opposed to consumer credit scoring. Hence, various studies focus on one of both. The focus of this paper will be on the Latter.

2 Overview of credit scoring

At times where the economy is flourishing, taking credit in form of services or products from a company, or as a loan from banks is relatively easy. But when the economy is tightened, things become more complicated. Clients start to delay payments, which affects the working capital. And since loans from banks are much harder to have in such conditions, companies delay payment to suppliers as a way of financing their working capital.

When small and medium companies are caught in this vicious cycle, their liquidity start to shrink heavily and their risk of bankruptcy become serious.

Incorporating automated credit scoring in credit management operations is a best practice to ensure better quality decisions and more effective risk management.
Automated credit scoring shortens the time of credit approval, which is a key factor in customers' services, also it cuts down costs associated to application review. Moreover, it can be used to establish a policy of risk level acceptability and ensures objectivity.

3 Advanced vs Traditional statistical methods

The use of conventional statistical techniques, such as linear regressing and linear discriminant analysis, has not always been of great benefit in term of modeling complex functions. Advanced statistical techniques, like neural networks and genetic algorithms provide a good alternative to those traditional linear techniques.

Although multilayer feed forward are excellent classifiers Irwin et al. (1995), Palisade Corporation (2005), probabilistic neural nets could be trained faster with equal classification results or even better.

Neural networks are best suited for large data sets while genetic algorithms perform well with both small and large data sets Nath et al. (1997).

In the following, we will discuss some of the techniques used in credit scoring.

3.1 Linear regression

To describe the relationship between a dependent variable and one or more independent variables in any data analysis, linear regression methods are essential. It has been used in Two-class problem related to credit scoring applications.

In case where the customer makes proportionate repayments, a Poisson regression model could be used instead, and they could be re-expressed as Poisson counts.

Factors like customers' historical payments, guarantees and default rates can be analyzed with linear regression to score each factor and then to compare this with the bank's cut-off score.

Regression analysis has been used for commercial loans Orgler (1970), and for evaluating outstanding consumer loans Orgler (1971). Orgler concluded that the predictive ability of information not included on the application form is of great potential compared to information on the original application form.

The use of regression analysis extended such applications to include further aspects Hand and Jacka (1998).

3.2 Discriminant Analysis

Discriminant analysis is a technique used to classify data into two categories or more. It is still one of the most established methods to discriminate between good credit and bad ones, and still broadly applied in the credit scoring applications.

Fisher (1936) was the first to propose Discriminant analysis as classification technique. And its early use in credit scoring belongs to Durand (1941) for car loans applications.

Bankruptcy prediction is also a field in which Discriminant analysis have been applied by Altman (1968), who developed a Z-score based on a linear combination of five financial ratios.

Though this technique is been proven to be of a good use in credit scoring Desai et al.(1996); Hand and Henley (1997); Caouette et al.(1998); Hand et al.(1998); Sarlija et al.(2004); Abdou and Pointon (2009), it was criticized by many authors. Eisenbeis (1978) revealed several problems in his work (1977) that should be considered when using discriminant analysis.

These problems could not prevent discriminant analysis from being one of the most broadly used techniques in credit scoring Greene (1998); Abdou et al.(2008).

3.3 Logistic regression

The main difference between linear regression and logistic regression is that in the latter, the outcome variable is dichotomous (0/1). After this difference is taken into account, the application in both linear and logistic regression follows the same general principles Hosmer and Lemeshow (1989).

The extension of the simple Logistic regression model from one to more independent variables can be easily done, but it become harder to get multiple observations of all variables at all levels. Hence, the maximum likelihood method comes in handy when more than one independent variable are involved.

Theoretically, it might be supposed that logistic regression outperform linear regression given that the 'good' and 'bad' credit classes have been defined Hand and Henley(1997).

The field of credit scoring have known other methods such as nonparametric smoothing methods, expert systems, Markov chain models, neural networks, genetic algorithms Hand and Henley (1997).

3.4 Decision trees

Another techniques used in credit scoring modeling are Decision trees, also known as classification & regression trees (CARTs) or recursive partitioning Hand and Henley (1997).

Although the first use of a classification and regression tree model belongs probably to Breiman et al. (1984), Rosenberg and Gleit (1994) noted that the first initiation of a decision tree model was by Raiffa and Schlaifer (1961). And Later, David Sparks in 1972 developed a credit scoring model based on decisions trees.

It is a nonparametric method in which categorical and/or dependent variables are analysed as a combination of continuous explanatory variables.

The built of a dichotomous tree is done by splitting the records at each node depending on a function of a single input. All possible splits are considered to select the best one based on the lowest cost of misclassification or the overall error rate Zekic-Susac et al. (2004).
Other applications of CARTs in credit scoring were listed by Hand and Jacka (1998), Henley and Hand (1996), Paleologo et al. (2010)

3.5 Artificial Neural Networks

Neural networks are the imitation of human brain in problem-solving techniques. It has been defined as 'an artificial intelligence problem solving computer program that learns through a training process of trial and error' Gately (1996: 147).

This type of models belongs to an advanced category of credit scoring techniques compared to other statistical techniques such as discriminant analysis and regression models.

The basic element of a neural network; a Neuron. It takes the weighted sum of various characteristics chosen, the result is then compared to a threshold value on which the decision of granting a credit or not is made.

Neural Networks are assembled from many layers of neurons to model unknown data relationships. Thus is can recognize complex patterns between the input and the output data.

Their application has proven to be successful in many financial fields in general, and banking in particular. Applications like fraud, bankruptcy prediction, mortgage application and others. Gately (1996)

3.6 Fuzzy rule-based system

The ultimate model should imitate the human expert judgment in a way that it model how the analysis of application are done. This is not the way statistical methods model this problem; it is purely based upon crisp values of decision variables. Thus the rise of the need of methods that take into account the uncertainty, incompleteness and imprecision of information.

Usually, human experts use linguistic terms to express their appreciations.

Models based on fuzzy logic concept have been used to tackle the uncertainty of input data based on human judgment.

The computation of fuzzy logic is handled within three steps:

1. Fuzzification of inputs,

It is the transformation of input values, in a numerical form, into membership functions in the form of linguistic terms.

2. Fuzzy inference,

In this step, inputs are used to identify the rule from the Rule Base, and then it computes fuzzy linguistic variables in the output. The outputs of several rules are aggregated to produce a single output. Cherkassky (1998)

3. Defuzzification,

It is the process of transforming the output variables of the Fuzzy Inference step, presented as a fuzzy set, to a crisp numeric values. Bobyr et al. (2017)

4 Credit scoring process

The credit scoring field suffers from the lack of data due to its private nature, unless the modeling process is backed by institutions that can provide the data needed.

Thus, there is a growing need for more data sets to be made public and for more cooperation between academic researchers and financial institutions.

Moreover, it is very important to state out the definition of default as it is the very aim of the model. It will have a direct impact on what the model will predict.

4.1 Data Preprocessing:

This step is valuable in building credit scoring models. It aims to derive a sample that best represents the problem to be modeled. Therefore, the effectiveness of the model is highly correlated to the sample used.

The questions to be answered in this step are:

- Proportion of defaults. Data that contains less than 5% of defaults may be a serious modeling challenge.
- Frequency of variables in the data. Pie charts and scatterplots are typical tools to use.
- Outliers in the data. Boxplots allow the detection of outliers very easily.
- Missing data. Often, real clients' records are incomplete for many reasons, which can reduce the quantity of data available since incomplete records cannot be used in building a prediction model. Several methods are to be used to preserve as much information as possible when dealing with missing data. Florez-Lopez (2010)

4.2 Variable selection:

Also referred to as feature selection, is the problem of selecting a subset of features that best

describe the targeted concept. At this stage the business intuition and the human judgment are crucial to the success of the model.
Reducing the formality of the model using a formal method may reduce its complexity and increase its accuracy.
Assessing variable significance can be done in a more qualitative manner using the Gini coefficient, Pearson's chi-squared test or the information value criterion.

4.3 Model performance:

There are two properties to be satisfied when evaluating the performance of a credit scoring model:

- Goodness of fit; how well the model fits the existing data.
- The predictive power; how well the model can perform on new data.

If the problem is reduced to one response and one explanatory variable, measuring the goodness of fit would be trivial. But since we have to deal with several explanatory variables in such modeling problems, it is more appropriate to use other methods.
There are several techniques to predict the behavior of a model on a new unseen data such as ROC curve or the confusion matrix.

5 Applications:

If there is one difficult step that you could go through as a researcher it would be acquiring the data needed, especially when you are aiming at critical services such as financial ones.

The data gathered from clients of banks is not public, due to its nature and maintain of its privacy by banks, and so having it is somewhat near impossible unless you're working for the banks themselves. This is been said, the collaboration between the scientific community and the professional one may only harness the effectiveness and the innovation of banking services.

In this paper, we used a set of data introduced by Kaggle in a competition named "*Give Me Some Credit*" (2011) which aims for predicting the probability that a client will have a financial difficulties in the next two years using information like monthly income, age, debt ratio etc.

Clients that experienced financial difficulties in the training data will be referred to as "class 1", and "class 0" for those who are not. 1 and 0 being the values of the binary response variable.

5.1 Data

In the following, we describe each of the variables in the dataset and their type:

- Response: Serious delinquency in the next two years (SeriousDlqin2yrs); Person experienced 90 days past due delinquency or worse, 1 for Yes and 0 for No.
- F1: Revolving Utilization Of Unsecured Lines; Total balance on credit cards and personal lines of credit, except real estate and no installment debt like car loans, divided by the sum of credit limits; Percentage.
- F2: Age; Age of borrower in years; Integer.
- F3: Number Of Time 30 – 59 Days Past Due Not Worse; Number of times borrower has been 30-59 days past due but no worse in the last 2 years; Integer.

- F4: Debt Ratio; Monthly debt payments, alimony, living costs divided by monthly gross income; Percentage.
- F5: Monthly Income; Monthly income; Real.
- F6: Number Of Open Credit Lines And Loans; Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards); Integer.
- F7: Number Of Times 90 Days Late; Number of times borrower has been 90 days or more past due; Integer.
- F8: Number Real Estate Loans Or Lines; Number of mortgage and real estate loans including home equity lines of credit; Integer.
- F9: Number Of Time 60 – 89 Days Past Due Not Worse; Number of times borrower has been 60-89 days past due but no worse in the last 2 years; Integer.
- F10: Number Of Dependents; Number of dependents in family excluding themselves (spouse, children etc.); Integer.

5.2 Data distribution and Imputations

The map below shows the missing data that is indicated by a yellow line in each row.

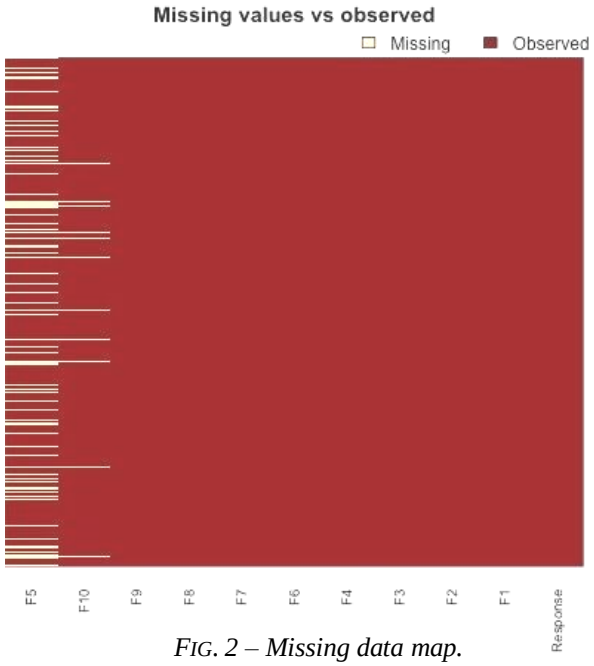


FIG. 2 – Missing data map.

It is no surprise to find missing values in any dataset. Dealing with them is a must before to go any further in the process.

Therefore, plotting missing values may be of good use. Hence The function missmap() in the Amelia package has been used to plot the data and distinguish missing values as shown in figure 2.

Here we have missing values in two variables F5 and F10. F5 has too many missing values presented as yellow horizontal lines, while F10 has rather few ones.

Now let's explore if there are any outliers or missing value and take the appropriate action.

- Age**:**

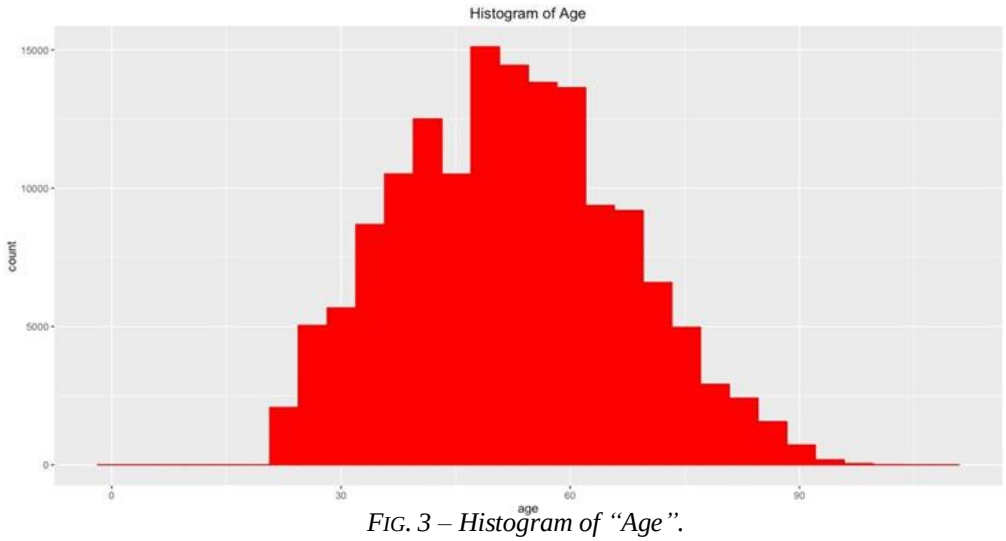


FIG. 3 – Histogram of "Age".

The distribution looks normal, with no missing values. This is a good indication that the sampling is from a normally distributed population.

- Revolving Utilization Of Unsecured Lines:

Revolving utilization, or "debt-to-limit ratio" measures the amount of the revolving credit limits that the client is using.

Below are the Minimum, 1st Quartile, Median, Mean, 3rd Quartile and the Maximum values of the variable.

Min: 0.00; 1st Qu.: 0.03 %; Median: 0.15, Mean: 6.05; 3rd Qu. 0.56; Max: 50710.

Though this ratio should be between 0 and 1, Table1 shows that there are some values greater than 1.

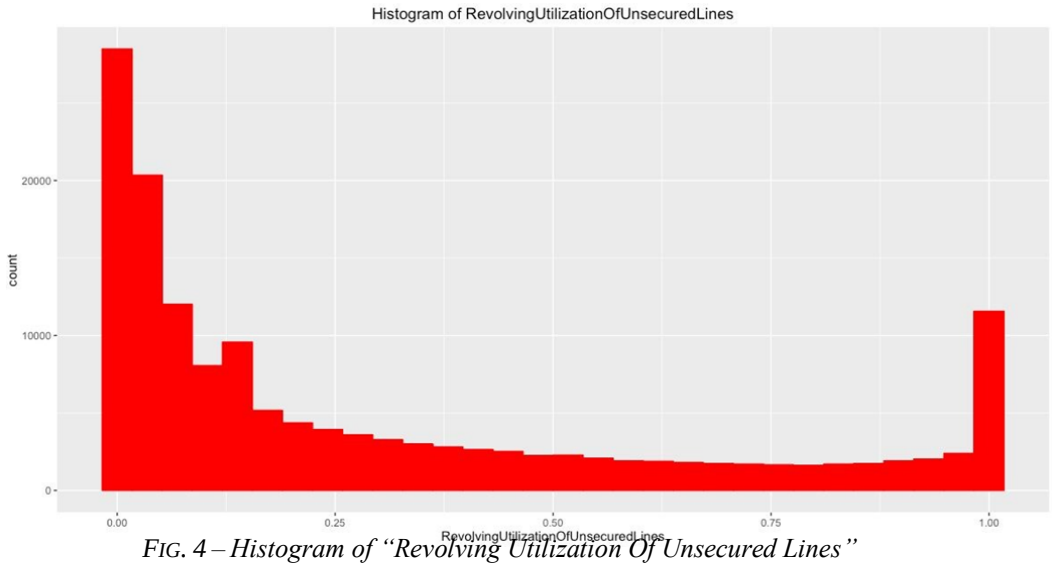


FIG. 4 – Histogram of "Revolving Utilization Of Unsecured Lines"

After imputing the Median to all entries higher than 1, the distribution looks reasonable as it can be seen in the Figure 4.

- Number Of Time 30 . 59 Days Past Due Not Worse:

The following table shows the number of records per each count, each count refers to how many times the client past the 30 – 59 days mark, and not more. It seems highly unlikely that anyone would have been 96 or 98 times past that mark, which means that the bank granted credit to the client 96 or 98 times, and in each time the client past the 30 – 59 mark. As we have enough data for the other counts, and for the sake of not dealing with missing data in a complex manner, we will be replacing them by 0.

Count	0	1	2	3	4	5	6	7	8	9	10	11	12	13	96	98
Number of records	126018	16033	4598	1754	747	342	140	54	25	12	4	1	2	1	5	264

TAB. 1 Number of records for each count of being 30 – 59 days past due.

The distribution function can be seen below.

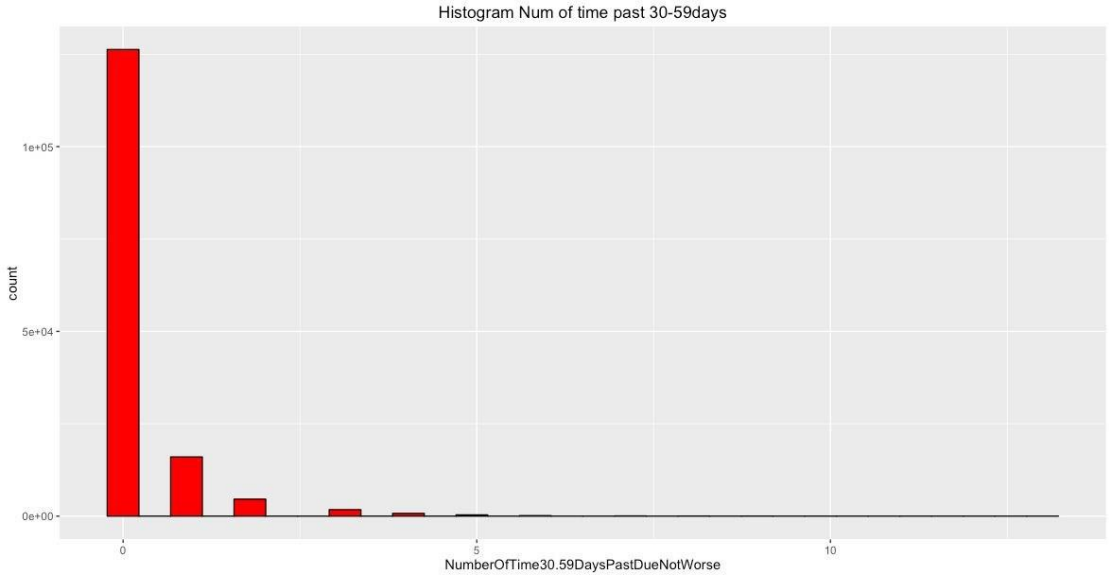


FIG.5. Histogram of "Number Of Time 30 - 59 Days Past Due Not Worse"

- Monthly Income:

The data related to this predictor has 29 723 missing values. Attributing the median to this value is sensible, for the mean is skewed by clients who are located in the higher side of the distribution.

- Number Of Open Credit Lines And Loans :

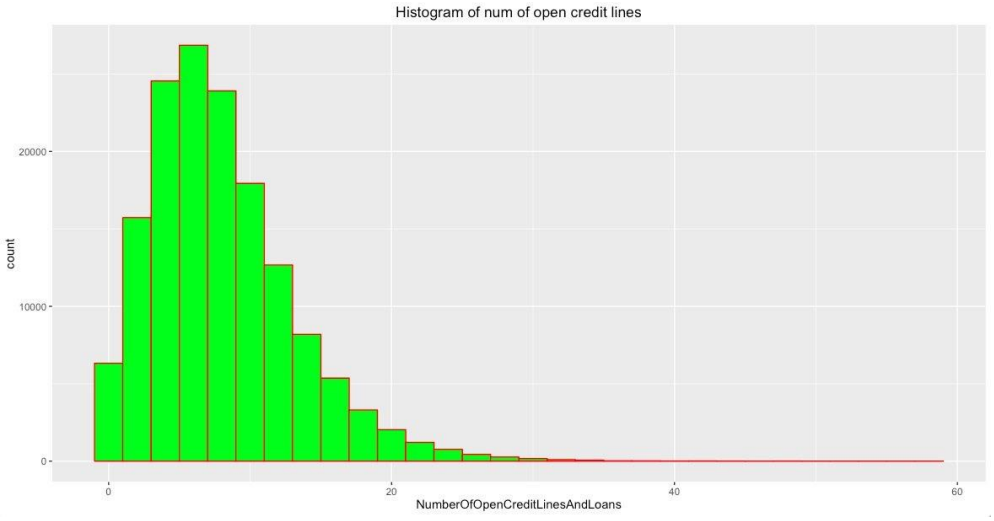


FIG.6. Histogram of "Number Of Open Credit Lines And Loans"

This predictor has no missing values, and its distribution seems reasonable.

- Number Of Times 90 Days Late:

This predictor has same issues as the "Number Of Time 30-59 Days Past Due Not Worse", so it will be treated in the same manner.

- Number Real Estate Loans Or Lines

All seem reasonable for this predictor except one record that has 54, we decided to drop it.

- Number Of Time 60 - 89 Days Past Due Not Worse

This predictor has same issues as the Number Of Time 30-59 Days Past Due Not Worse, so it will be treated in the same manner.

- Number Of Dependents

We found 3922 missing values, the obvious choice would be the attribution of 0.

- Debt Ratio

Some records presents a debt ratio greater than 100 000, which is kind of weird, we decided to drop these ones.

5.3 Data Imbalance:

The class "0" in the response variable represents 93.31% of the population, while the class "1" represents 6.685%. We could straighten up this imbalance by down-sampling the class "0" and so we will have new train and test data with almost equal proportions.

Figure 10 shows the distribution of each variable for the two classes of the response variable.

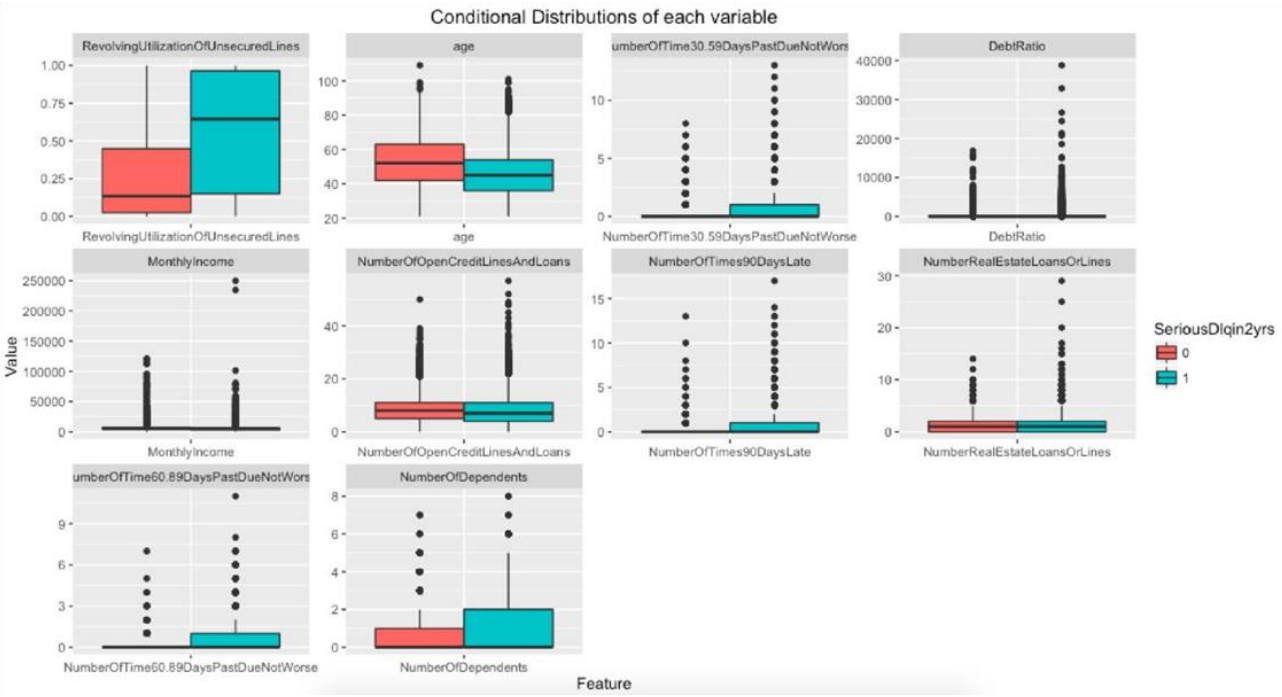


FIG.7. Features distribution across the classes "0" and "1".

This figure shows that there are two powerful predictors that are quite discriminatory in the prediction of the response variable, and their distributions are well separated across the values of the latter.

The two principal components explains 36% of the variance. Moreover, 48.42% of the variance is explained by the first three components. However, the Data still not that separable as it can be seen.

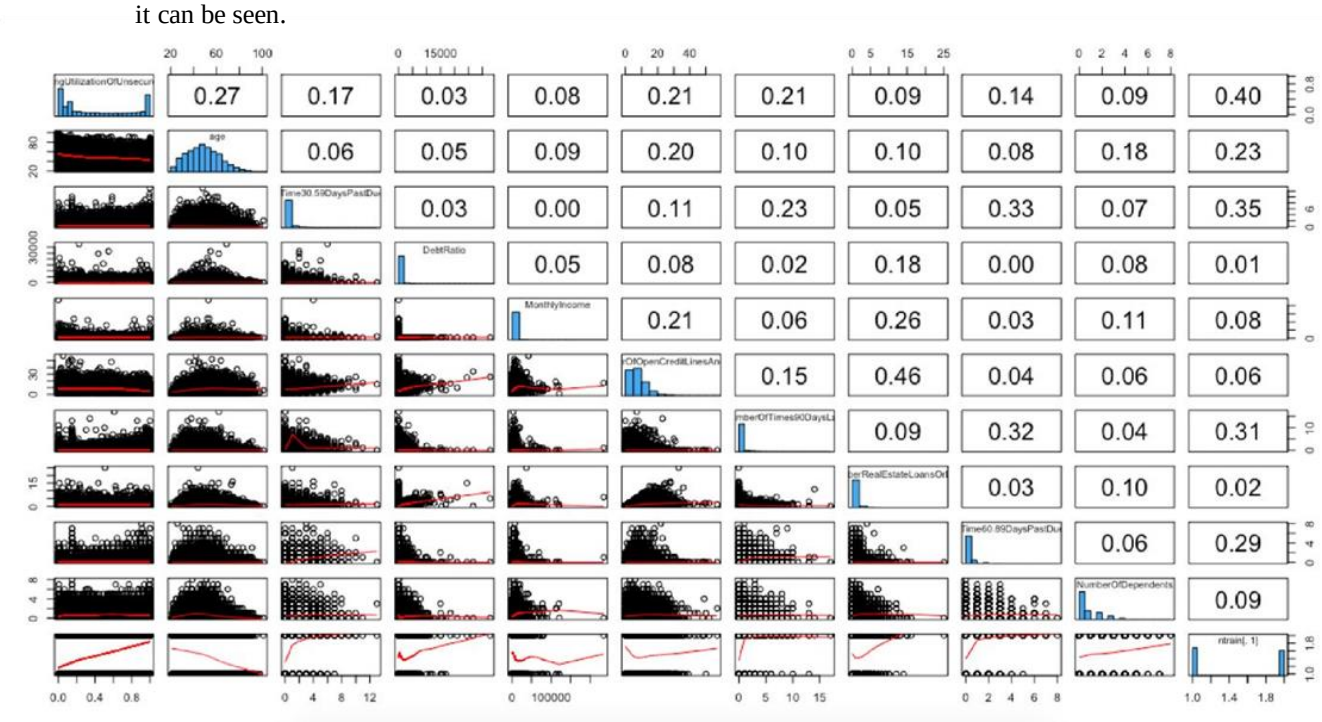


FIG.8. Scatterplot matrix;

Figure 13 shows that "Revolving Utilization Of Unsecured Lines", "Age", "Number Of Time 30-59 Days Past Due Not Worse", "Number Of Time 60-89 Days Past Due Not Worse".
are highly correlated to the response variable.

5.4 Prediction models:

- Logistic regression:

First, a main effects logistic regression model was fit. Figure 14 shows the effect plot of each predictor when others are maintained constant in order to figure out the effect of each predictor at the predictions.

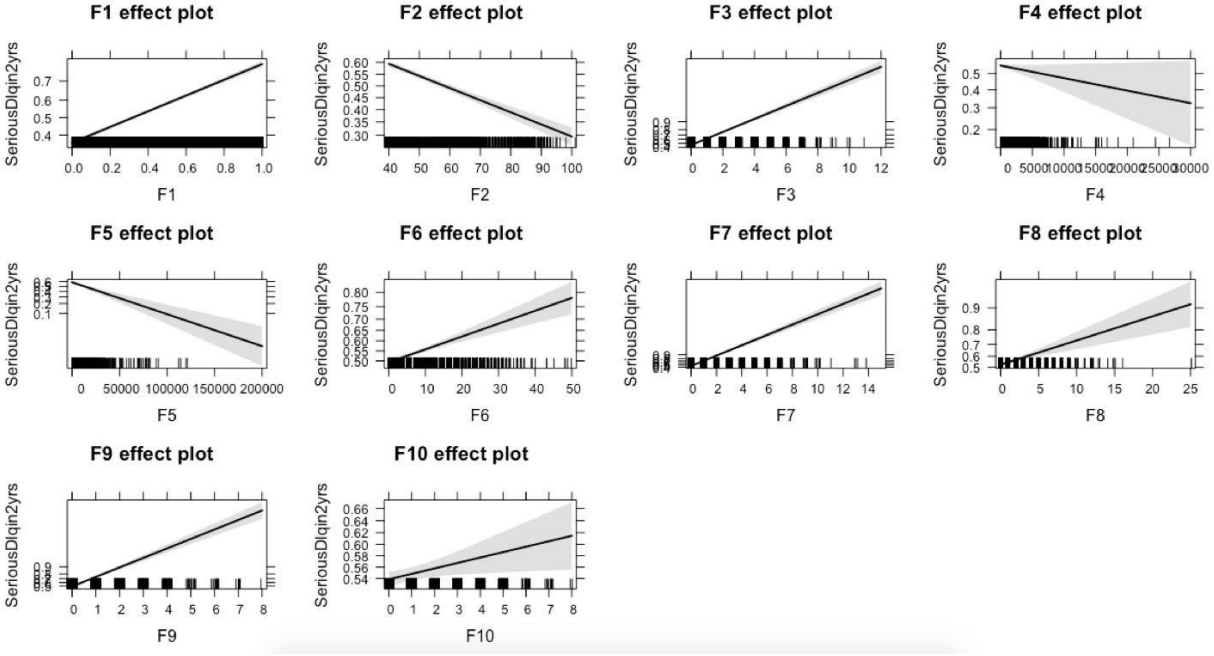


FIG.9. Effects of each predictor are shown in this fig. y-axis: Prediction, x-axis: Predictor.

- Random Forest:

This method require the tuning of the number of trees and the number of features to be considered at each node. With a trial-and-error method, the model converged to the following parameters: Trees: 5000, Predictors at each node: 2.

5.5 Evaluation:

Model	AUROC on test
Logistic Regression main effects	0.8458615
Logistic Regression step wise	0.8465
Random Forest	0.8526727

TAB.2 Accuracy of models using Area under Receiver Operating Characteristic

We see clearly that the Random forest model outperform logistic regression. Moreover, it has more flexibility in handling missing data than logistic regression.

References

Altman EI. 1968. *Financial ratios, discriminant analysis and the prediction of corporate bankruptcy*. The Journal of Finance XXIII(4): 589–609.

Abdou H, Pointon J. 2009. *Credit scoring and decision-making in Egyptian public sector banks*. International Journal of Managerial Finance 5(4): 391–406

Abdou H, Pointon J, El Masry A. 2008. *Neural nets versus conventional techniques in credit scoring in Egyptian banking*. Expert Systems with Applications 35(3): 1275–1292

Breiman L, Friedman JH, Olshen RA, Stone CJ. 1984. *Classification and Regression Trees*. Wadsworth: Belmont,CA

Bobyr et al. (2017), *A method of defuzzification based on the approach of areas' ratio*. Applied Soft Computing. Volume 59 Issue C, October 2017 .Pages 19-32

Caouette JB., Altman EI., Narayanan P. 1998. *Managing Credit Risk: The Next Great Financial Challenge*. JohnWiley and Sons Inc.: New York.

Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). *Recent developments in consumer credit risk assessment*. European Journal of Operational Research, 183, 1447-1465

Durand D. 1941. *Risk Elements in Consumer Instalment Financing, Studies in Consumer Instalment Financing*. National Bureau of Economic Research: New York.

Desai VS, Crook JN, Overstreet GA. 1996. *A comparison of neural networks and linear scoring models in the credit union environment*. European Journal of Operational Research 95(1): 24–37

Eisenbeis RA. 1978. *Problems in applying discriminant analysis in credit scoring models*. Journal of Banking and Finance 2(3): 205–219

Eisenbeis RA. 1977. *Pitfalls in the application of discriminant analysis in business, finance, and economics*. Journal of Banking and Finance 32(3): 875–900.

Fisher RA. 1936. *The use of multiple measurements in taxonomic problems*. Annals of Eugenics 7(2): 179–188.

Greene W. 1998. *Sample selection in credit scoring models*. Japan and the World Economy 10(3): 299–316.

Gately E. 1996. *Neural Networks for Financial Forecasting: Top Techniques for Designing and Applying the Latest Trading Systems*. John Wiley and Sons, Inc.: New York.

Hand DJ, Jacka SD. 1998. *Statistics in Finance*. Arnold: London

Hand DJ, Henley WE. 1997. *Statistical classification methods in consumer credit scoring: a review*. Journal of theRoyal Statistical Society: Series A (Statistics in Society) 160(3): 523–541

Hosmer DW, Lemeshow S. 1989. *Applied Logistic Regression*. John Wiley and Sons, Inc.: New York.

Henley WE, Hand DJ. 1996. *A k-nearest-neighbour classifier for assessing consumer credit risk*. The Statistician 45(1): 77–95

'Give Me Some Credit' Competition. 2011. https://www.kaggle.com/c/GiveMeSomeCredit

Irwin GW, Warwick K, Hunt KJ. 1995. *Neural Networks Applications in Control. The Institution of Electronic Engineers*: London.

Kumar, P. R., & Ravi, V. (2007). *Bankruptcy prediction in banks and firms via statistical and intelligent techniques - A review*. European Journal of Operational Research, 180, 1-28.

Nath R, Rajagopalan B, Ryker R. 1997. *Determining the saliency of input variables in neural network classifiers*. Computers and Operations Researches 24(8): 767–773

Orgler YE.1970. *A credit scoring model for commercial loans*. Journal of Money, Credit and Banking II(4): 435–445.

Orgler YE. 1971. *Evaluation of bank consumer loans with credit scoring models*. Journal of Bank Research 2(1):31–37.

Palisade Corporation. 2005. *Neural tools: neural networks add-in for Microsoft Excel*. Version 1.0. Palisade Corporation, New York

Paleologo G, Elisseeff A, Antonini G. 2010 *Subagging for credit scoring models*. European Journal of Operational Research 201(2): 490–499.

Raiffa H, Schlaifer R. 1961. *Applied Statistical Decision Theory*. Harvard University Press: Boston, MA.

Sarlija N, Bensic M, Bohacek Z. 2004. *Multinomial model in consumer credit scoring*. In 10th International Conference on Operational Research, Trogir, Croatia

Seafim Opricovic. (2007) *A fuzzy compromise solution for multi-criteria problems*. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 15:03, 363-380.

Thomas, L. C. (2010). *Consumer finance: Challenges for operational research*. Journal of the Operational Research Society, 61, 41-52.

V. Cherkassky, *Fuzzy Inference Systems: A Critical Review, Computational Intelligence: Soft Computing and Fuzzy-Neuro Integration with Applications*, Kayak O, Zadeh LA et al (Eds.), Springer, 1998

Zekic-Susac M, Sarlija N, Bensic M. 2004. *Small business credit scoring: a comparison of logistic regression, neural networks, and decision tree models*. In 26th International Conference on Information Technology Interfaces, Croatia

Summary

We present in this paper a picture of the state of the art in credit scoring models. We point out that efforts have to be joined between academics and professionals to improve existing techniques and provide a better way to predict financial distress.

It goes without saying that regardless of the technique used to build the model, a dataset that is well cleaned improve the predictive power of the model used afterwards, However any mistakes that left unhandled may heavily shrink the chance of having a good model.

The concept of generalisation and the issues of over-fitting and under-fitting have to be taken into consideration when dealing with credit scoring models, as a model may perform well in the data used to build it, but may lose its accuracy when confronted with new datasets.

At last, models that provide more understanding of the decision making process are to be improved and enhanced. With that in mind, a look into Generalized Intuitionistic fuzzy sets will be taken in future works.

Optimization Bigdata to Support Decision Making in Human Resources Management: A survey

Loubna Rabhi*, Noureddine Falih**
Lekbir Afraites***, Belaid Bouikhalene****

* Mathematics and Applications Laboratory
Faculty of Sciences and Technology, Sultan Moulay Slimane University
Rabhi.lubna@gmail.com
http://www.fstbm.ac.ma/
**Laboratory of Innovation, Applied Mathematics and Information Technologies
Polydisciplinary Faculty - Sultan Moulay Slimane University
nourfald@yahoo.fr
http://fp.usms.ac.ma
*** Mathematics and Applications Laboratory
Faculty of Sciences and Technology, Sultan Moulay Slimane University
lekbir.afraites@gmail.com
http://www.fstbm.ac.ma/
****Laboratory of Innovation, Applied Mathematics and Information Technologies
Polydisciplinary Faculty - Sultan Moulay Slimane University
b.bouikhalene@usms.ma
http://fp.usms.ac.ma

Abstract. Big Data in HR refer to the huge amounts of employees, customers, and transactional data in organizations. These data can come from social sites such as LinkedIn and Viadeo…that store the details of people who are signed up to the site. Analyzing these Big data by specific tools can help companies make decisions that affect recruiting and creating bigger benefits in order to increase productivity. In this paper, we will try to answer two fundamental fields: The contributions of Big Data in Human Resources Management and How company can process to HR Analytics.

Keywords: Human Resources (HR), Big Data, HR Management, HR Analytics, Social Big Data Analytics.

1 Introduction

The explosion of objects connected to the Internet (Internet of thing IOT) and data, processes, people connected to each other (Internet of everything IOE) (Ahmed and al., 2017), the world nowadays knows a huge volume of data exchanged called "big data".

According to (Arora Y, 2016), Big Data are defined by a huge amount of data coming from heterogeneous sources at a very high speed, which is not possible for the existing tools and techniques to analyze and extract value from it. That is means that traditional tools and

techniques cannot be able to analyze this large amount of data. So, advanced techniques of analyze must be required.

These sophisticated analytics cited above are called "Big Data Analytics". Its aim to extract pertaining information from this data in order to uncover valuable insights from the data, minimize risks, and improve decision making within the company.

In Human Resources, Big Data refer to data gathered about employees including skills, performance ratings, age, tenure, safety record, sales performance, educational background, manager, prior roles, behavior etc.

Many HR departments still try to learn what can they do with this massive amount of data and how they can process and interpret it. So, learning how to manage Big Data seem challenging for any HR department that wants to develop and advance. Big Data can actually make a big difference if used wisely: they can help HR department to find skilled employees as they can affect on existing staff by learning how to increase its productivity and how to adjust its work processes…

Big Data in HR help to evaluate and improve practices including talent acquisition, development, retention, and overall organizational performance. This involves integrating and analyzing internal metrics, external benchmarks, and social media data to deliver a more informed solution to the business problem facing company (Charles Henri Besseyre des Horts, 2014).

This paper tries to cover major aspects of big data Analytics and its application to Human Resources Management. It begins with a brief introduction of the topic. Important Aspects in Big Data is discussed in II. The issues of Big Data Analytics are discussed in III. Big Data Analytics at the service of Human Resources Management is treated in IV, while the Conclusion and future trends are drawn in V.

2 Big Data Definitions

Today "Big Data" draws a lot of attention in the IT world. They have gained great importance in the last years and are increasingly used in several contexts (Sagiroglu & Sinanc, 2013): Healthcare, Public sector administration, Manufacturing, Banking (Srivastava & Gopalkrishnan, 2015).... We mean by "Big Data", the huge volume of data collected from various sources like sensors, smartphones, social media and many others digital sources. These data can have many different types as videos, audio, images, text and so on. All of these data are generated in real time.

According to (J. Campos et al. 2017), big data are defined by its Volume, Velocity and Variety (3Vs definition). That it means that, data size is large, the data will be created rapidly and the data will be collected in multiple types and captured from different sources, respectively:

1. **Volume**: Denotes the large amount of data which is generating in every second from different sources.
2. **Velocity**: Means the speed at which data must be collected, analyzed and exploited. Various treatments based on data rate are:
   (a) Batch: It means running several queries in a sequential way without any intervention of human;
   (b) Real Time: It means delivering immediately the information after its collection. There is no delay to provide information;
   (c) Interactive: It means executing the tasks which require frequent user interaction;

(d) Streaming: It means the method of processing the data as it comes in. The insight into the data is required as it arrives;

3. **Variety**: Means that the incoming data can have different types. Data are classified into four categories as:
   (a) Structured Data: it concerns all data which we can store in table with rows and columns, e.g.: Relational database management systems.
   (b) Semi structured data: it not arranged into tables but it can be converted into structured data, e.g.: web server logs data, XML documents data …etc.
   (c) Unstructured data: This type of data is very difficult to store into database. It has no standard structure. For example: videos, audios, images, documents, emails and so on.
   (d) Multi Structured Data: Data which is a mix of Structured, semi structured and unstructured data. Example operating system logs…

In addition, studies of (Lakshen and al., 2016) added value and veracity to build the 5Vs definition of big data.

4. **Value**: It measures the usefulness of data from this huge amount of Big Data.
5. **Veracity**: Measures the correctness and accuracy of the data.

According to (Owais & Hussein, 2016), four others characteristics of Big data are added : Variability and Visualization. So, we can talk about 9V's instead of 5V's:

6. **Variability**: It refers to data whose meaning is constantly changing.
7. **Validity**: Means the data is correct and accurate for the intended use. Valid data is the key for making the right decisions.
8. **Volatility**: Means how long does company need to store data. In this world of real-time data, company needs to determinate at what point the data are no longer relevant to the current analysis.
9. **Visualization**: Once it's been processed, data must be presented in readable and accessible manner for better decision making.

To sum up, Big data are massive and rapidly-expanding, but it's also noisy, messy, constantly-changing, in hundreds of formats and virtually worthless without analysis and visualization.

3. Big Data Analytics

3.1 Big data lifecycle

To better understand the "Big Data Analytics", let's start with the description of the big data system lifecycle. This latter contains a suite of steps as shown [Fig.1]: data generation, data acquisition, data storage, data analytics and data visualization.

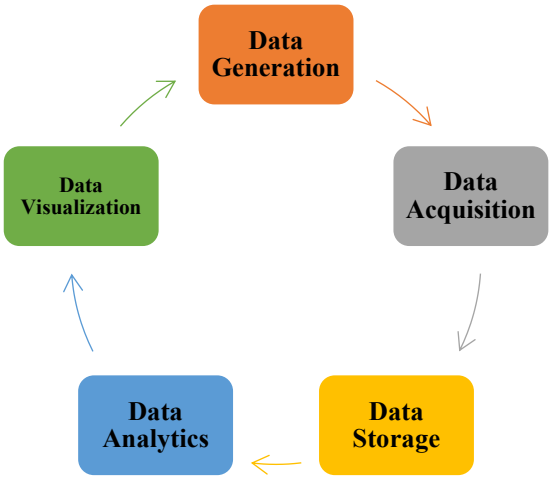


FIG.1: *Big Data lifecycle.*

1. **Data generation**: the question that arises is how data can be generated. Data can be gathered from various data sources (sensors, video, click streams and several other digital sources…).
2. **Data acquisition**: concerns the process of obtaining information from data. It contains consecutive steps:
   (a) Gathering required Data: After identification of problem, data needs to be gathered from a rich and varied data environment (Jadon and al., 2016);
   (b) Selection data: Pertinent data will be selected from the gathered data which will be useful for the analysis
   (c) Pre-processing the data: translate data in to fixed format before providing data to algorithms or tools. it aims at detecting, cleaning, and filtering the unnecessary, inconsistent, and incomplete data to make them the useful data.
3. **Data storage**: concerns persistently storing.
4. **Data analytics**: Analytics refers to the process of deriving actionable insights from data in order to help making decisions using qualitative and quantitative techniques (Tsai and al., 2015). Analytics can be performed using various algorithmic concepts such as regression, classification and so on.
5. **Data visualization**: used for displaying the output of data analytics. It is an interactive way to represent data insights.

3.2 Data Analytics process

According to what is discussed above, we can trace a process to build a complete data analytics system: it is necessary to gather data first and then find information from the data and display the knowledge to the user going through these steps known as Data analytics process [Fig. 2].

Transformation of data: Transform preprocessed data into data-mining-capable format using various methods such as: dimensional reduction, sampling…and so on.

Performing analysis over data: After transformation data, analysis can be performed using various statistical methods and data mining algorithms such as regression, classification, clustering … (Dave & Gianey, 2017)

Evaluation: Measure the results of data analysis;

Interpretation: Doing applicable decisions and displaying the output of data analysis by an interactive way to represent data insights/Knowledge.

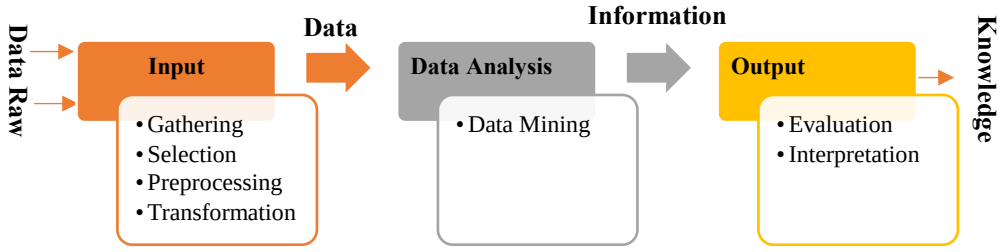


FIG. 2: *Data analytics process.*

3.3 Big Data Analytics Process

Nowadays, the data that need to be analyzed are big, contained heterogeneous data types, and even including streaming data which may change the statistical and data analysis approaches. Therefore, several new issues for data analytics come up, such as privacy, security, storage, fault tolerance, and quality of data (Lv and al., 2017). So, it's time to discuss new approach of big data called Big Data Analytics instead of Data Analytics.

A term "Big data analytics", is a set of advanced technologies designed to work with large volumes of heterogeneous data in order to further improve the traditional Data analytics process. We can cite three types of analytics techniques as shown [Fig.3]:

Descriptive analytics: Largely based on historic data. In this technique new insights are developed using probability analysis, trending, and development of association over data that is classified and categorized.

Predictive analytics: Used to predict the future outcomes based on historical and current data. It provides information on what will happen, what could happen, and what actions can be taken.

According to (Gandomi & Haider, 2015), Predictive analytics techniques are primarily based on statistical methods. So, it needs to develop new statistical methods for big data due to several factors. First, the notion of statistical significance is not relevant to big data. Indeed, conventional statistical methods are applied for a small sample and then generalized to the entire population. In contrast, big data samples are massive and represent the majority of population if not the entire. Second, many conventional methods for small samples do not scale up to big data. Third, the big data is known by its distinctive characteristics heterogeneity, noise accumulation, spurious correlations, and incidental endogeneity...

Prescriptive analytics: Helped to derive a best possible outcome by analyzing the possible outcomes. It is flexible and has the ability to improve with experience.

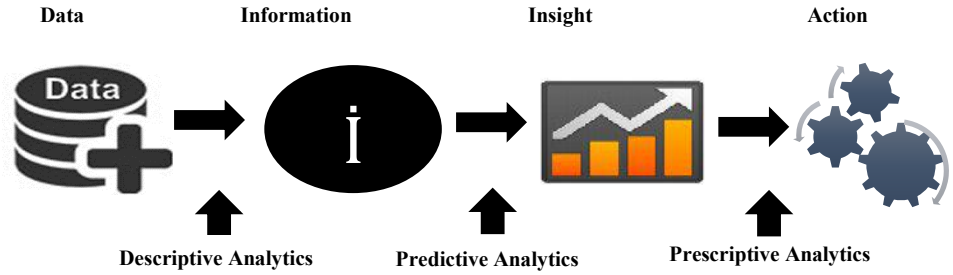


FIG. 3: *Types of Analytics Techniques.*

In (Sivarajah and al., 2017)], the authors are completed the classification of types for analytical techniques by Pre-emptive analytics and Inquisitive analytics [Fig.4]:

Pre-emptive analytics: Is about having the capacity to take precautionary actions on events that may undesirably influence the organizational performance, for example, identifying the possible risks and recommending mitigating strategies far ahead in time.

Inquisitive analytics: Is about probing data either to certify or reject business propositions.

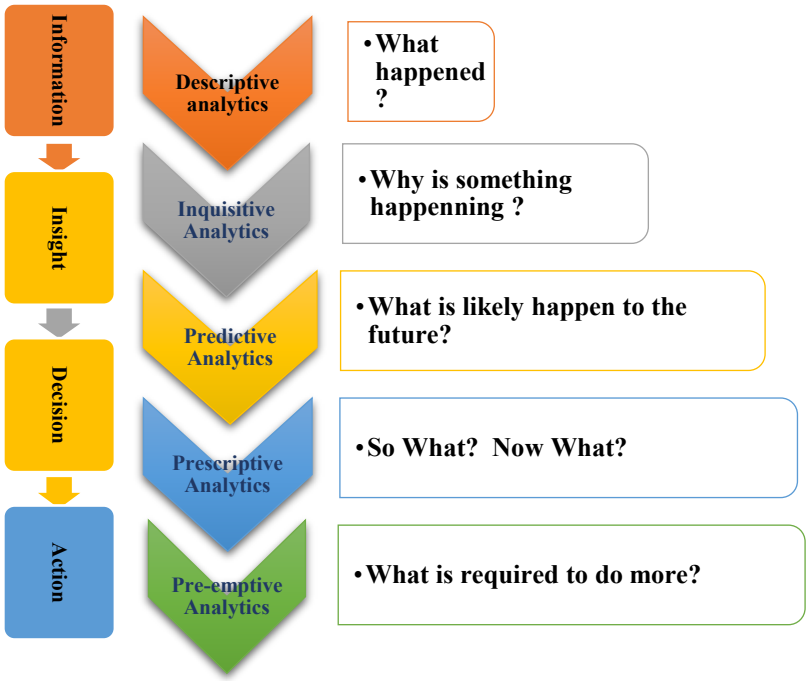


***FIG. 4**: Update of the Types of analytical techniques.*

-Using advanced analytics techniques like prescriptive analytics we can get answers for questions like why did it happen, when will it happen again, what caused it to happen, what can be done to avoid it?

-Traditional analytics is batch oriented, hence, before obtaining the required insight, it's necessary to wait for ETL jobs to complete. But In big data analytics, the data change is highly dynamic and requires to be integrated quickly for analysis using the support of software meant for it.

4. Big Data matters to HR Management

4.1 The Contribution of Big Data in HRM

4.1.1 Recruitment Process

Traditionally, recruitment process usually follows such steps: First, the organization begins by describing the position to be filled and the definition of the desired profile (Training, skills, experiences ...). Secondly, the recruitment message will be launched in the press, internet and others. Third, interested ones submit their resume. Fourth, HR manager would analyze these applicants' resume and select appropriate ones. Fifth, invitations will be sent to candidates for interviews and assessment tests in order to finally choose the best candidate. But the reality shows that the results of interviews are often biased. Because mot of the time, the interviewer can't have correct information about candidates that's leads to false results.

To contribute to organization's success, HR team must hire more relevant people for the job. Therefore, it needs data-driven knowledge to pick the right talent from the pool of candidates (Couaillier & Projet, 2016). Combining big data from social media, such as LinkedIn, and recruitment process can help recruiter search for potential talents and everything they would possibly want to know about them is on their profile (personal picture, living conditions, social relationships…etc.). HR manager can match between the candidate's skills and personal beliefs and the company's needs. Hence, company can avoid invest in bad hires.

4.1.2 Training Process

As it's known, talent training can lead to increase employees' level of knowledge and skill, it can also enhance their work performance. By traditional talent training organized by the company, professional trainers can be hired to ensure training which is spend a lot of material, human and financial resources. Usually such training takes traditional form of classroom instruction which can not meet all the needs of employees.

Using Big data context, any employee can easily search and access to the information that he need to know on Internet at any time and anywhere. He can also choose its favorite form of teaching either videos or courses… (Zang & Ye, 2015)

4.1.3 Employees Career Management Process

Employees career choice and planning are closely linked with data. By analyzing all of the gathered information of employees such as: interest on job, professional experience,

performance …, HR could find new ways to motivate employees and make them more engaged. Companies can combine traditional career management and career management of Big Data to make planning of new more effective talent-retention programs and to avoid employee turnover.

To sum up, Big Data can improve many aspects of HR department including recruiting, training and employees career management processes...and so one. Big data can help human resource to speed up the hiring process, improve productivity, understand employee turnover, manage talent career...

To make the most of big data, company needs to create a strong and reliable team of professionals who can leverage Big Data efficiently and use the results for future planning. It must therefore call on new skills, such as data analysts

4.2 HR Analytics

HR Analytics is a powerful tool of making decision that enables HR professionals to model the impacts of HR policies on business performance (Chavanne, 2015). The use of HR Analytics can then constitute a significant competitive advantage by creating synergies between different HR processes such as recruitment, remuneration, mobility, career management, etc. Therefore, the implementation of HR Analytics represents a business project that requires a significant investment in terms of costs and time.

The maturity level of an organization depends on its ability to implement solutions to different types of problems. The Maturity Model of a talent analytics system (Philippe Burger and al., 2016) is shown below (Fig. 5):

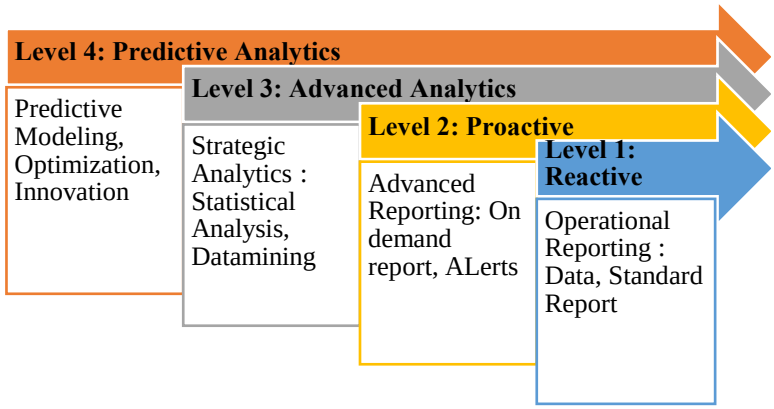


FIG. 5: *The Maturity Model of HR Analytics*.

Step 1: Operational reporting: there are no reporting tools for integrating and sharing data. The data are unreliable so it is difficult to make relevant decisions based on these data.

Step 2: Advanced reporting: It is a question of moving from a reactive posture to a proactive one that can guide HR policy decisions and model their impacts. This passage often proves to be complex for organizations that find it difficult to use a large amount of data to extract information and analyze it. However, few HR professionals know how to deal with this type of problem. They must therefore call on new skills, such as data analysts.

Companies consolidate their data between different departments. In order to make them more reliable, they set up indicators to ensure the quality of data during reporting. They also produce multidimensional dashboards that allow to visualize the current situation of the company in a panoramic way.

Step 3: At this third level, the data and indicators are therefore completely reliable and display a consolidated history of several years. Organizations then begin to use sophisticated and predictive statistical analysis on their data to extract information about their resources and guide strategic decision-making in terms of human capital (e.g. Reduction of turnover, Organization of recruitment, Staff motivation…).

Step 4: Finally, the final step and the ultimate goal of HR Analytics is predictive Analytics where Human Resources can be able to identify causal links and design them, no longer in a linear, but multidimensional way with multivariate analyzes. For example, the impact of the place of residence on absenteeism or overwork".

To sum up, HR Analytics is considered as a tool for decision-making and a prediction that can guide and anticipate strategic decision-making(Whitepaper, 2013).

According to what we have seen above, analyzing Big Data helps to get better insights, understand processes within the company, make important business decisions according to patterns found in research and analysis, and plan strategic business moves, So, Based on Big Data Analytics, (Soumyasanto Sen, 2017) illustrate the Maturity Model for HR Analytics as showing [FIG. 6]:

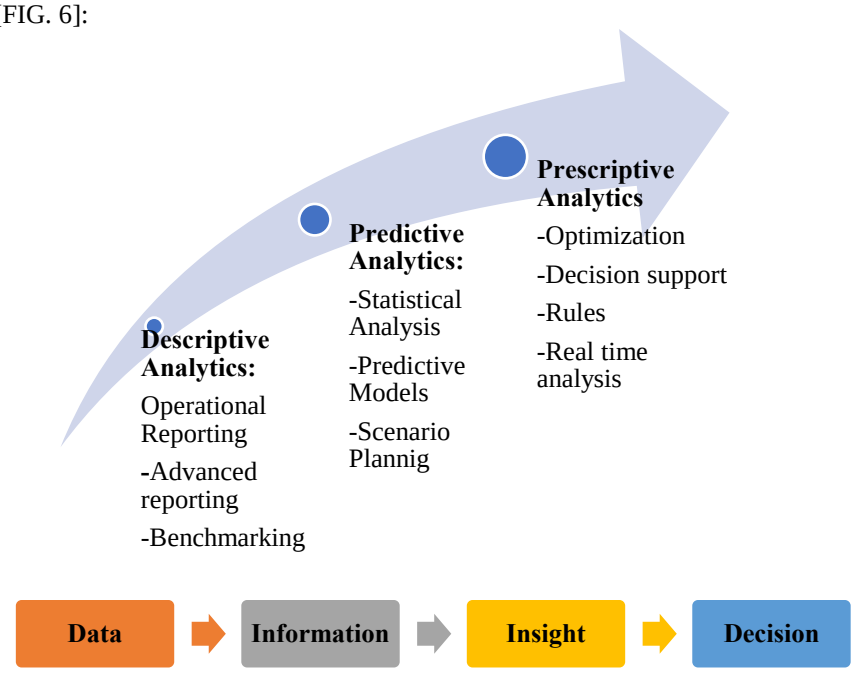


FIG. 6: *Update of Maturity Model Analytics.*

-**Descriptive Analytics**: uses operational reporting that focus on data exploration, data accuracy, and metrics analysis and uses Advanced reporting for benchmarking, decision making and to generate dashboards.

-**Predictive Analytics**: uses statistical analysis, forecasting, correlations, and development of the predictive models. It helps in making predictions and for taking smarter decisions like in talent management for the organizations such as employee benefits, their behavior, causes of delay in hiring etc. Moreover, analytics can also to be used for forecasting HR related metrics. For example, it can help to predict which employees will reach their target goals and why or what would be their possibility of leaving the company...

-**Prescriptive Analytics:** assists in this with optimization, strategic foresight, and real-time analysis. Prescriptive analytics not only anticipate what will be happen and when it will happen. But also tell why it will happen.

4.3 HR Analytics tools

We present below a set of the most used existing tools for HR Analytics:

RStudio: It is an open source HR Analytics software. It's great for statistics and visualization and it has a friendlier user interface. The interface contains a code editor of language R that has a very extensive library with R packages. RStudio can enables to work with much larger datasets(Lyndon Sundmark, 2016).

Python: Python is another programming language which can be used for HR Analytics. It has a faster learning curve, and it can be used interchangeably for R. Python can easily be integrated with other languages.

Comparing R and Python languages in (DeZyre, 2016), we can conclude [*TAB. 1]*:

Feature	R language	Python Language
Model building	X	X
Model Interpretability	X	
Production		X
Community support	X	
Libraries	X	X
Visualization	X	
Learning curve		X
Mastery of mathematics	X	X

TAB. 1: R Vs Python.

In short, R Language is great for statistical analyses and Python language is good for learning curve.

Microsoft Excel[1]: It is the most basic, intuitive and easy to use tool. Excel is able to work much more efficiently with tables. It offers tables, charts and graphics that help present data in fresh ways.

[1] https://products.office.com/en-us/excel

Excel enables to clean HR data easily by transforming a dataset into table and then check the data in each column if they contain outliers.

Excel present also sets of predefined tools that enable to do practically HR analytics in Excel such as:

-VLOOKUP: predefined function to merge data sets. It makes connecting two separate data sets very easy.

-Pivot tables: Technique of summarizing large quantities of data.

The limitation of Excel is that it not enables to work with much larger datasets compared to R.

Power BI[2]: It is a Business Intelligence reporting tool that makes the aggregation, analysis, and visualization of data very simple.

Power BI is a suite of business analytics tools that deliver insights throughout your organization. it can connect to multiple of data sources like SQL databases of employees' data, a live twitter feed, Excel spreadsheets and others. All these different data sources are then combined in one large database in Power BI.

Power BI can Produce beautiful reports and transform data into live dashboards then publish them for your organization to consume on the web and across mobile devices.

SPSS[3]: Is a Statistical software, it is one of the most commonly used HR analytics tools.

Its user-friendly interface enables to analyze data without having extensive statistical knowledge. In addition, SPSS is open source extensibility and easy to integrate with big data.

To sum up, the selection of the most appropriate HR Analytics tools depends on the aim of HR manager. For example:

-To create Dashboards, it is better to choose **Power BI** or **Excel because** Such tools make data aggregation and data visualization quite simple.

-To get some basic insights about employee (e.g. comparing employee performance in each department), it is good to choose simpler tool like **Excel** or **SPSS** because they require a low level of analytics skills.

-To deeply analyze HR data and make predictions about the future, it is better to select data analysis tools like **Python** or **R Studio** because they provide the capability to do the most advanced analysis…

In fact, Big Data is still in the development and its related techniques and tools are far from mature. So, HRM faces also challenges in the use of Big Data in terms of Storage, Analytics and Management.

5. Conclusion

The emergence of digital and Big Data have led companies to transform the way they use the data they hold to analyze the impacts of their investments and predict future performance. In the present time, Big Data Analytics and HR work together to create opportunities for business and make analytics-driven decisions.

[2] https://powerbi.microsoft.com/en-us/

[3] https://www.ibm.com/analytics/data-science/predictive-analytics/spss-statistical-software

In this paper, we have presented and examined the innovative topic of "Big Data" in order to describe, review, and reflect on big data analytics and its contribution in Human Resources Management.

This work first defined what is meant by big data to consolidate the divergent discourse on big data. Second, it puts the accent on "Big data analytics" where advanced analytic techniques are applied on big data in order to store, analyze and manage this large amount of enormous Data. Third and finally, it applies Big Data Analytics on HR Management which has a very important role to play, particularly in the management of individual performance, career plans, and so on. It aims to make decisions, to anticipate and correct HR actions and policies through the presentation of the most used HR Analytics tools.

Although, major innovations in analytical techniques for big data have not yet taken place. Our future work is planned in the sense of the contribution to the good governance of "Big Data Analytics Systems". For instance, real-time analytics will likely become a profitable field of research following the remarkable growth in location-aware social media and mobile apps.

References

Ahmed, E., Yaqoob, I., Hashem, I. A. T., Khan, I., Ahmed, A. I. A., Imran, M., & Vasilakos, A. V. (2017). The role of big data analytics in Internet of Things. *Computer Networks*. https://doi.org/10.1016/j.comnet.2017.06.013

Arora Y, G. D. (2016). Big Data : A Review of Analytics Methods & Techniques, 225–230.

Campos, J., Sharma, P., Gabiria, U. G., Jantunen, E., & Baglee, D. (2017). A Big Data Analytical Architecture for the Asset Management. *Procedia CIRP*, *64*, 369–374. https://doi.org/10.1016/j.procir.2017.03.019

Charles Henri Besseyre des Horts. (2014). "L'analytique RH" nouveau graal des DRH ? | RH info. Retrieved February 26, 2018, from https://www.rhinfo.com/thematiques/approche-globale-de-lentreprise/lanalytique-rh-nouveau-graal-des-drh

Chavanne, Y. (2015). L ' analytique pour une gestion RH sur mesure, 16–19.

Couaillier, J., & Projet, D. U. (2016). CATÉGORIE - ANALYTICS RH, 50–54.

Dave, M., & Gianey, H. (2017). Different clustering algorithms for Big Data analytics: A review. *Proceedings of the 5th International Conference on System Modeling and Advancement in Research Trends, SMART 2016*, 328–333. https://doi.org/10.1109/SYSMART.2016.7894544

DeZyre. (2016). Is Predictive Modelling easier with R or with Python? Retrieved April 7, 2018, from https://www.dezyre.com/article/is-predictive-modelling-easier-with-r-or-with-python/245

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Jadon, K. S., Bhadoria, R. S., & Tomar, G. S. (2016). A Review on Costing Issues in Big Data

Analytics. *Proceedings - 2015 International Conference on Computational Intelligence and Communication Networks, CICN 2015*, 727–730. https://doi.org/10.1109/CICN.2015.148

Lakshen, G. A., Vranes, S., & Janev, V. (2016). Big data and quality: A literature review. *2016 24th Telecommunications Forum (TELFOR)*, 1–4. https://doi.org/10.1109/TELFOR.2016.7818902

Lv, Z., Song, H., Basanta-Val, P., Steed, A., & Jo, M. (2017). Next-Generation Big Data Analytics: State of the Art, Challenges, and Future Research Topics. *IEEE Transactions on Industrial Informatics*, *13*(4), 1891–1899. https://doi.org/10.1109/TII.2017.2650204

Lyndon Sundmark. (2016). A Tutorial on People Analytics Using R - Employee Churn - Analytics in HR. Retrieved April 7, 2018, from https://www.analyticsinhr.com/blog/tutorial-people-analytics-r-employee-churn/

Owais, S. S., & Hussein, N. S. (2016). Extract Five Categories CPIVW from the 9V's Characteristics of the Big Data. *International Journal of Advanced Computer Science and Applications*, *7*(3), 254–258.

Philippe Burger, and al. (2016). Tendances RH 2016 Nouvelles organisations , nouveaux plans de vol.

Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. *2013 International Conference on Collaboration Technologies and Systems (CTS)*, 42–47. https://doi.org/10.1109/CTS.2013.6567202

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, *70*, 263–286. https://doi.org/10.1016/j.jbusres.2016.08.001

Soumyasanto Sen. (2017). Maturity of HR Analytics Demands Right Foundation – The HR Tech Weekly®. Retrieved February 26, 2018, from https://hrtechweekly.com/2017/01/25/maturity-of-hr-analytics-demands-right-foundation/

Srivastava, U., & Gopalkrishnan, S. (2015). Impact of big data analytics on banking sector: Learning for Indian Banks. *Procedia Computer Science*, *50*, 643–652. https://doi.org/10.1016/j.procs.2015.04.098

Tsai, C. W., Lai, C. F., Chao, H. C., & Vasilakos, A. V. (2015). Big data analytics: a survey. *Journal of Big Data*, *2*(1), 1–32. https://doi.org/10.1186/s40537-015-0030-3

Whitepaper, C. (2013). Les Ressources Humaines à l'heure du Big Data : progrès, perspectives et limites.

Zang, S., & Ye, M. (2015). Human Resource Management in the Era of Big Data. *Journal of Human Resource and Sustainability Studies*, *3*(March), 41–45. https://doi.org/10.4236/jhrss.2015.31006

Résumé

Les données massives au niveau des ressources humaines font référence aux quantités exponentielles d'employés, de clients et de données transactionnelles dans les organisations. Ces données peuvent provenir de sites sociaux tels que LinkedIn et Viadeo ... qui stockent les coordonnées des personnes inscrites sur le site. L'analyse de ces données volumineuses à l'aide d'outils spécifiques peut aider les entreprises à prendre des décisions qui affectent le recrutement et créent des avantages plus importants afin d'accroître la productivité. Dans cet article, nous allons essayer de répondre à deux questions fondamentales : Les contributions du Big Data dans la gestion des ressources humaines er Comment l'entreprise peut-elle traiter les analyses au niveau des département des Ressources humaines.

Mots clés : Ressources Humaines (RH), Big Data, Gestion des RH, Social Big Data Analytics.

Etude de l'impact des Fintech sur le système bancaire

Elhassane Belrhali and Moutahaddib Aziz

Adresse : Ecole Nationale de Commerce et de Gestion Kenitra,
B.P : 1420- Kénitra 14000. Université Ibn Tofail
elhassane.mb@gmail.com ; moutahaddib@gmail.com
www.encgk.ac.ma

Résumé. Le principe de la Fintech est d'offrir de multiples services financiers via la plateforme web et les smartphones (par application). Elle englobe plusieurs méthodes (Blockchain « Cryptomonnaie », crowdfunding, crowdlending, online banking etc).

Comme nous savons avec le développement stratosphérique que connait les technologies va donner naissance aux banques en ligne qui offre le même service et même amélioré par apport aux banque conventionnels. Ce qui a amené les banques à revoir nécessairement leurs business model et s'adapté aux nouvelles spécificités de leurs secteur, suivre/mettre une veille technologique. Le changement d'habitudes bancaires (consultations en ligne, virement en ligne, épargne en ligne, transfert en ligne), une dématérialisation presque totale du service bancaire classique.

Il est évident que l'impact des FinTech sur le domaine bancaire est important vu la course des plus grandes bancaires à racheter les start-up ou intégrant des cellules technologies d'incubation Fin-tech dans leurs structures.

1 Introduction

La digitalisation est le sujet phare de ces dernières années, le monde se numérise et les usages changent à chaque fois. En effet, en 2016 il y'avait presque 3.42 milliards d'internautes et 3.79 milliards de mobinautes (selon le magazine Digital insider, 2016), des chiffres faramineux montrant la grande expansion du numérique partout dans le monde.

Toutes les économies du monde sont intéressées par cette question digitale, des pays comme la Turquie, Le Canada, Malaisie ont instaurés même une éducation numérique dans les programmes de leurs écoles (par exemple 1 million de tablette en Turquie).

Plusieurs domaines sont intéressés par la digitalisation à savoir le domaine financier qui a vu l'apparition de nouvelle technologie financière dit (FinTech). De nombreuses start-ups ont vu le jour dans la Silicon Valley et dans de nombreux européens. Ces petites entreprises offrent un service financier digitalisé et viennent concurrencer les banques classiques.

Ces FinTech attire la frénésie des capital-risqueurs par leurs business juteux, en 2017 16.6 milliards de dollars ont été investis dans 1120 start-up (selon le site les echos.fr tirant son information du rapport de CB insight)

Les principales technologies financières existantes (Block-chain, Paiement mobile, Crypto-monnaie).

Les banques se sentent tantôt menacé mais aussi dans l'obligation de suivre le courant de la digitalisation.

La problématique de ce travail de recherche est de savoir quel est « **le réel impact des Fintech sur le secteur bancaire ? »**

Dans une première partie, nous allons définir les FinTech et leurs apports, les enjeux de la digitalisation, ensuite dans une deuxième partie mesurer l'impact de ces FinTech sur l'avenir du secteur bancaire et les changements engendrés par la digitalisation

2 L'évolution des technologies financière dans le domaine bancaire

L'évolution technologique a touché presque tous les domaines de notre vie, allant du domestique à la finance. Le numérique est la troisième révolution industrielle que connait l'humanité. La grande tendance est la digitalisation, tout est exécuté à travers des applications et à travers le canal révolutionnaire qu'est l'internet.

Plusieurs entreprises se sont vues obligés de réadapter leurs business model aux nouvelles spécificités sectorielles. Le consommateur est devenu hyperconnecté et très exigeant vu qu'il a accès à toutes les informations précises et abondantes surtout. Nous sommes dans une nouvelle ère où le consommateur sait ce qu'il veut et le veux plus rapidement possible.

Mais mutation la plus importante que connait le numérique est celle des technologies financières, principalement les banques sont les plus concernés par ce nouveau dénouement du Banking.

La finance digitale est un outil très en vogue pour faire parvenir les services financiers à toutes les populations, considéré comme un levier de développement socio-économique.

Les banques sont très conscientes que leurs continuités est conditionné par l'importance accordée au numérique (Fintech) dans leurs stratégies globales, la digitalisation est l'unique voie de développement et de « survie ». En effet, une refonte totale des métiers de la banque doit être opérer pour suivre le nouvel essor du secteur qui se voit impacté d'une manière énorme par les nouvelles technologies.

Nous assistons actuellement au changement 360° du modèle bancaire classique, les clients actuels sont des internautes ou bien mobinautes vu l'utilisation quotidienne et répétitive du smartphone, veulent seulement à travers une application faire toutes les transactions possibles de leurs téléphones sans devoir se déplacer en agence pour des opérations simples (virements, consultation du solde, transfert d'argent, etc).

Alors, le principe de la banque en ligne est apparu vers le début des années 2000 et avec la croissance d'utilisation d'internet, ce nouveau service s'est répandu en donnant naissance aux banques virtuelles (exemple d'Orange banque, ING Direct)

Les Fintech sont diverses (Blockchain, Big data, Clouding , Pee-to-peer, Paypal, payment mobile) , ces exemples sont les plus courus en terme d'utilisation.

Le big data est défini par Mc Kinsey « Le Big Data est défini par McKinsey, comme un, regroupement de données dont la taille ne permet pas aux logiciels classiques de les traiter (récupération, stockage, analyse).
On en distingue plusieurs types :

• les données interpersonnelles (principalement les données de communications électroniques du type mails réseaux sociaux, etc.),

• les données d'interaction homme-machine (archives de cartes bleues, historiques de navigation

WEB, etc.),

• les données inter-machines (échanges de données ,GPS, caméras de surveillance, etc.)

A quoi sert le « Big data » dans le secteur bancaire ? , c'est la question que se pose plusieurs data-scientist. La réponse est que le big data sert à rapatrier toutes les informations nécessaires ou pourraient être nécessaire dans le futur, ces données servent à identifier la typologie de la clientèle de la banque, leurs habitudes d'achats, les montants, les sites, la fréquence des achats.

Donc tous ces éléments donnent une vision assez claire sur le client, ce qui rends la tâche simple à la banque qui conceptualise des offres et des applications basés sur un besoin réel de la clientèle.

La définition du Blockchain « La Blockchain est une base de données transactionnelle distribuée, comparable à un grand livre dans lequel chaque nouvelle transaction est écrite à la suite des autres, sans avoir la possibilité d'effacer ces dernières. Cette technologie fonctionne sans intermédiaire : par exemple, dans le cas d'une transaction entre deux individus sans Blockchain , une banque va vérifier que le payeur a bien les fonds qu'il dit détenir et va accepter ou non la transaction. La banque joue le rôle d'intermédiaire et de tiers de confiance » *par Laurent Leloup Expert Blockchain auprès du Pôle de compétitivité Finance Innovation en France (Bpifrance, 2016)*

Le paiement mobile est aussi le moyen le plus répandu au monde en termes d'utilisation, facile et très efficace, plusieurs interfaces de paiements existent tels qu'avec Checkout, dans le Google Wallet) et Microsoft Tux Etats-Unis ou encore les télécoms (ISIS aux USA). Presque les deux quarts de la population mondiale possèdent un smartphone et 1/3 effectue des achats sur internet.

Les banques créent des cellules de suivis et d'incubations de Fintech Solutions, pour veiller sur les nouveautés technologiques.

Bien sûr une banque doit parfaitement connaitre son image auprès des consommateurs, et ceci dit elle doit voir comme elle perçue par les tiers ou clients. Du coup, une présence sur les réseaux sociaux est indispensable (Twitter, Facebook, Blogs, Instagram), c'est aussi un outil de collecte de la data. La banque est devenue digitalisé presque dans sa totalité, la mise en place de système informatique performant, des changements des moyens utilisés et un rajeunissement des effectifs. Des équipes de conduite de changement digital sont créées pour accompagner la banque dans tout le processus de digitalisation.

N'oublions pas un point très essentiel dans le numérique, c'est l'aspect sécuritaire qui lui est prône sur tous les éléments, puisque un consommateur ne donne sa confiance que lorsqu'il se sent assez protégé et avec un risque de piratage qui tends vers 0%, on parle (Security payment , security transaction, security services , security check). Les Fintech utilisent les données des banques pour détecter les fraudes de façon réactive, en recoupant en temps réelles comportements inhabituels (type et montant d'opérations, géolocalisation de smartphones).

Evidemment le risque dans le secteur bancaire est l'un des éléments à suivre de plus près, dès lors plusieurs solutions informatiques traitant ce volet ont fait leurs naissances, ces Fintech sont spécialisées dans l'application des Big Data à l'analyse du risque global et offrent aux banques des outils d'aide à la décision, notamment à travers la mise en place de plateformes de valorisation d'actifs financiers.

Quelques Fintech de ce type : QuantCube Technology, Scaled Risk

Pour résumer cette première partie, les banques sont « obligées » de suivre la transformation numérique que connait l'humanité, c'est soit disons la 4ème révolution industrielle après celle de

l'internet. Les banques ont le choix soit de créer leurs propres filiales Fintechs ou bien faire appel à des sociétés spécialisées dans ce type de technologie.

Dans la deuxième partie de cette communication nous allons voir l'impact global des Fintechs sur le secteur bancaire.

3 L'impact des FINTECHS sur le secteur bancaire

L'impact qu'a porté la technologie sur tous les secteurs est grandiose, tous les métiers sont assujettis à une digitalisation de leurs processus de fonctionnement ou bien même de leurs business.

L'appellation Fin Tech désigne « la technologie financière » ou bien les sociétés qui opèrent dans ce type de métier, il faut faire la distinction pour éviter une éventuelle confusion.

Une banque a le choix entre la création d'une filiale Fin Tech spécialisé dans les services bancaires, l'instauration du Big Data, de configurer des solutions de paiement en ligne, tout cela à l'interne.

La deuxième possibilité est le recours aux prestataires externes « Les Start-Ups Fin Tech » qui vont être comme un partenaire et pourront apporter de la technologie de pointe réadapté.

Les perspectives de rendement de certaines activités bancaires rendues accessibles grâce au numérique ont contribué à inciter de nouveaux entrants à conquérir ce marché.

La vulgarisation du smartphone et l'avènement des données mobiles ont ouvert des horizons aux Fin Tech qui ont su capitaliser sur le digital et l'accessibilité des données pour réinventer l'expérience client.

Les banques l'ont désormais bien compris et multiplient les initiatives pour recentrer leur business model sur le client, notamment via la collaboration avec des Fintech (prise de participations, partenariats, accélérateurs

Les banques sont concurrencées par les nouveaux entrants de la banque mobile, une sorte d'Ubérisation du secteur

FIG.1 : Source : Le Lab fnance (BPI 2017)

A cet effet, Quels sont les risques encourus par les banques après l'expansion des FinTech ?. Pour répondre à cette question plusieurs pistes peuvent être évoquées:

- Subir un déplacement de la clientèle, une perte de revenus et une diminution des marges.
- Perdre son rôle principal d'intermédiaire financier
- Voir la relation client s'intermédier et transformer les établissements financiers en « banques-usines » (assurant uniquement un rôle de cantonnement des fonds et de gestionnaires de compte) :
- Perdre l'accès direct aux données, à l'heure où ces données constituent un atout considérable pour améliorer la connaissance des clients (Le réel danger);
- Perdre les clients, puisque les agrégateurs qui pourraient devenir les principaux interlocuteurs des clients qui ne se rendraient plus sur les interfaces des banques, mais effectueraient l'intégralité de leurs opérations via les agrégateurs.

Le numérique n'est une menace que quand il touche aux intérêts directe de la banque certes, mais ces bienfaits sont énormes et change carrément la chaine de valeur bancaire :

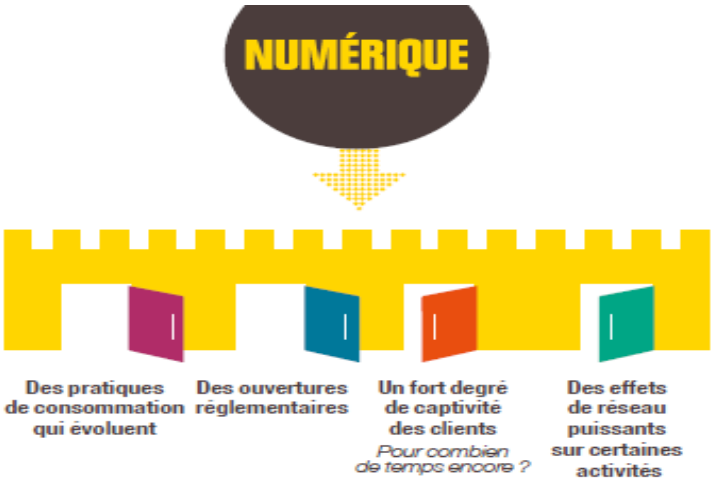


FIG.2 : Source : Le Lab fnance (BPI 2017)

Avec la digitalisation plusieurs barrières s'estompent et la taille n'est pas un critère d'évaluation pour supplier à un marché, le niveau de technologie utilisé lui seul permet de juger si la banque est performante ou pas.

L'autre impact de la digitalisation est le remaniement au sein de la banque, la restructuration s'avère une condition sine qua non pour réussir la transition numérique. Les fiches de métier changent vers plus souplesse, Ainsi, la filière Marketing sera imprégnée des flux de données que le numérique permet de collecter.

Des métiers de data-scientist apparaissent, à la frontière entre l'exploitation de données (aujourd'hui plutôt au niveau des DSI) et du marketing. Bâti sur les bases du business intelligence, on verra apparaître de nouveaux profils devant mixer appétence et sensibilité marketing et compétence technique de modélisation de données.

3.1 Les métiers des Ressources Humaines

De la même manière, la filière «RH» se déploie avec des consultants en conduite du changement. L'impact est tel sur l'organisation du travail, sur les méthodes de gestion de projet, sur les relations interpersonnelles (moins de frontières pro/perso) que la pression sur les collaborateurs s'accroît.

Ainsi, un des impacts du numérique est la nécessité de tenir compte de la qualité de vie au travail afin d'éviter des situations de stress de plus en plus avérées. Les DRH doivent s'équiper afin de pouvoir maîtriser ces risques psycho-sociaux.

3.2 Les métiers du risque

La filière «Risques» se doit également de s'adapter. La réputation de l'entreprise est plus facilement mise en jeu. Les réglementations deviennent plus compliquées à faire respecter. La mise en place d'une stratégie de sécurité réussie afin de faire face aux risques de l'automatisation des processus, de la gestion des données et de la cybercriminalité devient une nécessité primordiale. Sept banques sur dix ont déjà été touchées par la cyber-fraude (Etude iTPro, 2017) … Les fonctions légales doivent aussi évoluer.

3.3 Les métiers de la communication

Ces métiers vont connaître une révolution avec des possibilités d'exploitation multicanale rend toute action à forte visibilité. L'implémentation de plateformes collaboratives, de TV interne, d'applications mobiles permet de toucher un plus grand nombre et de personnaliser les messages. Que ce soit en interne ou en externe, la filière retrouve une seconde jeunesse, à condition de savoir s'adapter.

3.4 Les métiers des technologies de l'information

Bien entendu, les métiers des technologies de l'information évoluent également fortement. Des compétences en architecture IT, de gestion de données, en pilotage de projet sont renforcées.

La maîtrise de nouveaux outils se fait sentir. La DSI va devoir changer de paradigme pour répondre à cette évolution numérique. Enfin, l'organisation du travail et des projets évoluent vers des approches communautaires. Le métier de community manager explose. Toutefois, les contours de ce nouveau métier sont très variables d'une entreprise à l'autre.

4 Conclusion

La FinTech est l'avenir de la banque sans aucun doute, chaque banque doit investir le maximum et doubler ses efforts pour une digitalisation parfaite à travers les partenariats, la mise en place d'un service spécial du digital.

Le consommateur est lui aussi devenu exigeant, très informé, très mobile et prudent, il est surtout zappeur. Avec la croissance exponentielle que connait l'internet, les applications de paiement mobiles, les outils de placements en ligne, la banque mobile, tout est devenu facile.

Le fait d'avoir un compte bancaire n'est plus un problème, tout est fait de manière virtuelle.

La banque doit jouer sur l'aspect sécuritaire puisque c'est une boucle essentielle dans le processus de digitalisation.

Les banques doivent adoptés toutes les technologies financières existantes (Blockchain, Peer-to-peer, Mobile Banking) , seules ou combinés.

Les banques doivent faire très attention aux nouveaux entrants du secteur tel que les GAFA (Google, Amazon, Facebook, Apple) qui propose des solutions numériques eux aussi.

Pour conclure, la digitalisation doit être prise au sérieux et faire d'elle un projet d'investissement à long terme.

Références

Moysan Yvon Moysan (2016), FinTechs et plates-formes bancaires : faut-il copier la Chine (lecturer Digital Marketing (IÉSEG School of Management) article paru sur la revue (Revue de banque)

Ron Shevlin (2011), Smarter Bank: Why Money Management is More Important Than Money Movement to Banks and Credit Unions

Brett King (2012), Bank 3.0-: Why banking is no longer somewhere you go, but something you do

Roger Peverelli (2014), Reinventing Financial Services: what consumers expect from future banks and insurers

Susanne Chishti (2015), The Fintech Book

Le pôle de compétitivité mondial FINANCE INNOVATION Banque & Fintech, Enjeux d'innovation dans la banque de détail,

PWC (2017), Global FinTech_Report

Xerfi (2015), Les FinTech ou nouveaux entrants dans la banque et la Finance

Olivier Wyman (2016), Blockchain in Capital Markets

Accenture (2016), Fintech and the evolving landscape: landing points for the industry.

Summary

The principle of FinTech is to offer multiple financial services via the web platform and smartphones (by application). It encompasses several methods (Blockchain, crowdfunding, crowdlending, online banking, etc.).

As we know with the stratospheric development that technology knows will give rise to online banking that offers the same service and even improved by providing conventional banking. This has led banks to necessarily review their business model and adapt to the new specificities of their sector, monitor / put a technological watch. The change of banking habits (online consultations, online transfer, online savings, online transfer), a virtual dematerialization of the classic banking service.

It is obvious that the impact of FinTech on the banking sector is important given the race of the largest banks to buy start-ups or integrating cells tech-end incubation technologies in their structures.

Chapter

4

Internet of Things & Banking

ASD'2018

Content

Safety at level crossings: advanced statistical accidents analysis

Ci Liang*'**, Mohamed Ghazel*, El-Miloudi El-Koursi*,
Fouzia Boukour*, Olivier Cazier***

* University Lille Nord de France, IFSTTAR, COSYS, ESTAS,
20, Rue Elisée Reclus, F-59650 Villeneuve d'Ascq, France
ci-liang@railenium.eu
el-miloudi.el-koursi@ifsttar.fr
mohamed.ghazel@ifsttar.fr
Fouzia.boukou@ifsttar.fr
oliviercazier@hotmail.fr

http://www.ifsttar.fr

Abstract. Every year, more than 300 people are killed at road-rail level crossings (LXs) in the European Union. LXs have been identified as being a particular weak point in road/rail infrastructure, seriously affecting rail safety. This paper focuses on advanced statistical risk analysis on French LXs. Various kinds of impacting factors, namely, transport mode, geographical region and traffic moment, are analyzed by means of statistical techniques to dig out their statistical characteristics based on the accident data from SNCF, the French national railway operator. Then, we assess the effect of various factors on the risk level quantitatively, in such a way as to open the way for setting efficient solutions and consequently, reaching the point of improving LX safety.

1 Introduction

Level crossing (LX) safety involves various aspects: technical elements, operational procedures, human factors and environmental considerations (Berrado, 2011). In order to significantly reduce the accidents and their related consequences at LXs, it is crucial to carry out a series of thorough analyses and modeling to understand the potential reasons for these accidents and thus, enable the identification of practical design and improvement recommendations to prevent accidents at LXs (Ghazel, 2014). LXs have been identified as being a particular weak point in road/rail infrastructure, seriously affecting rail safety (Khoudour, 2009). In some case of railway transport level crossings can represent up to 50% of all fatalities caused by railway operations. A level crossing (LX) is an intersection where a railway line intersects with a road or a path at the same level. Level crossings constitute a significant safety concern.

Every year, more than 300 people are killed at road-rail level crossings (LXs) in the European Union. An average, every day, one person has been killed and close to one seriously

injured at level crossings in Europe. In Europe, the number of fatalities in all kinds of railway accidents has decreased, except those related to level crossing accidents. This can be partly explained by the continuous increase in road traffic across Europe, which may increase the likelihood of a level crossing accident. The level crossing safety is viewed as a road safety problem by railway infrastructure managers. It is viewed as a secondary problem by the road authorities. It appears that the concept of shared and delegated responsibility in road safety often fails to deliver the targeted results when it comes to level-crossing safety. As demon-strated by accident and incident statistics, LX safety is one of the most critical issues that railway stakeholders need to deal with. There are more than 118,000 LXs in the 28 countries of the European Union (E.U.) which correspond to an average of 5 LXs per 10 line-km. Ac-cidents at European LXs account for about one-third of the entire railway accidents [Liang et al., 2017].

In 2014, 506 significant level crossing accidents occurred in the EU-28 resulting in 282 fatalities and 287 serious injuries. Level crossing accidents represent 24.4% of all significant railway accidents and 26.8% of all fatalities on the railway, suicides excluded. There was stagnation in the number of level crossing accidents, with 506 accidents recorded on railways of the EU countries in 2014, compared to 510 accidents in 2013. However, since 2009, a slightly decreasing trend has been observed. The number of level crossing accidents has reduced a 3% per annum (ERA, 2016).

This paper focuses on advanced statistical risk analysis on French LXs. Various kinds of impacting factors, namely, transport mode, geographical region and traffic moment, are analyzed by means of statistical techniques to dig out their statistical characteristics based on the accident data from SNCF, the French national railway operator. Then, we assess the effect of various factors on the risk level quantitatively, in such a way as to open the way for setting efficient solutions and consequently, reaching the point of improving LX safety. In details:

- A general risk analysis of average accident frequency in terms of transport mode and geographical region is performed.
- Then, the normalized risk analysis, namely the average accident frequency normalized by the traffic moment, is performed with regard to various traffic moment groups.
- The normalized frequency distributed in different French regions is generated.

The objective is to make thorough analysis on various kinds of transport mode, different geographical regions and traffic moment to explore their influences on LX accidents. It should be noticed that the database from SNCF used in the analysis reported in this paper contains detailed information about LX accidents/incidents from 1974 to 2014.

2 Statistical analysis: an overview

In France, the railway network shows more than 18,000 LXs for 30,000 km of railway lines, which are crossed daily by 16 million vehicles on average, and around 13,000 LXs show heavy road and railway traffic (SNCF Réseau 2011). In 2013, 148 train/vehicle collisions occurred at French LXs, giving rise to 29 deaths. In 2016, 111 train/vehicle collisions at French LXs led to 31 deaths (Liang et al 2017). This number was half the total number of collisions per year at LXs a decade ago, but still too large. LX safety involves various as-

pects: technical elements, operational procedures, human factors and environmental considerations. Due to non-deterministic causes, complex operation background and the lack of thorough statistical analysis based on detailed accident/incident data, risk assessment of LXs remains a challenging task. There are four main LX types in France (SNCF 2015), namely SAL (signalisation automatique lumineuse) as shown in figure 1 [Fig. 1]:

(1) SAL4: Automated LX systems with four half barriers and flashing lights;
(2) SAL2: Automated LX systems with two half barriers and flashing lights;
(3) SAL0: Automated LX systems with flashing lights but without barriers;
(4) Crossbuck LXs, without automatic signaling.

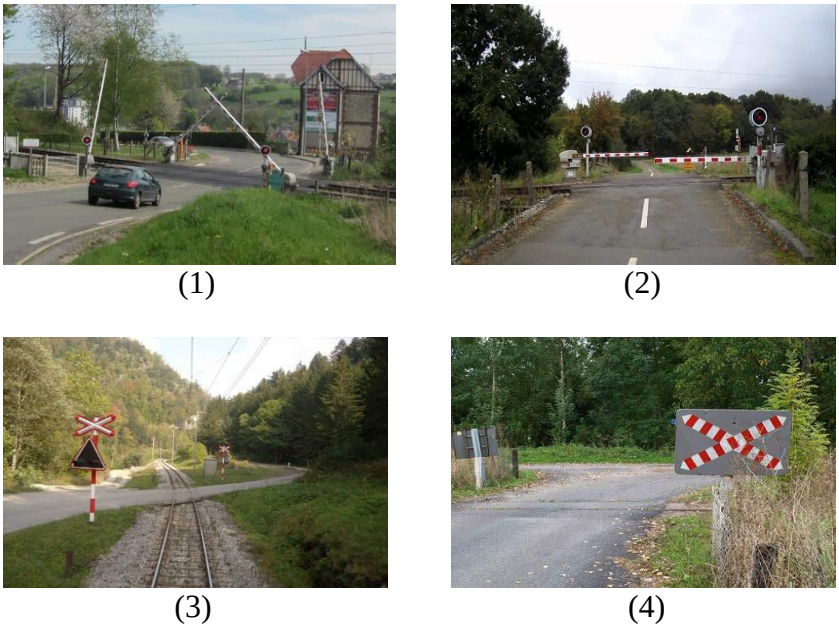


FIG. 1*: Four types of LXs in France.*

Level crossing SAL2 (more than 10,000) is the most widely used type of LX in France. Moreover, more than 4,000 accidents at SAL2 LXs contributed most to the total number of accidents at LXs from 1974 to 2014. In addition, according to the SNCF statistics, the accidents at SAL2 LXs can be considered as the most representative for LX accidents in general. Hence, we will focus on the analysis of collisions occurring at SAL2 LXs.

3 Statistical analysis

LX accidents involve the following transport modes: 1) motorized vehicle (MV), 2) pedestrian and bicycle (PB). There are 21 geographical administrative regions in mainland France, as divided in 2014. Accidents which are caused by main types of transport mode: 1) motorized vehicle (MV), 2) pedestrian or bicycle (PB), are considered respectively to allow for making statistical analysis in different regions.

3.1 General risk analysis

The general frequency of accidents occurring at each SAL2 per year are used to represent the general risk level involving total accidents, MV accidents, and PB accidents in different 21st French regions during the last 40 years. We can calculate the general frequency through the Eq. (1):

$$F_{G_i} = \frac{Nb\_acc_i}{Nb\_SAL2_i \times Nb\_year}, i = 1, 2, ..., 21; \tag{1}$$

Where F_{G_i} represents the general frequency in i^{th} region; Nb_acc_i represents the number of accidents occurring in i^{th} region, Nb_SAL2_i represents the number of SAL2 in i^{th} region, and Nb_year represents the number of years of the considered period. Three kinds of general frequency related to total accidents, MV accidents, and PB accidents will be calculated. Correspondingly, the number of total accidents, MV accidents, and PB accidents are presented as Nb_acc_i respectively when calculating these three kinds of general frequency.

Now that these three kinds of general frequency in each region are determined, maps of French regions with the frequency values presented are generated to show the frequency distribution in different regions. As shown in FIG. 2a, the general frequency value of total accidents in the red region (greater than 0.02) is the highest. The general frequency value of total accidents in the orange region (between 0.02 and 0.01) is the second highest, and the general frequency values of total accidents in the green region (less than 0.01) are the lowest. When we analyze the frequency figures in detail, we find that the risk is most serious in Île-de-France with a frequency of more than 0.02; Languedoc-Roussillon takes the second place with the frequency of about 0.017 followed by Provence-Alpes-Côte-d'Azur with the frequency of about 0.016. On the other hand, Limousin has the lowest risk with the frequency of about 0.005. Haute-Normandie and Basse-Normandie occupy the second and the third places of lowest risk successively. In FIG. 2b and FIG. 2c, the general frequency of accidents caused by motorized vehicles, pedestrians and bicycles in different regions is shown respectively. Considering the accidents caused by motorized vehicles, the distribution of frequency in different regions in FIG. 2b is relatively consistent with the distribution shown in FIG. 2a. The only exception is Champagne-Ardenne. However, in FIG. 2c, as for the accidents caused by pedestrians and bicycles, the distributions of frequency in different regions are very different from the distribution shown in FIG. 2a.

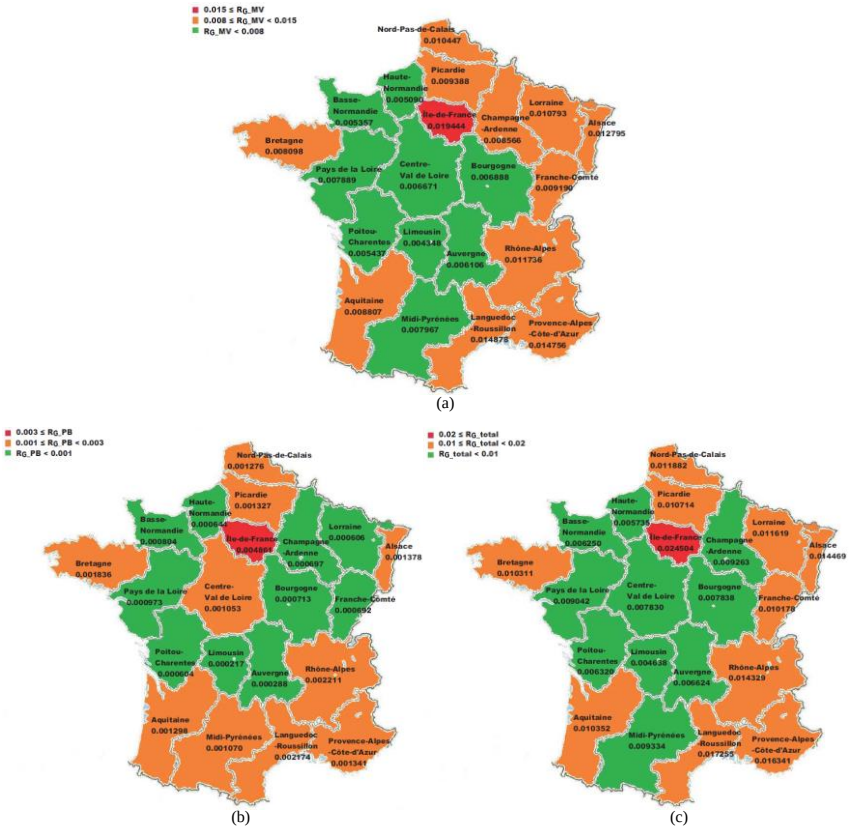


FIG. 2 : (a) *Frequency of total accidents distributing in different regions;* (b) *Frequency of MV accidents distributing in different regions;* (c) *Frequency of PB accidents distributing in different regions.*

The results shown in FIG. 2 indicate that motorized vehicle is the main transport mode causing LX accidents in France (Liang et al. 2018 b). Moreover, as the LX accident frequency caused by motorized vehicles increases, the entire LX accident frequency increases accordingly (Liang et al. 2017). Considering the train/motorized vehicle (train-MV) collisions, SAL2 LXs also have the major part of LX accidents according to the statistics shown in (Liang 2017). For all these reasons, in the following content, we will focus on the analysis of train-MV collisions occurring at SAL2 LXs. It should be noted that suicide scenarios are not in the scope of our global study.

3.2 Normalized risk analysis in terms of traffic moment

The motorized vehicle is the main transport mode causing accidents at SAL2 LXs. Therefore, in this section we will focus on the impact of traffic moment on the whole risk level. We suspect that this parameter is one of the main parameters impacting LX risk level. In-

deed, traffic moment gives the combined traffic (train/MV) at the LX and is defined as follows:

Moment = Road traffic frequency x Railway traffic frequency. (2)

Where Road traffic frequency represents the number of motorized vehicles per year at the LX, and Railway traffic frequency represents the number of trains per year crossing the LX. Here, we use "M" for short to denote the moment.

SAL2 LXs are classified by the category of M. In order to make the number of SAL2 in each M group to be as far as possible similar to any other group in each region, the moment groups have been defined in such a way that the number of SAL2 LXs in every group belongs to the interval [102, 155]. For example, there are three M categories in Auvergne, which are "0 < M < 750", "750 < M < 5000", "5000 < M <401850", with the corresponding numbers of SAL2: 143, 144, 147, respectively. Besides, 401850 is the maximum M in this region. In this way, we can make risk analysis with regard to these SAL2 groups, thus making it possible to highlight the risk level related to different categories of M. The average frequency normalized by M distributed over the different regions. It illustrates that the risk level is the highest in Île-de-France region with the normalized risk of 7:31x10-4; Alsace region takes the second place followed by Languedoc-Roussillon and Provence-Alpes-Côte-d'Azur. Through detailed analysis, we find that the general average frequencies of MV accidents in these 4 regions are also the highest. Moreover, according to the recorded statistics, more train-MV collisions happened at SAL2 with small M in Île-de-France than in the other 3 regions during the period considered.

4 Accident frequency prediction model

In this section, an advanced accident frequency prediction model is developed, which enable to rank risky LXs accurately and identify the significant impacting parameters efficiently. The parameters considered in this model are shown in TAB. 1.

Parameter	Data coding
Railway traffic characteristics	
Average daily railway traffic	Numerical, used directly;
Railway speed limit	Numerical, used directly;
Roadway traffic characteristics	
Average daily road traffic	Numerical, used directly;
Annual road accidents	Road accident factor: *National annual road accidents in a given year / National average road accidents per year over the period observed*;
LX characteristics	
Alignment	Alignment indicator: 0, 1 and 2 represent "straight", "curve" and "S", respectively;

Profile	Profile indicator: 0 and 1 represent "normal" and "hump or cavity", respectively;
LX width	Numerical, used directly;
Crossing length	Numerical, used directly;
Region	Region risk factor, highlighting the general LX-accident-prone region: *The number of SAL2 accidents over the observation period in the region considered / The number of SAL2 LXs in the region considered*;

TAB. 1 – *Parameters considered and data coding.*

4.1 Model development

The accident frequency prediction model (Liang 2018 a, Liang 2018 b) is developed as follows:

$$\lambda_{10Y} = K \times F_{RAcc} \times V^{0.354} \times T^{0.646} \times \exp(C_{Profile} \times I_{profile} + C_{Align} \times I_{Align} + C_{Wid} \times Wid + C_{Leng} \times Leng + C_{RSL} \times RSL + C_{Reg} \times F_{Reg}) \quad (3)$$

Where λ_{10Y} represents the annual accident frequency at a given SAL2 for a period of 10 years; K is the constant coefficient; F_{RAcc} is the road accident factor which reflects the variation of annual road accidents as time advances (a time-dependent variable); V is the average daily road traffic; T is the average daily railway traffic; $I_{profile}$ and $C_{Profile}$ are respectively the profile indicator and its corresponding coefficient; I_{Align} and C_{Align} are respectively the alignment indicator and its corresponding coefficient; *Wid* and C_{Wid} are respectively the LX width and its corresponding coefficient; *Leng* and C_{Leng} are respectively the crossing length and its corresponding coefficient; *RSL* and C_{RSL} are respectively the railway speed limit and its corresponding coefficient; F_{Reg} and C_{Reg} are respectively the region factor and its corresponding coefficient. We consider ($V^{0.354} \times T^{0.646}$) as an integrated parameter that reflects the combined exposure frequency of both railway and road traffic, which is called *CM* for short. Since the accident data are over-dispersed, here we use Negative binomial (NB) model (cf. Eq. (4)) to obtain regression coefficient and make further prediction.

$$P_{NB}(X = k) = \frac{\Gamma\left(k+\frac{1}{\alpha}\right)}{\Gamma(k+1)\Gamma\left(\frac{1}{\alpha}\right)}\left(\frac{1}{1+\alpha\lambda}\right)^{1/\alpha}\left(\frac{\alpha\lambda}{1+\alpha\lambda}\right)^{k},\ k = 0,\ 1,\ 2, \ldots \quad (4)$$

The coefficients as estimated as follows:

Parameter	**Coefficient**	**Estimated value**	***t* - statistic**	**Significant**
	K	-9.424	-20.615	×
F_{RAcc}	C_F	0.616	1.793	
CM	C_{CM}	0.006	16.493	×

$I_{profile}$	$C_{Profile}$	-0.107	-0.850	
I_{Align}	C_{Align}	0.298	4.159	×
Wid	C_{Wid}	0.199	7.173	×
Leng	C_{Leng}	0.031	3.201	×
RSL	C_{RSL}	0.010	7.034	×
F_{Reg}	C_{Reg}	1.508	4.294	×

TAB. 2 –*Regression results of* λ_{10Y}

4.2 Predictive accuracy validation

Further analysis to assess the predictive accuracy of the prediction model is carried out. As shown in TAB. 3, f_k denotes the percentage of samples of observed annual accident frequency with *k* accidents involved in a given year (f_k = the number of samples of observed annual accident frequency involving *k* accidents occurring in a given year / the total number of samples *n*). The estimated relative annual accident frequency reflected by estimated probabilities on average is computed as : $\hat{f}_k = \sum_{i=1}^{n} \hat{P}(X_i = k)/n$, where $\hat{P}$ $(X_i = k)$ is the estimated probability of k accidents occurring at a given SAL2 in a given year.

# Annual accidents considered (*k*)	Observed annual Accident frequency (f_k in percent)	NB- λ_{10Y} estimated Relative annual accident frequency ($\hat{f}_k$ in percent)
0	99.6313	99.2801
1	0.3616	0.5129
2	0.0071	0.0088
>2	0	0.0008

TAB. 3 – *The predictive accuracy validation.*

The results shown in Tab. 3 indicate that the λ_{10Y} model combined with the NB distribution shows a high predictive accuracy with regard to various annual numbers of accidents occurring at a given SAL2 during the 10 years from 2008 to 2017. In the case of more than 2 accidents occurring at a given SAL2 in a given year, the predictive accuracy of the λ_{10Y} model combined with the NB distribution shows a deviation of only 0.0001% compared with the actual $f_k = 0$. In fact, there are no SAL2 LXs showing more than 2 accidents in the same year during this 10-year period considered.

5 Bayesian network modeling

In this section, we will develop an integrated Bayesian network (BN) model that considers the factors in accident prediction model described in section 4 and some other motorist behavior factors to make accident/consequence prediction and cause diagnosis (Liang 2018 c). Firstly, data discretization is applied on continuous causal variables. Namely, the continuous causal variables, i.e., “Average Daily Road Traffic”, “Average Daily Railway Traffic”, “Railway Speed Limit", “LX Width", “Crossing Length" and “Corrected Moment", are divided into 3 groups that each group has the similar number of samples. As for the “Region Risk" factors corresponding to 21 regions in mainland France, they are divided into 3 groups as well, ranked according to the risk level in descending order, and each group contains 7 region risk factors. As for the finite discrete causal variables, i.e., “Alignment", “Profile", “Stall on LX", “Zigzag Violation", “Blocked on LX" and “Stop on LX", we allocate an individual state to each value of the variable.

The BN model is built as shown in FIG. 3. One can notice that the developed BN risk model shows two layers: 1) Layer 1 (in the bottom) is used for diagnosing influential factors; 2) Layer 2 (in the top) is used for evaluating the consequences of LX accidents. The “SAL2 MV Accident" node colored in yellow is the key node connecting the two layers, as well as the target node of accident prediction. In Layer 1, we split the network into 2 parts: the left-hand part contains the static factors (SF) and the right-hand part includes motorist behavior factors (MBF).

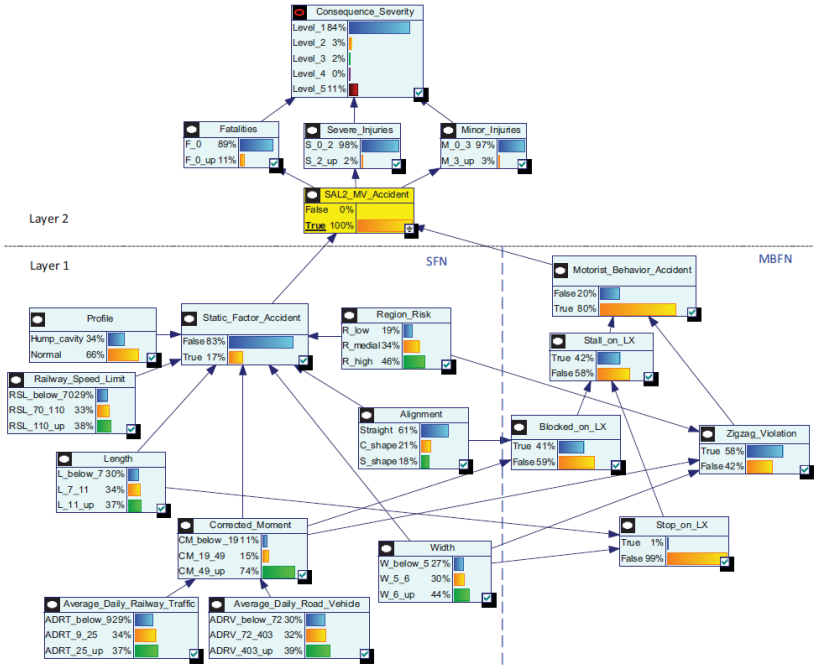


FIG. 3: *BN risk model.*

FIG. 3 shows that the “SAL2 MV Accident = True" state is configured as the targeted state. In this way, one can assess the contribution degree of each influential factor to train-MV accident occurrence through reverse inference. It is worth noticing that accidents caused by inappropriate motorist behavior contribute 80% to the entire train-MV accidents at SAL2 LXs, while accidents caused by static factors contribute only 17%. As for inappropriate motorist behavior, “Zigzag Violation" is more significant than “Stall on LX" in terms of causing train-MV accidents, because of the contribution of 58% (compared with 42% contribution of “Stall on LX"). On the other hand, in terms of static factors, when a train-MV accident occurs at an SAL2 LX, this LX has the probabilities of 74%, 38%, 44%, 37% and 46% respectively involved in the most risky situations that “Corrected Moment" in the “CM_49_up" group, “Railway Speed Limit" in the “RSL_110_up" group, “Width" in the “W_6_up" group, “Length" in the “L_11_up" group and “Region Risk" in the “R_high" group. These results indicate that more attention needs to be paid to LXs having the above risky characteristics. Moreover, special accommodation and/or technical solutions need to be implemented to prevent motorist zigzag violations. As for the consequences caused by accident, it is most likely to be 0 fatality (P (F = F_0) = 0.8875), less than 2 severe injuries (P(S = S_0_2) = 0.9789) and less than 3 minor injuries (P(M = M_0_3) = 0.9664). Thus, to a large extent, the consequence severity would be Level 1 (P (CS = $Level$_1) = 0.8396, P(CS = $Level$_2) = 0.0292, P(CS = $Level$_3) = 0.0181, P(CS = $Level$_4) = 0.0006 and P(CS = $Level$_5) = 0.1125).

6 Conclusions

This paper presents risk analysis relative to LXs based on recorded accident statistics. Various parameters have been taken into account in our study: the involved road transport mode, geographical regions and the traffic moment. The thorough statistical analysis allows us to identify the main risk factors and quantify their impacts on the overall LX risk. Although the analyses reported in this paper were based on the accident data of France and focus on the SAL2 LXs, The motorized vehicle is main transport mode causing accidents at SAL2 LXs, we focused on the studies related to train-MV accidents at SAL2 LXs using BNI-RR approach which offers an integrated modeling and analysis framework that allows for performing thorough risk analyses on a given LX or a set of LXs at a global level. The findings obtained through applying the BNI-RR framework on LX risk analysis offered a significant perspective on the major factors causing LX accidents and pave the way for identifying practical design. This work will serve as a basis for the ongoing SAFER-LC European project (www.safer-lc.eu), which aims to develop innovative solutions to improve LX safety.

Références

Berrado, A, EL-Koursi, EM, Cherkaoui, A, Khaddour, M, (2011). A Framework for Risk Management in Railway Sector: Application to Road-Rail Level Crossings, Open transportation Journal, Bentham Open, 19p, http://www.bentham.org/open/totj/,

ERA (2016). Railway safety performance in the European Union. Rapport technique, doi:10.2821/129870.

Ghazel.M, EL Koursi.E.M, (2014), Two-Half-Barrier Level Crossings Versus Four-Half-Barrier Level Crossings: A Comparative Risk Analysis Study, IEEE Transactions on Intelligent Transportation Systems, Vol5, issue4, p1123 - 1133, DOI : 10.1109/TITS.2013.2294874,

Khoudour.L, Ghazel.M, Boukour.F, Heddebaut.M, El Koursi.EM, (2009), "Towards safer level crossings: existing recommendations, new applicable technologies and a proposed simulation model", European Transport Research Review, Springer. (2009) 1: 35–45, DOI 10.1007/s12544-008-0004-z.

Liang.C, Ghazel.M, Cazier.O, El Koursi E. M. (2017). Risk analysis on level crossings using a causal Bayesian network based approach. Transportation Research Procedia 25, 2172-2186.

Liang.C, Ghazel.M, Cazier.O, El Koursi E. M. (2018a). Developing accident prediction model for railway level crossings, Safety Science 101, 48-59.

Liang.C, Ghazel.M, Cazier.O, El Koursi E. M. (2018b). A new insight on the risky behavior of motorists at railway level crossings: An observational field study. Accident Analysis and Prevention 108, 181-188

Liang.C, Ghazel.M, Cazier.O, El Koursi E. M. (2018c). Analyzing risky behavior of motorists during the closure cycle of railway level crossings, Safety Science. https://doi.org/10.1016/j.ssci.2017.12.008

Plesse .G (2017). Des détecteurs d'obstacles déployés aux passages à niveau. Rapport technique, France. From http://www.leparisien.fr/info-paris-ile-defrance-oise/transports/des-detecteurs-d-obstacles-deployes-aux-passages-a-niveau-02-06-2017-7011714.php.

SNCF Réseau (2011). World Conference of Road Safety at Level Crossings. Rapport technique, France. From http://www.planetoscope.com/automobile/1271-nombre-de-collisions-auxpassages-niveau-en-france.html.

SNCF (2015). Research on the material of level crossing in 2014. Rapport technique, France.

Résumé

Chaque année plus de 300 personnes sont tués aux passages à niveau (PN) en Europe. Les passages à niveau sont identifiés comme point faible pour la sécurité ferroviaire. L'objet de ce papier est l'analyse quantitative des risques et les techniques de modélisation dans le but d'améliorer la sécurité aux passages à niveau. Une analyse statistique quantitative sera présentée sur l'impact de divers facteurs (mode de transport, région géographique et le moment de trafic) sur le niveau de risque aux PNs. À travers cette analyse, le principal mode de transport (véhicule motorisé) responsable des accidents aux PNs. Un modèle bayésien a été proposé pour consolider l'analyse quantitative.

C-T-Engine : A Real time building engine of urban traffic congestion trajectories

Mohamed Nahri*, Azedine Boulmakoul*, Lamia Karim**

*LIM/IOS., FSTM, Hassan II University of Casablanca, B.P. 146 Mohammedia, Morocco
**Higher School of Technology EST Berrechid, Hassan 1st University, Morocco

Abstract. Urban traffic congestion poses big challenges for governments, industrials and scientists. These challenges concern various related aspects such as representing, detecting, analyzing and acting immediately to reduce its worse effects. Boulmakoul team has introduced an innovative meta-model representing congestion, considering it as a trajectory of a set of linked congestion complex events. This paper tends to extend this work by developing an engine architecture for building congestion trajectories. Thus, we highlight congestion trajectory meta-model in which we explain congestion event and congestion trajectory meaning and patterns. Furthermore, we propose a framework that we call C-T-Engine for real-time congestion trajectories building as well as their storage and their visualization. Finally, we present deployment and first tests of the developed framework by using events retrieved from road traffic simulator.

1 Introduction

Road traffic congestion in large cities represents a scourge characterizing citizen's daily life. This phenomenon impacts varied sectors such as the economy, healthcare, and environment. Thus, urban traffic congestion has been among must interests of the governments, industrials and scientists. In fact, traffic congestion occurs when vehicular flows exceed the road capacity, starting in a specific part of the road and spreading over a wide area. Different types of solutions have been adopted to deal with this problematic, among them we find the augmentation of road capacity, the invention of new transportation means and finally the use of artificial intelligence. Actually, the emergence of the internet of things (IoT) and Big-Data analytic ecosystems favor the adoption of new solutions for real-time detection of road traffic anomalies from microscopic events generated by connected objects including smart vehicles. Indeed, the evolution of the ubiquitous technologies of communication such as RFID, Zigbee, Bluetooth, WSN (David et al., 2017) and other practices have led to the emergence of the machine-to-machine communication. Moreover, the evolution of long-term communication technologies such as internet technologies, cellular technologies and MAN networks, particularly in smart cities, adding to the evolution concerning the connected vehicle, have led to the apparition of a new concept so-called internet of vehicles (IoV)(Sun, 2013). In fact, IoV supports different types of vehicular communication as well as real time vehicular data centralization, allowing performing further data traffic analytics thanks to the evolution of Big-Data analytics technologies.

All these technological evolutions have opened a new perspective for treating traffic congestion problematic in a centralized manner. In particular, these technologies allow detecting

and acting immediately for unlocking bottlenecks and regulating traffic circulation. However, any solution of this kind faces many challenges, including realizing an advanced representation of traffic congestion, aligned with the powerfulness of the mentioned technologies. The issue of representation congestion has been a big problem regarding the dynamic aspect of the phenomena and the difficulty to evaluate its magnitudes. In order to overcome this issue, Boulmakoul team (L. Karim et al., 2017) has introduced a new meta-model representing traffic congestion based on trajectory of events. Indeed, the proposed meta-model presents an advanced representation of congestion aligned with both the road traffic characteristics and the technological advances in IoT and Big-Data analytics. Furthermore, it allows tracking traffic congestion evolution in a detailed manner. In fact, the key idea behind this representation is to consider congestion as a set of successive events regrouped in a trajectory. Every congestion event represents traffic congestion occurring in a specific section of the road.

Moreover, another challenge to overcome concerns the design of a technical architecture for real-time traffic data collection and analysis. This architecture must justify high flexibility and openness to support the variety of information sources, high performance, low latency and support of large quantities of data. Furthermore, several technologies and architectures have been proposed for real time analytics, including stream processing and complex event processing (CEP) (Buyya et al., 2016). Thus, These solutions present a great support for real-time analytic systems allowing detecting patterns and targeted events.

Throughout this work, we propose (C-T-Engine) an engine for real-time congestion trajectories building from congestion events. The rest of this paper is organized as follow: Section 2 gives an explication of congestion events and congestion trajectories meaning and representation. Section 3 describes the developed C-T-Engine in its first version. First test and results of this solution are presented in section 4. Then Section 5 concludes this paper, presenting the advantages and the perspectives of the given work.

2 Congestion events and congestion trajectories

In this chapter, we explore congestion events and congestion trajectories detailed in (L. Karim et al., 2017) work tending to simplify their meaning and their representation, while converging to our purpose concerning the design and the development of an engine for building congestion trajectories from congestion events.

In fact, congestion event (CE) represents a traffic jam occurring at a given section in the road network in a given moment. Thus, congestion event is known for its location, start time and end time. Moreover, this event, occurred in a specific space and time, can have an effect on other related links in a different space and time. As can be see, the congestion phenomenon results from a trigger event and it propagates in space and time touching a wide road area. Indeed CE is characterized by the link in which it occurs and its start and ending time. Figure 1 shows an instance of linked congestion events occurred in Zektouni Boulevard in the city of Casablanca, Morocco in which CE1 represents the trigger event and CE2, CE3 and CE4 are immediate results of this one. The challenge that has been posted is to master and represent the whole of related events in unique model, knowing that events occur in delayed moments and different spaces.

FIG. 1 – *Linked congestion events occurred in Boulevard Zerktouni.*

The cited Boulmakoul team work has proposed an innovating representation mastering all events based on trajectories. Congestion trajectory (CT) contains a set of linked events including trigger and resulting events. For example, let consider events represented in figure 1. A simplified representation of congestion events (CEi) and CT is shown as follows: CE1(link1,ts1,te1), CE2(link1,ts1,te1), CE3(link1,ts1,te1) and CT(id,[CE1,CE2, CE3,CE4]). Note that this representation is just for simplifying and more detailed information characterizing CE and CT will be given in the rest of this paper. Therefore, CT results from a CE trigger CE1, that represents a head of the trajectory and it propagates causing other CE such as (CE2) located in the bottom of trajectory and (CE3,CE4) considered as the ends of trajectory.

However, congestion trajectories have a different concept with classical trajectories due to the dynamicity of elements that constitute it. In the most of trajectory models (Mazimpaka et al., 2016), changes are coming from adding new points with some static information to the set of items forming trajectory. Although, adding to this characteristic, all events in CT can change their state continuously. The later characteristic adds more complexity to the mentioned model. This complexity involves the CT representation and evaluation and the data update, storage and analysis.

For mastering these complexities, we consider that each CEi presents an active element in a dynamic CT. Therefore, CE can change its state at each moment. Effectively, CEi is created after a traffic jam showed in a specific link, integrating a specific CT and further it can change its state from congested to fluid and vice versa while staying in the same CT. Furthermore, we introduce a simplified life cycle of CT (see figure 2) presented as follows:

- Creation: corresponding to the birth of trajectory, occuring when a CE trigger has appeared representing the head.
- Update: we distinguish between two types of updates. The first concerns adding a new CE to the CT. The second concerns the update of an included event state when a change in state has occurred.
- Destruction: occurs when all included events become fluid.

FIG. 2 – *Trajectory lifecycle.*

For consolidating and schematizing these ideas, we present the activity diagram shown in figure 3. It should be noted that we assume that an event corresponding to a link situation is evaluated as congested or fluid only after detection.

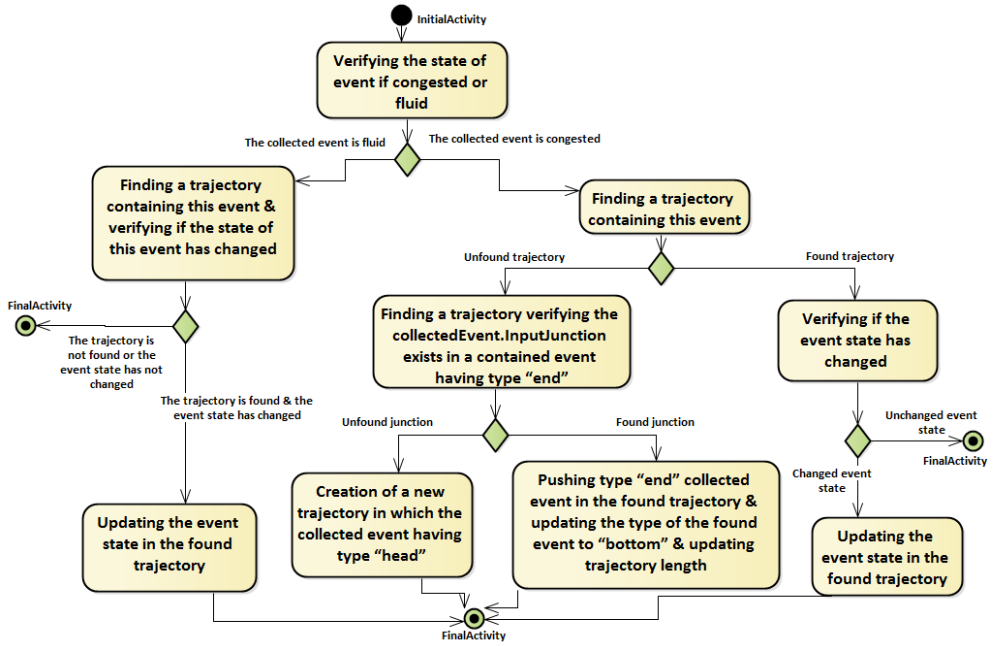


FIG. 3 – *Activity diagram for creating and updating congestion trajectory.*

Note that the creation and update of CT are performed according to activity diagram up cited. However, the destruction of CT is performed according to an interval batch verifying if all events in such trajectory are fluid.

We mention that the size of effect of the occurred CE can change depending on the type and the importance of the infected link. For instance, a link can be owned by an arterial, collector or a local road. This idea is more developed in (L. Karim et al., 2017) work. Moreover, two operators are defined for concatenating and assembling CTs according to their types. The first operator concerns concatenating two CTs with the same road type. The second consists of aggregating CTs with different road type. However, these operators are taken in this work with their simple form allowing concatenating CTs regardless of their types. Thus, two CTs can be concatenated if an end of one is linked to the head of the other.

In this work, we take the simplest definition of CE and CT regardless of the details about road category. Effectively, we consider CE as an event which occurred in a link between two junctions in which travel time (Falcocchio et al., 2015) is too high. We assume that a link is congested when TTI is over than 1.33. This means that congestion occurs in a link when its travel time is 33 percent longer than the travel time in free flow. We consider travel time in free flow equal the minimum travel time that can be shown in the link corresponding to the maximum speed.

3 Congestion trajectories building engine

In this section, we present the congestion trajectories engine C-T-Engine. The later reveals another delicate concept concerning Big-Data real time analytics. Moreover, we introduce some difficult points characterizing our specific issue.

3.1 Real time analytics generalities and issues

In Big-Data era, real time analytics of voluminous data presents a relevant and delicate concept. Indeed, it allows the detection of useful meaning in a real time after data reception, giving the possibility to act immediately to the target situations. Thus, added to voluminous and variety aspects of BigData, this field of data analytics takes in consideration the velocity aspect, considering the time of data analysis as an important factor. However, real-time factor remains confused depending on the application domain and the usefulness time of the meaning wished to achieve. Generally, in BigData field, this factor presents the ability to analyze data as they arrive (Buyya et al., 2016). Particularly, in road traffic management field, there is a great need to minimize time of the detection of traffic anomalies, and therefore reacting immediately. Moreover, in our context concerning detecting and building congestion trajectories from received traffic events, the analytics time has to be minimalist, if we consider the time spent in data communication and events detection.

Furthermore, in real time Big-Data analytics, two key concepts are distinguished which are stream processing and CEP. Stream processing concerns the analysis of unbounded data continuously as they arrive, and CEP focuses on detecting target patterns from correlated events (Flouris et al., 2017). In our context, the proposed architecture requires a combination between stream processing and complex event processing systems. In fact, the stream processing is needed for collecting, filtering, categorizing and windowing events arriving from event generators. In addition, the CEP is used for detecting and creating patterns referring to CT.

Moreover, Several technologies and architectures are proposed to perform stream processing and complex event processing in a Big-Data context, among them Lambda and Kappa architectures (Azarmi, 2016). Lambda architecture provides near real time processing architecture reposing in three main layers (batch layer, speed layer and serving layer) tending to ensure fault-tolerance and law latency. In fact, this architecture combine batch processing and real time processing working in parallel and performing the treatments on the same data with two different manners. It should be noted that this architecture constitutes an interesting transition between batch processing and real-time processing. Kappa architecture is an evolved form of lambda architecture eliminating batch processing and performing direct stream processing.

CT building architecture presents more difficulty referring to the continuous and high number of complex CT creation and update. Especially, the updates concern adding new CE to a related CT and changing the state of events in CT as well as verifying the state of CT for a potential destruction. It should be noted that we tend to minimize the number of creation of trajectories. Furthermore, every change in a CT must be stored for a further and advanced analysis purpose. Therefore, we propose to store an instance of every CT new version in a Big-Data database. Thus, the architecture and technological choices must present a high efficiency, low latency, high availability and scalability supporting large volumes of data.

3.2 Proposed architecture and technological choices

In this part, we propose the description of the proposed C-T-engine detailing its components and showing its interactions with other systems. Figure 4 provides a global vision over all components of C-T-Engine and other external systems such as IoV ecosystem and events detection framework.

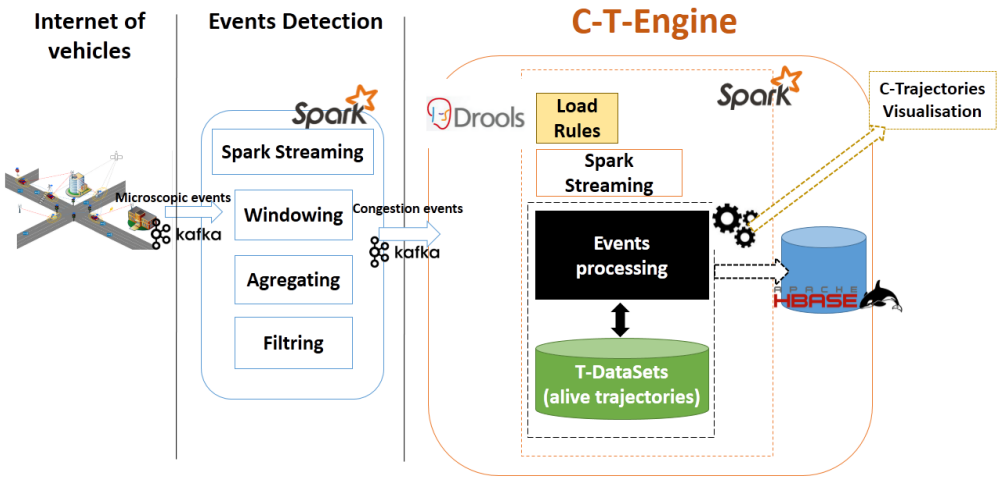


FIG. 4 – *Congestion trajectories building engine.*

Adding to C-T-Engine, the figure above shows an overview of the global architecture including IoV and a framework for event construction. The latter is not subject of this work. However, it role present a great importance. In fact, this framework allows receiving microscopic events from connected vehicle in IoV and performing real time analytics for finally providing congestion events presenting traffic situation in each road section. In the rest of this paper, we put more focus on C-T-Engine.

Comparing to Storm and Flume, Spark shows more latency in data treatment (Chintapalli et al., 2016).Furthermore, spark provides a full processing engine integrating streaming, hdfs, SQL engine and machine learning algorithms. In C-T-Engine, adding to spark streaming we excessively use SparkSQL, especially Datasets and Dataframes (Mazimpaka et al., 2016) providing a rapid access needed data CT building. Furthermore, Referring to the complexity of the rules governing CT construction, the use of rules management engine is essential.

Drools (Salatino et al., 2016) presents one of the most famous rules engines. Moreover, drools integrates time factor in reasoning process. Integrating rules engine like Drools in C-T-Engine allows us to have more clarity over the rules governing trajectory building process as well as efficiency. In addition, the developed rules are deployed and charged only one time at the beginning of the program having no influence in latency.

Kafka is a message broker supporting large data quantities. Comparing to other message brokers such as RabbitMQ and ActiveMQ, Kafka exceeds them in terms of the supported amount of messages (John & Liu, 2017). Moreover, Kafka works under zookeeper orchestrating distributed systems and supporting high performance and high availability.

Compared to the Big-Data Nosql databases(Corbellini et al., 2017), the choice of Hbase is justified by including a key characteristic that we excessively use consisting in managing versions of trajectories. Thus, it allows us to store all the history of trajectory from its creation until its distraction.

For performing first tests of C-T-Engine, events are retrieved directly from a road traffic simulator. An example of event is shown in the next part.

3.3 Events processing flow

In this part, we describe the processing steps starting from retrieving event from kafka broker after creation of congestion events until (see figure 4) creation, update and destruction of CT.

- Preprocessed events are retrieved from Kafka (see figure 4) informing about time, link characteristics and tti.
- These events are retrieved by spark streaming and filtered according to their tti for identifying the state of each event (congested or fluid). An instance of congested event is shown as follow:

{"id":"e- hassan2_10@20180219192045",
"time":"2018-02-19T19:20:45+00:00",
"tti":"1.98",
"state":"congested",
"tposition":"head",
"link":{
"id": "hassan2_10",
"length": " 126,93",
"juctionFrom": "hassan2-ruedesouss",
"juctionTo": "hassan2-coline",
}
}

- CTs are built and updated following the activity diagram shown in figure 3. It should be noted that each trajectory creation or update involve storing an instance of trajectory adding to the list of trajectory versions in trajectories HBase database. Moreover, the latest version of active trajectories is charged from database at the first of each batch performed by spark streaming. An instance of CT is shown as follow:

{ "id":"e- hassan2_10@20180219192045",
"length":"126,93",
"events":"[{"id":"e- hassan2_10@20180219192045",

```
"time":"2018-02-19T19:20:45+00:00",
"tti":"1.98",
"state":"congested",
"tposition":"head",
"link":{
        "id": "hassan2_10",
        "length": "101.23",
        "juctionFrom": "hassan2-ruedesouss",
        "juctionTo": "hassan2-coline",
       }
}]
 }
```


- The destruction of trajectories are performed according to a batch program launched periodically verifying if all events in each active trajectory are fluid.
- Another batch program is launched after each batch processing performed by Spark and it concerns the concatenation of trajectories.

We recall that C-T-Engine integrates an important point concerning saving all updates of each trajectory allowing to have all the story of each trajectory from creation to destruction. Thus, this functionality will allow performing a further deep analytics over the trajectories evolutions. However, the read option from database is less used, notably at the beginning of each batch processing and each destruction program.

4 Deployment, tests and first results

The components of the developed architecture have been deployed in separate machines i5 with 8Go of RAM working under Debian OS according to the deployment diagram shown in figure 5. Note that the solution is not yet clustered in its first version.

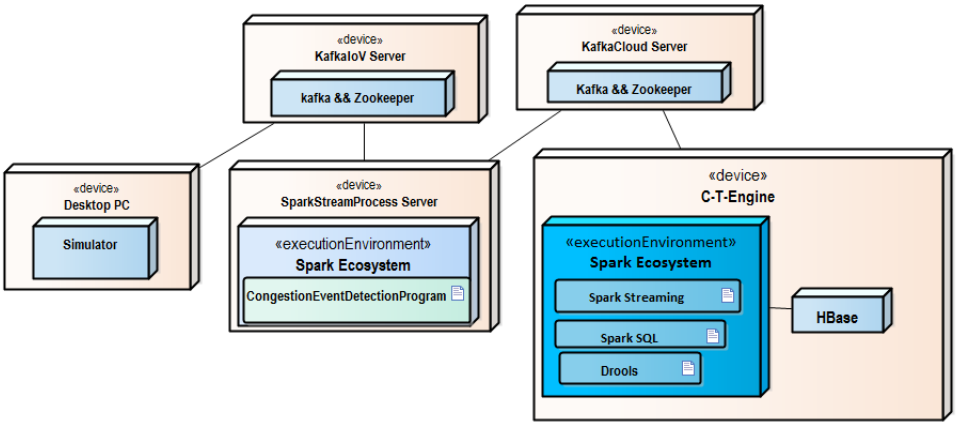


FIG. 5 – *deployment diagram.*

Figure 6 shows the variation of the length of three trajectories. As can be seen, CT1 is shown with all its cycle of life from creation to distruction. Moreover, CT2 is shown at the destruction phase and C3 is shown at the creation and evolution phases. Note that the length of CT represents the sum of events length with the congested state.

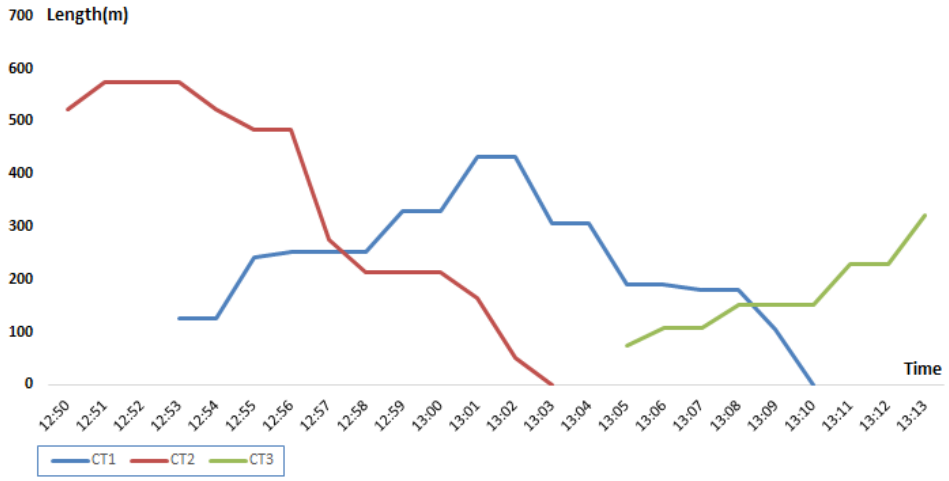


FIG. 6 – *Trajectories length evolution.*

5 Conclusion

This work presents C-T-Engine, as an engine for building congestion trajectories, on its first version. The developed engine is based on most recent technologies that have proven their performances. Moreover, we add an important aspect to the model of congestion trajectories concerning their lifecycle, allowing separating active trajectories from the passive ones. Furthermore, the architecture of C-T-engine permits the saving of all the history of each trajectory, allowing further and deep analysis of the evolution of congestion trajectories. However, to deduce the limits of the proposed architecture, we will test it in an intensive mode.

References

Azarmi, B. (2016). *Scalable Big Data Architecture*. https://doi.org/10.1007/978-1-4842-1326-1

Buyya, R., Calheiros, R. N., & Dastjerdi, A. V. (2016). *Big Data: Principles and Paradigms. Big Data: Principles and Paradigms*. https://doi.org/10.1016/C2015-0-04136-3

Chintapalli, S., Dagit, D., Evans, B., Farivar, R., Graves, T., Holderbaugh, M., … Poulosky, P. (2016). Benchmarking streaming computation engines: Storm, flink and spark streaming. *Proceedings - 2016 IEEE 30th International Parallel and Distributed Processing Symposium, IPDPS 2016*, 1789–1792. https://doi.org/10.1109/IPDPSW.2016.138

Corbellini, A., Mateos, C., Zunino, A., Godoy, D., & Schiaffino, S. (2017). Persisting big-data: The NoSQL landscape. *Information Systems*, *63*, 1–23. https://doi.org/10.1016/j.is.2016.07.009

David Hanes, Salgueiro, G., Grossetete, P., Barton, R., & Henry, J. (2017). IoT Fundamentals_ Networking Technologies, Protocols, and Use Cases for the Internet of Things-Cisco Press (2017). Cisco Press.

Falcocchio, J., & Levinson, H. (2015). *Road Traffic Congestion: A Concise Guide*. *Springer Tracts on Transportation and Traffic* (Vol. 7). https://doi.org/10.1007/978-3-319-15165-6

Flouris, I., Giatrakos, N., Deligiannakis, A., Garofalakis, M., Kamp, M., & Mock, M. (2017). Issues in complex event processing: Status and prospects in the Big Data era. *Journal of Systems and Software*, *127*, 217–236. https://doi.org/10.1016/j.jss.2016.06.011

John, V., & Liu, X. (2017). A Survey of Distributed Message Broker Queues.

L. Karim, A. Boulmakoul, A. L. (2017). Real time analytics of urban congestion trajectories on Hadoop-MongoDB cloud ecosystem. *Acm*, (August). Retrieved from https://www.researchgate.net/profile/Azedine_Boulmakoul/publication/315664262_Real_time_analytics_of_urban_congestion_trajectories_on_Hadoop-MongoDB_cloud_ecosystem/links/599b293a45851574f4ac6653/Real-time-analytics-of-urban-congestion-trajectories-on-Had

Mazimpaka, J. D., & Timpf, S. (2016). Trajectory data mining: A review of methods and applications. *Journal of Spatial Information Science*, *13*(13), 61–99. https://doi.org/10.5311/JOSIS.2016.13.263

Salatino, M., Maio, M. De, & Aliverti, E. (2016). *Mastering Jboss Drools 6*.

Sun, W. (2013). *Internet of Vehicles*. *Advances in Media Technology*. https://doi.org/10.1007/978-3-319-11167-4

Distributed and scalable framework for Smart city Real-time Complex Event Processing

Wadii BASMI*, Azedine BOULMAKOUL*

* LIM/IOS, FSTM, Hassan II University of Casablanca, Mohammedia, Morocco,
{wbasmi@gmail.com, azedine.boulmakoul@gmail.com }

Abstract. The large scale development in the information technology industry has inspired many researchers around the world to have it serve the daily life of citizens in civilized societies. Smart city is a term given to a city that make use of innovative approaches to help its citizens reaches its vital services, thus increasing their quality of life and the plus-values they give to society, leading to a positive demographic impact, relying on Internet of things technologies to help achieving the above objectives, by building infrastructure using both software and hardware such as electronic sensors and large servers to help managing *"things"*. The diversity of platforms within one city often leads to incompatible systems that are unable to communicate, due to the specific layer that interacts with hardware, therefore there is a need for a shared solution that extends these platforms by collecting data from things, additionally, by being a proxy between these platforms and any other IoT platform that make use of real life data. This paper provides an explanation about building an open platform for collection of what we refer to as a space time complex data; a data that represents all the events that might occur within the perimeter of a city.

1 Introduction

The scientific field research has achieved their pick in the late millennium, with the excessive number of experiences to extract data and to describe a phenomena with beautiful mathematical equations. Furthermore, science helped humanity to achieve a high level of civilization, in addition to a greater sense of well-being through the accessibility to data.

Likewise, governors, decision makers and the private sector in many cities realized the importance of upgrading the infrastructure, embracing new technologies and working together to surface solid strategies to bring out a diversity of data. In fact, in the late decade, it has become quite normal to track Internet users actions to control the ads they see, as result, the companies become able to reach their target audience with a minimal cost. Thus, in order to make a city "smarter", all the involved parties must invest in innovative solutions that gather data all over the places, for example, it related to citizens behavior, environment or its infrastructure, through, IoT (Internet of things) technologies.

Smart city IoT platforms use software and physical infrastructure (sensors, actuators, etc.) to solve specific use cases, such as resources optimizations or reducing health threats, easing daily life actions and uncountable problems that a citizen may encounter. Yet, these solutions are closed; their connectivity is exclusive to their inner components. In fact, almost all IoT solutions were not designed with the possibility to cooperate within themselves, by sharing their data and the results they have gathered, and thus adding more complexity for the decisions makers to gather data from different sources and shape them to act. Therefore, the lack of connectivity between IoT platforms causes unnecessary, considerable additional costs.

This paper provides an architecture of a central independent open bridge system between any potential IoT solution; *EventX* increases the connectivity within a smart city; it is a highly scalable distributed ecosystem that connects sensors and actuators with a high level of abstraction in its context, to exchange three-dimensional data (space-time events) collected and projected onto a city. Accordingly, electronic sensors and actuators, IoT platforms and any system that can take data as input or that outputs it, can utilize our solution to exchange data in ease. Hence, in the next sections, we discuss the benefits of IoT systems to make a city smart, then, we explain the concept of complex space time events and how we intercept them, later, afterward, we surface the EventX logical architecture, the platform that manage the interactions between all the things that generate and seek these events in a microservices ecosystem, and at last, we show how we used docker as a containerization solution to wrap the different microservices in an abstract scalable ecosystem.

2 The benefits of IoT platforms

Medellin in Colombia, Seattle in the United States or Milan in Europe are cities that brought *Smart city* concept to life, through many factors that involve technology, the social capital, governance, improvement of vital sectors (education, health care, etc.), economy, natural environment sustain and a quality infrastructure. In fact, the collaboration of many parties in these cities to come up with a diversity of strategies was the main reason of their success, that emphasis the usage of new technologies to serve as a backbone of the principles of their future projects and programs. In addition, the main concern was to involve public and private organizations into these projects, in order to speed up the trip toward a smarter city [Milla 2016].

Making a diversity of IoT platforms that targets vital sectors is a top priority for any decision maker. Accordingly, they planned, many researches, projects to cover health care, circulation traffic, environment monitoring and street surveillance, etc. Moreover, more private organizations started to invest in innovating more IoT solutions.

2.1 Smart health care

To illustrate, one of the many IoT platforms: *iHome health care IOT system*. In fact, it is an intelligent medicine box (iMedBox) that serves as a home healthcare gateway connected to medical devices. Additionally, the solution provides a body-worn Bio-Patch that detects and

transmits the user's bio-signals to the iMedBox in real time. Moreover, the iMedPack collects data and display them on the iMedBox. At last, as a major contribution, it virtually brings the whole hospital environment to the patient house. In other words, the patients have easy access to their health status data using the iMedBox G. Yang 2014.

2.2 Smart parking

Likewise, Parker is an IoT solution for real-time parking availability application that offer the citizens a map showing all the available parking places. Furthermore, Parker makes it possible to reserve a parking slot in advance, a highly desirable feature during peak hours. Importantly, it helps citizens to avoid wasting time in searching for unavailable slots. In summary, The research process includes the spot, the city, the reason behind reservation and its duration Elena 2013.

2.3 Smart security

As a last example, Fallcare+; an IoT system designed to detect falls recorded by homeboxes in form of a Web camera and Raspberry 2. Moreover, whenever a fall happens, the homebox sends a dedicated event to the system. Additionally, Fallcare+ comes with real-time video streaming and applies deep learning to predict falls beforehand using the data it collected so far under given circumstances Charles 2017.

The race toward innovative IoT solutions had already started and is still on Sergio 2016, however, researchers and companies development is constantly changing, but it proved that it was worth the investment in order to alter citizens daily routines and making their life smarter. No wonder, the biggest companies have already seen through the shell - Amazon and Microsoft - by introducing their own IoT solutions.

2.4 Big firms IoT solutions

Since 2015, Amazon and Microsoft were putting so much efforts as competitors to come up with their own IoT solutions. Amazon has now a product called AWS IoT services; it is a product presenting a set of services, each service serves a purpose. In details, it lets connected devices interact with cloud applications and a variety of devices securely. In addition, It is highly scalable and supports a big number of devices and messages. Furthermore, AWS IoT gives the possibility to monitor and manage IoT devices remotely. Also, it is possible to run analytics on the collected data. The best part about AWS IoT is that it comes with an operating system designed for micro-controllers that is shipped with the whole pack. At last, there's a service called AWS IoT 1-click that make it possible to tie some events to an action Sajee 2014.

A great amount of successful IoT platforms that changed the way we see life. However, unlike the precedent solutions, EventX adds another layer of abstraction to the concept of "sensor" and "actuator" as emitters and source of useful events that holds data, regardless of the nature - physical or virtual - of the "thing" they represent. Furthermore, it labels these events as complex space-time events, which will be detailed in the next section.

3 Complex space time events

3.1 Key concepts of complex space time events

The first step toward having a shared concept between all existing platforms, starts by defining the common properties of measurable events related to a phenomena in a smart city. As a result, the citizens quality life improves.

In EventX's context, a sensor observes a phenomena and emits data. For example, it can be an electronic Arduino board (temperature, air quality, traffic light), a computer or a service deployed on the cloud. Furthermore, the recording time, the geographic coordinates and the measurement - that is not necessarily related to a natural phenomena - are the identifier of the collected event. Therefore, it is three-dimensional. Thus, complex space-time events can be train stations' schedules, traffic jams' frequencies, air quality indicators, regional crime rates and others. To illustrate, figure 1 shows a variety of *things* interacting with EventX.

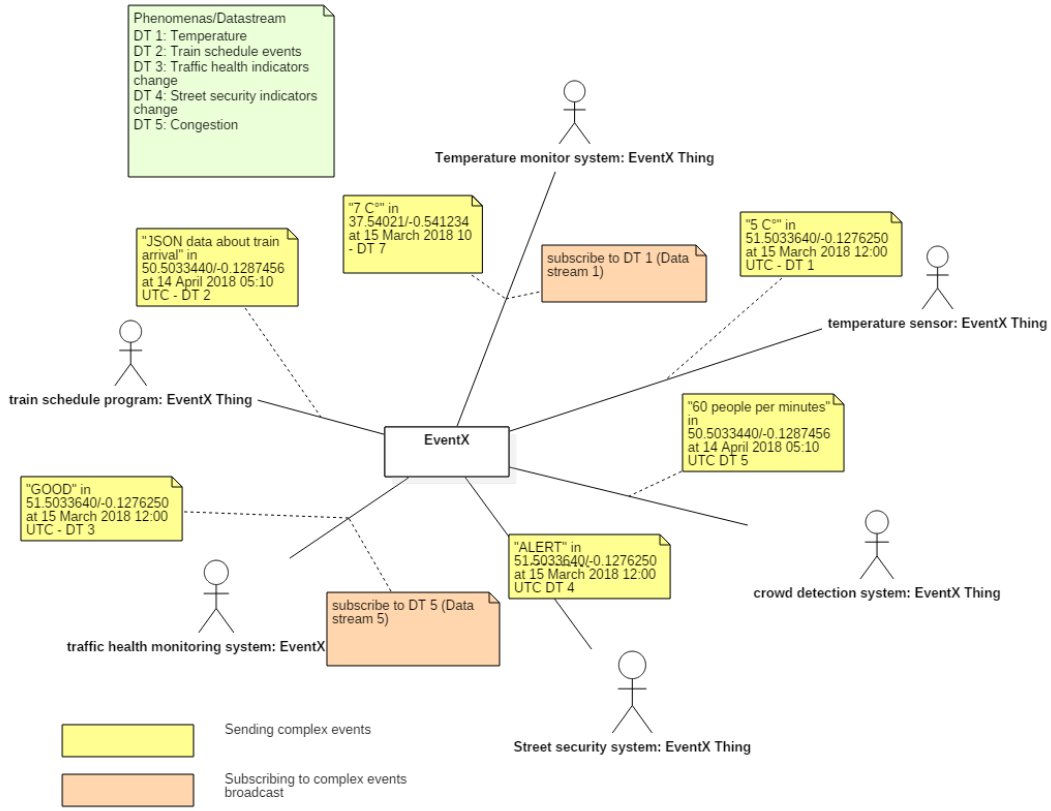


FIG. 1 – *External actors interacting with EventX system*

3.2 EventX data model

Figure 2 describes the data model of EventX database, it was inspired by OGC Sensorthings data model that is currently an OGC standard. On the other hand, the database is powered by MongoDB; it is a document store based database management system that guarantees consistency (C) and partitioning (P) tolerance through its replica-set model, where there is only one master node that writes data to documents, whereby all the others nodes called *secendary nodes* can only read it, in fact, when the primary node fails, MongoDB elects one of the secondary nodes to take its place. At the time the write node is down in the election process, any write requests are simply put aside, therefore, it drops availability (A) Kyle 2012.

The data model shows the classes of data stored in the database, a data stream is a set of observations that bring togheter the *EventX* things, in details, when a sensor – it can be a physical android board, a Raspberry device or an IoT system – located in a specific position sends an observation, that can be a measurement or a word describing an event, in addition, it has a content type to distinguish between plain text and formatted data, for example: JSON, XML or YAML data, the time when it has been recored, and the unit of measurement that is optional. Moreover, *feature of interests* are the subjects of data streams, it has a name and a geographic feature to keep the choices over its geometry open. Furthermore, the data stream describes a phenomena, for example: air quality, temperature, congestion, train schedule delays or frequency of accidents of some sort, again, the choices are endless as long as it can be source of events. On the other hand, a data stream object has two sets of subscribers:

— Actuators that listens on upcoming events;
— Sensors that emits events;

In this context, the data stream plays the role of a channel that pipes many observations, collected and requested by different sensors and actuators. Moreover, sensors and actuators are EventX things, they have a name and a description, in details, a sensor observes changes of a thing's state and can transform that event into a digital information, additionally, the type can be the family reference of an Arduino board or an IoT platform category. Whereas, an actuator is an entity that receives digital information and transform it into physical or digital form to perform a specific task.

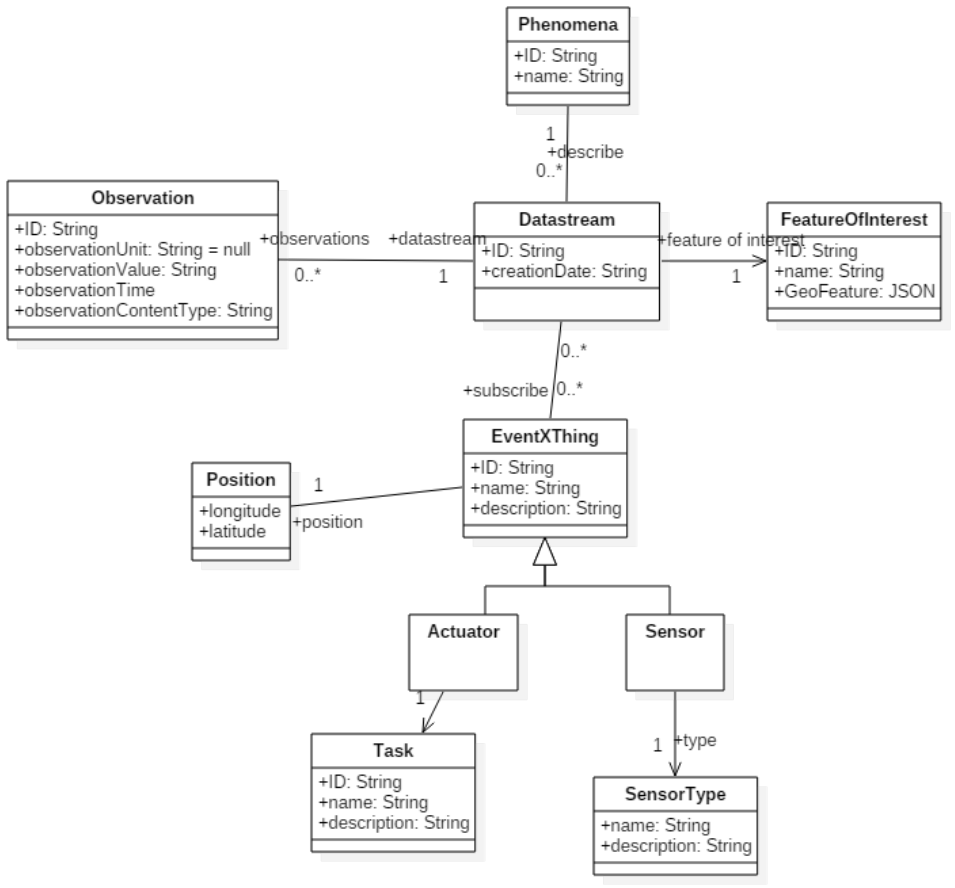


FIG. 2 – *EventX data model*

4 EventX logical architecture

4.1 Benefits of microservices over monolithic systems

Huge monolithic systems era has reached its ice age. In fact, software companies and the biggest firms are starting now to adopt the microservice approach to design their systems' architectures. However, it can be quite troublesome to transit from one way to another. Therefore, it is always necessary to draw a clear architecture of the system we want to build and forecast the out-coming challenges it may face in terms of efficiency and possible features Sam 2016.

Microservices are small, autonomous services that work together. Additionally, they help to set boundaries between business domains within the architecture. Moreover, they enriches the

technology heterogeneity, such as using *Node.js* for highly traffic consuming applications and Scala for intensive computing tasks in the same platform. In addition, one of the key benefits is resilience. In fact, if one microservice fails the others remain intact as they are independent entities. Above all, the system is less problematic to extend unlike monolithic architectures Sam 2016.

4.2 Core microservices of Eventx

Undoubtedly, Eventx is a built-on top of three microservices:
— Data gatherer;
— Data broadcaster;
— Authenticator;

These three microservices interact choreographically. In fact, they don't communicate directly, but instead, through a message broker. To explain, the role of a message broker is to act as a mail box for a microservice. Moreover, the interaction remains asynchronous unlike the traditional request-response model. To illustrate, figure 3 illustrate the message broker protocol *AMQP* implemented by RabbitMQ. Accordingly, producers sends data to exchanges, consumers subscribe to exchanges and get the data through named or unnamed queues. Moreover, RabbitMQ takes care of load balancing between consumers through Round-robin dispatching Phillipe 2017.

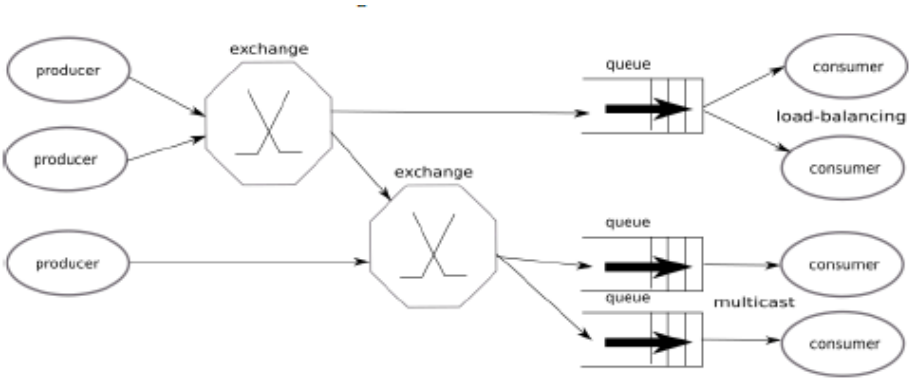


FIG. 3 – *RabbitMQ message broker Phillipe 2017*

Accordingly, EventX is a choreographic microservices system – see figure 4 – that uses RabbitMQ as a message broker. In details, each one is running independently of the others, and consider its inputs coming from anonymous instances. Furthermore, the *'data gatherer'* and *'data broadcaster'* microservices uses COAP and HTTP protocols to interact with incoming requests of EventX things. Moreover, the application server of both of them runs on *Node.js* process workers. In particular, Node.js uses an event driven architecture – it is a pattern where a system reacts to events – along it's asynchronous nature, in fact, a Node.js process is monothread, therefore, it is not possible to utilize a thread API to execute multiple tasks – it can be

achieved by rather having a pool of processes as workers – on multiple processors or cores, however, it reacts to the upcoming I/O events and execute an event handler on the main thread. Accordingly, EventX get these benefits, as it is technical challenge is to handle a huge number of requests in a short amount of time without latency, additionally, it doesn't have any high computation functionality (yet), therefore, it is a safe choice.

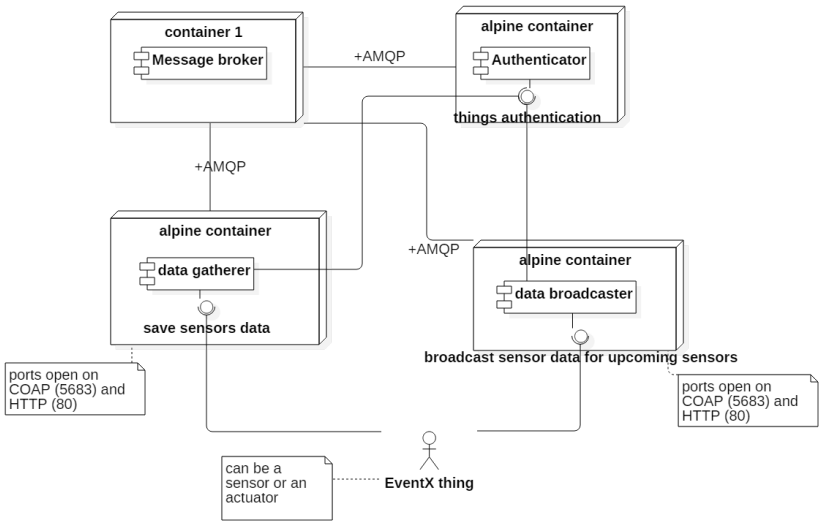


FIG. 4 – *EventX logical architecture*

The data gatherer and the data broadcaster are microservices that plays the role of public interface of EventX. In details, the gatherer collects data from EventX sensors that request saving data to a specific data stream, afterward, they hand it over to the broadcaster to publish it to all actuators subscribed to that data stream. In fact, they are implemented in Node.js, and listen on default ports of HTTP (80) and COAP (5683) each in their dedicated hosts. Additionally, in order for any of these two microservices to communicate, it is not recommended for any of them to access the other's data layer. In fact, it is important to remember that the microservices comes with a layer that hides all the complexity behind the exposed services provided, therefore, communication should go through the message broker.

On the other hand, we chose to have a request-response communication mode between the microservices and the authenticator. EventX uses JWT (Json Web Token Open standard RFC 7519) protocol to authenticate sensors and actuators; each time a sensors sends a request it needs to provide a *token* in the authorization header of the request, so it can access the resources with the rights they have.

To illustrate a scenario of how data circulates between different actors in EventX, the figure 5 describes the scenario that happens when a sensor tries to send its data from the first

time. In fact, the communication goes through the message broker to keep the microservices independent from each other.

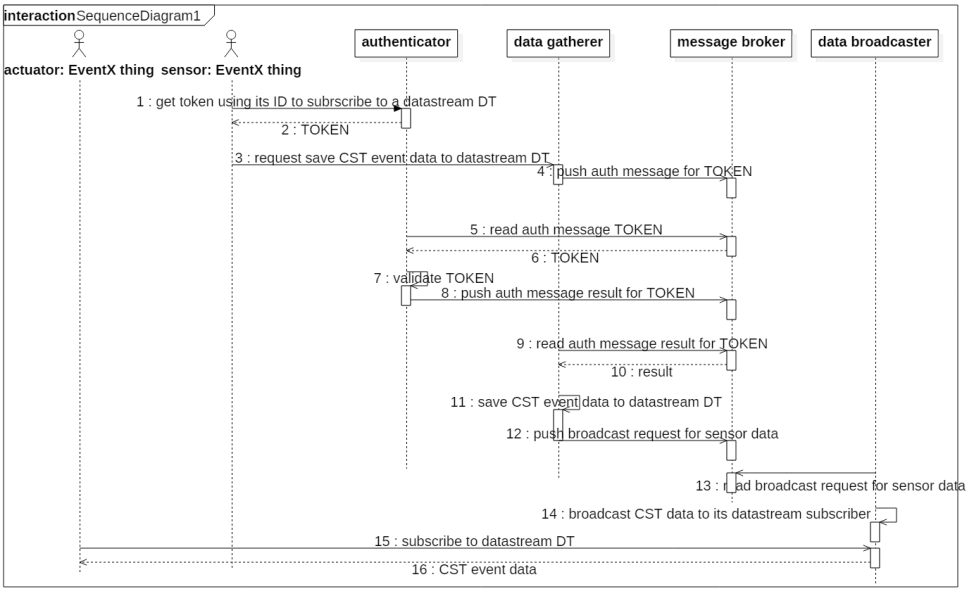


FIG. 5 – *Scenario for saving an event in EventX*

5 EventX artifacts deployment

5.1 Docker the containerization technology

Docker is one of the greatest revolutionary technologies ever made in the software development field. In particular, it is a platform that runs many applications packaged in so-called *containers*; they are created from *images* that are snapshots of other application in a known-state. For example, a software that uses a SQL database can run on-top of a container and store its data on a pre-configured MySQL database that exists in a container too. In addition, what is great about docker, the application software container can be - and strongly advised - separated from the data that it holds; containers can store their generated data in *volumes*, in details, a volume is a shared directory between the container and docker's host file systems.

5.2 EventX system deployement

Additionally, Docker has swarm mode; multiple docker hosts runs with this mode on and acts as workers called *nodes*, each node executes a service, on the paper, it is a command that runs a blueprint of containers called tasks in the context of swarm using an image with a

specified configuration. Furthermore, Docker swarm acts as a load balancer to distribute requests among services. Likewise, the instructions to configure the swarm, are always subject to changes, but Docker provides a great documentation. At last, after following the documentation, as illustrated in figure 6, we managed to achieve the technical architecture using 6 virtual machines hosted on a cloud service.

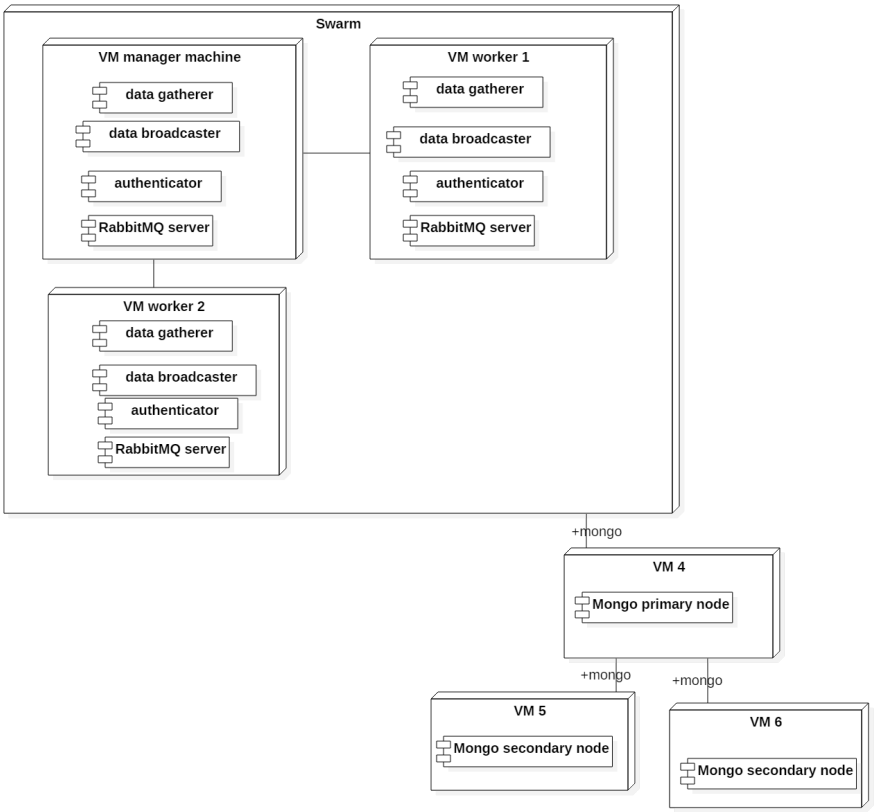


FIG. 6 – *EventX technical architecture*

Upon deploying a service in the swarm, it is possible to specify its number of services which will be distributed over the existing nodes equally. In general, it is not mandatory to have the same number of tasks running on each node, however, it increases the efficiency of the system. Additionally, it is possible to add another worker and join it to the swarm without restarting anything, therefore, it gives a strong availability to the system.

As for the database, MongoDB replica set can be deployed on its own ecosystem. Moreover, it is highly to have at least two secondary nodes, in addition to the primary node which does the write.

6 Conclusion

EventX is a scalable system that collects and broadcast digitalized three-dimensional events called **complex space time events**. Accordingly, these events are observed by sensors and requested by actuators. Furthermore, the notion of "sensor" and "actuator" doesn't stop in electronic micro-gadgets, but it is extended to any source of events that can communicate its state, and can benefits from the broadcast data.

EventX uses a microservice approach to organize its core components, that are implemented using Node.js to benefits from its speed in terms of handling big traffic. Moreover, EventX stores its data in a replica set of MongoDB as a NoSQL document store database system and it is distributed over multiple virtual machine behind Docker swarm that executes the microservices in containers as tasks in a way that guarantees load balancing, in addition, to the possibility of scaling the number of nodes - hosts that have Docker engine installed and joined to the swarm - without shutting down the whole system. EventX is still a system that can be improved by adding:

— Logging microservice that monitor the data collection/broadcast and tracks the connected things.
— Enhancing the security to maintain a healthy traffic and protect the system from corrupted data.
— Introducing an eventual consistency model to speed up the process.
— Caching highly demanded data streams.
— Supporting MQTT as a machine-to-machine connectivity protocol along side COAP and HTTP.

References

Charles C.-H. Hsu, Michael Y.-C. Wang, Hsien C.H. Shen, Ranma H.-C. Chiang, Charles H.P (2017). Wen, *"FallCare+: An IoT surveillance system for fall detection"*, Applied System Innovation (ICASI), 2017 International Conference.

Elena Polycarpou, Lambros Lambrinos, Eftychios Protopapadakis, (2013). *"Smart parking solutions for urban areas"*, IEEE 14th International Symposium on "A World of Wireless, Mobile and Multimedia Networks".

G. Yang et al., *"A Health-IoT Platform Based on the Integration of Intelligent Packaging, Unobtrusive Bio-Sensor, and Intelligent Medicine Box,"* in IEEE Transactions on Industrial Informatics, vol. 10, no. 4, pp. 2180-2191, Nov. 2014. doi: 10.1109/TII.2014.2307795.

Kyle Banker (2012), *"MongoDB in action"*, Oreilly.

Mila Gascó ESADE Business & Law School (2016), Ramon Llull University, *"What makes a city smart? Lessons from Barcelona"*, 49th Hawaii International Conference on System Sciences.

Mircea Eremia, Lucian Toma, Mihai Sanduleac (2017), *"The Smart City concept in the 21st Century"*„ 10th International Conference Interdisciplinarity in Engineering INTER-ENG 2016.

Philippe Dobbelaere, Kyumars Sheykh Esmaili (2017) - *Kafka Versus RabbitMQ*.

Sajee Mathew, AWS Principal Solutions Architect (2017) - *Overview of Amazon Web Services*. AWS Whitepaper.

Sam Newman (2016), *"Building Microservices, designing fine-grained systems"*, Oreilly.

Sergio Trilles, Andrea Calia, Joaquín Torres-Sospedra, Raúl Montoliu, Óscar Belmonte, Joaquín Huerta (2016), *"Deployment of an open sensorized platform in a smart city context"*, Future Generation Computer Systems.

Résumé

Le développement à grande échelle dans l'industrie des technologies de l'information a inspiré de nombreux chercheurs dans le monde entier à faire en sorte qu'il serve la vie quotidienne des citoyens dans les sociétés civilisées. La ville intelligente est un terme donné à une ville qui utilise des approches innovantes pour aider ses citoyens à bénéficier de ses services vitaux, augmentant ainsi leur qualité de vie et les plus-values qu'ils donnent à la société, par objectif de créer un impact démographique positif, en utilisant des technologies pour aider à atteindre les objectifs ci-dessus, en construisant des infrastructures en utilisant à la fois des logiciels et du matériel tels que des capteurs électroniques et des serveurs de grande taille pour aider à gérer *"choses"*. La diversité des plates-formes dans une ville conduit souvent à des systèmes incompatibles qui sont incapables de communiquer, en raison de la couche spécifique qui interagit avec le matériel. Par la suite, il est nécessaire d'une solution partagée qui étend ces plates-formes en collectant des données, mais aussi être un proxy entre ces plates-formes et toute autre plate-forme IoT qui utilisent des données réelles. Ce document fournit une explication sur la construction d'une plate-forme ouverte pour la collecte de ce que nous appelons des données complexes spatiotemporelles; une donnée qui représente tous les événements qui pourraient survenir dans le périmètre d'une ville.

Electronic ADR Transport Document Management Micro-Service for Hazmat Transportation

Ghyzlane Cherradi*, Adil El Bouziri*, Azedine Boulmakoul*

* LIM Lab. IOS, Computer Sciences Department, Faculty of Sciences and Technology, Mohammedia, Morocco

Abstract. Hazmat transportation involves various stakeholders such as shippers, transporters, regulators and emergency responders, etc. Sharing data, such as transport documents between these multiple stakeholders represent a crucial element for shipping dangerous goods by road. Moreover, in the event of an accident, it may be impossible to retrieve the transport documents. Thus, dematerialized documentation accessible via portable electronic devices can be useful for minimizing the control time.

In fact, when dangerous goods are transported, the consignment must be accompanied by a transport document, declaring the description and the nature of the goods. Documentation must be in accordance with the specifications set by the dangerous goods regulations applicable to the chosen mode of transport. This paper represents the microservice that focuses on the dematerialization of transport documents in order to provide in time the appropriate documentation for appropriate authority, which can dramatically reduce the time and error rate to a minimum.

1. Introduction

The definition of hazardous materials includes those materials designated by the secretary of the department of transportation as posing an unreasonable threat to the public and the environment (ADR 2017). Due to its nature, every production, storage, and transportation activity related to the use of HAZMAT has many risks for both society and the environment. HAZMAT are transported throughout the world in a great number of road shipments. While HAZMAT accidents are rare events, the commercial transport of HAZMAT could be catastrophic in nature and poses risks to life, health, property, and the environment due to the possibility of an unintentional release. In order to avoid the risks turning into real events, it is necessary to integrate risk mitigation and prevention measures into the transport management.

Indeed, a multi-purpose system of identification, management of dangerous vehicles and providing various communication and control functions related to the hazmat transportation, could decrease their risks and might effectively provide detailed information on hazardous cargo, vehicle status, incidents, and routes. Therefore such information allows to better understand the causes and consequences of hazardous material transportation incidents, and would also enhance safety measures for emergency personnel called to handle an accident of hazardous materials. As well as this, the transportation documents are a vital piece of information for emergency response to incidents involving dangerous goods.

This paper aims to contribute to the safe transportation of the hazardous materials by providing a real-time microservice-based management system that can offer specific services, particularly the electronic ADR transport document management microservice. The latter Use electronic information processing and electronic data interchange techniques to facilitate the preparation, exchange, and replacement of documents. The rest of the paper is organized as follows: Section 2 presents the architecture of the proposed system and provides an overview of its microservices and functionalities. Section 3 describes the detailed architecture of the main microservice of this paper. We conclude the paper in Section 4 and address areas of future work.

2. System Architecture

Smart hazmat transportation system is advanced applications of information, communication technologies, and management strategies in order to optimize the movement of people and goods, improve public safety, and the environment, and provide highly efficient services related to hazmat transportation. The idea is to utilize advanced and emerging technology in such fields as computer technology, information technology, electronic communication and control over the field of transportation to build an integrated system of people, roads, hazardous materials, and vehicles. Thus, the proposed system can be viewed as a collection of various services, integrated with each other to form a cohesive and flexible system. Accordingly, the microservices-based architecture is a best-practice approach for realizing such system requirements. It can be defined as an approach to developing a single application as a suite of small services; each running in its own process and communicating with lightweight mechanisms (Newman, 2015). Furthermore, Due to distributed nature of microservices, the system will be developed as a suite of tiny services aligned with the risk-management process; each service will be running in its own logical machine or container technology such as docker (Cherradi et al. 2017).

The key characteristics of the microservice architecture relevant to the context of this work are described below.

- **Componentization via Services:** software componentization is a practice that breaks software system down into smaller easily identifiable pieces. Microservice-based architecture achieves the componentization via breaking systems down into microservices, which are capable of running independently and are responsible for its own data. These microservices can talk to each other but are not dependent on others. A microservice is independently replaceable and independently upgradable (Newman, 2015).
- **Decentralized Data Management:** The proposed system databases are heterogeneous; they are spatial, relational and document database. Therefore, there is no unifying schema in a central database. Microservices architectures are distributed systems with decentralised data management. This means that every service has its own, independent, storage subsystem that is isolated from other services (Namiot et al., 2014).
- **Technology Heterogeneity:** With a system composed of multiple, collaborating services, we can decide to use different technologies that are most suitable for each one (Abbott et al., 2009). This use of multiple languages, frameworks and technologies for individual service gives benefits such as scalability and interoperability.

- **Evolutionary Design:** The change has historically been difficult to anticipate and expensive to retrofit; consider that over time, more capabilities can be migrated. The use of microservices-based architecture will permit to add new user experiences and new business capabilities (Familiar, 2015).

A global view of the different microservices of the system is shown on figure 1.

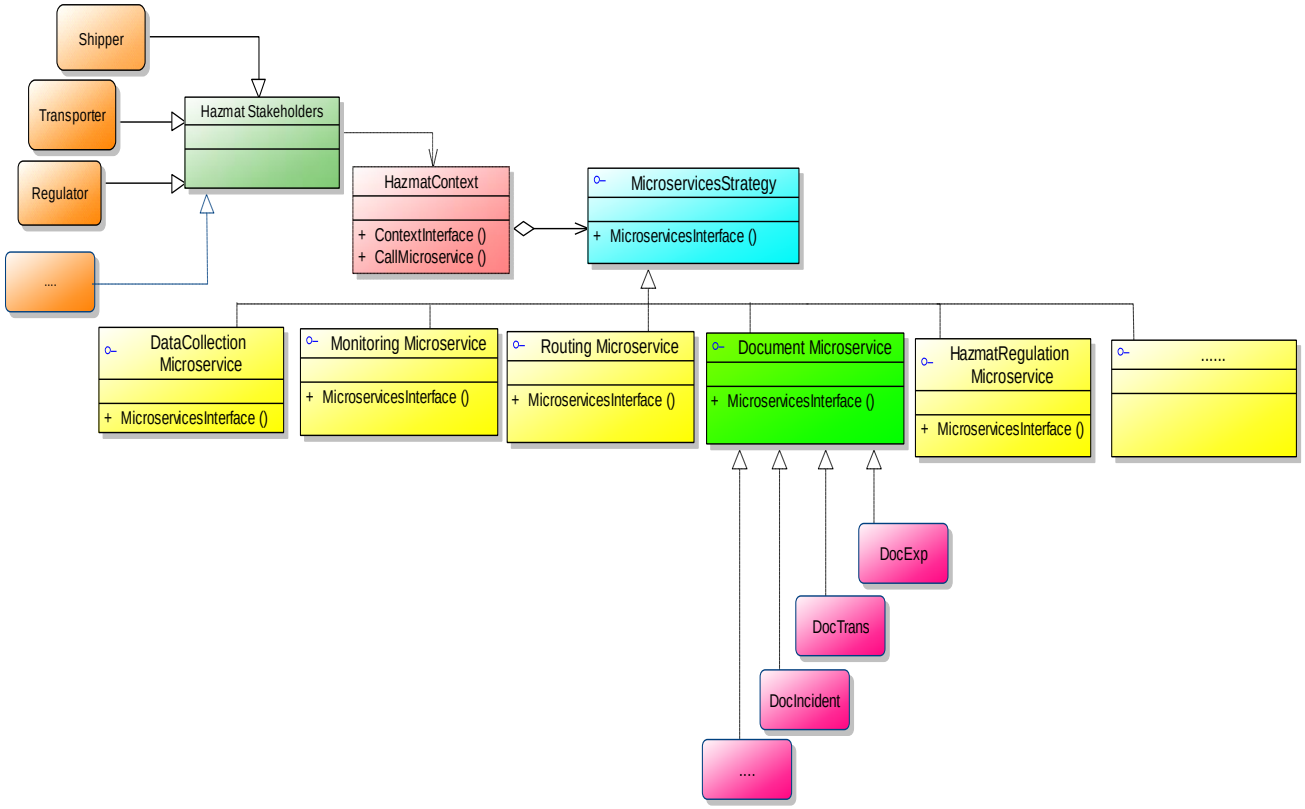


FIG. 1 – *Abstraction of services offered by the system*

The application is deployed as a set of microservices in the form of containers. Client apps can communicate with those containers through an API Gateway, which represents the single entry point for all clients. The API Gateway (Gateway, I. 2016) encapsulates the internal architecture of the system and provides a suitable API for each client. It may have other responsibilities such as authentication, routing and load balancing. It is placed in the demilitarized zone (or DMZ for short) to guarantee a level of security to the internal ecosystem. As shown in Figure 2, the internal ecosystem is composed of many microservices each microservice is implemented in a different way. Each can have a different architecture model and use different languages and databases depending on the nature of the application, busi-

ness needs and priorities. The units of deployment for microservices are Docker containers (Vohra, D. 2016) and the application is a multi-container application that embraces microservices principles.

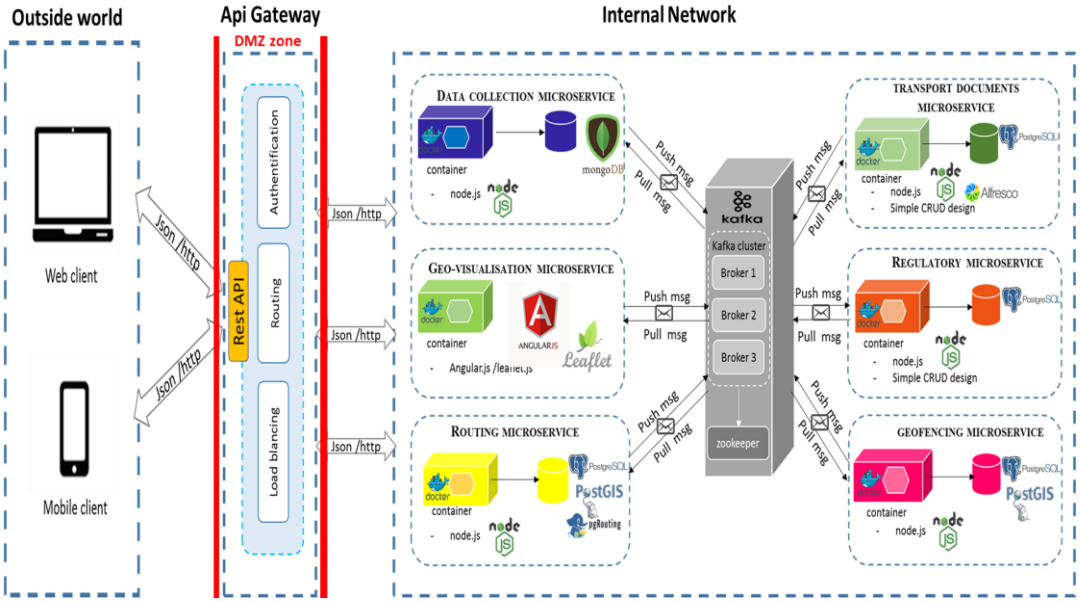


FIG. 2 – *Technical architecture of the system*

The internal ecosystem is composed of many microservices each microservice is implemented in a different way. Each can have a different architecture model and use different languages and databases depending on the nature of the application, business needs and priorities.

- **DataCollection Microservice**: This microservice communicates with in-vehicle sensors and collects sensor data (such us real-time information about goods, driver and vehicle, etc.). Once the data received, the microservice stores the row data onto the database.
- **Monitoring Microservice**: based on GIS web map and GPS, the system can visualize the vehicle moving on the map with all the important data and parameters right alongside it.
- **Geo-visualization microservice**: based on GIS web map and GPS, the system can visualize the vehicle moving on the map with all the important data and parameters right alongside it. Moreover, through techniques of spatial query, it allows users to explore, synthesize, present (communicate), and analyze the meaning of any given layer.
- **Routing microservice**: This microservice is based on a combination of the geographical information system (GIS) with the spatial database and the pgRouting algorithms. It aims to manage, treat, and represent geographical data related to the

evaluation of the risk of transport on a road network, in view of finding the least-risky routes for HAZMAT shipments.

- **Regulatory microservice:** This microservice makes it possible to use the ADR regulations in electronic format, to know the important administrative texts, and to recognize the regulatory modifications made.
- **Document Microservice**: The documentation is crucial for shipping dangerous goods by road. Therefore, providing in time the appropriate documentation for appropriate authority can dramatically reduce the time and error rate to a minimum. The Transport documents microservice is based on the alfresco solution, which is an open source content management system allows dematerializing, classify, search, store and distribute documents. Represents the main microservice in this paper, which will be present in detail in the following section.

The client application provides a way to interact with the application through intuitive screens that use techniques to make the application easier to learn and use. The graphical user interface (GUI) targets the most popular desktop environments and supports the business requirements of the system. AngularJs will be used to create the GUI screens. These screens will contain a minimum of processing logic to more effectively separate logical architectural levels.

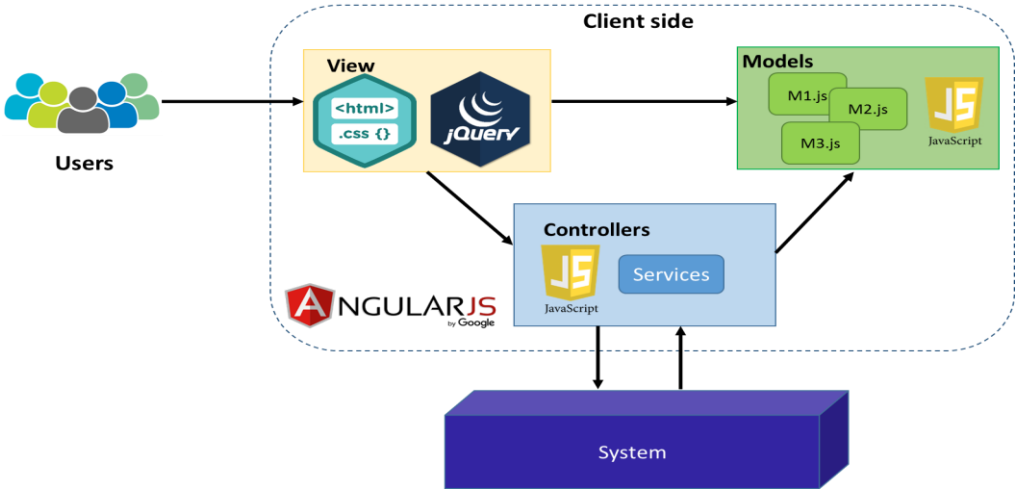


FIG. 3– *External ecosystem architecture*

3. Transport documents microservices

All transport of goods, regulated by ADR, must be accompanied by documentation (ADR 2017). To facilitate the viewing of transport documents, we call the transport documents microservice, which will make it possible to download the various documents required from the registration number of the vehicle. Then we use the regulation microservice for

provide extensive regulatory information, thus, this can increase the level of speed, accuracy and reliability of road transport regulatory authority's decision.

The main capabilities of the proposed microservice are as follow

- Automatically scan license plates using a mobile cameras.
- The system provides according to the role of each authority; data collected on the vehicle corresponding to the license plates. As well as all required documents. By document, here mainly means "file" (type Word, Excel, Adobe PDF ...), plus associated metadata (title, author, description ...). The documentation is always kept up to date in accordance with the ADR regulations.
- The system can help the supervisory authority to carry out a visual inspection.
- The system can help shipper to create the shipper's declaration for the transport of dangerous goods by road, in accordance with the ADR regulations.
- The system allows to download any type of document and include a description of it to inform about its contents and also allow to search in all defined documents. Such shipping document, transport document, incident document, written instructions, safety data sheets, certificates (driving, chemicals, vehicles ...), etc. In the appendix, there are examples concerning the each required transport document
- In the case of an accident, the system provides an incident report document, which is documentation of an event that has disrupted the normal operation. The data providing in this report is fundamental to hazardous material transportation risk analysis and risk management. It allows welling understand the causes and consequences of hazardous material transportation incidents. It helps to demonstrate the effectiveness of existing regulations and to identify areas where changes should be considered. This type of document is accessible to all, any person in possession of a hazardous material during transportation, including loading, unloading, and storage incidental to transportation, must report if certain conditions are met.
- The transport documentation can be exported to various formats, including PDF and Excel. And may change if needed.

The transport documents microservice is based on the alfresco solution, which is an open source content management system allows dematerializing, classify, search, store and distribute documents. Figure 4, shows the transport document microservices architecture.

Alfresco is a Document management system, also known as enterprise content management system. While products like google docs and live office making a lot of noise in the document editing market and products like dropbox helping to save and share documents and other type of files. Alfresco offers comprehensive document management services including document repository, versioning control on documents, and support for multiple file formats such as text, audio and video files. These documents can be integrated with internal approval workflows and business processes (Shariff M. 2007).

Alfresco not only allows the electronic management of documents but also other features such as:

- Compliance Management
- Business Process Management
- content management
- Records management

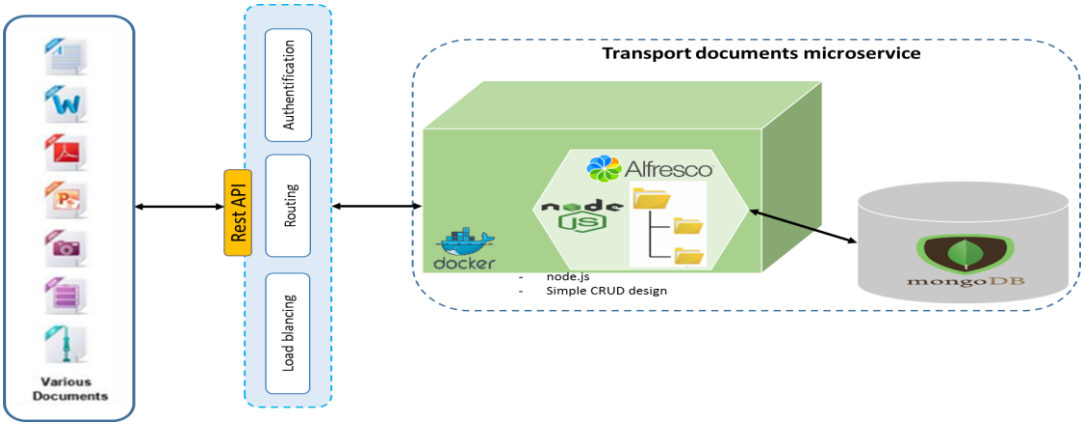


FIG. 4– *Transport document microservice architecture*

The system is developed using NodeJS (Holmes, S. 2015), as a server-side runtime environment with the integration of Alfresco JavaScript API, and AngularJs as a front-end framework. For data server side, we use MongoDB (Fowler, M. et al. 2012), which represent a highly scalable and flexible storage solution. This combination allows to quickly control the content repository and obtain content for a large number of customers.

The figure below shows transport document microservice output.

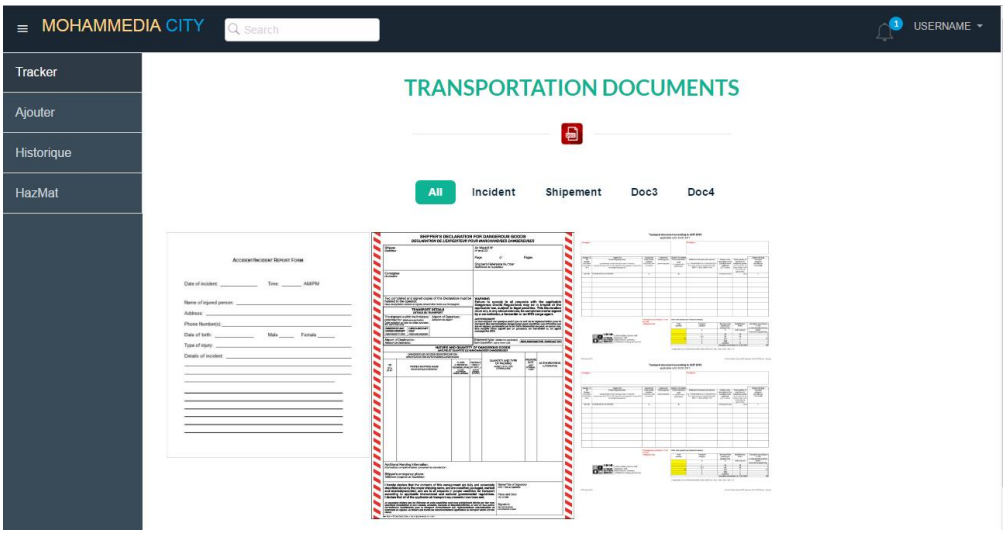


FIG. 5– *Screenshot of transport document microservice output*

4. Conclusion

The transport of dangerous goods by road represents a major technological risk, particularly in urban areas where a large population lives and works near roads on which circulates a growing number of hazardous substances. In order to build a planning aid system for reducing the risks of transporting hazardous materials, there is a need to visualize, analyze and evaluate the affected areas. As well as, known the transportation documents, which represent a crucial piece of information for an emergency response in case of an incident involving hazardous material. This paper presents a real-time microservice-based management system that can offer various services, especially the electronic ADR transport document management microservice, which can provide in real time the appropriate documentation for appropriate authority, which can dramatically reduce the control time and error rate to a minimum.

Acknowledgements
This work was funded by the CNRST project in the priority areas of scientific research and technological development "Spatio-temporal data warehouse and strategic transport of dangerous goods».

References

Newman, S. (2015). Building microservices. "O'Reilly Media, Inc.".

Cherradi, G., Bouziri, A. E., Boulmakoul, A., & Zeitouni, K. (2017) "Real-Time Microservices Based Environmental Sensors System for Hazmat Transportation Networks Monitoring." Transportation Research Procedia 27: 873-880.

Namiot, D., & Sneps-Sneppe, M. (2014). On iot programming. International Journal of Open Information Technologies, 2(10).

Abbott, M. L., & Fisher, M. T. (2009). The art of scalability: Scalable web architecture, processes, and organizations for the modern enterprise. Pearson Education.

Familiar, B. (2015). Microservices, IoT and Azure: Leveraging DevOps and Microservice Architecture to deliver SaaS Solutions. Apress.

Shariff, M. (2007). Alfresco enterprise content management implementation. Packt Publishing Ltd.

Gateway, I. (2016). IoT Gateway.

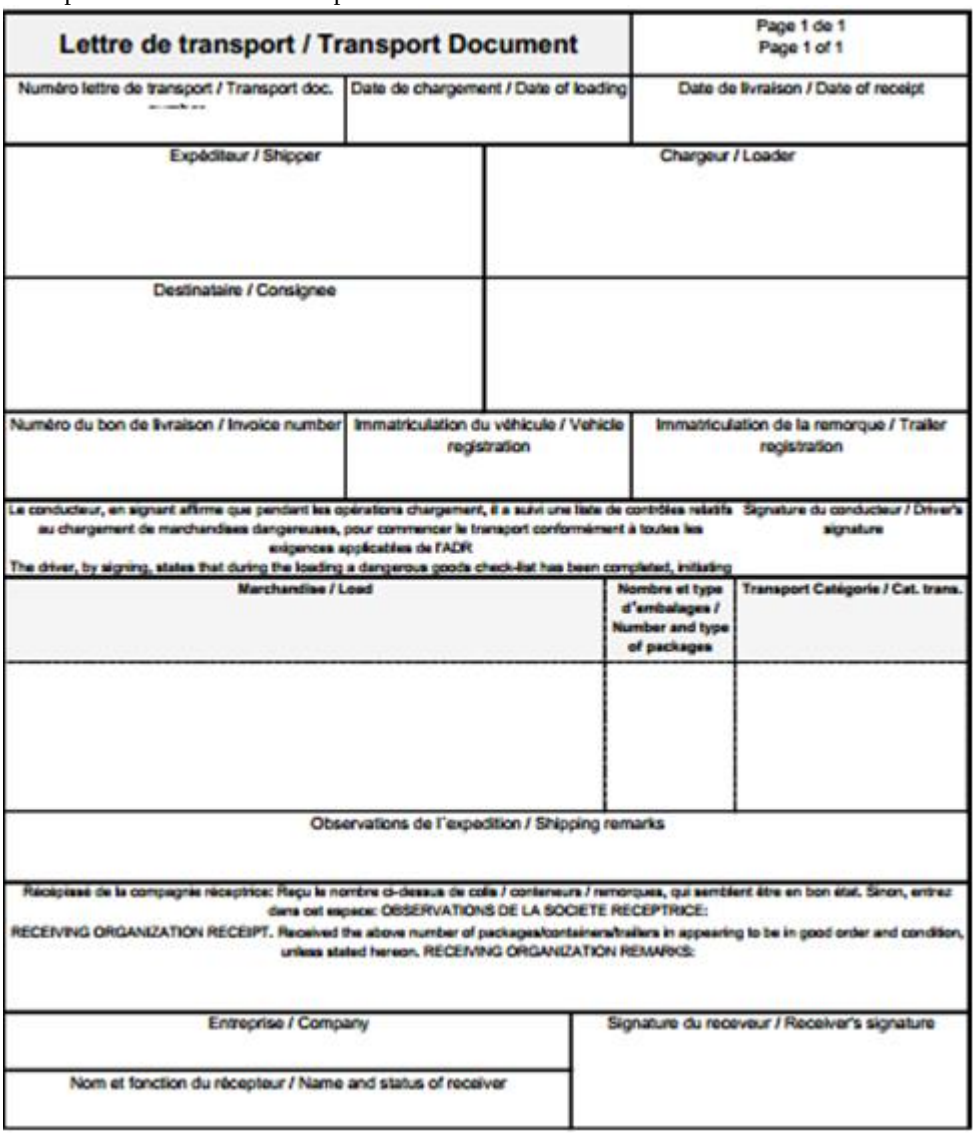
Vohra, D. (2016). Kubernetes microservices with Docker. Apress.

Holmes, S. (2015), Getting MEAN with Mongo, Express, Angular, and Node, Packt Publishing.

Fowler, M. and Sadalage, P. (2012). NoSQL distilled: a brief guide to the emerging world of polyglot persistence. Addison-Wesley.

Appendix

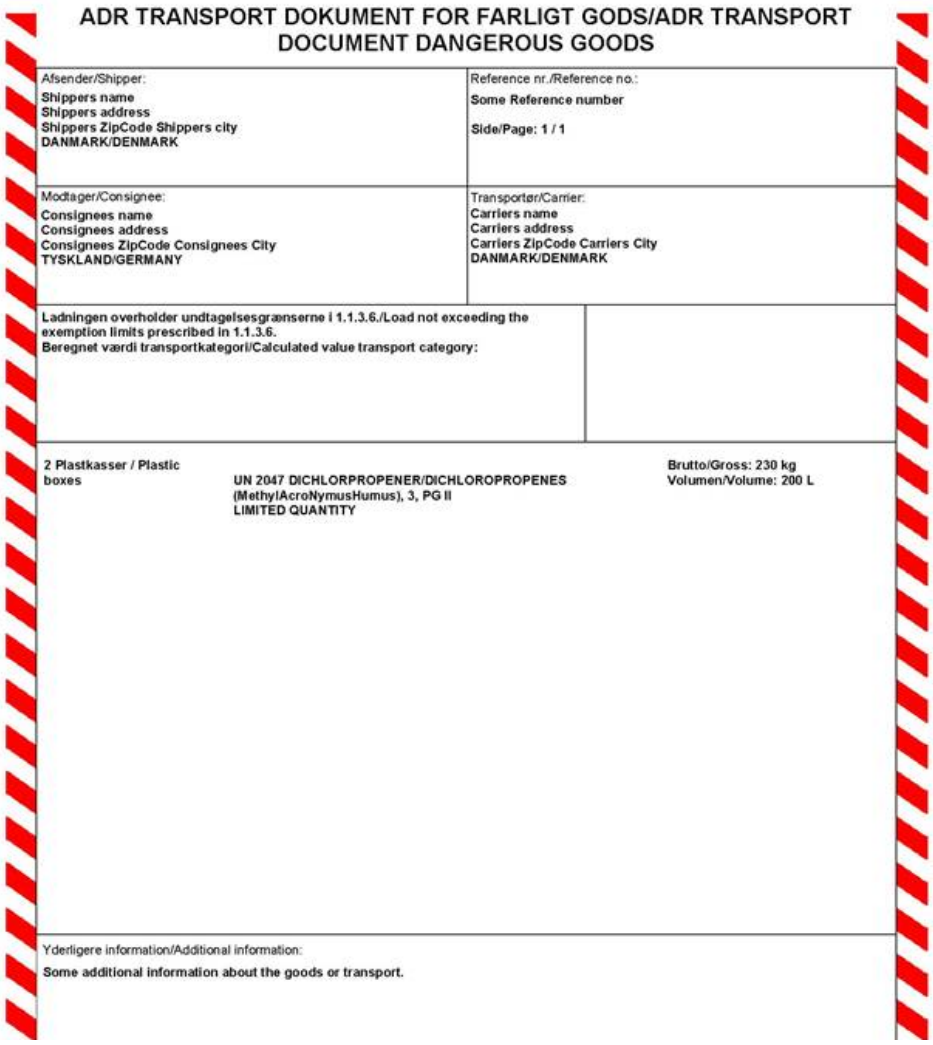
Transportation Document Template ADR 2017



Lettre de transport / Transport Document		Page 1 de 1 Page 1 of 1
Numéro lettre de transport / Transport doc.	Date de chargement / Date of loading	Date de livraison / Date of receipt
Expéditeur / Shipper		Chargeur / Loader
Destinataire / Consignee		
Numéro du bon de livraison / Invoice number	Immatriculation du véhicule / Vehicle registration	Immatriculation de la remorque / Trailer registration
Le conducteur, en signant affirme que pendant les opérations chargement, il a suivi une liste de contrôles relatifs au chargement de marchandises dangereuses, pour commencer le transport conformément à toutes les exigences applicables de l'ADR The driver, by signing, states that during the loading a dangerous goods check-list has been completed, initiating		Signature du conducteur / Driver's signature
Marchandise / Load	**Nombre et type d'emballages / Number and type of packages**	**Transport Catégorie / Cat. trans.**
Observations de l'expedition / Shipping remarks		
Récépissé de la compagnie réceptrice: Reçu le nombre ci-dessus de colis / conteneurs / remorques, qui semblent être en bon état. Sinon, entrez dans cet espace: OBSERVATIONS DE LA SOCIETE RECEPTRICE: RECEIVING ORGANIZATION RECEIPT. Received the above number of packages/containers/trailers in appearing to be in good order and condition, unless stated hereon. RECEIVING ORGANIZATION REMARKS:		
Entreprise / Company		Signature du receveur / Receiver's signature
Nom et fonction du récepteur / Name and status of receiver		

XML schema for Shipper's Declaration for Dangerous Goods

```
<DOCUMENT>
 <DGMCOMMANDS>
  <CMDDOCUMENT>9</CMDDOCUMENT>
  <CMDTRANSPORTEMERGENCYCARDS>8</CMDTRANSPORTEMERGENCYCARDS>
 </DGMCOMMANDS>
 <DGMPRINTERS>
  <DEFAULTPRINTERNAME>HPLaserJ</DEFAULTPRINTERNAME>
 </DGMPRINTERS>
 <DGHEADER>
  <ADDITIONALINFO>Some additional information about the goods or transport.</ADDITIONALINFO>
  <CARRIERADDRESS1>Carriers address</CARRIERADDRESS1>
  <CARRIERCITY>Carriers City</CARRIERCITY>
  <CARRIERCONTACTPERSON>Carriers contact person</CARRIERCONTACTPERSON>
  <CARRIERCONTACTPHONE>Carriers Contact phone</CARRIERCONTACTPHONE>
  <CARRIERCOUNTRYCODE>DK</CARRIERCOUNTRYCODE>
  <CARRIERNAME>Carriers name</CARRIERNAME>
  <CARRIERZIPCODE>Carriers ZipCode</CARRIERZIPCODE>
  <CONSIGNEEADDRESS1>Consignees address</CONSIGNEEADDRESS1>
  <CONSIGNEECITY>Consignees City</CONSIGNEECITY>
  <CONSIGNEECONTACTPERSON>Consignees contact person</CONSIGNEECONTACTPERSON>
  <CONSIGNEECONTACTPHONE>Consignees contact phone</CONSIGNEECONTACTPHONE>
  <CONSIGNEECOUNTRYCODE>DE</CONSIGNEECOUNTRYCODE>
  <CONSIGNEENAME>Consignees name</CONSIGNEENAME>
  <CONSIGNEEZIPCODE>Consignees ZipCode</CONSIGNEEZIPCODE>
  <DOCMANCUSTOMERID>SomeUniqueCustomerId</DOCMANCUSTOMERID>
  <REFERENCE>Some Reference number</REFERENCE>
  <SEADOCHIDEAIRINFO>1</SEADOCHIDEAIRINFO>
  <SHIPPERADDRESS1>Shippers address</SHIPPERADDRESS1>
  <SHIPPERCITY>Shippers city</SHIPPERCITY>
  <SHIPPERCONTACTPERSON>Shippers contact person</SHIPPERCONTACTPERSON>
  <SHIPPERCONTACTPHONE>Shippers contact phone</SHIPPERCONTACTPHONE>
  <SHIPPERCOUNTRYCODE>DK</SHIPPERCOUNTRYCODE>
  <SHIPPERNAME>Shippers name</SHIPPERNAME>
  <SHIPPERZIPCODE>Shippers ZipCode</SHIPPERZIPCODE>
  <SHIPPINGCOMPANY>Shipping Company</SHIPPINGCOMPANY>
  <SHIPPINGPLACE>Shipping place</SHIPPINGPLACE>
  <SHIPPINGSIGNATURE>Shipping signature</SHIPPINGSIGNATURE>
 </DGHEADER>
 <DGITEM>
  <!-- required (not including class 7) -->
  <DGMID>1631</DGMID>
  <NUMBOFPACKAGE>2</NUMBOFPACKAGE>
  <QUANTITY>200</QUANTITY>
  <PACKAGETYPEID>110</PACKAGETYPEID>
```



ADR TRANSPORT DOKUMENT FOR FARLIGT GODS/ADR TRANSPORT DOCUMENT DANGEROUS GOODS

Afsender/Shipper: **Shippers name** **Shippers address** **Shippers ZipCode Shippers city** **DANMARK/DENMARK**	Reference nr./Reference no.: **Some Reference number** **Side/Page: 1 / 1**
Modtager/Consignee: **Consignees name** **Consignees address** **Consignees ZipCode Consignees City** **TYSKLAND/GERMANY**	Transportør/Carrier: **Carriers name** **Carriers address** **Carriers ZipCode Carriers City** **DANMARK/DENMARK**
Ladningen overholder undtagelsesgrænserne i 1.1.3.6./Load not exceeding the exemption limits prescribed in 1.1.3.6. **Beregnet værdi transportkategori/Calculated value transport category:**	

2 Plastkasser / Plastic boxes	**UN 2047 DICHLORPROPENER/DICHLOROPROPENES (MethylAcroNymusHumus), 3, PG II** **LIMITED QUANTITY**	**Brutto/Gross: 230 kg** **Volumen/Volume: 200 L**

Yderligere information/Additional information:

Some additional information about the goods or transport.

Mobile Sensor Driven Exposure Analysis to Air Pollution: A Comprehensive Survey

Rafiqul Haque*, Yehia Taher*,
Karine Zeitouni*

*University of Versailles Saint-Quentin-en-Yvelines, Paris-Saclay University
rafiqul.haque@cognitus.fr, yehia.taher@uvsq.fr, karine.zeitouni@uvsq.fr

Abstract. The advent of the new generation of low-cost lightweight, and connected sensors makes a paradigm shift in environmental studies. In particular, nomadic sensors allow for a very precise personalized measurement, by continuously quantifying the individual exposure to air pollution components. Moreover, a broad dissemination among volunteers of theses devices, or their deployment on vehicle fleets, is becoming a credible scenario. Another major interest of such sensor deployment is to densify the air quality monitoring network, which is today restricted to sparse nodes, providing only averaged measures per hour. However, this high spatio-temporal resolution raises several issues related to their analysis. Among them, modeling, quality consideration, interpretability and crosss-correlation with other data sources, scalability of query processing and data analytics against big sensor data streams. After an overview of the projects relying on this technology, this article points out the remaining challenges to be addressed in the ongoing project Polluscope.

1 Introduction

Environmental pollution has been a critical concern for many decades. Pollution engenders various diseases. In some cases, it poses enormous threat to human health and claims lives. The world health organization (WHO) estimates that about a quarter of the disease facing mankind today occur due to prolonged exposure to environment pollution (see Afrifa et al. (2013)). The detrimental effects of different types of pollution has given the rise to a significant question *how to control pollution*? Although, some pollutants have fallen sharply in the last two decades, yet studies as Bentayeb et al. (2015) show that air pollution remain a worrying problem which reduces life expectancy by several months. An investigation conducted by Amos (2016) shows that more than 5.5 million people worldwide are dying prematurely every year as a consequence of air pollution.

A plethora of research activities have been conducted over the years which aims at preventing pollution, including creating awareness among people, reducing industrial waste, *etc*. Nowadays, Internet of Things (IoT), in particular, sensors has become critically important in pollution monitoring (see. Xia et al. (2012)). Sensors enable easier and faster collection of pollution data and cover both indoor and outdoor environments. They can be devised into *Mobile Sensors* and *Fixed Sensors*. The *mobile sensors* are mostly mounted on transportation

mediums to collect data, because most of today sensors are still large or heavy. However, more compact models are continuously introduced, which opens the way to carry them by humans to measure their own exposure. For such hand-held sensors, data collection process may either task the users to observe some zones or actively post some informations, which is called *participatory sensing*; however, in several occasions, the data collection is entirely automatic – this is known as *opportunistic sensing* (Ganti et al. (2011)).

In addition to sensor technology, various smart systems have been developed within the scope of different projects including personalized alert systems, analytics, and visualization tools. In concert with sensor technology, these smart systems are leading a major change in pollution monitoring and paving the foundation of high-end next generation city-wise air quality observatory targeting both citizens and decision-makers. This paper presents a survey of different projects related to air quality monitoring based on the emerging low-cost mobile sensors. We consider the architectural and functional aspects of the proposed solutions within the scope of our study. We provide the results of our comparative analysis among different solutions and discuss the challenges and future research directions. We end by presenting our proposal the current collaborative project Polluscope.

2 A Survey of Related Projects

In the following subsections, we survey related projects and report our findings in particular, the solutions proposed within the scope of these projects.

2.1 OpenSense

The OpenSense project – led by the Swiss Federal Institute of Technology (ETH) – aims at addressing the key research challenges in the domain of information and communication systems related to community-based sensing using wireless sensor network technology in the context of air pollution monitoring [1]. The core challenges that are dealt with include the following: (i) heterogeneity and variety of sensor equipments, measurements and data analysis, (ii) supporting and exploiting mobility of sensors, and (iii) involving the community in a trusted, fair and transparent manner into the monitoring activity. Different goals were defined to address these challenges and to perform the pollution monitoring task efficiently.

The major scientific goal of OpenSense project is, as stated in Riahi et al. (2013), to efficiently and effectively monitor air pollution using wireless and mobile sensors by adopting complex utility driven approaches towards sensing and data management . The optimization goals include: (i) reducing resource consumption, and thus cost of the infrastructure, while increasing accuracy and value of the information produced, (ii) optimizing accuracy of data considering application demands, (iii) minimizing transmission, analysis, and storage of measurement data, (iv) reducing the latency of real-time information delivery against perturbation factors and uncertainties.

From architectural perspective, the OpenSense system is straightforward. It relies on conventional distributed architectural paradigm consisting of data sourcing entities (sensors), data repository, and application server containing different applications for doing various jobs. The

1. http://www.nano-tera.ch/pdf/sheets/OpenSense.pdf embedded

traditional client-server architecture was adopted between the external and internal components, in particular: the sensor nodes and the data repository. However, interestingly, OpenSense relies on decentralized controlling system and thus, the sensorbox at each station (in both mobile and fixed stations) are controlled autonomously without the influence of external systems. The OpenSense system involves different sensors which are packaged in a single box called *sensorbox* dedicated to collect data. All OpenSense sensorboxes are based on a custom-manufactured platform (Buchli et al. (2011)). All the functionalities including data collection, data processing and visualization in OpenSense system rely on different models that are developed within the scope of this project. These models are explained in the following:

— *Sensor deployment Model*: In OpenSense, the deployment model is hybrid and relies on both mobile sensors (mounted on trams and buses) and fixed station sensors (installed on crowded locations of the city such as bus stops). Sensorboxes are geolocalized. They monitor various parameters: particulate matter (PM), Carbon Dioxide (CO2), Nitrogen Dioxide (NO2), temperature and relative humidity. It is worth noting that indoor air pollution monitoring was out of the scope of the project.

— *Data Collection*: In OpenSense, data are collected over micro-windows of time. Two different time models for two different pollutants were reported in by Li et al. (2012): PM are recorded every 5 seconds while ozone and CO2 are observed each 20 seconds. The measurements are transmitted to a data server running GSN (Aberer et al. (2006)) and are publicly available [2]. GSN is a platform which enables building a scalable infrastructure for integrating heterogeneous sensor network technologies using a small set of powerful abstractions. As reported by Aberer et al. (2006), GSN provides a logical view on sensor networks through the virtual sensor abstraction. The virtual sensors model sensor data as temporal streams of relational data, and allow to represent derived views on sensor data streams, possibly from different sources.

— *Sensing Model*: Unlike traditional sensing, where the primary focus is on optimally sampling the environment, the sampling policies in OpenSense system is driven by the application-layer requirements and projected utility of data being sampled (Riahi et al. (2013)). The main reason of such sensing model is to reduce energy consumption. A two-tier optimal mobile sensing model called *OptiMos* was developed by Yan et al. (2012). The lower tier, called *optimal segmentation*, focuses on sensor data segmentation. The objective is to find the optimal (or near-optimal) segmentation based on data modeling on these raw readings. For each segment, the objective is to find the best sampling from the mobile sensor readings, i.e., to select only a subset of sensor readings. This subset can keep enough modeling information for regression of the whole segment and for prediction of non-selected sensor readings (refer to Yan et al. (2012) for more detail) .

— *Semantic Data Enrichment*: a solution called SeMiTri was developed by Yan et al. (2011) for annotating the trajectory data. The goal is to support semantic enrichment of trajectories exploiting both the geometric properties of the stream and the background geographic and application data. Examples of annotations include inferring and recording the means of transportation used by a moving person. Another example is to identify specific trajectory segments. Each of these segments called *episode* corresponds to a maximal sub-sequence of the trajectory that complies with a given

2. http://data.opensense.ethz.ch

predicate. For instance, a trajectory may be segmented into episodes of *stop* and *move*, according to the predicates speed $< \delta$ for *stops* and speed $\geq \delta$ for *moves*.

— *Creating Probabilistic Database*: Imprecise and uncertain data are difficult to deal with in sensor driven technical ecosystem. In order to deal with uncertainty, in OpenSense, a framework is introduced by Sathe et al. (2011) to create probabilistic database. It consists of two key components that are *dynamic density metrics* and the Ω – View builder that efficiently creates probabilistic views by processing a probability value generation query (executed using SQL-like syntax). A dynamic density metric is a system of measure that dynamically infers time- dependent probability distributions of imprecise raw values. It takes as input a sliding window that contains recent previous values in the time series. Various dynamic density metrics were introduced in this solution including: Naive Dynamic Density Metrics, Garch Metrics, and enhanced Garch metrics (refer to Sathe et al. (2011) for more detail). For efficient processing, two different types of caching methods were introduced: α – caching and context-aware caching along with the probabilistic database creation process.

— *Data Management*: a data management framework called *ConDense* was introduced by Cartier et al. (2012). The objective was to efficiently manage the generated environmental data. It provides a multi-model based abstraction by condensing information generated by a CGSN (Community-driven Mobile GeoSensor Networks). ConDense takes into account the unique properties of CGNSs and treats the underlying sensor network as a disconnected component, which is collecting data using local policies and principles. The authors introduced the notion of *model covers* to allow users to use multiple models which, according to them, is a feasible option since data modeling covers large geographical areas of a CGSN. Adaptive strategies were introduced in this framework. These strategies discover spatial areas that can be modeled using single or multiple models. They were implemented by extending two popular clustering algorithms. Additionally, the strategies adapt to the changing nature of the sensed phenomenon by adjusting the geographical granularity of the models to capture the phenomena with high fidelity (see Cartier et al. (2012)). Furthermore, the authors claimed that the strategies have user-defined approximation error thresholds, which can be used for adjusting the level of geographical granularity and quality of the models.

— *Processing Queries on Time Series Data*: A framework called AFFINITY was introduced within OpenSense project by Sathe and Aberer (2013). Its goal is to ensure efficient computation of statistical measures by exploiting the concept of *affine* relationships. Affine relationships can be used to infer statistical measures for time series from other related time series, instead of computing them directly; thus, reducing the overall computational cost significantly. In addition to affine relationship, the framework contains an index structure to improve the processing of the statistical queries.

2.2 OpenSense II

OpenSense II is an extension of OpenSense project. Its goal is to leverage and improve methods developed in the framework of OpenSense, particularly on: mobile monitoring of air pollution, sensor and communication platforms, calibration methods, sensor data gathering and visualization, statistical modeling, activity recognition, and personalized health recommendations. In OpenSense II, the dimension of crowdsourcing and human-centric computation were

introduced to study possibilities to incentivize users to make available states based on physical measurements, such as location, motion and pollution, through their mobile personal devices or monitoring assets that they can install in their homes or on their cars. They introduced a *dispersion model* to compute high-resolution air pollution maps for the cities of ZÃ¼rich and Lausanne (based on Land-use regression models). The main objective is to provide independent information on air pollutant distributions. First of all, however it is necessary to evaluate the quality of the sensor data and their suitability to measure city-scale air pollution levels. OpenSense II also aims at measuring the impact of long – or medium – term exposure to air pollution on human health. Lastly, it intends to evaluate the potential of crowdsourcing for providing feedbacks to users.

2.3 CITI-SENSE

CITI-SENSE project is co-funded by the EU FP7 CITI-SENSE (2016). It aimed at developing *Citizens Observatory* to engage mass population to contribute to and participate in environmental governance, and enable them to support and influence community and societal priorities and associated decision making CITI-SENSE (2016). Citizens observatory is defined as a platform for observing and understanding environment related problems, and more particularly as reporting and commenting on them Liu et al. (2014). The objectives of this project entail: raise environmental awareness through user participation and providing feedback on the impact that citizens had in decisions. To that end, a citizen observatory toolbox (COT) has been developed to provide tools and services to the citizens. The toolbox offers monitoring and sensor platform, mobile applications, widgets, and an interface for data browsing and download.

The sensor platform of COT called *AQMesh sensor pods* is used to capture the observations (i.e., *measurements*) of chemical compounds found in outdoor air include: NO, NO_2, O_3, CO, PM, temperature, humidity, noise, pressure in outdoor air. Atmospheric Sensors have been used to measure and collect data of gases and particulate matter (PM) in indoor air. Data fusion is used to combine these data with high-resolution data collected from model. It is worth noting that the model information was derived from the urban air pollution dispersion models. Two different dispersion models are used based on the locations: EPISODE model Slørdal et al. (2003) was used for Oslo and statistical land-use regression model used for all other sample locations within the scope CITI-SENSE project. EPISODE is a 3-D Eulerian/Lagrangian dispersion model that provides urban-and regional–scale air quality forecasts of atmospheric pollutants. The *universal Kriging* methodology is used to combine observation with the model data by predicting concentration at unknown locations by simultaneously interpolating the observations and using the model data to provide information about the spatial patterns (CITI-SENSE (2016)). Universal Kriging–similar to ordinary Kriging – with external drift allows the overall mean to be non-constant throughout the domain and to be function of one or more explanatory variable.

The CITI-SENSE data management framework comprises three different platforms:

— *Sensor application platforms*, comprise standard technologies provided by the sensor providers, typically compliant to W3C and OGC [3] standards. The platforms are bundled with mobile sensors and applications. The data are collected using sensors and

3. https://www.w3.org/ and http://www.opengeospatial.org/

stored using a NoSQL engine. SensApp –an open source mobile application– enables registering sensor measures and notifying clients with newly arrived data. SensApp also offers Web Services to upload data in central repository for storage and retrieval.

— *Spatial Data Service Platform*. The GO loader and publisher are core components of this platform. The loader supports configuration of enterprise relational data and supports loading of data delivered in XML and GML. It supports data provided in a number of standard data specification adopted in CITI-SENSE project such as SenML (W3C), SensorML (OGC), Observation and Measurements, and INSPIRE environmental and Monitoring facilities. The GO publisher enables to make data available via several web based interfaces such as WFS (Web Feature Services), REST (Representational State Transfer), RSS (Rich Site Summary), *etc*. The publisher takes the responsibility to transform data from SQL format to JSON, XML, and GML.

— *Linked Data Platform* uses CITI-SENSE ontology for annotation, and TripleStore to publish data in the Linked Open Data cloud.

In addition to the data management platforms, CITI-SENSE proposes a personal air monitoring toolkit which enables to measure personal air quality. This toolkit consists of three different tools: mobile sensor unit which captures measurements of NO, NO_2, and O_3. Furthermore, CITI-SENSE framework provides a visualization interface for the various information.

2.4 CITI-SENSE-MOB

The Citi-Sense-Mob is is partly funded by EMMIA: The European Mobile and Mobility Industries Alliance. It is focused on mobile air quality sensing. It was deployed in the city of Oslo. It is worth noting the tight collaboration of CITI-SENSE-MOB with CITI-SENSE in terms of sharing resources, and technology for data handling. Here, the sensors are mounted on vehicles, such as the city buses or electric bicycles. The argument is discomfort of the sensors due to their heavy weight and their large size. The citizens remains involved via a mobile app, and social media (see Castell et al. (2015)). The main interest of this project is to encompass the short-term air quality effect and the long-term effect on climate change. Indeed, the sensors monitor the CO2 emissions related to traffic along with the driving style in order to promote eco-driving behavior. However, this project relies on a basic design of the data service architecture. Also, it does not measure the actual daily exposure of individuals since only vehicles (or bicycles) are equipped by sensors.

2.5 INTASENSE

The INTASENSE concept is to integrate a number of micro- and nano-sensing technologies onto a common detection platform to produce a low-cost miniaturized system that can comprehensively measure air quality, and identify the nature and form of pollutants (INSTASENSE-A (2013)). The objective of this project is to develop a smart air quality monitoring system that can intelligently interface with existing ventilation and air treatment systems to maximize their energy efficiency and effectiveness. So, the project focuses on indoor pollution only. In order to achieve the objective, several systems have been developed within the scope of the project: (i) a particulate matter detector; (ii) produce a smart miniaturized high performance

sampling and pre-conditioning support platform for use with gas sensors and particle detector; (iii) combustion gas and VOC detector module comprising advanced structured sensors capable of detecting a range of gaseous pollutants to better than 1ppm; and (iv) a wireless sensor network system allowing effective incorporation into building HVAC (Heating, ventilation, and air conditioning) control systems.

The INTASENSE air quality monitor is wirelessly linked to air-handling and pre-conditioning infrastructure allowing air circulation to be managed in an energy efficient way while maintaining a healthy environment (see INSTASENSE-A (2013)).

2.6 hackAir

hackAir is a collective awareness platform for outdoor air pollution. It is an open technology platform that anyone can use to access, collect, and improve air quality information. The main purpose of open platform is to exploit user crowd as source and enrich air quality information. hackAir is funded under European Union's H2020 program and will be ended in 2018. The core objective of hackAir is to develop different collective sensing approaches. The approaches include the following: (a) collecting measurements from existing air quality stations and open data on the Web, which includes environmental related web pages and services, (b) collecting and analyzing sky-depicting images including publicly available geo-tagged and time-stamped images posted through social media platforms (e.g. Flickr[4]), images captured by the users of the hackAIR mobile app, and webcams, and c) crowdsourcing measurements via low-cost open hardware devices (HackAir (2016)).

A platform for collection, fusion, and visualization of air quality has been developed within the hackAir project. As mentioned earlier, it is an open platform; it enables communities of citizens to easily set up air quality monitoring networks and engage their members in measuring and publishing outdoor air pollution levels, leveraging the power of social networks, mobile and open hardware technologies and engagement strategies HackAir (2016). Within hackAir, different data collection methodologies have been developed to extract data from multiple sources that have been classified mainly into text-based sources, image based sources, and hardware based sources. The text based sources include environmental websites, web services; the image sources include flickr and webcam; and the hardware-based sources are essentially sensors. An empirical study was carried out with these resources to identify the best practice best practice techniques for extracting data from these sources. For instance, a domain specific search technique has been proposed to retrieve data from environmental nodes on the Web. The collected data are stored the data layer the hackAir platform. The data layer includes data persistence mechanisms that are used for storing and retrieving data. The data layer consists of traditional (MySQL) storage and NoSQL (MongoDB) storage for storing and managing data. Additionally, there is a knowledge-base for storing semantically enriched information.

The data processing module resides within *business logic layer* of hackAir platform. It processing image data using image processing techniques to locate/identify the best portion (i.e., image quality) of the sky image collected from Flickr. Within the same layer, there is a fusion module which is used to combine data collected from different sources. The fusion module uses geostatistics to combine scattered point based observations of air quality with spatially exhaustive output from a chemical transport model or a statistical air quality model. New

4. https://www.flickr.com/

values are added in observation by spatially interpolating in a mathematical objective function. For special queries, the data retrieval module contained in data layer will enable access and retrieval of requested information. The Knowledge Base module provide support for accessing data, as it contains the semantics information. hackAir solution integrates a recommendation and decision support system which reads data from knowledge base and perform reasoning and provide recommendation to the user. The application layer contains four modules which are mainly used for user interactions in visualization, communication, personalization, and profile management. Each of these modules has specific purpose. For instance, profile management module for user profiling, AQ visualization module is used for visualizing results.

2.7 AirSensa

AirSensa is a privately funded project aimed at delivering sensors for capturing pollution data of NO2 and PM2.5. The objectives of this project is to assist people to understand the quality of air through real-time smart applications which perform analysis with data. AirSensa is a package of solutions including pollution avoidance journey planning, high pollution alerts particularly, particularly for vulnerable groups (see AirSensa (2014)). AirSensa system is founded on a cloud based software platform called STORRM Cloud. The platform enables gathering data from every location where sensors are deployed and then prepares extracted data which are used by the applications that perform analysis and visualization. Unfortunately, no technical details of AirSensa components is provided.

2.8 expAIR

The exposure to urban AIR pollution (expAIR) is a project launched by Brussels Environment (expAIR (2011)) with twofold objective: (i) assessing the individual exposure to air pollution of the people of Brussels both in indoor and outdoor environments, and (ii) raising their awareness to urban pollution to make them change their behavior. Among the various harmful pollutants resulting from human activities, black carbon (BC) particles and Volatile Organic Compounds (VOC) are considered the reference pollutants, respectively, for outdoor and indoor environments. Therefore, BC and VOC data are collected by the participants wearing devices including aethalometers and radiellos for five days, then submitted for analysis to reveal the exposure level, and mapped. The technical description of expAIR solution components has not been found. Specifically, data processing and analysis techniques were not detailed in literature.

2.9 EveryAware

EveryAware is a European Union funded project under FP7 framework. The key notion of this project was that the citizens should be involved not only as passive receivers of prepackaged environmental information, but also as active producers of it, by means of the networking possibilities allowed by mobile devices, pervasive Internet access, Web 2.0 and the mobile Web tools that support sharing and annotation of geo-localised contents. EveryAware aimed at integrating all crucial phases in the management of the environment in a unified framework, by creating a new technological platform combining sensing technologies, networking applications and data-processing tool (see EveryAware (2007)). An integrated plat-

form was developed within EveryAware project to handle both subjective and objective data. The former comprises reactions of humans faced with particular environmental conditions, and the latter stems from sensors.

The EveryAware platform is a modular system based on two hardware components: a smartphone controlling the data acquisition and a modular sensor box with several pluggable sensors. With a software application, the smartphone acts both as data gateway (using standard mobile data connection) and as a local system and user interface. Sensorbox consists of custom hardware and firmware that allows the integration of the air quality sensors and communicates with the smartphone through a Blue-tooth connection. Since Sensorbox comprises various sensors, calibration of sensors is a critical need. Calibration means performing simultaneous measurements with the SensorBox and a reference device, and then train a model that is able to map the values measured by sensor array with the values recorded by the reference. The fundamental idea of SensorBox calibration relies on supervised regression techniques to train sensor array to a target pollutant concentration values gathered by a more reliable and consequently more expensive monitor device. The field calibration is treated as multivariate supervised learning regression approach, where the inputs are the readings from the gas sensor array. To be more specific, Artificial Neural Network (ANN) was used to perform calibration and micro-aethalometers [5] was used for reference. Two types of calibration strategies were adopted in EveryAware: *stationary calibration* and *dynamic calibration*. The former is applied when sensorbox is deployed in a fixed station whereas the latter is applied when the sensorbox is in motion. It is worth noting that EveryAware supports measuring both sound and air pollution using WideNoise and AirProbe applications respectively.

EveryAware has a backend platform which performs collection, storage, and processing of data. The platform comprises conceptual layer and implementation layer. The conceptual layer defines the basic entities and features the EveryAware system supports. The core concepts are data points with descriptions, sessions and feeds. The implementation layer realizes the conceptual layer based on advanced storage and application structures which consists of: the storage itself, the web application for receiving and retrieving data, and data processor which processes and enhances inbound data. MySQL engine has been used for implementing data storage, Apache Tomcat Servlet container has been used for implementing web application which also offers different REST endpoints such as WideNoise endpoint. It is worth noting that EveryAware supports measuring both sound and air pollution using WideNoise and AirProbe applications respectively.

The basic building block of the data storage is data pipeline which is is divided into several logical nodes where one is master node and several are worker nodes. The data processor is responsible for parsing the received data, resolving extensions, apply knowledge discover processing steps, and augmenting them with additional semantic information from various sources. The data processor consists of several components. The module selector component selects processing module which extracts actual data from raw contents. The storage handler stores resulting data in dedicated content table, the output table, and possibly semantics table. The application endpoints WideNoise and AirProbe use these data.

5. micro-aethalometers are used to measure the black carbon

3 A Comparative Study

We performed capability comparison and technical comparison among the solutions proposed in projects discussed in the sections earlier. We considered four critical criteria to compare these solutions. These include the sensing model (Indoor/Outdoor), the coverage of pollutants, the functional capabilities, and finally the architectural Strength (e.g., scalability).

3.1 Sensing Model

The Table below shows the sensing model supported by the state of the art solutions of existing air quality monitoring systems developed within the scope of above projects. Table 1 shows that 3 out of 9 solutions: OpenSense II, CITI-SENSE, and EveryAware provide supports for monitoring and measuring both indoor and outdoor air quality. Vividly, both sensing model is critical for measuring individual exposure level. Only good air quality in outside is not sufficient for preventing health hazard as a significant period of time people spent in home and also in micro-environment such as kitchen in home and bus while traveling. Therefore, measuring air quality in such environments is inevitable and hence, the technologies should be able to provide support for both environment.

Name of the Project	Indoor / Outdoor
OpenSense	outdoor
OpenSense II	both
CITI-SENSE	outdoor / indoor [6]
CITI-SENSE-MOB	outdoor
hackAir	outdoor
INTASENSE	indoor
expAIR	indoor
EveryAware	both
AirSensa	outdoor

TAB. 1 – *The sensing models used in different projects*

Such capability will help to develop more efficient solutions which will increase situational awareness significantly.

3.2 Pollutant Coverage

There is no *all-in-one* sensor which alone can capture the pollution levels of all pollutants that are carcinogenic or else that lead to severe damage of human health. Therefore, several sensors are assembled together in a box and deployed in a fixed station or carried by human or vehicles. Each of the sensors capture data of a specific pollutant. All the projects used this approach to collect data of different pollutants. Table 2 shows the pollutants covered within the projects surveyed in this paper.

Evidently, the highest number of sensors was tested in EveryAware. In fact, almost all major health hazard chemical compounds can be monitored by the sensors contained in the

Name of the Project	Pollutants
OpenSense	CO, NO2, O3, PM
OpenSense II	CO, NO2, O3, PM, SO2, VOC
CITI-SENSE	NO, NO2, O3, CO2, SO2, PM1, PM2.5, PM10
CITI-SENSE-MOB	CO2 (Note: The system is able to use the sensors of CITI-SENSE)
hackAir	CO2
INTASENSE	PM, VOC
expAIR	CO, CO2, SO, NO, VOC, PM10
EveryAware	O3, VOC, CO, NO_x, NO_2, Black Carbon, PM2.5, Temperature, Humidity
AirSensa	PM and NO2

TAB. 2 – *The list of pollutants covered by different projects*

SensorBox. CITI-SENSE, expAIR, and OpenSense II platforms cover a fair number of pollutants. hackAir is the weakest platform as it is able to measure only CO2.

3.3 Functional Capabilities

According to our study, most of the solutions share common functional capabilities including data collection, data calibration, data processing, analysis, and visualization. However, the underlying techniques which drive these functionalities is different. It is worth noting that no detail of functional capacities of the projects include INTASENSE, AirSensa, and expAIR is available. According to our study, the two most functionally rich solutions are offered by OpenSense II and EveryAware. Additionally, the underlying technologies used in these function is advanced. For instance, deep neural network is used in EveryAware platform for sensor calibration. Since deep deep neural network enables to define multiple hidden layers, data calibration can be done efficiently and effectively.

3.4 Architectural Strength

The architectural style of all the proposed solutions are almost the same. They follow conventional multi-layered software architectural paradigm. Also, most of the solutions adopt service oriented paradigm (as some of the services are loosely coupled and rely on REST principles). The major drawback of these architectural paradigms is scalability at physical level. More specifically, it is rather impossible to add more nodes *on the fly* or *on demand*. Notably, the conventional architectural styles are adequately efficient if the dataset is small or number inbound request to application server or data processing engine is small. However, continuous streaming of miniaturized flow can lead failure of servers and hence, the availability could be highly challenging.

4 Polluscope: The New Technical Horizon for Air Quality Monitoring

In this section, we describe the on-going Polluscope project [7], an ANR project which envisaged to address the limitations of state of the art technologies of air quality monitoring.

Taking advantage of the technological evolution of wearable and lightweight environmental sensors, the interdisciplinary Polluscope project aims at bringing together experts from environmental, metrology, epidemiological, and data sciences while providing methodologies, techniques, and tools - expected to drastically change the way individual's exposure and exposure variability are measured, perceived, and evaluated. The first objective of Polluscope is to improve the knowledge of individual exposure anywhere at anytime. It suggests a new concept, namely a *community-based participatory observatory for pollution and exposure* where citizens contribute data to the system with the purpose of sharing events of interest within the community. The measurements will consider gaseous pollutants (Ozone, NO2), and particulate matter among which black carbon, and VOC, which provides a representative overview of the air pollution. Gaining such enriched insights into individual's exposure will contribute towards reducing individual risks of some diseases by changing their behavior. This will end up in a solid, invaluable, and vital societal impact namely, saving life and improving the individual well-being.

To achieve the aforementioned objectives, a novel infrastructure for real individual's exposure data acquisition, processing, and analysis will be developed. For this to be done, several scientific and technical challenges come into the picture. The data are collected at a high frequency and might be massive and noisy. Therefore, the system must be able to process them efficiently, while taking into account both their velocity and their uncertainty. More importantly, it has to offer microenvironment and user's activity recognition, through integration with external spatio-temporal resources. An efficient data collection and analysis will provide an insightful knowledge on individual's exposure over his/her daily life activities, and will enable conducting analytical queries, novel risk assessment modeling, mining and comparing profiles of pollution exposures, and so on. Therefore, it is evident that a robust, efficient, and powerful data science technology is crucial.

Lastly, Polluscope will be evaluated under real-world use cases. In particular, several type of population will be targeted by the data acquisition campaign. Both diseased and healthy subjects will be involved to conduct an epidemiological study relating air pollution exposure to health on the one hand, and volunteer participants for the crowd sensing on the other hand.

5 Conclusion

In this paper, we provided a comprehensive study of different projects aimed at providing technology enabled air pollution prevention ecosystems. We discussed the technologies and tools proposed in these projects. We provided a comparison of the proposed solutions with a special focus on outlining their strength and weaknesses. Also, we briefly discussed the Polluscope project that we have been working on aiming at addressing the shortcoming of state of the art. In the future, we plan to develop different solutions within Polluscope ecosystem.

7. http://polluscope.uvsq.fr/

References

Aberer, K., M. Hauswirth, and A. Salehi (2006). A middleware for fast and flexible sensor network deployment. In *Proc. of the 32nd int. conf. on Very Large Data Bases (VLDB)*, pp. 1199–1202. VLDB Endowment.

Afrifa, C. G., F. G. Ofosu, S. A. Bamford, D. A. Wordson, S. M. Atiemo, I. J. Aboh, and J. P. Adeti (2013). Heavy metal contamination in surface soil dust at selected fuel filling stations in accra, ghana. *American Journal of Scientific and Industrial Research 4*(4), 404–413.

AirSensa (2014). Airsensa. `http://www.airsensa.org/how.php`. Accessed: February 18, 2018.

Amos, J. (2016). Polluted air causes 5.5 million deaths a year new research says. `http://www.bbc.com/news/science-environment-35568249`. February 19, 2018.

Bentayeb, M., V. Wagner, M. Stempfelet, M. Zins, M. Goldberg, M. Pascal, S. Larrieu, P. Beaudeau, S. Cassadou, D. Eilstein, et al. (2015). Association between long-term exposure to air pollution and mortality in france: A 25-year follow-up study. *Environment international 85*, 5–14.

Buchli, B., M. Yuecel, R. Lim, T. Gsell, and J. Beutel (2011). Demo abstract: Feature-rich platform for wsn design space exploration. In *Information Processing in Sensor Networks (IPSN), 2011 10th International Conference on*, pp. 115–116. IEEE.

Cartier, S., S. Sathe, D. Chakraborty, and K. Aberer (2012). Condense: managing data in community-driven mobile geosensor networks. In *Proc. of the 9th Annual Comm. Society Conf. on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*, pp. 515–523. IEEE.

Castell, N., M. Kobernus, H.-Y. Liu, P. Schneider, W. Lahoz, A. J. Berre, and J. Noll (2015). Mobile technologies and services for environmental monitoring: The citi-sense-mob approach. *Urban climate 14*, 370–382.

CITI-SENSE (2016). Citisense - development of sensor based citizen's based observatory community for improving quality of life in cities. `http://www.citi-sense.eu/Default.aspx`. February 16, 2018.

EveryAware (2007). Everyaware: Enhancing environmental awareness through social information technologies. `http://www.everyaware.eu/wp-content/uploads/2011/04/EveryAware.pdf`. February 19, 2018.

expAIR (2011). The expair project: Assessing the individual exposure of the people of brussels. `http://document.leefmilieu.brussels/`. February 16, 2018.

Ganti, R. K., F. Ye, and H. Lei (2011). Mobile crowdsensing: current state and future challenges. *IEEE Communications Magazine 49*(11).

HackAir (2016). Hackair: Environmental node discovery, indexing and data acquisition. `http://www.hackair.eu/wp-content/uploads/2016/12/d3.1_environmental_node_discovery_indexing_and_data_acquisition_1st_.pdf`. February 19, 2018.

INSTASENSE-A (2013). Instasense: Integrated air quality sensor for energy efficient environment control. `http://www.intasense.eu/index.php/9-uncategorised/`

24-welcome. February 19, 2018.

Li, J. J., B. Faltings, O. Saukh, D. Hasenfratz, and J. Beutel (2012). Sensing the air we breathe - the opensense zurich dataset. In *Proceedings of the National Conference on Artificial Intelligence*, Volume 1, pp. 323–325.

Liu, H.-Y., M. Kobernus, D. Broday, and A. Bartonova (2014). A conceptual approach to a citizens' observatory–supporting community-based environmental governance. *Environmental Health 13*(1), 107.

Riahi, M., T. G. Papaioannou, I. Trummer, and K. Aberer (2013). Utility-driven data acquisition in participatory sensing. In *Proceedings of the 16th International Conference on Extending Database Technology*, pp. 251–262. ACM.

Sathe, S. and K. Aberer (2013). Affinity: Efficiently querying statistical measures on time-series data. In *Proc. of the 29th Int. Conf. on Data Engineering (ICDE)*, pp. 841–852. IEEE.

Sathe, S., Hoyoung, and K. Aberer (2011). Creating probabilistic databases from imprecise time-series data. In *Proc. of the 27th Int. Conf. on Data Engineering (ICDE)*, pp. 327–338. IEEE.

Slørdal, L., S. Solberg, and S. Walker (2003). The urban air dispersion model episode applied in airquis2003. technical description. *Norwegian Institute for Air Research, Kjeller (NILU TR 12/03)*.

Xia, F., L. T. Yang, L. v, and A. Vinel (2012). Internet of things. *International Journal of Communication Systems 25*(9), 1101.

Yan, Z., D. Chakraborty, C. Parent, S. Spaccapietra, and K. Aberer (2011). Semitri: a framework for semantic annotation of heterogeneous trajectories. In *Proc. of the 14th int. conf. on extending database technology*, pp. 259–270. ACM.

Yan, Z., J. Eberle, and K. Aberer (2012). Optimos: Optimal sensing for mobile sensors. In *Proc. of the 13th Int. Conf. on Mobile Data Management (MDM)*, pp. 105–114. IEEE.

Résumé

L'émergence de nouveaux capteurs environnementaux à bas coût, légers et connectés amène à un changement de paradigme dans les études environnementales. En particulier, grâce à ces capteurs nomade, les mesures personnelles en continu de différents polluants permettent de quantifier l'exposition individuelle aux risques sanitaire de la pollution de l'air avec une précision jamais égalée auparavant. Par ailleurs, une large diffusion auprès de contributeurs volontaires (en statique et en mobilité) ou sur des flottes de véhicules devient un scénario crédible. Elle présente un intérêt majeur de massification du éseau d'observations jusque là limitées à des stations éparses de mesures retournant des moyennes horaires. Cette haute résolution spatiale et temporelle soulève néanmoins des questions quant au traitement de données. Les principales concernent la modélisation, la prise en compte de la qualité, la sémantique et le croisement avec des sources de données traditionnelles, et finalement le passage à l'échelle de l'analyse de ces données face à des volumes et de flux de capteurs conséquents. Après un tour d'horizon des projets liés, cet article met en évidence les défis qui seront adressÃ©s au sein du projet Polluscope en cours.

Chapter

5

Big data & Business Intelligence

ASD'2018

Content

A Clustering-based Approach To Build Distributed Data Warehouse Using a Column family NoSQL Database

Mohamed Boussahoua*, Omar Boussaid*, Fadila Bentayeb*, Nadia Kabachi **

*University Lumiere Lyon 2, ERIC EA 3083,
5, avenue Pierre Mendês 69676 Bron-France
{Mohamed.Boussahoua, Omar.Boussaid, Fadila.Bentayeb}@univ-lyon2.fr
**University Claude Bernard Lyon 1,
43, boulevard du 11 novembre 1918, 69100, Villeurbanne-France
Nadia.Kabachi@univ-lyon1.fr

Abstract. The column family NoSQL databases offer storage techniques that are well adapted to data warehouses. Several scenarios are possible to develop the data warehouse on these databases. In this paper, we propose a new method to build a distributed data warehouse using a column family NoSQL DBMS. Our method is based on an attribute-grouping strategy to define the column families that constitute the logical warehouse schema, which allows the most appropriate physical data model. For this purpose, the *Particle Swarm Optimization algorithm* is used to group different attributes that are accessed together to create the column families. To evaluate our method, we adopt the TPC-DS benchmark. We then carried out several tests to show the effectiveness of this algorithm to build a data warehouse in NoSQL HBase database on a Hadoop platform. Our experiments suggest that defining a good data grouping on HBase database during the implementation of a data warehouse increase significantly the performance of the decisional queries.

Keywords: NoSQL databases, Data warehouses, Data partition.

1 Introduction

In the Business Analytics Platform, the Data Warehouse (DW) provides a quick and cost-effective way for reporting and data analysis. DWs are often implemented in relational database management systems (RDBMS). However, the unusual volume of data becomes an issue when faced with the limited capacities of traditional systems, especially when data storage in a distributed environment. To address this issue, optimization techniques such as: vertical and horizontal partitioning (Navathe et al., 1984) and (Bellatreche and Boukhalfa, 2005), materialized views (Gupta, 1999), and indexes (Golfarelli et al., 2002), have been used in traditional RDBMS. However, these

techniques are usually expensive, in these instances, the data warehousing relational systems would require processing overhead to maintain a database schema that could accommodate all the information. In recent years, other optimized data warehouse solutions were developed to improve query performance. For example, some companies are focusing on using NoSQL systems [1]. These new approaches constitute an interesting way of constructing data warehouses able to support large masses of data. However, the problematic thing with NoSQL systems is that the data implementation process is not trivial (de Freitas et al., 2016) because each NoSQL database model has specific data structures and concepts (e.g. Key-Value stores, Column families databases, Document databases, Graph databases, and other models NoSQL databases). Consequently, these approaches require revisiting the principles of traditional data warehouses modeling process, especially at the logical and physical design level.

In this paper, we address the storage and implementation process of data warehouses with column family NoSQL databases. So, to take advantage of these types of databases, the main question that arises when trying to accommodate the data structures is: how to organize data in column families to serve effectively OLAP queries? To answer this question, we studied the benefits of grouping techniques on column families' creation within the context of data warehousing. In this case, we propose the application of a clustering technique, based on the meta-heuristic *Particle Swarm Optimization (PSO)*, to determine which attributes frequently used by queries should be grouped together. In order to obtaining a better design of column families schema leads to a more optimized physical data schema for data distribution in a multi-nodes cluster by using column family NoSQL DBMS. Several tests were made to evaluate the effectiveness of the proposed method. We adopt the TPC-DS data benchmark. To design the columnar NoSQL data warehouse (*CN-DW*) for TPC-DS benchmarking database, we used 3 different methods, first our method, and then 2 other methods that have been already tested and implemented successfully in (Dehdouh et al., 2015) and (Chevalier et al., 2015). We proceed to perform a query workload on different schemas upon *CN-DW* built over TPC-DS. It has been found that the application of clustering techniques for designing a data model *CN-DW* effectively improves the query execution time, compared to the other two other methods.

The remainder of the paper is organized as follows. Section 2 presents the related works and the problem we address in this paper. Section 3 presents the proposed approach. Section 4 evaluates our approach. Conclusion and future works are given in section 5.

2 Related work and Problem statement

2.1 Related Work

Using column family NoSQL databases for data warehousing solutions, has been debated within the scientific community. Several works, like (Li, 2010), (Dehdouh et al.,

1. http://nosql-database.org/

2015), (Chevalier et al., 2015) and (Yangui et al., 2016), have treated the problem of modeling and implementing the data warehouse according to these models. These works can be classified into two main categories (Figure 1):

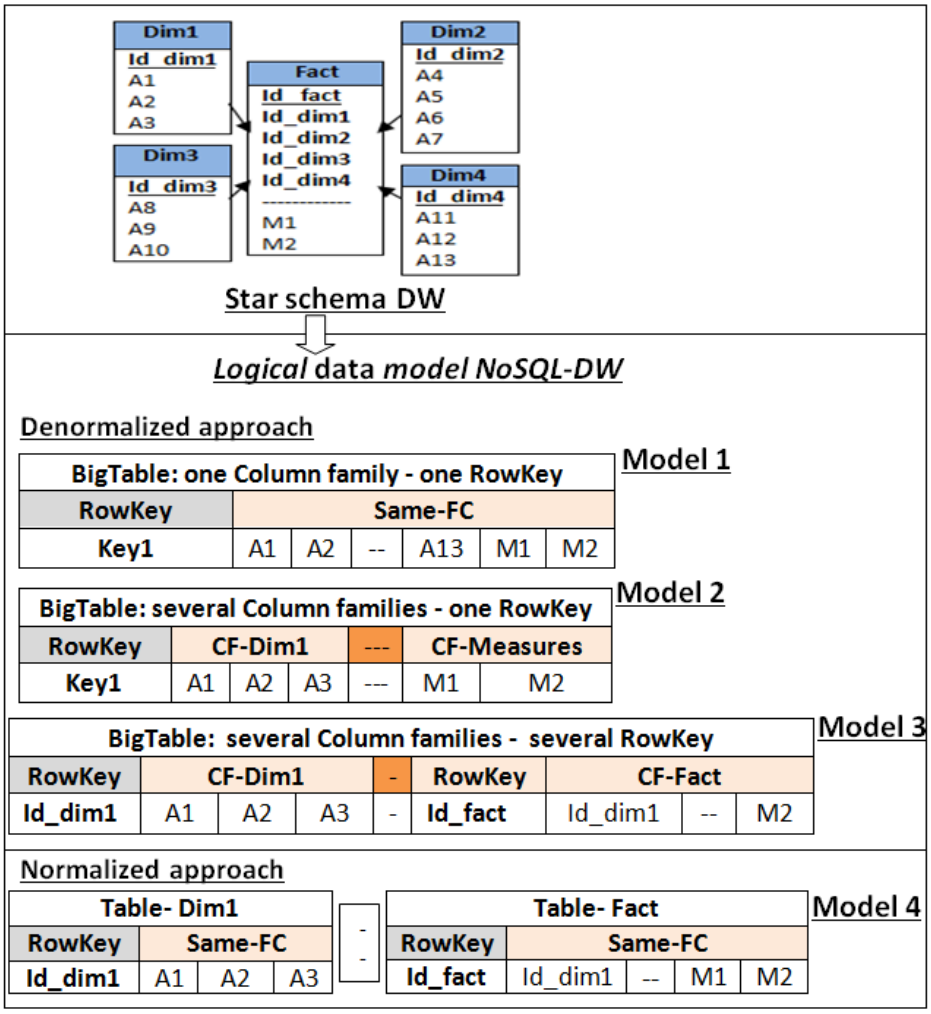


FIG. 1 – *Data warehouse- columnar logical models*

A) The first one (Denormalized approach): the aim of these works is to propose a storage schema that combines the fact and dimension tables into one table. 3 solutions are commonly applied for building the column family schema:

1. The first logical data model *(model 1)*: all attributes of fact and dimension tables are combined in one column family;

2. The second logical data model *(model 2)*: has one column family for each dimension table and one column family dedicated for the fact table. *All column families are referenced by one RowKey*;
3. The third logical data model *(model 3)*: one column family for each dimension table and one column family dedicated for the fact table. But, *each column family is referenced by specific Rowkey.*

b)The second approach (Normalized approach): it uses different tables *(Separate Tables)* for storing fact and dimension tables at a physical level. The fact table is stored into one table with one column family, each dimension table is stored into one table with one column family *(model 4)*.

Note: the column family NoSQL databases provide variable-width tables that can be partitioned vertically and horizontally across multiple nodes of a cluster. Moreover, the RowKey ensures a data horizontal partitioning mechanism, the Column family ensures a data vertical partitioning mechanism.

We have noticed that, these methods have certain shortcomings:

Solution	Disadvantages
Model 1	- Loss of the benefits of the vertical partitioning
Model 2	- Imbalance between column families - No control the number of column families can be generated
Model 3	- Imbalance between different column families - No control the number of column families can be generated - Loss of the benefits of the horizontal partitioning - Create a special join *(java codes)* between column families
Model 4	- Create a special join *(java codes)* between column families - Loss of the benefits of horizontal and vertical partitioning

TAB. 1 – *certain disadvantages in Naive solutions.*

Other research efforts tried to enhance the performance of the data schema by optimizing column family schema. In (Romero et al., 2015), to speed searching and have direct access to data blocks in column families structure, the authors propose a promising approach, in which they use composite indexes on the HBase table. In (Scabora et al., 2016), to solve the problem of distributing attributes between column families, they implemented, by intuition, the data warehouse in an HBase table with two column families. The first one groups the attributes of fact and dimension tables more frequently interrogated. The second column family contains the attributes of the other dimensions. But the authors do not take into consideration the clustering techniques to create the column families that contain the required data for processing a query or multiple queries. In (Yang et al., 2015), the authors propose an automatic approach based on a Genetic Algorithm to optimize the column family schema in HBase. The authors did not focus on the data warehouse implementation process on

column family NoSQL databases. They evaluate their approach by using simple data sets and basic queries.

2.2 Problem Statement

The implementation of a data warehouse that incorporates the best features of the column family NoSQL systems (scalability, aggregation capabilities and data partition options like the vertical and horizontal partitioning) is the goal of several research works. We have seen that implementation processes of the data warehouse based on these systems usually use denormalized approaches. In addition, these works are based essentially on only one input parameter: the conceptual model or the relational logic model. These are used as input parameters in the phases of the column family schema design process based on certain rules of transformation between the relational schemas to NoSQL schemas. In this case, the NoSQL modeling remains dependent on the relational modeling of the data warehouse, that can be suggested as the more generalized design process. It should be noted that the column family schema in HBase, Cassandra and BigTable can be considered a physical mechanism used to vertically distribute the writing and query load across the cluster's nodes. Therefore, an appropriate column family schema design can help in tuning the data warehouse's performance.

In this work, we tried to find new data models to improve the response times to the queries complex. We looked at the impact of other column family specific parameters on the performance of a data warehouse. To do this, we propose a clustering-based approach for column family schema construction to improve query gain performances. Our goal is to minimize the total amount of data scanned while performing OLAP query by optimizing:

1. the number of column families with a vertical pre-partitioning of data warehouse schema before its implementation;
2. the number of columns in column families.

3 The proposed Approach

Our strategy, to create a data warehouse in a column family NoSQL system, is subdivided into 2 steps that are detailed in the following sub-sections.

3.1 Building the Attribute Usage Matrix (AUM)

This step consists in processing the set of attributes relating to the initial query workload. To build AUM, we taken account all the attributes present in each query *(those that appear in the Select and Where clauses, except for the attributes of the join predicates)*. Let the workload consists of a set of most frequent queries $Q = \{q_1, q_2, .., q_N\}$, that access the set of attributes $R = \{a_1, a_2, ..., a_M\}$. The matrix AUM represents couples $((q_i)_{i=1,..,N}, (a_j)_{j=1,..,M})$, where general term AQ_{ij} equal to 1 if a_j appears in query q_i and to 0 otherwise. To illustrate, let $T = \{Fact, Dim1, Dim2, Dim3\}$ be the set of the warehouse tables, the workload consists of a set of queries $Q =$

$\{q_1, q_2, q_3\}$ that access the set of attributes $R = \{A1, .., A7, M1, M2, Id\_dim1\}$. Figure 2 shows the AUM corresponding to Q.

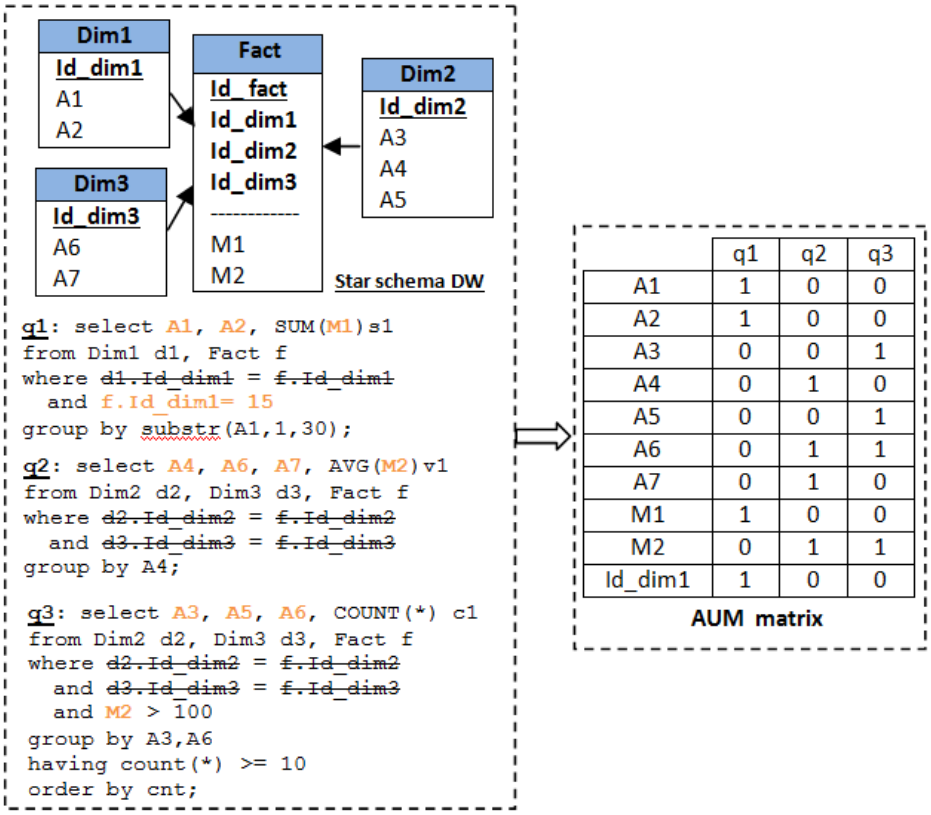


FIG. 2 – *example Attribute Usage Matrix AUM*

3.2 Constructing Column Families

In this second phase, our goal is to generate the column family schema, that optimizes data access for query workload. Our solution is to implement a process of grouping the attributes that are frequently queried together. This grouping will form set of column families that make up the logical schema of the columnar NoSQL data warehouse*(CN-DW)*. We chose to use the meta-heuristic *Particle Swarm Optimization (PSO)* algorithm (Eberhart et al., 1995). This algorithm is inspired by the swarms of insects or animals and their displacement in groups to find needs. *PSO* processes a population (called *swarm*) of ($N \geqslant 2$) *particles*. The process starts with a random initialization of the swarm in the search space. During the optimization process, at each iteration, each particle is moving according to a velocity. To do this, it linearly combines three fundamental information: - its current speed; - its best position (until

to the i^{th} iteration); - the best position of its neighbors in the entire swarm.

Our choice to use *PSO* is motivated by the fact:

1. It is a dynamic approach;
2. It allows us to define two input parameters:
   (a) the maximum number of groups to be constructed,
   (b) the maximum number of individuals in a group.

This proves to be an advantage as long as we want to control, on the one hand, the number of column families that can be created, on the other hand, the number of columns in column families. This helps to build well-balanced column families. To make it simple, to improve the design of the column family schema corresponding to the query workload Q, the optimization problem can be defined as follows:

— $Q = \{q_1, q_2, ..., q_N\}$: set of N query;
— $R = \{a_1, .., a_M\}$: set of M attributes used by Q;
— fq_l: access frequency related to a query $(q_l)_{l=1,..,N}$;
— W: Max number of column families $(2 \leq W)$,
— B: Max number of attributes in a column family $(Int(\frac{M}{W}) \leq B \leq Int(\frac{M}{W}) + 1)$,
— $S = \{S_1, S_2, ..., S_z\}$: the set of all realizable solutions, where z is the iterations number of the *PSO* algorithm, $(S_i)_{i=1,..,z} = \{CFi_1, ..., CFi_W\}$, $(CFi_j)_{j=1,..,W}$ are subsets of attributes, such as:
   1. $\forall CFi_j \in S_i : CFi_j \subset R$,
   2. $\forall a \in R : \exists CFi_j \in S_i : a \in CFi_j$,
   3. $\forall CFi_j, CFi_h \in S_i : CFi_j \cap CFi_h = \varnothing$.
— F : An objectif function that takes its values on S.

The problem is to find a solution $S^* \in S$ witch optimizes the value of the objectif function F such as: $\forall (S_i)_{i=1,..,z} \in S : F(S^*) \leq F(S_i)$.

The objectif function F allows to measure the quality of S_i solutions obtained after each iteration of *PSO*. Our cost function based on the works of (Derrar et al., 2015). Initially, this function is computed using the *Square Error (E^2)*, taking account of the access frequency of queries. The *(E^2)* of the attribute groups schema (S_i) is calculated as follow:

$$E^2_{S_i} = \sum_{j=1}^{W} \sum_{l=1}^{N} [(f_{q_l})^2 \times \alpha_j^{q_l} (1 - \frac{\alpha_j^{q_l}}{\beta_j})] \tag{1}$$

$(\alpha_j^{q_l})$ is the number of attributes in CFi_j appearing on a schema S_i accessed by the query q_l, (β_j) is the total number of attributes in CFi_j. Also, more the $E^2_{S_i}$ value approaches 0, more optimum is this grouping schema.

To illustrate this point, Let's return to our previous example (AUM in Figure 2), we consider the frequency $(f_{q1} = f_{q2} = f_{q3} = 1)$, $(W = 4)$ thus $(2 \leq B \leq 3)$. We assume that in the i^{th} iteration of the *PSO* the set of attribute groups are: $CF1 : (A1, A2, Id\_dim1), CF2 : (A3, A4, A5), CF3 : (A6, M2), CF4 : (A7, M1)$. The *Square*

Error is calculated as follows:

	E_{CF1}	E_{CF2}	E_{CF3}	E_{CF4}	Sum_{qi}
q1	$(3*(1-3/3))$	$(0*(1-0/3))$	$(0*(1-0/2))$	$(1*(1-1/2))$	0.50
q2	$(0*(1-0/3))$	$(1*(1-1/3))$	$(2*(1-2/2))$	$(1*(1-1/2))$	1.16
q3	$(0*(1-0/3))$	$(2*(1-2/3))$	$(2*(1-2/2))$	$(0*(1-0/2))$	0.66
$E^2_{S_i}$	0.00	1.32	0.00	1.00	2.32

TAB. 2 – *example of how to calculate the sum of Square Error*

In this iteration, we see that the group $CF2$ *(whose value is* 1.32 *is the largest value of the four outcomes)* contain the attributes that are least queried together. In algorithm 1 we present a general version of the *PSO*, which exploits the objective function F to build the column families. For this version, we consider each of these particles represents one attribute. The PSO algorithm considers as input : - the Attribute Usage Matrix (AUM) for a set of queries Q; - the maximum number W of groups and the maximum number B of attributes by group; - *PSO algorithm* parameters. For each iteration i of the *PSO*, one grouping attributes schema (S_i) is generated, which will be evaluated according to a cost function $F(E^2, S_i)$. At the end, the best grouping output will form the column families of the *CN-DW*. Our method is summarized in Figure3 (we do not have space to present all part of the *PSO*)

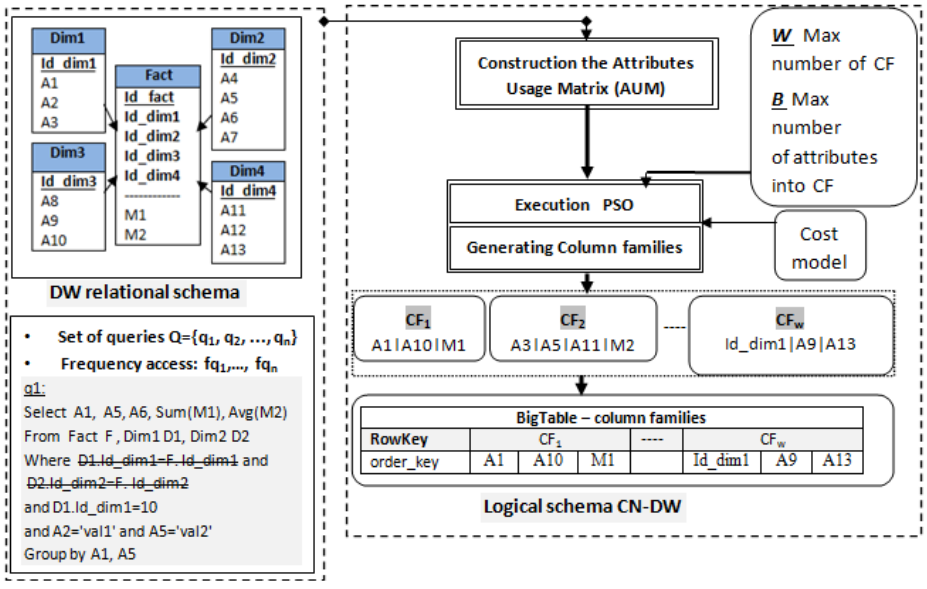


FIG. 3 – *PSO method to design columnar NoSQL data warehouse*

Algorithm 1 PSO-CF algorithm

Input:
AUM : *Attribute Usage Matrix*; AFM : *Access Frequency Matrix*
W: *maximum number of column families*
B: *maximum number of attributes in a column family*
VPSO: *Set of the variants of PSO*
Notation:
S_{init}: *initial attribute groups,* S_i*:* i^{th} *solution*
F : *PSO objectif function*
S_{opt} : *optimal attribute groups schema*
begin
Randomly initialize the W attribute groups: S_0 = Swarm(S_{init})
while Stop criterion is not satisfied **do**
1- Determine the best position of all or part of swarm
2- Particle displacement depending on the adopted strategy
3- Structural adaptations by executing Split and Merge functions
4- Update the velocity and position of the particles
5- Evaluate the objectif function $F(E^2, S_i)$
6- If Best Solution : $S_{opt} = S_i$
end while
return (S_{opt}) */* Optimal column family schema */*
End

4 Implementation, experiments and results

To validate our PSO method for designing the column families, we developed a software tool named (RDW2CNoSQL: Relational Data warehouse to Columnar NoSQL) with Java programming language.

1. Dataset: To evaluate our approach, we used the TPC-DS benchmark [2]. The TPC-DS uses a constellation schema which consists of 17 dimension tables and 7 fact tables. In our case, we used the STORE_SALES fact table and its 9 dimension tables (CUSTOMER, CUSTOMER_DEMOGRAPHICS, CUSTOMER_ADDRESS, ITEM, TIME, DATE, HOUSEHOLD_DEMOGRAPHICS, PROMOTION, STORE). The DSDGEN data generator of TPC-DS allows to generate data files in a (*file.data*) format with different sizes according to a *Scale Factor (SF)*. We set *SF* to 100 which produces in STORE_SALES fact table (287.997.024 tuples).

2. Query workload: The TPC-DS benchmark offers 99 queries. We selected 19 separate queries (Table 3) that access 67 attributes, which exploit the entire schema of the STORE_SALES fact table and its dimension tables, using the operations *(selection, join, aggregate, projection)*. These queries compute the OLAP cubes with a gradually increasing number of dimensions. The degree of this dimensionality is divided into 3 levels: (*small: SD)*, (*medium: MD)* and (*large: LD)*, according to: (1) The number of tables used by a query; (2) The number of attributes and predicates for each query. It

2. Benchmark (TPC-DS) v2.0.0, http://www.tpc.org/tpcds/.

should be noted that our objective is to use TPC-DS benchmark to evaluate the performance of our technique when forming column families. Due to, some requirements are not feasible with Apache Phoenix [3] (on query read capabilities) and HBase databases, these queries would require some modifications (syntax changes).

	(SD)			(MD)									(LD)						
	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10	q11	q12	q13	q14	q15	q16	q17	q18	q19
Tables	1		2	3							4		5						6
Attributes	4	9	9	4	4	6	6	6	6	9	6	7	8	10	10	11	11	12	14
Predicates	4	3	10	4	5	12	7	7	6	6	5	7	21	13	10	8	23	15	15

TAB. 3 – *Queries characteristics*

3. Experimental configuration: To achieve our evaluation goals, we setup two storage environments. The first one is relational non-distributed with intel-core machine TMi7-4790S CPU@3.20 GHZ with 8 GB of RAM, and a 500 GB disk. It runs under the 64-bit Ubuntu-14.04 LTS operating system, which is used as a PostgreSQL server dedicated to the storage of the relational data warehouse. The second is a distributed NoSQL storage environment. It is a cluster of computers consisting of 1 master server *(NameNode)* and 3 slave machines *(Data Nodes)*. The *(NameNode)* has an Intel-Core TMi5-3550 processor CPU@3.30 GHZx4 with 16 GB RAM, and a 1TB SATA drive. Each of the *(DataNodes)* has an Intel-Core TMi5-3550 processor CPU@3.30 GHZx4 with 16 GB RAM and 500 GB of disk space. These machines run on 64bit Ubuntu-14.04 LTS and Java JDK 8. We used Hadoop (v2.6.0), MapReduce for processing, HBase (v0.98.8), ZooKeeper for track the status of distributed data in the *Region-Servers (DataNodes)*, Phoenix (v4.6.0) and SQuirreL SQL Client to simplify data manipulation and increase the performance of the HBase.

4.Tests and results: We choose three different methods, 2 already existing approaches in addition to our method, for implementing the *TPC-DS* in HBase system: (1)in the first one, the STORE_SALES fact table and its 9 dimension tables are stored into one HBase table with only one column family for all attributes (called *Flat schema*); (2) in the second, the STORE_SALES fact table and its 9 dimension tables are stored into one HBase table with 9 column families, each of the tables would correspond to a column family (called *Naïve schema*); (3) the last one consists of *CN-DW*, built according to our method with in addition three different schemas, i.e all data are stored in one HBase table, we varied the number of column families ($W = 5, W = 10, W = 12$) corresponding to *(schema W=5 with* $E^2_{w=5} = 3.94$*)*, *(schema W=10 with* $E^2_{w=10} = 1.61$*)* and *(schema W=12 with* $E^2_{w=12} = 1.93$*)*. We executed all queries presented in (Table 3), on the five schemas described above. Note that, in this experiment, we do not want to make a performance comparison between the relational DBMS and NoSQL databases. Our primary focus is to seek the main elements that have an impact on query execution time in a Columnar NoSQL Data warehouse.

5. Discussion: We discus our results in this subsection.

a) Impact of the data size on query execution time:

3. https://phoenix.apache.org/

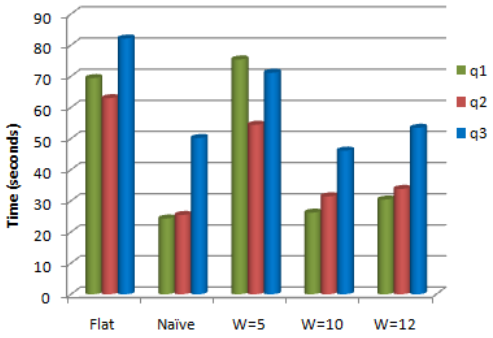


FIG. 4 – *SD Queries execution time*

As shown in Figure 4, queries execution time increases significantly, when using the *(Flat schema)* or the *(schema W=5)*. These schemas record poor results compared with other schemas *(Naïve, schema W=10, schema W=12)*. In the *(schema W=5)*, these results are due to its poor quality *(is caused by bad choices of the number of column families, having a greater value of the* ($E^2_{w=5} = 3.94$)*)*. But, in the *(Flat schema)*, these results are due to the pressure on the memory caused by the large amounts of data coming from the same column family *(in Flat method: all attributes of the fact and dimension tables are combined in one column family)*. Indeed, In HBase, the data from a single column family are stored in a set of *HFiles* (*the number of HFiles depends on the data size in a column family)*. For reading data in *Flat schema*, HBase will automatically solicit and loaded into memory a large number of $HFiles$, this offers the possibility of performing multiple processing at the memory, which results in increased execution time and decreased system performance. On the other hand, we observe a slight variation between the queries execution times, run on the schemas *(Naïve, schema W=10 and schema W=12)*. In the*(Naïve schema)* the queries frequent 1 to 2 column families, in the *(schema W=10* and *schema W=12)*, the queries use 2 to 3 column families. To respond to $q1$, $q2$ and $q3$ in these 3 schemas, HBase exploits column families having small data sizes *(the number of columns in each column family of the schema W=10 and the schema W=12 , not exceed 7 and 6, respectively)*. This, allows it to considerably reduces the number of *HFiles* in the memory *(fewer HFiles)* of data are scanned during the query execution.

b) Impact of the number of column families on query execution time:
The objective of this experiment is to examine the scalability of the *PSO method*, when faced with variations in the number of dimensions. To do this, we executed 16 queries ($q4$ to $q19$) presented in (Table 3). These queries compute the OLAP cubes with a gradually increasing number of dimensions. In Figures 5 and 6, we observed that the proposed method gives better performance for all queries, whatever the number of dimensions involved, when using the *(schema W=10, schema W=12)*, in these schemas the queries can be recalled (2 to 4 column families). On the other hand, from Figure 6 it can be seen that *(Naïve schema)* and the *(Flat schema)* are equivalent

in terms of response times for some complex queries *(LD queries)*, this was foreseeable. Indeed, to respond of these queries in the *(Naïve schema)*, HBase system solicits a very large number of column families (5 to 6 column families), which generates a high cost of combinations and reconstruction of the intermediate results. Recall that the the *Naïve approach* constructs the column families according to the principle where each dimension of the relational model must be transformed into a column family.

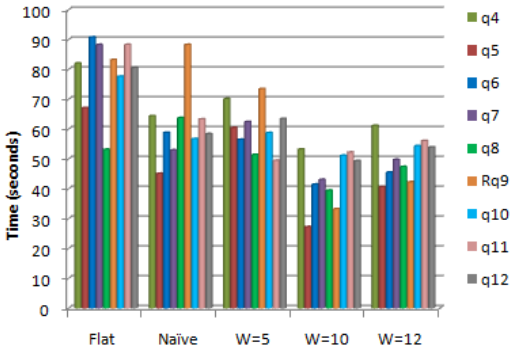


FIG. 5 – *MD Queries execution time*

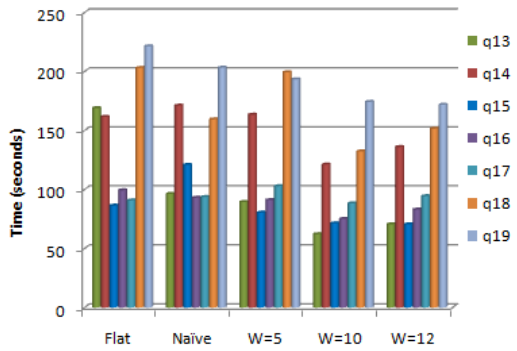


FIG. 6 – *LD Queries execution time*

Finally, by analyzing these preliminary results, we observe the query runtime is dependent on the way to model the form of column families. Figure 7 shows the *PSO method* in the *(schema W=10)* has lowered global query execution time, up to 23.2% and 37.7% compared to *Naïve approach* and *Flat approach*, respectively. In general, to improve query run time on the data warehouse implemented on HBase database, It's readily apparent that: 1-limit the number of columns in column families; 2- define the good number of column families, it is not advisable to create too many column families.

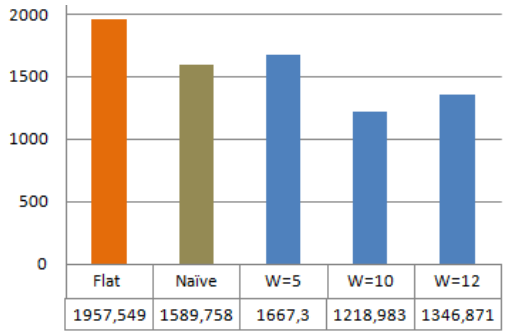


FIG. 7 – *Global queries response time*

5 Conclusion

In this paper, we presented our approach to modeling the data warehouses with column family NoSQL Databases, we resort to the clustering techniques to obtain a better design of column families. To increase the query performance, our method is based on the end user's needs and the characteristics of their interactions with the data warehouse. To evaluate our proposal, some experiments are carried out using the TPC-DS benchmark and HBase database, several tests are made to evaluate the effectiveness of our method. The obtained results confirm the benefits of grouping techniques for the column family's creation. Note that, in this paper, we did not take into account change in query workload. In future work, it would be useful to consider all changes and configuration parameters related to the data warehouse environment.

References

Bellatreche, L. and K. Boukhalfa (2005). An evolutionary approach to schema partitioning selection in a data warehouse. In *International Conference on Data Warehousing and Knowledge Discovery*, pp. 115–125. Springer.

Chevalier, M., M. El Malki, A. Kopliku, O. Teste, and R. Tournier (2015). Implementation of multidimensional databases in column-oriented nosql systems. In *East European Conference on Advances in Databases and Information Systems*, pp. 79–91. Springer.

de Freitas, M. C., D. Y. Souza, and A. C. Salgado (2016). Conceptual mappings to convert relational into nosql databases. In *ICEIS (1)*, pp. 174–181.

Dehdouh, K., F. Bentayeb, O. Boussaid, and N. Kabachi (2015). Using the column oriented nosql model for implementing big data warehouses. In *Int. Conf. on Parallel and Distributed Processing Techniques and Applications (PDPTA)*, pp. 469–475.

Derrar, H., O. Boussaid, and M. Ahmed-Nacer (2015). An objective function for evaluation of fragmentation schema in data warehouse. In *Encyclopedia of Information*

Science and Technology, Third Edition, pp. 1949–1957. IGI Global.

Eberhart, R. C., J. Kennedy, et al. (1995). A new optimizer using particle swarm theory. In *Proceedings of the sixth international symposium on micro machine and human science*, Volume 1, pp. 39–43. New York, NY.

Golfarelli, M., S. Rizzi, and E. Saltarelli (2002). Index selection techniques in data warehouse systems. In *Proc. of the International Workshop on Design and Management of Data Warehouses DMDWâĂŹ02*, pp. 33–42.

Gupta, H. (1999). *Selection and maintenance of views in a data warehouse*. Ph. D. thesis, stanford university.

Li, C. (2010). Transforming relational database into hbase: A case study. In *2010 IEEE Int. Conf. on Software Engineering and Service Sciences*, pp. 683–687. IEEE.

Navathe, S., S. Ceri, G. Wiederhold, and J. Dou (1984). Vertical partitioning algorithms for database design. *ACM Transactions on Database Systems (TODS) 9*(4), 680–710.

Romero, O., V. Herrero, A. Abelló, and J. Ferrarons (2015). Tuning small analytics on big data: Data partitioning and secondary indexes in the hadoop ecosystem. *Information Systems 54*, 336–356.

Scabora, L. C., J. J. Brito, R. R. Ciferri, C. D. d. A. Ciferri, et al. (2016). Physical data warehouse design on nosql databases olap query processing over hbase. In *Inter. Conf. on Enterprise Information Systems, XVIII*, pp. 111–118. Inst. for Systems and Technologies of Information, Control and Communication-INSTICC.

Yang, F., D. Milosevic, and J. Cao (2015). An evolutionary algorithm for column family schema optimization in hbase. In *Big Data Computing Service and Applications (BigDataService), 2015 IEEE First International Conference on*, pp. 439–445. IEEE.

Yangui, R., A. Nabli, and F. Gargouri (2016). Automatic transformation of data warehouse schema to nosql data base: Comparative study. *Procedia Computer Science 96*, 255–264.

Résumé

Le modèle NoSQL orienté famille de colonnes offre une grande flexibilité de modélisation et permet la gestion de gros volumes de données dans des environnements distribués. Dans cet article, nous proposons une méthode d'implémentation d'un entrepôt de données dans un système NoSQL orientée famille de colonnes. Notre méthode est basée sur une stratégie de regroupement des attributs issus des tables de faits et de dimensions, sous forme de familles de colonnes. Nous utilisons un algorithme de regroupement *Optimisation par Essaim Particulaire (OEP)*. Pour évaluer notre méthode, nous avons effectué différents tests sur le benchmark TPC-DS au sein du SGBD HBase, avec une architecture de type MapReduce sur une plateforme Hadoop.

Spatial-Sampling-Based Clustering For Data Lake

Redha BENAISSA[*,**,***], Omar BOUSSAID[*]
Aicha MOKHTARI[**], Farid BENHAMMADI [***]

[*]University Lumiere Lyon 2, ERIC EA 3083, 5, avenue Pierre Mendes 69676 Bron-France,
{redha.benaissa, omar.boussaid}@univ-lyon2.fr
[**]RIIMA Laboratory, USTHB University, Algiers, Algeria,
amokhtari@usthb.dz
[***]AI Laboratory, Polytechnic School of Algiers, Algiers, Algeria
fbenhammadi2008@gmail.com

Abstract. A data lake is a central repository that can store multi-structured data. Clustering is an important task in data lake analysis. In order to obtain valuable insights from the massive volume and variety in data lake, efficient and effective modified clustering methods are emerged as machine learning tools due to the limitations of conventional clustering algorithms. These methods are based on sampling, parallelization and reduction techniques to improve the clustering task within reasonable accuracy and time. Nevertheless, some sampling problems still exist when facing some complex datasets. To tackle the scalability issue of these conventional clustering methods, this paper develops a spatial-sampling technique for clustering approach. Rather than attempting to construct a random sampling, this work proposed a novel sampling strategy that computes plausible centroid points of the final clusters based on their data spatio-structure. Moreover, it suggests a parallelization issue in conjunction with a Spark platform. An analysis of proposed approach to evaluate performance gains with respect to the most popular $k$means algorithm is presented which reveals that our spatio-sampling technique is more effective in terms of both quality and stability.

1 Introduction

The growth in volume of data has also spawned advancements in techniques for data lake analysis. Conventional clustering are the commonly used techniques and have been widely used for these analysis (He et al., 2014). The objective of clustering involves the task of dividing mass objects into separate groups such that each groups share similar features and are different from objects belonging to other groups. Depending on the data properties, there are many conventional clustering techniques, such as partition-based, hierarchical-based, density-based, grid-based, Model-based and evolutionary-based. However, conventional clustering techniques cannot cope with data lake because of their high complexity, scalability and computational cost to handle a huge volume and variety of data (Sreedhar et al., 2017). Thus, the

performance of these conventional clustering techniques are unable to meet the needs of data lake management. An issue related to data lake concerns the data dimension reduction, processing parallelization or data sampling. In the recent years, several conventional clustering techniques have been modified for big data in order to scale up and speed up these techniques with minimum sacrifice to the clustering quality. These modifications allow to classify clustering technique for data lake in three different categories: single machine clustering, multiple machine clustering and hybrid clustering (Saha, 2017). The reduction-based and sampling-based techniques are generally used for the first category to provide a simpler representation of data. The reduction methods attempt to accelerate clustering algorithms and also to improve their accuracy by eliminating noisy and redundant data using principal component analysis method. CLIQUE (Yadav and Kumar, 2014) is an example of the reduction methods proposed to find clusters within subspaces of data set. This method combines density-based and grid-based clustering. However, the sampling-based clustering techniques perform on sample of dataset and then generalizes it to whole data set to find the final clusters. Most of the clustering based on the sampling techniques are partition-based algorithms (Zhang et al., 2013; Ng and Han, 2002; Wang et al., 2011; Rajasekaran and Saha, 2013). Unlike single-machine techniques, the second clustering category operates by first dividing the clustering task into a number of independent sub-tasks that can be performed simultaneously, and then efficiently merging these solutions into the final clustering solution. Thus this category divide the huge amount of data into small pieces that will which are distributed on different machines and the huge clustering problem can be solved using processing power of these machines (Chen et al., 2017). In this category, we find two parallelism paradigms such as standard parallel-based clustering and MapReduce-based clustering (Zhao et al., 2009; Sreedhar et al., 2017; Ferreira Cordeiro et al., 2011; Chen et al., 2017; Xia et al., 2016; Lu et al., 2018; Wang et al., 2017).

The present work proposes MapReduce-based clustering approach based on a novel spatial-sampling technique that can be used to speed up any conventional partition-based clustering algorithm in conjunction with a Spark framework. Inspired from density clustering our spatial-sampling scheme uses a number of sub-area of n-dimentional space determined by a circular tessellations surrounding medoid center (real data point). Each tessellation is clustered independently into k clusters by a partition-based clustering algorithm to select its representative points. Thereafter, we put the retained representatives points together and partitioned those points into k clusters to identify final medoid centers. These latter are used to assign the remaining points to that cluster whose medoid centers is the closest to these points. Second we address the parallelization issue of our spatial-sampling-based clustering method in conjunction with a Spark platform parallel processing implementation to reduce computational times produced by our clustering algorithm. We demonstrate the efficacy of our clustering algorithm using some well-known large benchmark data sets. Experimental results show that the proposed parallel clustering algorithm results in a speed-up of more than that of $k$means algorithm for clustering data lake sets with a similar accuracy.

The rest of the paper is organized as follows. This paper starts with a brief overview of MapeReduce-based clustering approaches in Section 2. The concept of data lake constitutes the topic of Section 3. Our spatial-sampling technique in conjunction with it's parallelization on Spark platform is treated in Section 4. Experimental results are discussed in Section 5. Finally, Section 6 provides final conclusions with our findings and future research directions.

2 Related Works

Big data needs the immense volume and makes operations such as data clustering hugely time-consuming. One way to meet such difficulties is to parallelize these operations. MapReduce-based clustering approaches are the most efficient solutions which process large scale data in parallel with many low end computing models (Wang et al., 2017). Hence substantial efforts are involved to rewrite conventional clustering algorithms in a scalable manner for the distributed versions based on MapReduce or spark framework. In this section, we provide a survey on MapReduce-based clustering approaches when one has massive volume of structured, unstructured or heterogeneous data such as data lake. Chu et al. (Ng et al., 2006) applied a big data parallel programming technique involving K-means clustering through the MapReduce framework. PKMeans (Zhao et al., 2009) is a MapReduce-based clustering that partitioned the data in the distributed file system. The local clustering is performed in map operation using Kmeans algorithm. Results show that the proposed algorithm has almost linear speed up and a good scale up. Another mapreduce-based clustering has been proposed by Ene et al. (Ene et al., 2011). They develop partition-based clustering algorithm based on sampling technique to decrease the data size and run in a time constant number of Map/Reduce rounds. Ferreira et al. (Ferreira Cordeiro et al., 2011) proposed an approach for data partitioning that leverage data localities in MapReduce to cluster large datasets. Thier algorithm minimizes I/O costs and reduces costs related to networks to generate efficient clustering. Xia et al. (Xia et al., 2016) proposed a MapReduce-based $k$means optimization algorithm. Their algorithm used the Euclidean distance for cluster centroids. The proposed algorithm shows improvements in terms of execution time. Recently, a more flexible method has appeared to extend the conventional $k$means clustering in (Lu et al., 2018). Their approach uses the parallel tabu search clustering algorithm on Spark platform. The authors utilize the centroid-driven orientation of the $k$means algorithm under the guidance of a simple version of tabu search. This strategy facilitates the parallel implementation of the $k$means in the Spark environment. Computational experiments disclose that the proposed approach can generate better solutions than the $k$means algorithm in terms of both quality and stability. A recent, interesting proposal for MapReduce-based clustering is presented in (Sreedhar et al., 2017). They have developed an improved version $k$means algorithm by making modifications to the clustering distance metric. The proposed approach produces high-quality clusters with high levels of intra-cluster similarity and with low levels of inter-cluster similarity relative to the proceeding clustering algorithms. An approach targeted at speed-up density-based DBSCAN using MapReduce framkork (He et al., 2014). Their algorithm partitions all spatial data into different maps and then performs conventional DBSCAN in each mapper and merges the bordering spaces in Reduce step. A major drawback of this MapReduce-based clustering is the limitation of load balancing among parallel clustering tasks because some procedures are not designed for shared-nothing environments. Finally, sampling-based methods permit to reduce the clustering computation time by first choosing a subset of the big data set and then using this subset to find the final clusters. The sampling process picks a subset of the given input and makes inferences on the original dataset. The key idea behind all sampling-based clustering methods is to obtain the cluster representatives, using only the sampled subset, and then assign the remaining data points to the closest representative. The popular methods to efficiently cluster big data sets of this category employ random sampling such as CURE (Guha et al., 2001) and CLARANS (Achlioptas, 2003). Other techniques are based on intelligent sampling scheme such as corset sampling (Har-Peled and

Mazumdar, 2004; Wang et al., 2011) or multi-levels sampling (Rajasekaran and Saha, 2013). This later approach uses randomly sampling levels that can be used to speed up any clustering algorithm. In the first level the input data is partitioned into several splits. Each split is clustered using hierarchical clustering. For each clustered split the authors choose a number of cluster representative points. Thereafter the choosen representatives of each cluster of each split are put together and are moved to the next level for a new deterministic sampling process. The determiste sampling process continues until the number of points is "small" enough in the final level. Finally, these representative points in the prior level are clustered into k clusters to identify some centroid points. As results, these centroids are used to assign each input point of dataset according to the closest centroid. Empirical results show that this technique results in a speed-up of more than an order of magnitude over conventional hierarchical clustering algorithms. This strategy is particularly attractive but it requires the numbers of levels to construct the final centroids. Also, this sampling depends on the local split clustering best solutions only without the global sample points of the dataset. One shortcoming of all these clustering approaches is the choose of subset of data samples randomly. In addition, these approaches do not use MapReduce framework to speed clustering processing.

3 Data lake Concept

The concept of a data lake is emerging as a popular way to build the next generation of systems to master new big dat[width=12cm,,height=4cm]a challenges. One of the primary motivations of a data lake is to provide as large a pool of data without losing any data that may be relevant for analysis (Terrizzano et al., 2015). Data lakes thus store all the data deemed relevant for analysis. Fig. 1 shows the simplified architecture that we propose for data lakes based on spark environment. A data lake is a central data repository that can store multi-structured (i.e., structured, semi-structured and unstructured) data in raw format and supports data transformations by integrating Big Data processing frameworks such as Apache Hadoop and Apache Spark. This concept is accepted as a way to describe any large data pool in which the schema and data requirements are not defined until the data is queried. So the data lake loads all the data and defines the structure of the data at the time it is used with powerful programming framework, such as MapReduce. For this reason, the data lake has some capabilities such as (Fang, 2015) :

— To capture and store raw data at scale for a low cost.
— To store many types of data in the same repository
— To perform transformations on the new data processing
— To define the structure of the data at the time it is used.
— To perform single subject analytics based on very specific use case.

4 Parallel Clustering Approach

4.1 A New Sampling Strategy

As mentioned in related works section, an important factor that influences the clustering algorithms for big data or data lake is the representative points selection (sampling strategy) employed to construct the centroids or core points for final clustering of the remaing points of

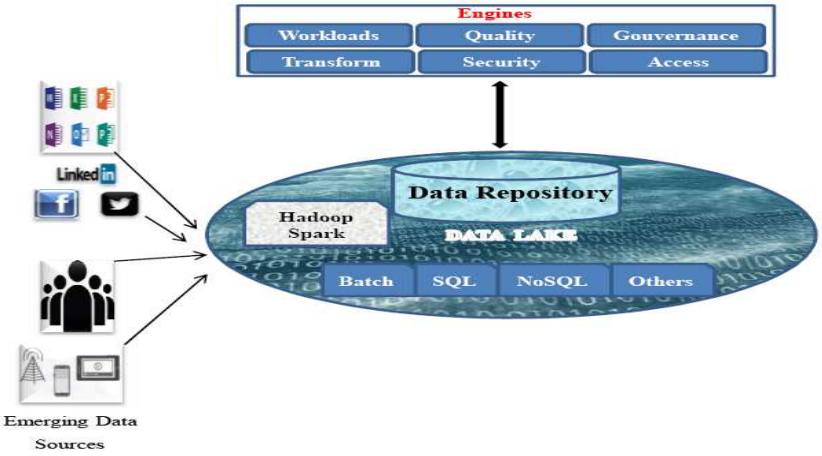


FIG. 1 – *Overall Data Lake Architecture.*

the datasets. To address the random point selection problem, we propose a spatial-sampling strategy based on a sample points selection using a circular tessellation by bands. The key idea behind our spatial sampling-based clustering technique is to obtain the cluster representatives, using only the sampled subset of each band, and then assign the remaining data points to the closest representative.

First, we introduce the definitions of related terms of our spatial-sampling strategy. Let $\mathfrak{S} = \{x_1, x_2, \ldots, x_n\}$ be a given set of n objects where each object (point) $x_i \in \mathcal{X}$ and $\mathcal{X} \subseteq \mathfrak{R}^d$ for some dimension d.

Definition 1 *The point x_c is considered a medoid center point if*

$$\forall x_i \in \mathfrak{S}, \sup_{i \neq c}\{\|x_c - x_i\|_2^2\} - \inf_{i \neq c}\{\|x_c - x_i\|_2^2\} \text{ is minimal.} \tag{1}$$

Definition 2 *Let $x_c \in \mathfrak{S}$ be the medoid center point of a d-dimensional object's space $\mathfrak{S}$. For any set $\mathfrak{S}$, its minimal radius r^- is defined as*

$$r^-(\mathfrak{S}) = \inf_{i \neq c}\{\|x_i - x_c\|_2^2 \,|\, x_i \in \mathfrak{S}\} \tag{2}$$

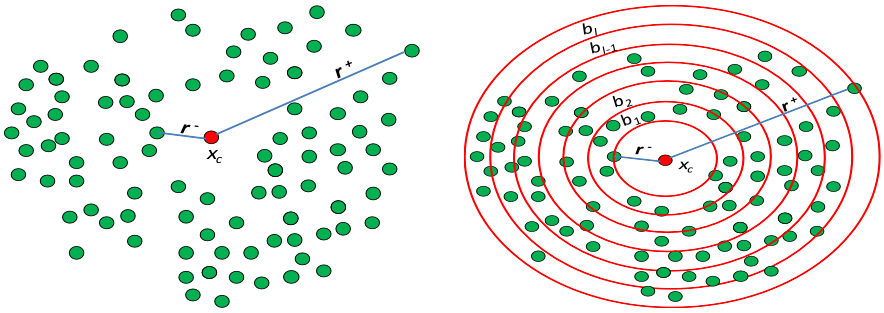
Definition 3 *For any set $\mathfrak{S}$, its maximal radius r^+ is defined as*

$$r^+(\mathfrak{S}) = \sup_{i \neq c}\{\|x_i - x_c\|_2^2 \,|\, x_i \in \mathfrak{S}\} \tag{3}$$

For illustration, Fig. 2a shows an example of medoid center object for 2-dimensional region space according to minimal and maximal radius.

Based on the aforementioned definitions, the idea of our spatial-sampling strategy is as follows. The d-dimensional region space is first tessellated into b_1, b_2, $\cdots$ b_l d-dimensional bands. This tessellation is determined by a variable radius r_l which is defined as follows:

$$r_l = r_{l-1} + \frac{r^+(\mathfrak{S}) - r^-(\mathfrak{S})}{n_b} \text{ with } r_0 = r^-(\mathfrak{S}) \tag{4}$$



(a) 2-dimensional space showing medoid center point of objets.

(b) Tessellation process for 2-dimensional space of objets.

FIG. 2 – *Tessellation process.*

where n_b is the number of bands.

The tessellation process begin from r^- to r^+ radius. All points in a d-dimensional space band b_l are delimited by r_{l-1} and r_l radius (see Fig.2b). The choice of the widthband is central to our partitionned-based algorithm. Such widthband can in general depend the configuration of the neighborhood points. Note that each band must contains a minimal number of points that ensures the partitionned-based clustering.

There are several samples according the number of bands in our spatial-sampling strategy and merge all points of the b_{l-1}, b_l and b_{l+1} bands to generate a sample S_l which is defined as follows:

$$S_l = \{x_i \,|\, r_{l-1} \leq \|x_i - x_c\|_2^2 < r_{l+1}\} \tag{5}$$

Fig. 3 shows an example in 2-dimensional space which has been sampled according to 5 bands. Then we cluster those S_l points into k clusters using $k$means clustering algorithm and we identify the centers of these clusters. Thereafter, we pick representative points from each such cluster of each sample S_l. These representative points (real points) are the closest points to the cluster centers generated by $k$means algorithm. Finally, to generate the representative medoid centers set $\mathfrak{R}$, all representative points are put together as follows:

$$\mathfrak{R}(\mathfrak{S}) = \cup_l S_l \tag{6}$$

This set allows to get a final representative medoid center set using also $k$means algorithm. Then this set is used to assign each input point $x_i \in \mathfrak{S}$ (the remaining points) to that cluster whose each representative medoid center is the closest to x_i. The Fig. 4 shows an example of representative points aggregate and the final medoid centers in 2-dimensional space. In fact in this example we have used only one representative medoid center of spatial sampling.

Clearly, the above spatial sampling technique can be employed in conjunction with any clustering algorithm but in the proposed work we adopt $k$means clustering algorithm.

4.2 Parallel clustering Algorithm

Due to the virtue of simplicity and scalability of MapReduce framework, we parallelize our spatial-sampling-based clustering algorithm for data lake. A few parallel models of data

Algorithm 1 : Spatial-sampling-based clustering

Input: A set $\mathfrak{S}$ of n data points , integers number of bands n_b and c number of clusters.
Output: The best c clusters.

1: Calculate all pair-wise point distances and place them in a $n \times n$ matrix M. This matrix is called the distance matrix.
2: Calculate $r^-(\mathfrak{S})$ and $r^+(\mathfrak{S})$ and the corresponding using the matrix M.
3: Tessellate the set $\mathfrak{S}$ using the band radius r_k and n_b.
4: **for** $r_k = r^-(\mathfrak{S})$ to $r^+(\mathfrak{S})$ **do**
5: Generate the sets S_l.
6: Cluster the sets S_l in c clusters using $k$means.
7: Find the representative medoid center of S_l sets.
8: **end for**
9: Put all of the representative medoid centers S_l together: $\mathfrak{R}(\mathfrak{S}) = \cup_l S_l$.
10: Cluster the set $\mathfrak{R}$ in c clusters using $k$means to calculate c final representative medoid centers.
11: Assign each remaining point $x_i \in \mathfrak{S}$ to that cluster whose final representative medoid center is the closest to x_i.

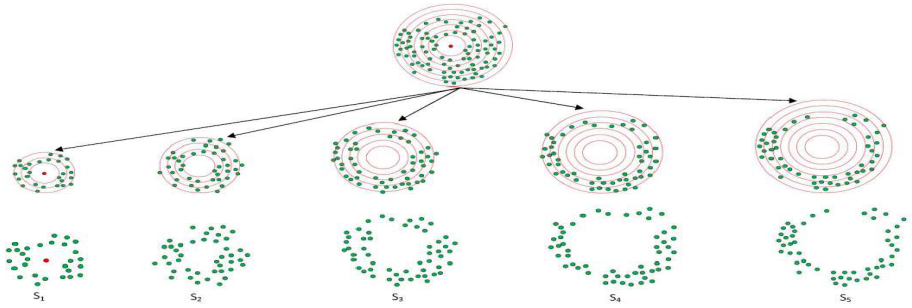


FIG. 3 – *A spatial-sampling example of d-dimensional space of objets.*

lake have been introduced in the literature based on spark environment (Sreedhar et al., 2017). In this section we present parallelizations of our algorithm using this environment. We have parallelized spatial-sampling and clustering processes. In the first process, we divide the data points into p equal sized parts (splits). The method of data split is very important for the implementation of the parallel algorithm. The data points are divided into several data slices with a given band range. Then all the points that fall into a given three adjacent bands belong to one data slice (one split), and points within this slice are assigned to the same node processed by Map process. Then we generate the final representative medoid centers computation. Since the data points can be treated independently in this process, during the pair distance matrix calculation (map process) each map computes the distances between data points of two splits in parallel. The reduce side relies on the output of map to compute the global distance matrix M and generate the partial representative medoid centers using $k$means algorithm and our selection procedure. After this generation by each local computing node, we merge all these representative medoids from different bands (reduce side) together. The last step of the spatial sampling process is to run $k$means the clustering on the whole partial representative medoid centers to generate the final representative center points in order to assign the dataset points.

The second process we assign data points according to final representative medoid centers. We first store the representative medoid centers on each node with the $k$means algorithm. Since

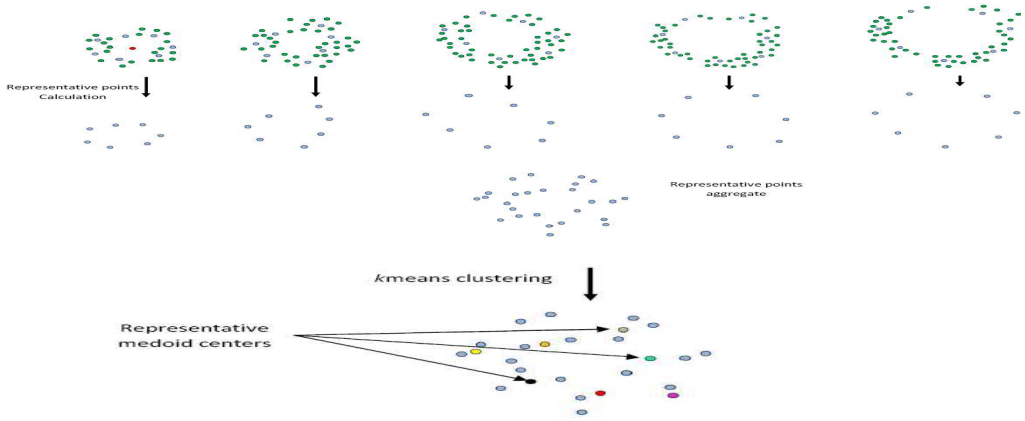


FIG. 4 – *Representative medoid points generation.*

the data points can be treated independently, during the clustering each data split points can be treated independently (map operation) where each map operation assigned each point into the closest representative medoid center in parallel. Algorithm 2 presents a flowchart of our parallel spatial-sampling-based clustering.

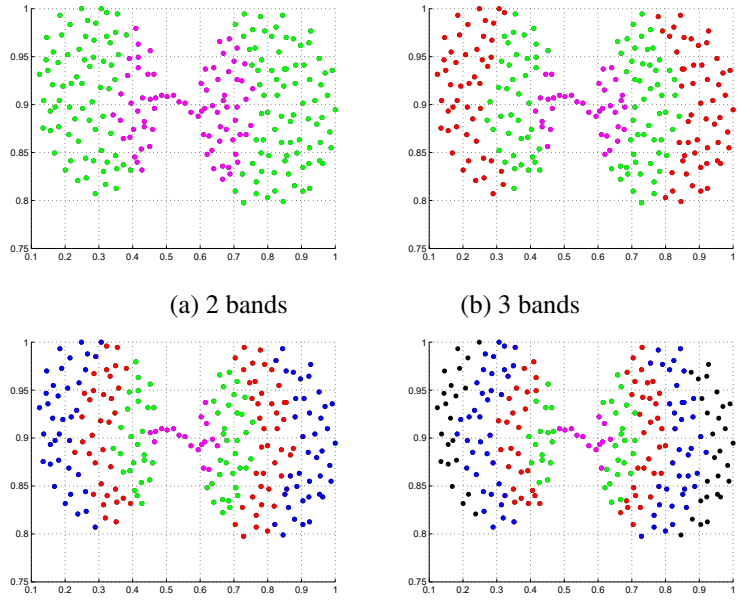
Algorithm 2 : Parallel Spatial-sampling-based Clustering.

Input: A set $\mathfrak{S}$ of n data points , integers number of bands n_b and c number of clusters.
Output: The best c clusters.
in parallel
1: Split the data points into p equal sized splits: $p_1, p_2, \ldots, p_p$ using data.mapPartitions (Spark method).
2: Assign each pair of splits (S_i, S_j) to a single processor.
3: Calculate the pair distance matrix M_{ij}.
4: Repeat $\frac{p\times(p-1)}{2}$ times steps 3 through 4.
5: Put all of the $\frac{p\times(p-1)}{2}$ distance matrix together: $M = \cup_{ij} M_{ij}$.
6: Generate the sets S_l.
7: Run the $k$means algorithm on each set S_l with the number c of clusters.
end parallel
8: Determine representative medoid centers.
in parallel
9: Split the remaining data points into p equal sized splits: $p_1, p_2, \ldots, p_p$ using data.mapPartitions (Spark method).
10: Assign each point x_i of each split p_i to that cluster whose final representative medoid center is the closest to x_i.
end parallel

5 Experimental evaluation

To evaluate the performance of our spatial-sampling clustering algorithm, we compared it to $k$means for the same parameter settings for all data split datasets. First we give some example of our tessellation process and we compare our sequential spatial sampling-based

clustering algorithm with $k$means one. Second, we give the complexity measure for the parallel version for our algorithm.



(a) 2 bands (b) 3 bands

(c) 4 bands (d) 5 bands

FIG. 5 – *Tessellation examples for bridge dataset.*

For the purpose of illustration, we have used some simple datasets. Figures 5 show our tessellation process for 2, 3, 4 and 5 bands using bridge dataset according to the medoid center point. Figures 6 compare our spatial sampling-based algorithm with $k$means when we have the same parameter settings. for example, figures 6(a) and 6(b) show a 2-dimensional bridge dataset containing 207 points from two separate clusters. The two clusters are identified perfectly with 99% when our spatial sampling-based and $k$means clustering algorithms are executed on this dataset. However, in Figures 6(c) and 6(d) the clusters of the 2-dimensional aggregate dataset is completely identified using 8 bands but $k$means algorithm failed. Note that clustering results for the run of $k$means algorithm are computed as the average of 20 times.

As shown in the confusion matrices in Tab.1, the errors, in terms of the number of points that are grouped into the wrong cluster for Bridge and Jain datasets, is about 0.86%, 12.06% and 51.21%, 0.86%, 5.63% and 49.27% for $k$means and our approach respectively. Generally, the majority of the data points are assigned to their corresponding cluster except for the spiral dataset. However, for Jain dataset the confusion matrix shows that the accuracy of our clustering algorithm is higher than that of the $k$means algorithm.

For data lake, we first divide the data points into p equal sized splits and hence each split has a size equal to $\ell = n/p$ where n is the total number of data points. Assume that we have m virtual machines. In the spatial sampling strategy, we maps each split's pair (p_i, p_j) to each virtual machine processing. The most computationally intensive operations in this

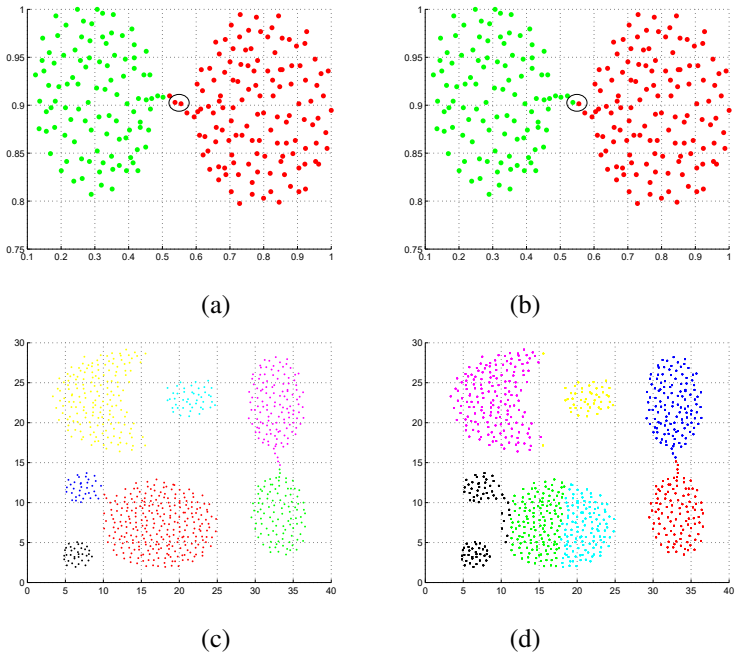


FIG. 6 – *Simple examples to illustrate our spatial-sampling-based clustering.*

Dataset	$k$means		Our approach	
	Class 1	Class2	Class1	Class2
Jain	96	1	77	20
	44	232	1	275
Bridge	104	2	104	2
	0	126	0	126
Spiral	44	62	55	51
	44	576	51	50

TAB. 1 – *Comparison of the confusion matrices of $k$means and our approach algorithms for 5 bands.*

step is the computation of the distance matrix for n points with the complexity time $\mathcal{O}(n^2)$ and the $k$means clustering of each split. According to the number of machines, this process requires $\frac{(p\times(p-1))}{2m} \cdot \ell k$ map operations to calculate all partial distance sub-matrix and partial representative medoid center points. In the reduce slice, we use operations to group these partial points for the final representative medoid centers. Thus the map and reduce operations generate these final representative points for the clustering the remaining points. The runtime complexity of the parallel spatial sampling stage is $\mathcal{O}(\frac{p\times(p-1)\cdot\ell k}{2m})$. For the parallel clustering

algorithm, we need the representative medoid data centers to be replicated in all the machines and randomly split the remaining data points into p splits of ℓ size. To assign n points to that cluster whose center is closest among k cluster representative medoid centers, the complexity can be done in $\mathcal{O}(nk)$. Proceeding in a similar manner, we need $\mathcal{O}(\ell k)$ time to assign each split in map operations and hence the total time spent in the second stage is $\mathcal{O}(\frac{p(p-1)}{2m} \cdot \ell k)$.

In summary, the overall running time complexity of the proposed sampling-based clustering is $\mathcal{O}(\frac{p\times(p-1)}{m} \cdot \ell k)$.

6 Conclusion

Sampling technique has played a major role in the design of efficient clustering algorithms for data lake. We have proposed a spatial-sampling-based clustering method that employs circular data points tessellation to generate a subset of representative medoid centers points. These latter allow to cluster the original dataset. In addition, we have proposed the parallelization of our clustering algorithm to manage data lake. As comparison, computational experiments disclose that our method can generate clustering quality similar than the $k$means algorithm for the same parameter settings with a stable acceleration. We thus feel that this strategy is a very effective sampling technique to achieve our objective of clustering data lake efficiently and accurately.

The future direction to explore concerns the band generation based on the points density because we believe that this is suitable to cluster data lake sets more efficiency.

References

Achlioptas, D. (2003). Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of computer and System Sciences 66*(4), 671–687.

Chen, M., S. A. Ludwig, and K. Li (2017). Clustering in big data. pp. 333–346.

Ene, A., S. Im, and B. Moseley (2011). Fast clustering using mapreduce. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 681–689. ACM.

Fang, H. (2015). Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem. In *Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), 2015 IEEE International Conference on*, pp. 820–824. IEEE.

Ferreira Cordeiro, R. L., C. Traina Junior, A. J. Machado Traina, J. López, U. Kang, and C. Faloutsos (2011). Clustering very large multi-dimensional datasets with mapreduce. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 690–698. ACM.

Guha, S., R. Rastogi, and K. Shim (2001). Cure: an efficient clustering algorithm for large databases. *Information Systems 26*(1), 35–58.

Har-Peled, S. and S. Mazumdar (2004). On coresets for k-means and k-median clustering. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pp. 291–300. ACM.

He, Y., H. Tan, W. Luo, S. Feng, and J. Fan (2014). Mr-dbscan: a scalable mapreduce-based dbscan algorithm for heavily skewed data. *Frontiers of Computer Science 8*(1), 83–99.

Lu, Y., B. Cao, C. Rego, and F. Glover (2018). A tabu search based clustering algorithm and its parallel implementation on spark. *Applied Soft Computing 63*, 97–109.

Ng, A. Y., G. Bradski, C.-T. Chu, K. Olukotun, S. K. Kim, and Y.-A. Lin (2006). Mapreduce for machine learning on multicore. *NIPS, December*, 281–288.

Ng, R. T. and J. Han (2002). Clarans: A method for clustering objects for spatial data mining. *IEEE transactions on knowledge and data engineering 14*(5), 1003–1016.

Rajasekaran, S. and S. Saha (2013). A novel deterministic sampling technique to speedup clustering algorithms. In *International Conference on Advanced Data Mining and Applications*, pp. 34–46. Springer.

Saha, S. (2017). Novel algorithms for big data analytics.

Sreedhar, C., N. Kasiviswanath, and P. C. Reddy (2017). Clustering large datasets using k-means modified inter and intra clustering(km-i2c) in hadoop. *Journal of Big Data 4*(1), 27.

Terrizzano, I. G., P. M. Schwarz, M. Roth, and J. E. Colino (2015). Data wrangling: The challenging yourney from the wild to the lake. In *CIDR*, pp. 4–7.

Wang, J., A. Zelenyuk, D. Imre, and K. Mueller (2017). Big data management with incremental k-means trees–gpu-accelerated construction and visualization. In *Informatics*, Volume 4, pp. 24. Multidisciplinary Digital Publishing Institute.

Wang, L., C. Leckie, R. Kotagiri, and J. Bezdek (2011). Approximate pairwise clustering for large data sets via sampling plus extension. *Pattern Recognition 44*(2), 222–235.

Xia, D., B. Wang, H. Li, Y. Li, and Z. Zhang (2016). A distributed spatial–temporal weighted model on mapreduce for short-term traffic flow forecasting. *Neurocomputing 179*, 246–263.

Yadav, J. and D. Kumar (2014). Sub space clustering using clique: An exploratory study. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3*, 372–378.

Zhang, J., G. Wu, X. Hu, S. Li, and S. Hao (2013). A parallel clustering algorithm with mpi-mkmeans. *J Comput 8*(1), 10–17.

Zhao, W., H. Ma, and Q. He (2009). Parallel k-means clustering based on mapreduce. In *IEEE International Conference on Cloud Computing*, pp. 674–679. Springer.

Résumé

Afin d'extraire des informations précieuses dans un data lake, des méthodes de Clustering sont émergées comme des outils d'apprentissage automatique. Ces méthodes sont basées sur des techniques d'échantillonnage, de parallélisation et de réduction. Pour palier au problème de l'évolutivité de ces méthodes de clustering conventionnelles, cet article développe une technique d'échantillonnage spatial pour l'approche de clustering. Ce travail propose une nouvelle stratégie d'échantillonnage qui calcule des points centroïdes plausibles en fonction de la structure spatiale des données, dont le traitement en parallele est réalise avec la plate-forme Spark. L'évaluation de l'approche proposée montre des gains de performance, qui révèle que notre technique d'échantillonnage spatio-données est plus efficace en termes de qualité et de stabilité.

Optimize Star Join Operation for OLAP Queries in Distributed Data Warehouses

Yassine RAMDANE*, Omar BOUSSAID*
Nadia KABACHI**, Fadila BENTAYEB *

*University Lumiere Lyon 2, ERIC EA 3083, 5, avenue Pierre Mendês 69676 Bron-France,
{Yassine.Ramdane, Omar.Boussaid, Fadila.Bentayeb}@univ-lyon2.fr
**University Claude Bernard Lyon 1, 43, boulevard du 11 novembre 1918,
69100, Villeurbanne-France
Nadia.Kabachi@univ-lyon1.fr

Abstract. The partitioning and the distribution of the data warehouses are optimized techniques which applied to improve OLAP queries performances in distributed systems. However, the schemas defined by some of these techniques, implemented on Hadoop ecosystem, are based on the workload. As the imposed query workload can change, these schemas should be redefined. A costly operation which can disrupt the system. In this paper, we propose a strategy for partitioning and distributing a big data warehouse (*DW*) upon a cluster of homogeneous nodes, independent of the query workload; taking into account the physical characteristics of the cluster and the data load balancing. With our approach, we can execute star join operation locally, in map side with low communication cost. To evaluate our contribution, we have done some experiments where we improve the execution time of OLAP queries up to 50 %.

1 Introduction

Hadoop become the standard platform for big data process. Many big companies, such as Facebook, Yahoo, etc., use it to store and manage their massive data. The main components of Hadoop V-2.x are: (1) HDFS, which is designed and optimized for storing very large files, and (2) YARN, which supports a more flexible execution engine than MapReduce, like Spark (Zaharia et al., 2010), Tez (Saha et al., 2015), and other frameworks. Hadoop uses load balancing technique to parallelize processing and to improve application execution time. However, the random distribution of Hadoop blocks would harm the system and cause network bottleneck. Moreover, current systems like Hive (Thusoo et al., 2009) and Spark-sql (Armbrust et al., 2015) use MapReduce or Spark to execute queries. OLAP queries are composed of several operations such as selection, projection, join, and aggregation. Each operation is performed in map or reduce phase. Thus, each operation generates an execution cost, e.g. I/O cost or CPU cost. The join operation is the most expensive one and often involves a high rate of communication cost. To optimize the join processing, several partitioning and distribution techniques implemented on Hadoop ecosystem have been proposed, especially in the context of relational *DW* (Wang et al., 2015). We can distinguish two types of data load balancing techniques, static

and dynamic techniques. In static one, we do the load balancing before proceeding with the processing. Static techniques can be divided into two categories. Those partitioning and distributing data depending on the type of processing (Valvåg et al., 2013; Eltabakh et al., 2011), and other with workload driven (Arres et al., 2015). For dynamic techniques, we do the load balancing at the moment of processing (Kwon et al., 2012; Gao et al., 2017). Static algorithms that use workload optimization (Arres et al., 2015) are typically based on attribute affinities of a given query workload. If the workload is changed, we need to re-run these algorithms to obtain a new partitioning and distribution schema. Some static algorithms (Valvåg et al., 2013) are not based on the workload, but aren't adaptable to the *DW* context.

On the other hand, while some partitioning and data load balancing techniques (Taniar et al., 2008; Benkrid et al., 2014) are suitable in many systems implemented on a massively parallel processing relational database (*MPPRD*), however, we can't apply straightly these paradigms to Hadoop ecosystem. For example, as far as we know, the current version of Hive doesn't support range partitioning, so, we must adapt hash partitioning function to obtain roughly balanced fragments. Moreover, although Hadoop and *MPPRD* like Teradata and Netezza platform [1] each support MapReduce paradigm, and may these tools, i.e. Teradata and other, achieve better performance than Hadoop for some applications, However, Hadoop is different to them in two main points: (1) Hadoop is cheaper open source while some of *MPPRD* is costly; (2) Hadoop is more scalable than almost all *MPPRD* technology. The system HadoopDB (Abouzeid et al., 2009) attempts to integrate MapReduce and *MPPRD* technology. Some improvements have been made on both scalability and efficiency. However, the results are still not satisfactory for *DW* applications. Especially when a join operation involves multiple join attributes, such as a star join, HadoopDB can lose its performance advantage. If we replicate each dimension table to all the database nodes, it will incur high space cost. If we adopt a partitioning method, we can only partition a dimension table and the fact table based on one join key. In this case, the access on other dimension tables will be very expensive. Spark-SQL is a different beast sitting between the MapReduce and *MPPRD* over Hadoop approaches [2]. Spark-SQL try to combine between the two techniques. Similarly to MapReduce, it splits the job into a set of tasks scheduled separately giving better stability. Like *MPPRD*, it tries to stream the data between execution stages to speed up the processing.

In this paper, we define a partitioning and distributing schema of a relational big *DW* upon a cluster of homogeneous nodes, using a static balancing technique such we take into account: the volume of data, the distribution of foreign and primary keys of the fact and dimension tables, and the physical characteristics of the cluster. Our idea is to fragment horizontally the fact and dimension tables, then we distribute these fragments equally upon the different nodes in which we can perform star join operation locally, in map phase with low communication cost. We developed and evaluated our approach using the Scala language on a cluster of homogeneous nodes, using the distributed system Hadoop-Yarn and Spark, the Hive system, and the TPC-DS benchmark.

The remainder of this paper is structured as follows. Section 2 summarizes related work on MapReduce model, the sources of imbalanced data loads in distributed systems and some join types in MapReduce. In Section 3 we detail our approach. We present our experiments in Section 4 and we concluded in Section 5.

1. availlable from sites https://www.teradata.com/ and https://www.ibm.com/software/fr/data/netezza
2. https://0x0fff.com/hadoop-vs-mpp/

2 Background and Related Work

2.1 The MapReduce Model

MapReduce is a parallel computing paradigm proposed by Google and the default processing engine of Hadoop platform. Today many companies are using this paradigm to process massive data like Yahoo, Facebook, and other. For more flexibility, other frameworks have been developed. They use the same paradigm but with different techniques, such as Spark and Tez. There are two main functions in MapReduce, the map function: ($K1$,$V1$) $\mapsto$ List[($K2$,$V2$)] and the reduce function: (K2, List[($V2$)]) $\mapsto$ ($K3$,$V3$). At the result of each mapper, a partitioner takes the intermediate pairs (key: $K1$, value: $V1$) and divide them into subsets, one for each reducer, such that all values associated with the same key are grouped and assigned to the same reducer. The partitioner uses a hash function to distribute the keys between the map and reduce phase. Many algorithms for parallel processing use this paradigm. However, the main challenge is how to optimize the load balancing for it, especially in the context of *DW* where the execution of OLAP queries requires several MapReduce cycles.

2.2 Sources of imbalanced data loads for MapReduce processing

According to (Hefny et al., 2014), we can distinguish four sources of imbalanced data loads in distributed systems: (1) *imbalanced data loads related to the split input* (or availability of data). The origin of this type of imbalancing is the unequal distribution of the data loads on the cluster. Some nodes finish the processing of their local data and they try to process the data of the neighboring nodes, which cause network bottlenecks (Valvåg et al., 2013). (2) *imbalanced data loads related to partition sizes*. Some nodes examine few partitions that have big size, which often involves memory overflow, other can run huge number of partitions of small sizes where I/O operations are increased (Vernica et al., 2012). (3) *imbalanced data loads related to the heterogeneity of the nodes*. In a cluster of heterogeneous nodes, the fast nodes finish their processing before the slow nodes. In this case the application seems not parallel (Sharafi and Rezaee, 2016). (4) *imbalanced data loads related to computations*. In some computational domains, we may use non-linear functions that can cause imbalance at the processing level, even with partitions of the same sizes (Gufler et al., 2012).

2.3 Types of join in distributed systems

Both of existing join algorithms rely with dynamic techniques of partitioning and data load balancing, such as repartition and broadcast join (Blanas et al., 2010), replication join (Afrati and Ullman, 2011), and other. Spark also, used Hash-broadcast-join as default join which is more efficient when we join large table A with small table B, such the small one can broadcasted and fit in memory, otherwise, it might overflow the memory. Few algorithms used static techniques, such as trojan join (Dittrich et al., 2010). This kind of algorithms requires prior knowledge of table schemas and join conditions.

Our idea is close to trojan join. We implemented a strategy for partitioning and distributing a big *DW* upon a cluster of homogeneous nodes. Using a static balancing technique, independent of the workload. We deal with problems (1) and (2) of imbalanced data loads (see Section 2.2). We used Spark as a query execution engine for Hadoop-YARN platform. We used new type of join existed in HIVE and Spark called *Sort-Merge-Bucket (SMB) join*, which allow to perform star join operation in only one MapReduce cycle.

3 Proposed Approach

Our approach consists to build horizontal fragments of the fact and dimension tables by using a hash-partitioning function. The idea to use this function is to get roughly balanced fragments. Then, we distribute these fragments equally upon the different nodes of the cluster, such we can perform star join operation locally, and in map side. We assume that the *DW* has a star schema and the cluster has homogeneous nodes. Before detailing our approach, we define some notations as shown in TAB. 1. Our approach is composed of two phases: (A) build the set of the $datasets$ (i.e. CF, and $CDd,\ (d \in \{1..k\})$), and (B) placing all the $datasets$ that share the same join key, i.e. $group$, in the same node. The FIG. 1 summarize the two phases of our approach.

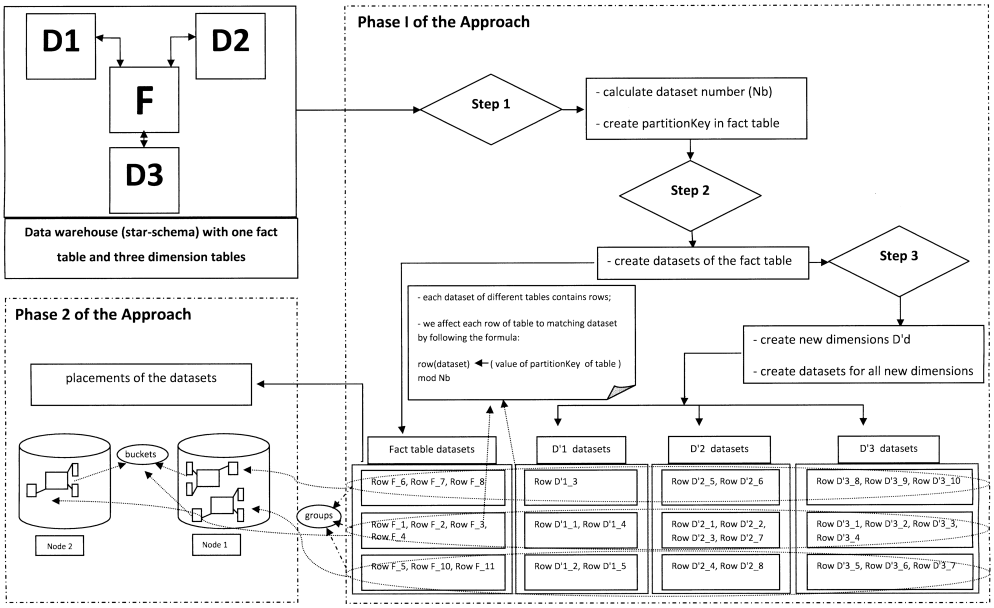


FIG. 1 – *the steps of our approach*

3.1 Building the set of datasets

This phase is composed of three steps: (1) determinate "Nb" and "$partitionKey$"; (2) building "CF"; and (3) building "$CDd,\ (d \in \{1..k\})$" of all dimension tables.

3.1.1 determinate Nb and $partitionKey$

Selecting the right "Nb" and "$partitionKey$" to guide partitioning is critical. To realize this, we address some technical challenges as outlined below.

1. **calculate** Nb: we select Nb from the interval [$lower\_Nb$,..,$upper\_Nb$], such as:

Symbols	Description
$dataset$	a $dataset$ is a set of rows of the fact or dimension tables. We denote $datasets$ the set of $dataset$.
$group$	a $group$ is a set of $datasets$ that share the same join key. It's composed of one $dataset$ of the fact table F and one $dataset$ of each dimension Dd. We denote also by $groups$ a set of $group$.
$partitionKey$	the $partitionKey$ is a key in the fact table F (i.e. new key added or a foreign key), where we partition the fact table and all dimension tables by this key.
E	$= \{F, D1, D2, .., Dk\}$, is a *DW* in stars schema with the fact table F and the dimensions Dd, $(d \in \{1..k\})$.
N	$= \{n_1, n_2, .., n_m\}$ the set of all cluster nodes.
n_c, n_t	n_c is the total number of CPU cores of all slave nodes; n_t is the number of CPU cores affected to the tasks. Note that $n_c > n_t$.
Nb	the ideal number of the $datasets$ of the fact and dimension tables that we want to build.
CF	$= \{datasetF_0, datasetF_1, .., datasetF_{Nb-1}\}$ the set of distinct $dataset$ of the fact table.
S_{CF}	$= \{\|datasetF_0\|, \|datasetF_1\|, .., \|datasetF_{Nb-1}\|\}$. the set of fact table $dataset$ size.
CDd	$= \{datasetDd_0, datasetDd_1, .., datasetDd_{Nb-1}\}$ a set of distinct $dataset$ of Dd $(d \in \{1..k\})$
S_{CDd}	$= \{\|datasetDd_0\|, \|datasetDd_1\|, .., \|datasetDd_{Nb-1}\|\}$, the set of $dataset$ size of the dimension Dd $(d \in \{1..k\})$.
V_E	the volume of the data warehouse E
V_M	the total memory size in all slave nodes.

TAB. 1 – *Notations*

(a) since our processing should be parallelized we put $lower\_Nb$=n_t, i.e. all CPU cores affected to Spark $executors$ [3] are used. Our aim is to assign at least for each CPU core node one Spark $partition$ (in our case, a $partition$ is a $group$)

(b) the choice of a large number of Nb ($Nb \gg n_t$) can disrupt the distributed system, as the result of the increase of the I/O operations. Since our processing is "In-Memory" using Spark, the $upper\_Nb$ is computed by the following formula:

$$upper\_Nb = \lfloor \frac{V_E}{V_M} \times lower\_Nb \rfloor \quad (1)$$

We argument this as follows: if the total memory size of the cluster is large ($V_M \approx V_E$), the $upper\_Nb \approx lower\_Nb$, in this case, we can process large $partition$. However, if the memory size is small ($V_M \ll V_E$), the $upper\_Nb$ increases, in this case, processing small $partition$ (i.e. $group$) is preferable.

(c) we choose Nb such as $Nb\ modulo\ n_t = 0$. The reason is to assign in each wave of Spark stage [4] the same number of $partitions$. This is an approximative solution. In fact the tasks may not finish at the same time due to various factors such as $partitions$ data skew, and differences in computing capability of different nodes.

2. **determinate** $partitionKey$: Our aim to select the appropriate $partitionKey$ is constructing the CF set, and all the sets CDd $(d \in \{1..k\})$ which have the minimum

3. an $executor$ is a worker node's process in charge of running individual tasks in a Spark job. We assign to an $executor$ a couple of resources (α cores, β of memory space) that are required to execute the tasks, where α is the number of CPU cores and β is the memory size allocated for executing the tasks. each node can contain more than one $executor$. By default Spark uses one CPU core per task.

4. In Spark, a stage is a set of consecutive operators that can be grouped and executed together, per $partition$ (i.e. one task by $partition$). In each wave we have n_t tasks executed in parallel.

standard deviation of its corresponding sets S_{CF} and all the sets S_{CDd} ($d \in \{1..k\}$) respectively. Since the volume of the dimension tables are negligible compared to the fact table, we focus our study to minimize the standard deviation of the set S_{CF}. Selecting the appropriate $partitionKey$ allow to obtain roughly balanced $datasets$ of the fact table (i.e. the minimum standard deviation of the set S_{CF}) is a particularly challenging task. If we partition the fact table with a foreign key where its values are uniformly distributed, we can resolve the problem. However, the transformation that we apply to the join condition (see Section 3.3) can increase the CPU cost. So, we think that the best solution is to add a new key, i.e. $partitionKey$, in the fact and all dimensions. The second challenge is how to choose the values of the attribute $partitionKey$. In the next Section we detail how to choose the values of this key.

3.1.2 Building fact table datasets

As we have shown in the Section 3.1.1, the construction of the CF set is based on the two parameters "$partitionKey$" and "Nb". To calculate the values of the key $partitionKey$ we propose the Algorithm 1. In our method, we have used the Round Robin fashion to calculate the values of this key. We start by affecting "0" value Nb times, then "1" value Nb times,.., until "Nb-1" value Nb times. We restart the operation with the same way, until the last row of the fact table. We have chosen Nb as the value of repetition since Nb value is negligible compared to $|F|$. Our method, not only allow to get "F $datasets$" almost equal in size, i.e. get minimal value of the standard deviation of the set S_{CF}, but to obtain the minimal size of the new dimensions created as we will explain in Section 4.2. To build the set CF, We assign each row of the fact table F to the corresponding $datasetF_i$ by using this formula:

$$rows\ of\ datasetF_i \equiv rows\ of\ F\ that\ have\ the\ same\ value\ of \\ (value(partitionKey)\ modulo\ Nb). \tag{2}$$

Algorithm 1 Calculate the values of the attribute $partitionKey$

Input: E, Nb
Output: new fact table with $partitionKey$

```
1: create the new key partitionKey to the fact table F
2: maxline ⟵ |F|; nbBucket ⟵ Nb; i ⟵ 0; j ⟵ 1; nbline ⟵ 1; bucketSuivant ⟵ 0;
   rangeInterval ⟵ Nb; /* we have chosen Nb as value of round robin cycle since Nb ≪ maxline */
3: while (nbline ≤ maxline) do
4:    if (j ≤ rangeInterval * nbBucket) then
5:       if (j mod rangeInterval = 0) then
6:          i ⟵ bucketSuivant;
            partitionKey[nbline] ⟵ i; /* we add the value i to the column partitionKey*/
            bucketSuivant++;
7:       end if
8:       if (j mod rangeInterval ≠ 0) then
9:          i ⟵ bucketSuivant; partitionKey[nbline] ⟵ i;
10:      end if
11:   else
12:      j++; nbline++; j ⟵ 1; i ⟵ 0; partitionKey[nbline] ⟵ i; bucketSuivant ⟵ 0;
13:   end if
14: end while
```

3.1.3 Building dimension datasets

After adding the new attribute $partitionKey$ to the fact table and create the CF set, we build the CDd sets, $d \in \{1..k\}$, of the dimensions. However, in order to create a $group$ (see notations in TAB 1) whose $datasets$ share the same join key "$partitionKey$", we must extends the dimensions Dd and create the new dimensions, denoted by $D'd$, such that the last ones contain the new attribute $partitionKey$. To do this we follow these steps:

First, we create an intermediate table IDd which corresponds to the dimension Dd, then we join Dd with IDd to obtain the new dimensions $D'd$. This method allow to create a $group$ of the $datasets$. The intermediate table IDd is composed of two attributes, fk and $partitionKey$, such that: (1) fk has the same value as the foreign key of the fact table F, coming from the dimension Dd; and (2) $partitionKey$ is the same attribute "$partitionKey$" added to the fact table F. The table IDd has the same number of rows as the fact table F (we denote by N_IDd=|IDd|). Before doing the join between Dd and IDd, we delete all duplication rows in table IDd. i.e. for each numbers i and j, such $j \in \{1..N\_IDd\}/\{i\}$, we delete all rows such that: $fk[i] = fk[j]$ and $partitionKey[i] = partitionKey[j]$. Then we extend IDd by adding some rows ("fk","$partitionKey$") such as: the value(fk) $\in$ (Dd/IDd) and value($partitionKey$) $\in$ 0..Nb-1, i.e. we select distinct values of $partitionKey$.

After that, we build the final $CD'd$, such as $CD'd$ is the same set CDd defined in TAB 1 correspond to the new dimension $D'd$. To create the $datasets$ of $D'd$, i.e. $CD'd$, we applied the same formula 2 such we replace F by the new dimensions $D'd$. The new dimensions $D'd$, $d \in \{1..k\}$, are created just to replicate some rows of Dd through the attribute $partitionKey$. We summarize the phase 1 of our approach in Algorithm 2.

Algorithm 2 Building the datasets

Input: E, n_c, n_t, V_E, V_M /* see TAB.1*/
Output: Nb, CF, and all $CD'd,\ d \in \{1..k\}$
1: determinate the Nb value
2: add $partitionKey$ to the fact table F /* see the Algorithm 1*/
/* first step: create CF*/
3: build the final CF ($datasets$ of the fact table F).
/*we create physically the $datasets$ using the *dataframes* of Spark and the *buckets*.*/
/*second step: build all $CD'd,\ d \in \{1..k\}$.*/
4: **for all** $(Dd, d \in \{1..k\})$ **do**
5: build the intermediate table IDd of dimension Dd. /* IDd has two attribute fk and $partitionKey$*/.
6: delete duplicate rows from the table IDd then extended IDd by adding some rows("fk","$partitionKey$")
7: build the new dimensions $D'd = Dd \bowtie IDd$ and create the final $CD'd$ for each new dimension $D'd$.
8: **end for**

3.2 Placement of the datasets

In this phase, see the FIG. 1, we distribute the $groups$ created equally upon the cluster nodes with round robin fashion. As described in Tab.1 a $group$ is composed of one $dataset$ of the set CF and a $dataset$ of each $CD'd,\ d \in \{1..k\}$, that share the same join key $partitionKey$. Formally, we can denote by $group_i = datasetF_i \uplus_{d=1}^{k} datasetD'd_i$, $i \in$ 0..Nb-1. Thus, we start by placing the $group_0$ in node 1, $group_1$ in node 2,..,and the $group_{p-1}$ in the node m, such as m=p $modulo$ Nb and $p <= Nb$. We restarted the operation in the same way, we put $group_p$ in node 1, $group_{p+1}$ in node 2,..., until the $group_{Nb-1}$ (see Algorithm 3).

Algorithm 3 Datasets placement

Input: CF, $CD'd$ ($d \in \{1..k\}$), N=$\{n_1, n_2, .., n_m\}$, N_disp the set of available nodes.
Output: new distribution schema of the data warehouse E

```
1: N_disp ⟵ N; i ⟵ 0; // i ∈ 0..Nb-1
2: while (CF ≠ ∅) and (CD'd ≠ ∅) do
3:     if (N_disp = ∅) then
4:         N_disp ⟵ N;/* we scan the nodes with round robin fashion*/
5:     end if
6:     put ((datasetF_i) and (each datasetD'd_i, d ∈ {1..k})) in the node N_disp[0]; /* N_disp[0] is the first node
       of the set N_disp. Physically we moved all HDFS block of each dataset to the targeted node N_disp[0]*/
7:     delete datasetF_i element from the set CF;
8:     for all (datasetD'd_i, d ∈ {1..k}) do
9:         delete datasetD'd_i from the set CD'd;
10:    end for
11:    delete the node N_disp[0]; i ⟵ i + 1; /* go to the next node.*/
12: end while
```

3.3 Query transformation

In our approach, we must make some changes to the join condition. If we consider the data warehouse E= $\{F, D1, D2, .., Dk\}$ and we denote by FK_{Dd} the foreign key of dimension tables Dd in the fact table F, and PK_{Dd} the primary key of dimension Dd. We must change join condition as follow: for each $d \in \{1..k\}$, $F.FK_{Dd} = Dd.PK_{Dd}$ become $F.partitionKey = D'd.partitionKey$ ($D'd$ is the new dimension). If we selected one of the foreign key of the fact table as a $partitionKey$ (see Section 3.1), we transform join condition as follow: for each $d \in \{1..k\}$, $F.FK_{Dd} = Dd.PK_{Dd}$ become $F.FK_{Dd}\ modulo\ Nb = D'd.partitionKey$. This is a heavy transformation which can increase CPU cost in the process.

3.4 Partitioning cost

We have seen that our method of creating the sets CF and $CD'd$, $d \in \{1..k\}$, is based on the size of the $dataset$ (i.e. the cardinality of the $datasets$ rows) and we haven't include in our calculus the volume of the $dataset$. In other words, if we take two dimensions $D'i$ and $D'j$, we can have $|datasetD'i_0| \geqslant |datasetD'j_0|$ but the volume($datasetD'i_0$)$\leqslant$ volume($datasetD'j_0$), since each dimension $D'd$ $d \in \{1..k\}$ can has few or many attributes. This can increase the volume of the new dimensions $D'd$. However, with the column storage as *"Parquet"* or *"ORC"*[5], only the attributes solicited by the queries are loaded into memory, and not all the $dataset$ of the set $CD'd$. Also, with the new compression and coding techniques in HDFS, the attribute added, i.e. $partitionKey$, occupies a negligible disk space since all $value(partitionKey) \in 0..Nb\text{-}1$.

4 Implementation and Experimentation

4.1 Implementation

In this Section, we present the implement steps of our approach. First of all, we generated the *DW* using the TPC-DS benchmark, where we store the data directly in HDFS using Parquet

5. Parquet (https://parquet.apache.org/.) and ORC (https://orc.apache.org/.) are column storage formats.

format. After that, we implement the algorithms 1, 2, and 3. To do these, we have used four machines characterized by CPU Pentium I7 (with 8 CPU cores), memory 8 GB and hard drive size 600 GB. We have installed in all nodes the last versions of Hadoop-YARN, Hive, the processing engine Apache Spark, the TPC-DS benchmark, Java, Scala language. We have added in the master node Scala Build Tool "SBT" for compiling and create packages of Scala and Apache Maven to compile and create packages of Java. For all our experiments we keep the default HDFS blocks size 128 MB and 2 the number of replication. Note that there are more than 150 parameters (Petridis et al., 2016) in Spark which can influence to the response time of queries. Thus, in our experimentations we focus only to the candidates parameters such the number of the $executors$ in the cluster, the number of $partitions$ or $buckets$, CPU cores and the amount of memory assigned to the $executors$.

Generation of the data. We have adapted the *spark-sql-perf* application developed by Databricks [6] using Scala language and Spark, where we generate a part of the data warehouse composed of the fact table $store\_sales$ and nine dimensions $customer$, $customer\_address$, $customer\_demographics$, $item$, $time\_dim$, $date\_dim$, $household\_demographics$, $store$ and $promotion$. The data is stored directly in HDFS using Parquet format.

Implementation of the approach. In order to not mixed between concepts and implementation, we now designate $dataset$ by $bucket$, $group$ by $partition$, "Nb by NBk, $lower\_Nb$ by $lower\_NBk$, $upper\_Nb$ by $upper\_NBk$. In our implementation, we have used Hive just to store the meta-data of tables created by Spark-SQL. To create the $buckets$ of the fact and dimension tables we used the instruction:

```
DF.write.bucketedBy($NBk$, "$index$").sortBy ("$index$").
format("parquet").mode
("overwrite").saveAsTable("DB.tablename").
```

Such that *DF* is a $dataframe$, $index$ is the $partitionKey$. We can create $bucket$ in Spark with only Parquet and ORC format, *DB* is the database created in HIVE and *tablename* is the name of the table bucketed.

Before creating the $buckets$, we have update the fact table $store\_sales$ such we replaced some null values of the foreign key, adding the new attribute $partitionKey$ then we compressed the table by $snappy$ codec.

To implement the algorithm 1 and 2, we have used three essential components: $dataframe$, $DataSet$ of Spark (don't be confused with $dataset$ of our approach) and $ArrayBuffer$. Spark bucketing not like Hive, it creates many files for each $bucket$. To implement the algorithm 3, we haven't modified the block policy placement of HDFS as (Eltabakh et al., 2011). To implement our method, the framework API of Hadoop V-2.x would need heavy modifications. Our strategy of placement has currently implemented as an external balancer application. This has the advantage of keeping safe the code of the HDFS default block placement policy.

6. Available from https://github.com/databricks/spark-sql-perf.

4.2 Experimentation

To evaluate our approach, we selected and adapted 4 queries of TPC-DS benchmark (see TAB. 2), such as:

— in query 3, we join two dimensions with the fact table, and we selected few attributes with few filters (i.e. predicates).
— in query 4, we join two large dimensions with the fact table, and we selected some more attributes than the other queries without used filters.
— the query 6 is composed of two subqueries and we selected only two attributes.
— in query 7, we join four dimensions with the fact table as in query 6 and we used more filters than the other queries.

We deleted from both queries aggregate operations, since the last ones performed in reduce phase. So, our idea is to shown how to perform star join operation in only one Spark stage with different type of queries. Since we used three slaves nodes, we generated 100 GB of the *DW* (we can generate 10 TB of data if we use more nodes). In our experiments, for each node, we created 2 $executors$ with 3 CPU cores and 3 GB of memory and we kept 2 GB of memory and 2 CPU core for operating system and $executors$. So, we configure Spark as follow: "$master = yarn$"; "$mode = client$"; "$driver - memory = 5g$"; "$num - executors = 6$"; "$executor - memory = 3g$"; "$executor - cores = 3$". With this configuration we can run 18 tasks (i.e. $3CPU \times 2\ executors \times 3\ slaves = 18$) in parallel (i.e. in each Spark wave). By using formula 1 of Section 3.1.1, we obtain [$lower\_NBk$,..,$upper\_NBk$]=[18,..,75], such $upper\_NBk$=$\lfloor lower\_NBk \times \frac{V_E}{V_M} \rfloor$=$\lfloor 18 \times \frac{100}{24} \rfloor$=75. We executed the queries of TAB. 2 with two the approaches. We denoted by BBH to the default partition and distribution of Spark with Hadoop (baseline approach), and $SMBS_{NBk}$ to our partitioning and distribution approach with different number of buckets, i.e. NBk. We executed the queries with five values of NBk=10, 36, 54, 75, 180 (see FIG. 2). In BBH approach we put "$spark.sql.shuffle.partitions = 50$" to obtain the optimal queries execution time. Note that this parameter in BBH approach equivalent to the parameter NBk in our approach. Also, in baseline approach, Spark execute by default the Broadcast Hash join. In our Approach $SMBS_{NBk}$, since we bucketed all the tables with the same join key $partitionKey$ and deactivated Broadcast Hash join (*Hivecontext.conf.set("spark.sql.autoBroadcastJoin Threshold",0)*) we can execute SMB join in the right way, such we can perform star join operation in only one Spark stage. The FIG. 3 shows the impact of the NBk values to the queries execution time. In FIG. 4 we compared the volume of the dimensions and the new dimensions created with different value of NBk.

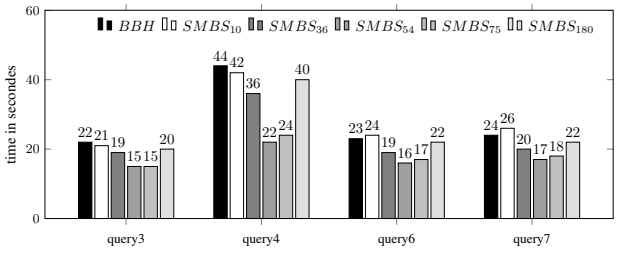


FIG. 2 – *Queries execution time with differences approaches*

name	code query
query 3	select dt.d_year, item.i_brand, item.i_brand_id, from date_dim dt, store_sales, item where dt.d_date_sk = store_sales.ss_sold_date_sk and store_sales.ss_item_sk = item.i_item_sk and item.i_manufact_id = 128 and dt.d_moy=11 limit 100;
query 4	select c_customer_id, c_first_name, c_last_name, c_preferred_cust_flag, c_birth_country, c_login , c_email_address, d_year from customer, store_sales, date_dim where c_customer_sk = ss_customer_sk and ss_sold_date_sk = d_date_sk limit 100;
query 6	select a.ca_state state, d.d_month_seq from customer_address a, customer c, store_sales s, date_dim d, item i where a.ca_address_sk = s.ss_addr_sk and c.c_customer_sk = s.ss_customer_sk and s.ss_sold_date_sk = d.d_date_sk and s.ss_item_sk = i.i_item_sk and d.d_month_seq = (select distinct (d_month_seq) from date_dim where d_year = 2000 and d_moy = 1) and i.i_current_price > 1.2 limit 100;
query 7	select i_item_id, d_month_seq, cd_dep_count, p_cost from store_sales, customer_demographics, date_dim, item, promotion where store_sales.ss_sold_date_sk = date_dim.d_date_sk and store_sales.ss_item_sk = item.i_item_sk and store_sales.ss_cdemo_sk= customer_demographics.cd_demo_sk and store_sales.ss_promo_sk = promotion.p_promo_sk and customer_demographics.cd_gender = 'M' and customer_demographics.cd_marital_status = 'S' and customer_demographics.cd_education_status = 'College' and (promotion.p_channel_email = 'N' or p_channel_event = 'N') and date_dim.d_year = 2000 limit 100;

TAB. 2 – *Queries selected*

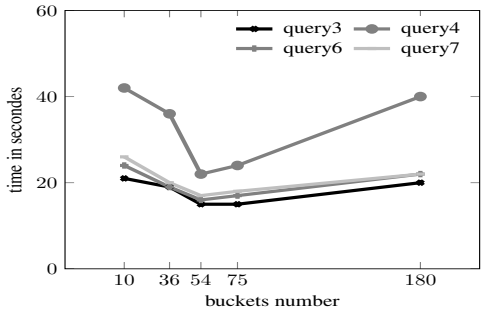


FIG. 3 – *The impact of the buckets' number to the response time of queries*

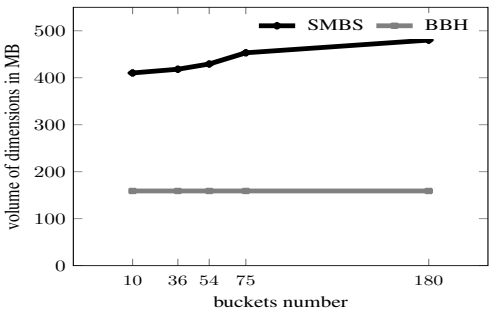


FIG. 4 – *Data volume of dimensions with the two approaches*

4.3 Discussion

First, we interpret the results of the FIG. 2. We have improved the query execution time between 30 to 50 % with $SMBS_{54}$ and $SMBS_{75}$. The reason is: Spark-sql has an optimizer called "Catalyst" which has the capability to choose the best physical plan of queries. Spark generates DAG (Diagram Acyclic Graph) plan for each query execution. The execution plan is composed of several stages. The number of stages is calculated depending of narrow and wide transformations that Spark is doing (e.g. operation "filter" is narrow transformation since it doesn't generate a Shuffle). Spark can group multiple narrow transformations in one stage.

So, with our strategy of partitioning and since we have activated "SMB" join, the queries 3, 4, and 7 are executed in only one Spark stage while are performed in 6, 7, and 10 stages respectively with BBH technique. The 6^{th} query, since it contains nested queries, is performed in 5 stages with our approach $SMBS_{NBk}$ while it necessitates 14 stages with the baseline approach. Also, although, the queries 6 and 7 seem sophisticated, however, we have obtained roughly response time in both approaches, i.e. BBH and $SMBS_{NBk}$, The reason is since we selected few attributes of small dimensions and as we stored data in parquet format, only the selected attributes loaded into memory, thus, we obtain such result. We noticed also that the best result obtained in query 4^{th} (50 %) because, we selected many attributes of the largest table "$customer$", and star join operation is performed with two larges dimensions ($customer$ and $customer\_address$). This, confirmed that our approach is more better when the dimension tables are larges.

We noticed, as shown in FIG. 2 and 3, that the number NBk has a significant impact to the queries performances. For both queries, the best results are obtained when $NBk \in [54, .., 75]$, and this confirm the reliability of our method to select NBk values. Though, as shown in FIG. 4, the number of NBK have an impact to the new dimensions volume, however, our approach don't incur high disk space cost since the volume is increased 2.5 to 3x compared to original dimensions. The reason that causes to increase the volume of new dimensions when the NBK rise is the size of the intermediate table IDd (see Section 3), created from the fact table. The size of the table IDd changes according to the values of "$partitionKey$". In other words, the probability to obtain more duplicated rows in table IDd is rising when we use small value of NBk and vice versa.

Moreover, we can note that, although we have roughly balanced the buckets upon the cluster. However, within each stage, some transformation (like $filter$ and $select$ operations) can incur imbalanced $partitions$ due to the type of query used. This involved imbalanced response time of tasks executed in the same Spark wave. Thus, to handle this issue, we can improve more our balanced technique, taking into account the workload used.

5 Conclusion

In this article, we have implemented a strategy for partitioning and distributing a relational big data warehouse implemented on hadoop ecosystem upon a cluster of homogeneous nodes. Our experiments has demonstrated that our approach allows to execute a star join operation locally, in map side with low communication cost. Though managed to find a good interval for choosing the ideal number of buckets. However, we have seen that it is quasi-impossible to predict the perfect number of buckets to get the best response time for all types of OLAP

queries. Many parameters can influence to the Spark job, such the data volume, data skew, type of the treatment (i.e. queries used) and the physical characteristics of the cluster. In other words, we can improve our approach if we taking into account the workload used.

There are a number of areas for future work. The most important: (i) validate our approach with large cluster and with big DW (1 TO, 10 TO) (ii) we extend and evaluate our work with cost model to predict response time of some OLAP queries; (iii) we adapt more our partitioning and load balancing schema through a well-determined workload; and (iv) we propose an approach piloted by Multi-Agent-System to reduce smartly the I/O cost, such we persist dynamically in memory the buckets most frequently used.

References

Abouzeid, A., K. Bajda-Pawlikowski, D. Abadi, A. Silberschatz, and A. Rasin (2009). Hadoopdb: an architectural hybrid of mapreduce and dbms technologies for analytical workloads. *Proceedings of the VLDB Endowment 2*(1), 922–933.

Afrati, F. N. and J. D. Ullman (2011). Optimizing multiway joins in a map-reduce environment. *IEEE Transactions on Knowledge and Data Engineering 23*(9), 1282–1298.

Armbrust, M., R. S. Xin, C. Lian, Y. Huai, D. Liu, J. K. Bradley, X. Meng, T. Kaftan, M. J. Franklin, A. Ghodsi, et al. (2015). Spark sql: Relational data processing in spark. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pp. 1383–1394. ACM.

Arres, B., N. Kabachi, and O. Boussaid (2015). Optimizing olap cubes construction by improving data placement on multi-nodes clusters. In *Parallel, Distributed and Network-Based Processing (PDP), 2015 23rd Euromicro International Conference on*, pp. 520–524. IEEE.

Benkrid, S., L. Bellatreche, and A. Cuzzocrea (2014). A global paradigm for designing parallel relational data warehouses in distributed environments. In *Transactions on Large-Scale Data-and Knowledge-Centered Systems XV*, pp. 64–101. Springer.

Blanas, S., J. M. Patel, V. Ercegovac, J. Rao, E. J. Shekita, and Y. Tian (2010). A comparison of join algorithms for log processing in mapreduce. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pp. 975–986. ACM.

Dittrich, J., J.-A. Quiané-Ruiz, A. Jindal, Y. Kargin, V. Setty, and J. Schad (2010). Hadoop++: making a yellow elephant run like a cheetah (without it even noticing). *Proceedings of the VLDB Endowment 3*(1-2), 515–529.

Eltabakh, M. Y., Y. Tian, F. Özcan, R. Gemulla, A. Krettek, and J. McPherson (2011). Cohadoop: flexible data placement and its exploitation in hadoop. *Proceedings of the VLDB Endowment 4*(9), 575–585.

Gao, Y., Y. Zhou, B. Zhou, L. Shi, and J. Zhang (2017). Handling data skew in mapreduce cluster by using partition tuning. *Journal of Healthcare Engineering 2017*.

Gufler, B., N. Augsten, A. Reiser, and A. Kemper (2012). The partition cost model for load balancing in mapreduce. In *Cloud Computing and Services Science*, pp. 371–387. Springer.

Hefny, H. A., M. H. Khafagy, and M. W. Ahmed (2014). Comparative study load balance algorithms for map reduce environment. *International Journal of Computer Appli-*

cations 106(18), 41–50.

Kwon, Y., M. Balazinska, B. Howe, and J. Rolia (2012). Skewtune: mitigating skew in mapreduce applications. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pp. 25–36. ACM.

Petridis, P., A. Gounaris, and J. Torres (2016). Spark parameter tuning via trial-and-error. In *INNS Conference on Big Data*, pp. 226–237. Springer.

Saha, B., H. Shah, S. Seth, G. Vijayaraghavan, A. Murthy, and C. Curino (2015). Apache tez: A unifying framework for modeling and building data processing applications. In *Proceedings of the 2015 ACM SIGMOD international conference on Management of Data*, pp. 1357–1369. ACM.

Sharafi, A. and A. Rezaee (2016). Adaptive dynamic data placement algorithm for hadoop in heterogeneous environments. *Journal of Advances in Computer Engineering and Technology 2*(4), 17–30.

Taniar, D., C. H. Leung, W. Rahayu, and S. Goel (2008). *High performance parallel database processing and grid databases*, Volume 67. John Wiley & Sons.

Thusoo, A., J. S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P. Wyckoff, and R. Murthy (2009). Hive: a warehousing solution over a map-reduce framework. *Proceedings of the VLDB Endowment 2*(2), 1626–1629.

Valvåg, S. V., D. Johansen, and Å. Kvalnes (2013). Cogset: a high performance mapreduce engine. *Concurrency and Computation: Practice and Experience 25*(1), 2–23.

Vernica, R., A. Balmin, K. S. Beyer, and V. Ercegovac (2012). Adaptive mapreduce using situation-aware mappers. In *Proceedings of the 15th International Conference on Extending Database Technology*, pp. 420–431. ACM.

Wang, H., X. Qin, X. Zhou, F. Li, Z. Qin, Q. Zhu, and S. Wang (2015). Efficient query processing framework for big data warehouse: an almost join-free approach. *Frontiers of Computer Science 9*(2), 224–236.

Zaharia, M., M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica (2010). Spark: Cluster computing with working sets. *HotCloud 10*(10-10), 95.

Résumé

Les techniques de partitionnement et de distribution des entrepôts de données sont appliquées pour améliorer le traitement des requêtes OLAP. Cependant, les schémas définis par certaines de ces techniques, sont basés sur une charge des requêtes. Comme celle-ci peut changer, ces schémas doivent être redéfinis, une opération coûteuse qui perturbe le système. Dans cet article, nous proposons une stratégie de partitionnement et de distribution d'un entrepôt de données volumineux, sur un cluster de nœuds homogènes, indépendamment de la charge de requêtes. Elle tient compte des caractéristiques physiques du cluster et de l'équilibrage de la charge des données. L'approche proposée permet d'exécuter la jointure en étoile localement dans la phase Map et avec un faible coût de communication. L'évaluation expérimentale de cette approche montre l'amélioration jusqu'à 50 % du temps d'exécution des requêtes OLAP.

Towards an Ontology-Based Data Access System for Aggregated Search

Ahmed Rabhi*, Hassan Badir**, Amjad Ratrrout***

*National School of Applied Sciences, Abdelmalek Essaâdi University, Tangier, Morocco
rabhi.ahmed.1992@gmail.com,
**National School of Applied Sciences, Abdelmalek Essaâdi University, Tangier, Morocco
hbadir@gmail.com
***Arab American University, Jenin, Palestine
amjad.ratrrout@aauj.edu

Abstract. Data sets in the web of data are accessed via data sources providing information from different domains, these data sources, also called SPARQL endpoints, are heterogeneous and managed independently, which complicates their querying, besides, the response to certain queries needs to aggregate information from several sources. Our work aims to set up a system to process queries over a large number of distributed data sources and this system should have a good knowledge base about sources and their domains of interest to improve answering user's queries. This paper presents a study about the usefulness of the Ontology Based Data Access paradigm to improve query answering by providing a conceptual unified view of several data sources to the user.

1 Introduction

The Web is evolving from a "Web of linked documents" into a "Web of linked data", the latter provides better opportunities for sharing and searching information. In fact, The Web of Data architecture enables access to data by making it available in machine-readable format and linking it based on a set of design principles provided by Linked Data standards and, consequently, enabling people and machines to interchange data, which greatly evolve the possibilities of searching information and data.

The Web of Data is a large collection of data from different domains (Life sciences, Government, Geography, Social networks, ...) related to each other forming data graphs. These data are stored in RDF datasets, and are accessed through data sources called Endpoints. The worry is that the number of data sources has recently increased on a large scale. Actually, the basic principles underlying the structure of these datasets include the provision of decentralized data, thus, certain queries can only be answered by retrieving information from several data sources. Moreover, these sources are distributed, heterogeneous and managed independently. Hence, searching information in the Web of Data means searching information in a number of data sources that have no relationship between them. Consequently, querying the sources of the Web of Data becomes a complicated task.

It is possible that a sought information cannot be found by querying a single source, which means that the user must collect the data from several data sources in order to find a response to his query, which leads us to aggregated search, this aggregated search involves looking for the response to a query by aggregating data from several sources, it actually collects fragments of information from more than one source, thus, to have the potential of giving more possibilities to access to decentralized data by creating associations between pieces of information that are published separately and related to the same entity.

In this paper, we will present a query processing engine aimed at solving the problem of querying distributed and independent data sources. The main idea behind this project is to look for answers to a user's query by aggregating the results from different sources besides having a good knowledge about queried data sets and their domains of interest. The proposed solution is based on different approaches such as indexing that was used to optimize source selection, and Ontology-Based Data Access paradigm (OBDA) to enrich query processing based on an ontology that provides domain knowledge and additional vocabulary for query formulation and expose data in a conceptually clear manner.

2 Overview & Basic concepts

The Web of Data. According to Heath and Bizer (2011), the web of data is a global space where individuals and organizations have adopted Linked Data standards to publish their data. Hence, the Web of Data forms a giant graph consisting of billions of RDF data distributed on a large number of Datasets covering all domains such as geography, politics, life science, social networks and others domains. An RDF Dataset is a data structure consisting of nodes and organized in a graph, in deed, a set of RDF triples forms an RDF dataset, and these Datasets are accessed via sources called Endpoints and queried using SPARQL query language that provides its own syntax, actually, a SPARQL query is composed of triple patterns each one presenting a triplet (Subject, Predicate, Object).

The main issue. Due to the distribution and independence of data sources, a SPARQL query may require interrogating multiple sources to find a good result reporting to the user's expectation. Thus, the heterogeneity of data sources poses a difficulty to answer certain queries which represents a major obstacle for search engines. Therefore, it is necessary to have an aggregated search engine able to collect the appropriate data that responds to a query by aggregating information from several data sources taking into account the distribution and heterogeneity of sources.

Aggregated search. Sushmita et al. (2010) defined aggregated search as an approach to access largely distributed information. It aims to produce responses to queries by integrating fragments of information from several sources from different domains into a single result interface. These queries look for objects that do not exist entirely in one of the queried sources but are constructed from different sources. Looking for aggregated information has the potential of giving more possibilities to access to distributed information by creating associations between pieces of information that are published separately and related to the same entity. The results are valuable objects that can be used in different domains.

Ontology-Based Data Access & Indexing. According to Calvanese et al. (2017), a system using the approach of OBDA is composed of three components: an ontology describing the domain of interest expressed in terms of relevant concepts and logical assertions characterizing the domain knowledge, a set of data sources and the mapping between the ontology and data sources which is a precise specification of the correspondence between the data contained in data sources and ontology's elements. as stated by Kharlamov et al. (2015); Baader et al. (2016), with OBDA, datasets querying is enriched with an ontology that provides domain knowledge and additional vocabulary for query formulation and expose data in a conceptually clear manner.

The task of selecting relevant sources is based on the indexing approach, the advantage of using an index is that it helps the search engine to access specific data sources and consequently optimize planning the execution of requests which decreases data transfer and leads to less of unnecessary executions.

3 Related work

3.1 Query processing over SPARQL Endpoints

Several systems were set up to query distributed data sources in the Web of Data. The main task of these systems is to allow SPARQL queries execution on multiple Endpoints as if it took place on a single large dataset and join the results from queried sources to return a single final answer, taking into account the large number of data sources and the process optimization. To this end, each system has different characteristics and proposes different techniques and approaches to process SPARQL queries, and these systems adopt different solutions for specific problems or situations.

DARQ is proposed by Quilitz and Leser (2008) as a federated search engine providing transparent query access to multiple SPARQL services it gives the user the impression to query one single RDF graph despite the real data is distributed on the web of data. The problem is that access to multiple distributed and autonomous RDF data sources makes query formulation hard and lengthy. The system offers a single interface for querying distributed SPARQL endpoints and makes query federation transparent to the client. Using service descriptions DARQ provides a powerful way to dynamically add and remove endpoints in a manner that is completely transparent to the user.

Due to the decentralized architecture of the interlinked data, several datasets contain duplicated data. Saleem et al. (2013) pay attention in there solution, to the effect of duplicated data on federated querying. The main innovation behind there solution is to avoid querying sources that would lead to duplicated results. The system proposes an index-assisted approach in his process to estimate the overlap between different sources results. To identify relevant sources for each triple pattern of the query, the system goes through two steps: the first step is triple pattern-wise source ranking, in this step, the system ranks sources based on number of expected results. The second step is triple pattern-wise source skipping in order to skip sources that contribute with little or no new results.

Cosmin Basca (2014) introduced Avalanche to find up-to-date answers to queries over SPARQL Endpoints. It first gets online statistical information about potential data sources and their data distribution. Then, it plans and executes queries in a concurrent and distributed

manner trying to quickly provide first answers. The main contributions of Avalanche are: a querying approach over Web of Data, without fine-grained prior knowledge about its distribution, and a novel combination of interleaving cost-based planning (with a simple cost-model) with concurrent query plan execution that delivers first results quickly. Avalanche is dynamically adaptive to changing external network conditions, provides up-to-date results, and is flexible since it makes few limiting assumptions about the structure of participating triple stores. Avalanche discovers sources using VoID stores, then collects statistics of the cardinalities (number of instances) for each triple pattern after querying selected sources, finally, Avalanche executes queries according to a plan matrix to finally return results

Görlitz and Staab (2011) presented a distributed query processing strategy in there work, called SPLENDID, and it consists of executing queries on distributed data sources and aggregates returned results. However, query planning requires a priori knowledge about data sources to judge whether a data source is relevant or not. As avalanche, SPLENDID uses VoID descriptions in his index to discover datasets and getting statistics, finally, it uses ASK queries to discard irrelevant datasets.

To enable a transparency in querying multiple datasets over the Web of Data, Akar et al. (2012) propose a solution called WoDQA (Web of Data Query Analyzer), it is a federated query engine that discovers relevant datasets in an automated manner using VoID documents as metadata. WoDQA focuses on powerful dataset elimination by analyzing query structure with respect to the metadata of datasets. In his architecture, WoDQA does not use either the index or ASK queries, it is only based on VoID description to select relevant data sources. The system is composed of three main modules: DataSetAnalyzer to discover relevant sources, QueryReorganizer to rewrite queries (depending on DatasetAnalyzer results), QueryExecutor to execute queries.

Wang et al. (2013) presented LHD as a parallelism-based distributed SPARQL engine. Several technics are used by LHD to minimize transferred data size through the network as well as to maximize the data transfer rate. First, data source selection techniques to decrease transfer and response time, then, the engine uses an optimization algorithm that quickly produces an optimal query plan for parallel execution, and a parallel execution system that fully exploits capacity of bandwidth and data sources. To select data sources, LHD uses 2 main tasks: the first one is based on the VoID description to obtain metadata of the data sources and analyses the predicate partition information in VoID files and identifies data sources having the same predicate as relevant candidates to a query triple pattern. Then ASK queries, enclosing the triple pattern, are sent to these candidates to refine selected sources to accurate coast estimation.

Discussion. According to this study, a sophisticated SPARQL query processing requires three major mechanisms: a query decomposing mechanism whose purpose is decomposing user's query and process every component of it, a source selection mechanism that optimizes data sources exploration and a result preparing mechanism to join the query results and return the final answers.

It can be noted that source selection is a task of great interest. Indeed, communication with external sources and data transfer will make the processing engine dependent on external conditions such as the quality of the connection and risks of server failures, which led to the implementation of various techniques and approaches that have been adopted by these systems in order to ensure a favorable and appropriate sources selection for different conditions. The

main techniques are: indexing to facilitate access to the sought resources, using ASK query to verify the existence of resources, and finally the VoID description that allows discovering sources automatically and provides statistics about datasets.

3.2 Ontology-Based Data Access

3.2.1 Definition and Notions

According to Bagosi et al. (2014), Ontology-Based Data Access (OBDA) is an important approach to access data through a conceptual layer. This paradigm is based on an ontology that plays the intermediary's role between the user and data sources, as stated by Kharlamov et al. (2015). Baader et al. (2016) reported that the general idea is to add an ontology which supplies knowledge of the domain of interest and enriches the necessary vocabulary for queries formulation. The main functionality of OBDA systems is query answering, actually, OBDA is a data integration approach that allows users to query data sources through a unified conceptual view. Thus, the user can look for information without having to know the structure of the data contained in sources. De Giacomo et al. (2017) noted that the ontology must not carry out updates on the data, because their modification presents a high risk which can deeply impact the content used by other users of the same sources. Calvanese et al. (2017) affirm that this paradigm provides an integrated view and a semantic description of the basic concepts in the data domain, as well as the relationships between these concepts and the logical modeling characterizing the domain knowledge, thus, information consumers can have a semantic access to data sets, as declared by Kogalovsky (2012). An Ontology-Based Data Access system uses ontology as a conceptual schema, whose implementation must be supported by a user interface, describing the relevant concepts of the subject domain in a set of data sources.

As reported by Calvanese et al. (2017), The main purpose of an OBDA system is to allow information consumers to query data using elements in the ontology. OBDA, actually, enriches datasets querying with an ontology that provides domain knowledge of different concepts provided by data sources to expose data in a conceptually clear manner. A system adopting OBDA approach provides a standard vocabulary for the target application domain (life science, geography, social network). It is true that in such systems, only a small part of the ontology's vocabulary will appear in the data layer. However, this small part plays a major role in the formulation of queries since ontology's axioms are linked to the data vocabulary. Thus, as cited by Calvanese et al. (2017), the user may be able to pose a query by referring only to the ontology without considering data sources that contribute to the response. So, thanks to OBDA, an aggregated search system may provide a clear and unified view of several sources in one single user interface.

According to Kogalovsky (2012) and Calvanese et al. (2017) OBDA provides an advanced formal expressive means for database representation and query specification, and the distinction between the ontology and data sources reflects the separation between the conceptual layer presented to the user, and the data layer, with the mapping that acts as an adapter between the two layers. De Giacomo et al. (2017) affirm that an interesting advantage of OBDA is that it ensures independence between data sources and the ontology, the two levels are only coupled using declarative mappings. This independence is of great importance since it prevents data sources to be modified by users. As reported by Calvanese et al. (2017), axioms in the ontology can be seen as semantic rules that are used to complete the knowledge given from data in

the sources. These axioms allow one to derive new facts from the source data, which leads to a new set of answers that are computed logically deriving from the combination of incomplete knowledge from sources and the ontology axioms.

3.2.2 OBDA system structure

As shown in figure 1, the main components of an OBDA system are: a conceptual layer O describing formally the domain of interest by expressing relevant concepts of the domain knowledge and logical relationships between them, a schema S of the data structure and the mapping M between the ontology and data sources.

The conceptual layer is a formal description of the domain of interest to a specific community of users, expressed in terms of relevant concepts, attributes of concepts, relationships between concepts and logical assertions characterizing the domain knowledge. This logic allows to specify a domain by defining classes and by providing a structure to the knowledge base about classes using a rich set of logical operators. The domain of interest is described in an ontology that provides "a single point of semantic data access", and allows queries to be formulated in terms of a user-oriented conceptual model that abstracts away complex implementation-level details. Knowledge is generally represented using Description Logics (DLs) that are widely recognized as appropriate logics for expressing ontologies and are at the basis of the W3C standard ontology language OWL. DL ontologies represent knowledge in terms of concepts, denoting sets of objects, and roles, denoting binary relationships between objects. All in all, the ontology's axioms provide a unified view of several data sources to the user and allow one to enrich the retrieved information.

The data layer contains repositories that are accessible by users where data concerning the domain are stored, the schema S represents data structure. In the general case, such repositories are numerous, heterogeneous and each one managed independently from the others.

The third component is the mapping M, its main purpose is to make the correspondence between the data and the ontology. Indeed, M consists of a set of mapping assertions, the terms in the ontology are mapped to the data layer using mappings which associate to each element of the conceptual layer a query over data sources. Thus, queries posed over the conceptual layer are translated into a query language that can be handled by the data layer. So that user's query can be independent of the conceptual data representation in sources. in other words, the mapping allows to automatically translate queries posed over the ontology into data-level queries that can be executed by data sources.

4 Our suggested solution

4.1 An Ontology-Based Data Access system for aggregated search

Due to the large number of data to be queried and the complexity of the user's requests, query processing must be effective in such a way as to return the maximum possible results, taking into account the large number of data sources, diversity of data to be handled and the diversity of domains of data contained in sources. In this context, we present a solution to aggregate data from multiple data sources. The idea is to build a search engine designed to query datasets in the Web of Data supporting SPARQL queries processing and it seeks

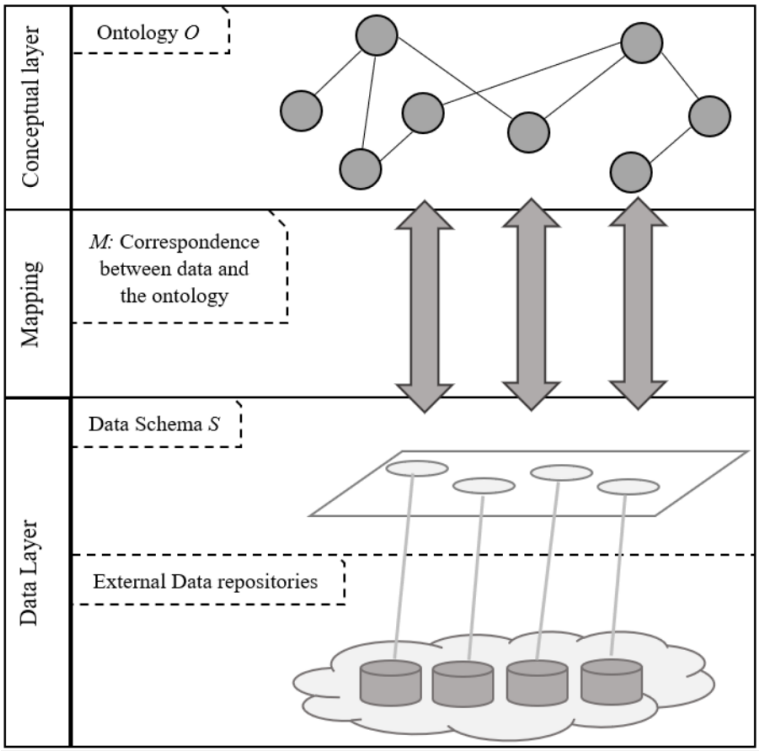


FIG. 1 – *OBDA system's structure*

answers over distributed, heterogeneous and independent data sources. the engine aims to provide a unified view of multiple data sources to the user via a single interface and to manage adaptation of queries since not all sources handle SPARQL 1.1.

As shown in figure 2, our solution is composed of four main Modules: The first one is the Query Analyzing module, the system uses this module to prepare the query, its purpose is to analyze different parts of the user's query and to decompose it into sub-queries in order to obtain a query for each triple pattern. The second Module, Sources Manager, manages the index and statistics about sources besides exploring new data sources over the web of data. Ontology-Based Data Access module focuses on representing domain knowledge of data sources in a conceptually clear way, unifying several sources in one single view and adapting user queries to be handled by sources. The last module is called Answers Processing, its objective is to evaluate sub-queries in the appropriate data sources, to collect results and statistics of the executions to finally return results to the user.

4.2 The system architecture

4.2.1 Query Analyzer module

The main functionalities of this module are: parsing the user's query, decomposing it and building sub-queries. Parsing the query is the first step of the process, it consists of reading the SPARQL query which is initially in String format and verifying its syntax, then transforming it into a SPARQL model. Next task is the Query Decomposing, this is an important task in the distributed search, indeed, its aim is to decompose the user's query in order to define the parts of sought information, and identify the different query's parts (triple patterns, variables, blank nodes, select statements, filters and aggregate functions). The third task of this module is sub-queries building, the purpose of this task is to create a sub-query for each triple pattern, thus, querying data sources returns more results and we'll get more variety of data, which enhance the possibility of responding the user's expectation.

4.2.2 Sources Manager module

To process a query, the search engine must first select candidate data sources that may return a beneficial result. This makes the step of selecting sources a very important step in search engines to minimize communication between the computing node and data sources which is expensive in terms of memory cost and execution time. Our system performs this step to select relevant Endpoints in order to plan the execution of sub-queries. There are several techniques and approaches for source selection, our engine is based essentially on the indexing approach, taking into consideration the couple (Predicate, Object) of each triple pattern. Actually, the use of indexing approach helps the search engine to access to specific data sources and consequently optimizes planning the execution of requests which leads to decreasing process time and less of unnecessary executions by avoiding querying irrelevant sources.

Another task performed by this module is source discovering. Indeed, the Source Manager module is also aimed to discover relevant data sources automatically on the Web using VoID descriptions that provide information about the contained data and also Endpoints that may be available for a dataset. VoID documents store metadata of datasets and we can access them trough VoID stores.

Finally, the module manages statistics about data sources, since the improvement and development of the search engine requires a good knowledge about sources to be queried, their availability and efficiency, in addition, the fact of having statistics of the execution history represents a very good benchmark in order to enrich the system over time.

4.2.3 OBDA module

Since Endpoints are independent and may return a great variety of data, we have adopted OBDA as a solution to enrich datasets querying with an ontology that provides a formal representation of the domains knowledge (Geography, life science, social network) as well as additional vocabulary of different concepts, and relationships between them, provided by the data sources to expose data in a conceptual manner. Hence, the user will be able to query several, independent, distributed and heterogeneous sources through a single interface providing a unified view of data repositories over the web of data.

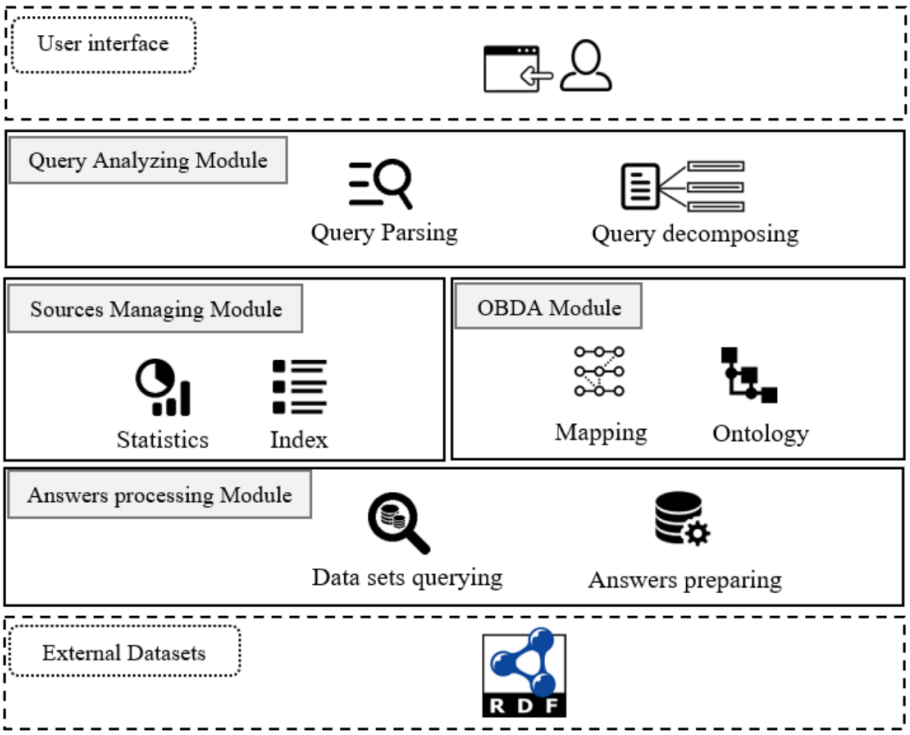


FIG. 2 – *The system architecture*

Terms of the ontology must be mapped to data layer using mapping assertion to make the correspondence between contained data in data sets and ontology axioms, this mapping associates a query over the conceptual layer (the ontology) to a query over the data layer (external sources), thus, the system will be able to manage the queries formulation in order to be handled by every targeted data source taking into account the query language handled by each source.

The data layer presents data sets to be queried, the physical location of the data. In our project we focus on processing SPARQL queries over the web of data, so, as shown in figure 2, the data layer presents external repositories reachable via SPARQL Endpoints.

4.2.4 Answers processing module

Using this module, the system evaluates sub-queries by executing them in data sources according to an execution plan established by the Sources Manager in order to return answers to the user. In addition to querying data sources and processing final results, this module carries out statistics recording and index updating with Sources Manager module.

5 Conclusion and Future work

In this paper we presented a solution for aggregated search over several data sources in the Web of Data. Our main purpose is to gather pieces of information by querying distributed and heterogeneous data sources, this solution focus on providing to the user a unified view of independent data sets as if it is a single one. To this end, we highlighted the usefulness of the Ontology-Based Data Access paradigm and its importance to present a knowledge description of data and to mediate between the user's expectations and different data sources structure. Next, we exposed the system architecture that may subsist to the need. The main objective of the system is to look for answers over several data sources by providing to the user a single interface, this engine is composed of four modules each one performs different tasks and functionalities, beginning with query decomposing, then, selecting data sources and query reformulation, ending with answers preparing.

In future, we hope to implement the OBDA module and improve the implementation of other module, we will extend the index to discover relevant data sources automatically on the Web using VoID descriptions. Next, the architecture will be extended by distributing the process on a cluster of machines, instead of a single one, thus, the aggregated search will be processed by a cluster of working nodes and will increase the parallelism in the queries processing. Besides that, we will make a data sets survey in order to define the domain knowledge of each data sets and build axioms of the ontology.

References

Akar, Z., T. G. Halaç, E. E. Ekinci, and O. Dikenelli (2012). Querying the web of interlinked datasets using void descriptions. *LDOW 937*.

Baader, F., M. Bienvenu, C. Lutz, and F. Wolter (2016). Query and predicate emptiness in ontology-based data access. *Journal of Artificial Intelligence Research 56 (JAIR) 56*, 1–59.

Bagosi, T., D. Calvanese, J. Hardi, S. Komla-Ebri, D. Lanti, M. Rezk, M. Rodríguez-Muro, M. Slusnys, and G. Xiao (2014). The ontop framework for ontology based data access. *Chinese Semantic Web and Web Science Conference*, 67–77.

Calvanese, D., G. De Giacomo, D. Lembo, M. Lenzerini, R. Rosati, and G. A. Ruberti (2017). Ontology-based data access and integration.

Cosmin Basca, A. B. (2014). Querying a messy web of data with avalanche. *Journal of Web Semantics 26*, 1–28.

De Giacomo, G., D. Lembo, X. Oriol, D. F. Savo, and E. Teniente (2017). Practical update management in ontology-based data access. *International Semantic Web Conference*, 225–242.

Görlitz, O. and S. Staab (2011). Splendid : Sparql endpoint federation exploiting void descriptions. *Proceedings of the Second International Conference on Consuming Linked Data 782*, 13–24.

Heath, T. and C. Bizer (2011). *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool publishers.

Kharlamov, E., D. Hovland, E. Jiménez-Ruiz, D. Lanti, H. Lie, C. Pinkel, M. Rezk, M. G. Skjæveland, E. Thorstensen, G. Xiao, et al. (2015). Ontology based access to exploration data at statoil. *ISWC*, 93–112.

Kogalovsky, M. R. (2012). Ontology-based data access systems. *Programming and Computer Software 38*(4), 167–182.

Quilitz, B. and U. Leser (2008). Querying distributed rdf data sources with sparql. *European Semantic Web Conference (ESWC)*, 524–538.

Saleem, M., A.-C. N. Ngomo, J. X. Parreira, H. F. Deus, and M. Hauswirth (2013). Daw: Duplicate-aware federated query processing over the web of data. *ISWC*, 574–590.

Sushmita, S., H. Joho, M. Lalmas, and R. Villa (2010). Factors affecting click-through behavior in aggregated search interfaces. *Proceedings of the 19th ACM international conference on Information and knowledge management*, 519–528.

Wang, X., T. Tiropanis, and H. C. Davis (2013). Lhd: Optimising linked data query processing using parallelisation.

Résumé

Les datasets dans le web des données sont accessibles via des sources de données fournissant des informations de différents domaines, ces sources de données, également appelées Endpoints, sont hétérogènes et gérées indépendamment, ce qui complique leur interrogation, de plus, la réponse à certaines requêtes doit agréger des informations provenant de plusieurs sources. Notre travail vise à mettre en place un système de traitement des requêtes sur un grand nombre de sources de données distribuées et indépendantes et ce système doit avoir une bonne base de connaissances à propos des sources de données et leurs domaines d'intérêt pour améliorer la réponse aux requêtes des utilisateurs. Cet article présente une étude sur l'utilité du paradigme d'accès aux données basé sur l'ontologie (OBDA) afin d'améliorer la réponse aux requêtes en complétant les connaissances extraites des sources de données en se basant sur les axiomes d'une ontologie.

Community Detection in Social Context based on Optimized Classification

Lamia Berkani*, Sara Madani**
Soumeya Mekherbeche**

*Laboratory for Research in Artificial Intelligence (LRIA), Dep. of Computer Science,
Faculty of Computing and Electrical Engineering,
USTHB University, Bab Ezzouar, Algiers, Algeria
lberkani@usthb.dz
** Dep. of Computer Science, USTHB University,
Bab Ezzouar, Algiers, Algeria
{smadani; smekherbeche}@usthb.dz

Abstract. The development of web2.0 and social media has generated an important volume of data on the web. New challenges and issues are araised regarding the management of this data. We focus in this paper on the identification of groups of members who share similar tastes and preferences in the context of social networks. We address the need of community detection in this context based on an optimized classification algorithm. Our approach applies the K-means algorithm as an initial classification solution, and then optimizes this classification by using two different meta-heuristics: the Tabu Search (employing local search methods to explore the solution space beyond local optimality) and the Bee Colony Optimization algorithm (a population-based search algorithm). Experiments and comparisons carried out on different datasets show that the results obtained are promising.

1 Introduction

With the development of Web 2.0 and social media technologies, social network sites have attracted millions of users, and many of them have integrated these sites into their daily practices. A social network can be represented by a graph consisting of a set of nodes and edges connecting these nodes. The nodes represent the users, and the edges correspond to the interactions among them. However, given the complexity of these networks, new challenges have emerged, requiring a restructuring of these networks to facilitate the efficient interaction of users. According to Bedi and Sharma (2016), the tendency of people with similar tastes and preferences to get associated in a social network leads to the formation of virtual communities/clusters. The experience and practice on social networks show that the detection of these communities can be beneficial in several ways and for many applications such as finding likeminded users for marketing and recommendations or a common research area in collaboration networks. This could strengthen links / create new links between people with similar interest profiles.

Community detection is one of the research fields of social network analysis. A large number of research works and algorithms have been proposed. A lot of surveys on social

community detection have been published such as (Fortunato, 2009; Plantié and Crampes, 2013; Wang et al., 2015; Bedi and Sharma, 2016).

The state of the art shows that some works are graph-oriented based on the modularity function (Newman, M., and Girvan, 2004), while other works use classification techniques. Our goal in this research is to propose a new approach taking advantage of the graph structure and using a classification technique. However, in order to optimize the detection of communities, we combine the classification algorithm with two different meta-heuristics: the Tabu Search and the Bee Colony Optimization algorithm.

The remainder of this paper is organized as follows: Section 2 provides some basic definitions and concepts on community detection and presents some related work on social community detection. Section 3 presents our approach for the detection of communities in a social context. The results of the experiments are given in section 4. Finally, the conclusion summarizes the most important results and describes some future perspectives.

2 State of the art

2.1 Background and Related Work

The most commonly used definition of the term community is that of Yang (2010): "a community is a group of network nodes, within which the links connecting nodes are dense but between which they are sparse". According to Fortunato (2009), communities, also called clusters or modules, as "groups of vertices that probably share common properties and/or play similar roles within the graph". Papadopoulos (2011) defines communities as: "groups of vertices that are more densely connected to each other than to the rest of the network".

A panoply of community detection algorithms exists in the literature. The first idea using static networks was proposed by Newman and Girvan (2004). The proposed method is based on a modularity function, aiming to obtain the optimum partitioning of communities. In the same direction, Blondel et al. (2008) have proposed Louvain algorithm to detect communities using the greedy optimization principle to optimize the gain of modularity. According to Babers and Hassanien (2017), traditional methods such as graph partitioning used to detect community within networks by dividing it based on predefined size. Spectral clustering method is based on similarity matrix and integrating similar groups used by hierarchical clustering technique.

Other researchers explored the dynamic aspect of networks to identify communities structure and their development over time. Hopcroft et al. (2004) have proposed the first work on dynamic community detection which decomposes the dynamic network into a set of snapshots (each snapshot corresponds to a single point of time). The authors applied an agglomerative hierarchical method to detect communities in each snapshot and then they matched these extracted ones in order to follow their evolution over time.

Recently, some works are using optimization algorithms for the community detection: Babers and Hassanien (2017) presented a cuckoo search optimization algorithm with Lévy flight for community detection in social networks. According to the authors, the experimental demonstrated that the proposed algorithm can define the structure and detect communities of complex networks with high accuracy and quality. Sharma and Annappa (2016) have introduced modularity metrics and Hamiltonian function combined with meta-heuristic optimization approaches of Bat algorithm and Novel Bat algorithm.

2.2 Discussion

The review of related works shows the limitations of the graph-based methods, and the current trend to apply optimization techniques for identifying communities in social networks. In this same direction, we attempt in this work to apply other meta-heuristics and to explore other methodologies of research for communities' detection in social networks.

Our goal is to explore other optimization algorithms, in addition to those already used in the literature, for this same problem. Our approach will be based on the use of an unsupervised classification algorithm that we will optimize thanks to the use of meta-heuristics. Indeed, the random effect of classification algorithms, including K-means, gives different results from one execution to another. The result expected by our approach would be obtaining the best classification.

3 Our Approach to Community Detection

In this section, we present our community detection approach in a context of social networks. A network is described by a graph where the nodes are the users and the edges are the social links between the users. In order to detect user communities we proposed two approaches: (1) a classification based on the K-means algorithm; and (2) an optimized classification using two different meta-heuristics, the Tabu search and the Bee Colony Optimization algorithm.

3.1 K-means-based Classification Approach

The k-means classification is one of the unsupervised classifications. The principle of the algorithm is the initial definition of the number of clusters. The algorithm assigns to each cluster a vertex, extracted randomly, from all the vertices.

In order to assign each vertex to the most appropriate community, a similarity criterion is developed indicating the nearest center of a given vertex. The similarity function used is the geodesic distance μij. This function calculates the number of edges of the shortest path connecting a vertex i and another vertex j. In order to elect the new centers we calculate the average centrality of each summit with respect to its community, the summit with the lowest value of centrality of intermediacy will be elected as new center. The average centrality being the average of the distances from the summit to all the others, as indicated by the following formula:

$$C_{AVG}(V_i) = \frac{1}{n-1} \sum_{i \neq j} \mu_{i,j} \quad (1)$$

where:

Vi: is the i^{th} vertex.
n: is the number of vertices.
μij: is the geodesic distance.

The K-means algorithm adapted to our case is stated as follows:

Algorithm1: K-means Algorithm

Input: A set of vertices $V_1 \ldots V_{n;}$ The number of clusters K

Output: A set of communities {C1, C2...C_k}

1. Choose k initial centers $C_1, \ldots, C_k$
2. Assign each individual to the nearest center
3. Recalculate the center of each cluster using the centrality function.
4. If no item changes group then stop and exit groups otherwise go to (2) until cluster stability.

3.2 Optimized Classification-based Approach

In this section, we present the two meta-heuristics that we have adapted to the problem of community detection.

3.2.1 Tabu Search-based Classification

The Tabu search is an optimization meta-heuristic used to solve complex and / or large problems (Glover, 1986). Taboo search overcomes the local optimum problem encountered in local search using a taboo list. The taboo list is represented by a hash table such that the key is the concatenation of the identifiers of the communities of the solution. The taboo list will contain all the solutions explored in the search. The implementation steps are as follows:

1. Generate an initial solution using K_means
2. Generate the neighborhood of this solution by using local search.
3. Return the best non-tabu neighbor solution
4. If the best neighbor does not exist then diversification.
5. Return the least worst non-tabu neighbor.
6. If the least worst non taboo neighbor does not exist then return the best solution taboo if the number of chances is greater than 0.
7. Otherwise generate a solution from k-means preserving the contents of the taboo list.
8. Repeat the process from Step 2 with the new neighbor returned

Local search is based on the exploration of the neighborhood of a solution, which consists of moving from one solution to another solution until a solution considered optimal is found, or the number of iterations be completed. The neighborhood of a solution *S* is a set of solutions where each solution is formed from the solution S by varying the communities of the users each time. The implementation steps are as follows:

1. Generate an initial solution from K-means.
2. Generate the neighborhood of this solution.
3. Return the closest solution.
4. Repeat the process from Step 2 with the new neighbor returned.

3.2.2 Bee Colony-based Classification

Colony optimization of bees manipulates a set of bees where each bee is a feasible solution to a given problem. In order to make the best use of meta-heuristics and to prove their effectiveness, it would be necessary to consider a codification that makes it possible to model the problem. In our case, each possible partitioning represents a solution, where each bee corresponds to a feasible partitioning. The solution can be represented by a vector containing the different existing communities. The vector indices represent the keys of each user in a hash table. Each box of the vector contains the identifier of the community to which the user belongs, where each user having as key the index of this box.

Illustrative example: Let's consider the following three communities:

- Community 1: u5, u6, u7and u10
- Community 2: u3, u2 and u11
- Community 3: u1, u8 and u9

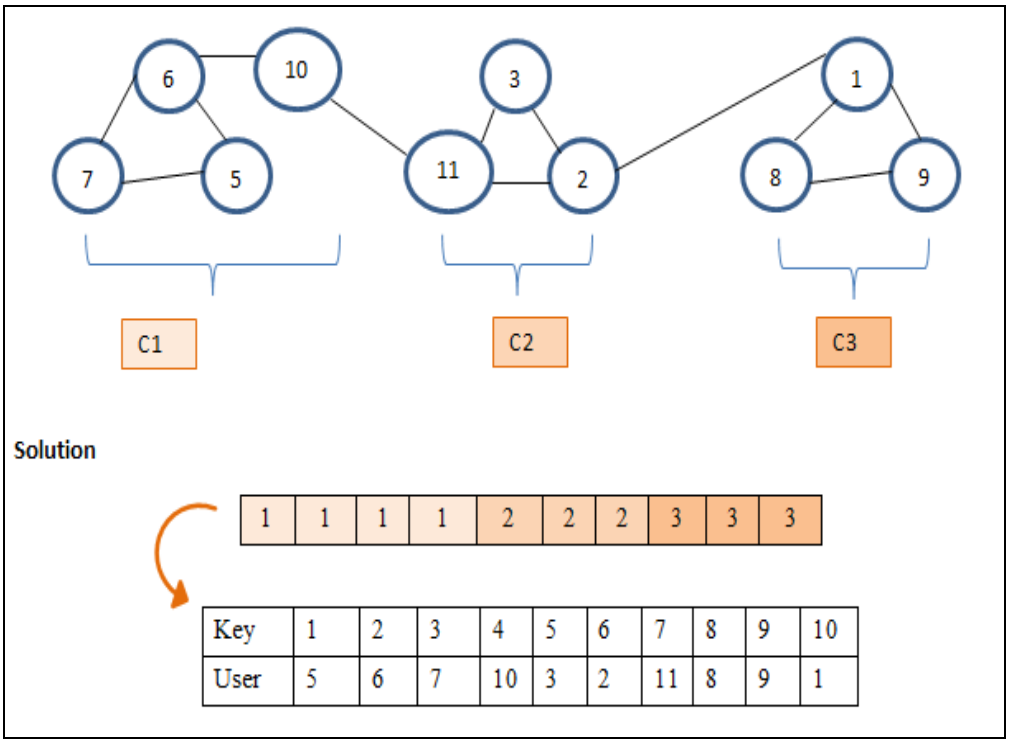


FIG. 1 – *Example of applying the BCO algorithm on a social network.*

The general operating process of Bee Colony Optimization is as follows:

Algorithm2: BCO Algorithm

Begin

Determine the search areas.

For each bee

1. Reach a peak of the search area
2. Perform a local search.
3. Communicate the local optimum (solution returned by the local search)

End.

Initialization of BCO

For the initialization of the BCO algorithm we used the K-means algorithm. For the latter, we give as input a set of vertices and we obtain at the output K communities representing an initial partition that represents an initial solution for the BCO algorithm.

Diversification strategy

In order to determine the different search areas, continuous overlap flip was used. The latter being part of the parameters of the BCO algorithm, allows diversification from a feasible solution. The principle is to change *m* values, where flip = *m*, to the vector representing the solution and to keep the rest of the vector values as they are, starting from the beginning while allowing the overlap.

Illustrative example: Let's consider the flip=3. We consider the following partitioning and we obtain the following areas:

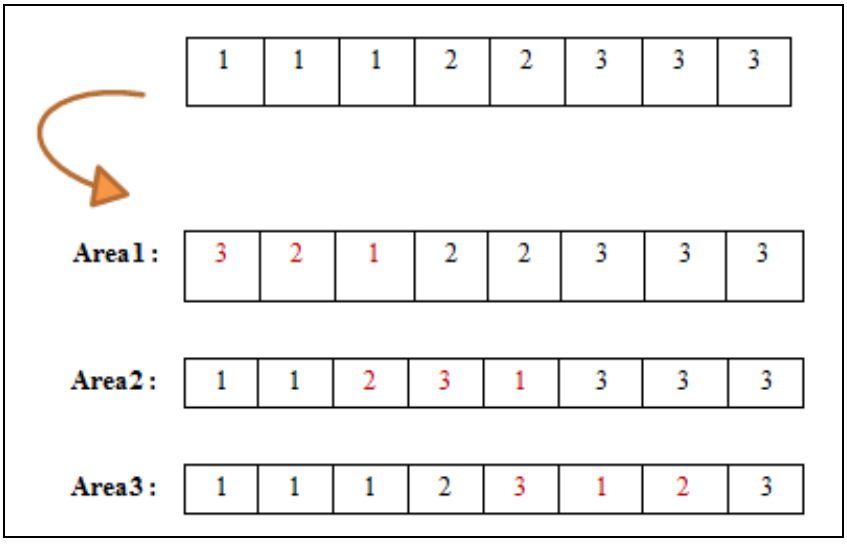


FIG. 2 – *Example of determining search areas using BCO algorithm.*

After generating the search areas, each bee conducts a local search and communicates its local optimum to the other bees. We synthesize the communication process between bees in the following algorithm:

Algorithm 3: BCO Bee-Communication Algorithm
Begin
Best := Best local optimum at time t;
Sref := Best solution at the moment (t-1);
Δf:=Modularity (Best) – Modularity (Sref);
If Δf> 0 then
Sref := Best;
Nb_chance = Max_chance;
If Nb_chance> 0 then
Nb_chance = Nb_chance-1;
Sref = Best;
Else Sref := new solution (generated by K-means);
End If
End If
Nb_chance := Max_chance;
End.

The quality of a solution is expressed using the modularity of Newman. At the end of the process, the solution with the highest value of modularity will be considered as the final solution.

4 Experiments

4.1 Datasets

To evaluate our community detection approach in a social context, we used the following datasets:

- Zahary Karate Club (Zackary, 1977), including 34 members who have relationships with each other. This network has two groups.
- The Lusseau Dolphins network (Lusseau, 2003), containing 62 nodes and 159 links and represents the frequent associations between these dolphins. This network consists essentially of two communities.
- A network of books on American politics (Krebs, 2008), sold by the Amazon online bookstore. It consists of 105 books and 441 links expressing the books purchased together.
- The American Football Network, consisting of 115 nodes and 613 links (Newman, 2005). The nodes represent the teams and the edges the games played between each two teams during the year 2000. This network consists of twelve communities.
- The Rich Epinions Dataset (RED), which contains reviews from users on items, trust values between users, items category, categories hierarchy and users expertise on categories (Simon et al., 2014). We randomly selected two data samples (number of nodes n = 50 and n = 100).

4.2 Evaluation Metrics

To evaluate our approaches, we were particularly interested in the modularity of Newman (Newman and Girvan 2003), widely used in the work on community detection. The idea of this function is that a random graph is not supposed to have a community structure so the existence of communities is revealed by the comparison between the actual density of the edges in a sub-graph and the density that it could have in the sub-graph if the vertices of the graph were attached independently of the community structure.

$$Q = \frac{1}{2m} * \sum_{j,k} \left(a_{V_j,V_k} - \frac{d(V_j)d(V_k)}{2m} \right) * \delta(V_j, V_k), j = 1 \ldots n ; k = 1 \ldots n \qquad (2)$$

where:

- m: number of links in the graph.
- n: number of nodes in the graph.
- *d(vj)*: the number of neighbors of the node v_j
- $\delta(vj, vk)$: is equal to 1 if v_j and v_k belong to the same community, and 0 otherwise.
- $avj\ , vk$: is equal to 1 if the nodes v_j and v_k are linked, and 0 otherwise.

4.3 Experimental Results

4.3.1 Evaluation of the K-means-based classification approach

Table 1 presents the modularity results obtained with the application of the K-means algorithm, where we varied the value of the number of clusters from 2 to 6 for each dataset.

Datasets	**K=2**	**K=3**	**K=4**	**K=5**	**K=6**
Zackary Club Karate (n=34,m=78)	0.27	0.29	0.25	0.24	0.26
Lusseau dolphin (n=62, m=159)	0.38	0.37	0.37	0.35	0.37
Politics Books (n=105, m=441)	0.28	0.27	0.14	0.23	0.24
American Football (n=115, m=615)	0.13	0.12	0.13	0.11	0.13
Red-50 (n=50, m=147)	0.17	0.14	0.20	0.17	0.18
Red-100 (n=100, m=468)	0.11	0.10	0.10	0.13	0.10

TAB. 1 – *Modularity obtained with the K-means Algorithm*

4.3.2 Evaluation of the Tabu Search-based approach

We tested the Tabu Search approach using the result of the K-Means algorithm as the initial solution. We set the value of the number of iterations to the best value found experimentally which is equal to 100. Then we varied the number of clusters. The results obtained are shown in Table 2.

Datasets	K = 3	K = 4	K = 5
Zakary Karate Club (n=34,m=78)	0.170	0.190	0.340
Lusseau Dolphins (n=62, m=159)	0.044	0.210	0.190
Politics Books (n=105, m=441)	0.290	0.100	0.150
Football américain (n=115, m=615)	0.096	0.110	0.170
Red-50 (n=50, m=147)	0.015	0.092	0.120
Red-100 (n=100, m=468)	0.047	0.087	0.051

TAB. 2 – *Modularity obtained with the Tabu Search Algorithm*

4.3.3 Evaluation of the BCO-based approach

The experiment is based on two series of tests that allow the modularity computation according to the number of iterations, the flip and the number of chances.

In the first experiment, we set the number of chances to 3, and we varied the flip from 2 to 5 and the number of iterations from 100,300 to 500. We used the Zackary base to set the best values for the parameters. Table 3 presents the modularity values obtained:

Flip	# Iteration = 100	# Iteration =300	# Iteration =500
2	0.358	0.354	0.216
3	0.371	0.352	0.354
4	0.371	0.355	0.352
5	0.371	0.358	0.371

TAB. 3 – *Modularity with BCO according to the flip and the number of iterations*

From the previous experiment, we can notice that from the value 3 of the flip and for a number of iterations greater than or equal to 100, the algorithm gives optimal values of modularity. In this experiment, the number of iterations is fixed at 100, the flip is varied from 2 to 5 and the number of chances from 3 to 5. The results of this experiment concerning the modularity obtained with the Zackary base are summarized in the following table:

Flip	# Chances = 3	# Chances = 4	# Chances = 5
2	0.354	0.196	0.161
3	0.226	0.371	0.177
4	0.319	0.162	0.196
5	0.187	0.201	0.235

TAB. 4 – *Modularity with BCO according to the flip and the number of chances*

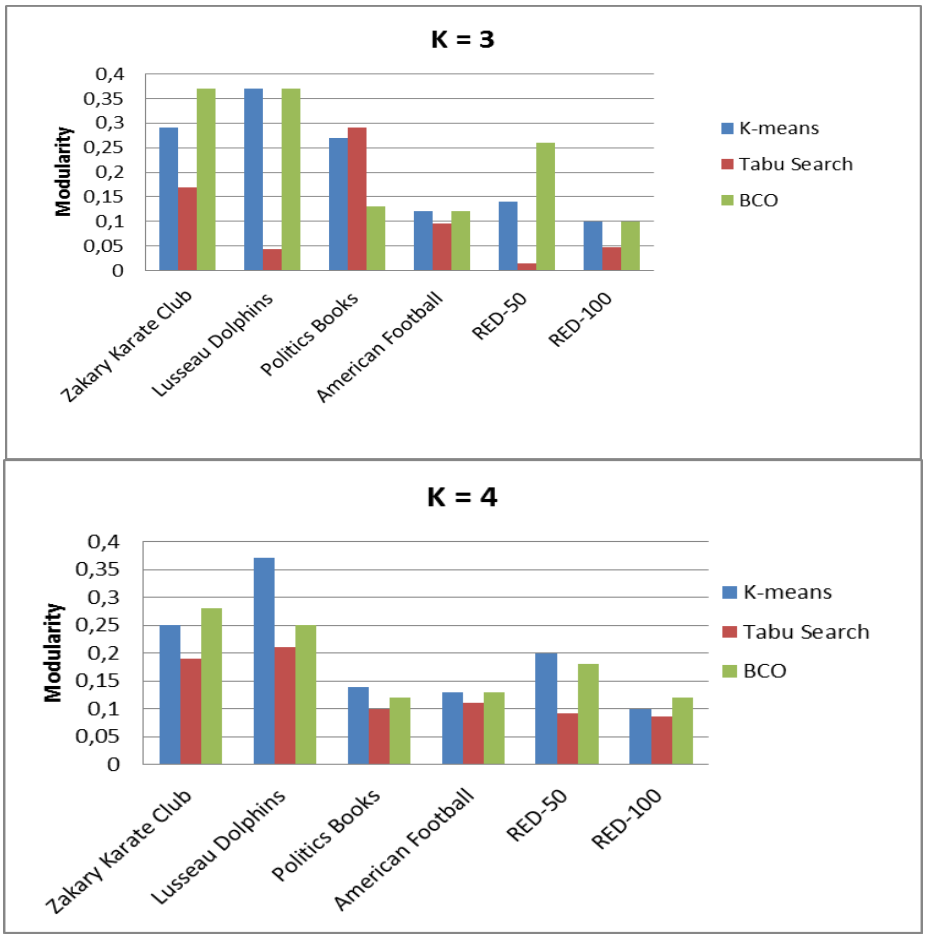
We can see that the best results are obtained for a number of chances between 3 and 4 considering that the best modularity values are obtained for flip values between 3 and 4.

After performing the previous tests, we chose the combination of the parameters of the BCO algorithm that gave the best results: flip = 3, number of iterations = 100, number of chances = 4. The modularity results obtained on the different datasets are shown in Table 5.

Datasets	K=2	K=3	K=4	K=5	K=6
Zakary Karate Club (n=34,m=78)	0.37	0.37	0.28	0.11	0.22
Lusseau Dolphins (n=62, m=159)	0.31	0.37	0.25	0.43	0.37
Politics Books (n=105, m=441)	0.16	0.13	0.12	0.24	0.26
American Football (n=115, m=615)	0.13	0.12	0.13	0.12	0.11
Red50 (n=50, m=147)	0.31	0.26	0.18	0.21	0.27
Red100 (n=100, m=468)	0.11	0.10	0.12	0.10	0.08

TAB. 5 – *Modularity values obtained with BCO algorithm*

As presented in the previous tables and illustrated by the following figure, the results of the experiments show that BCO algorithm has given a better quality of communities partitioning by using several datasets and with a variation of the number of clusters.



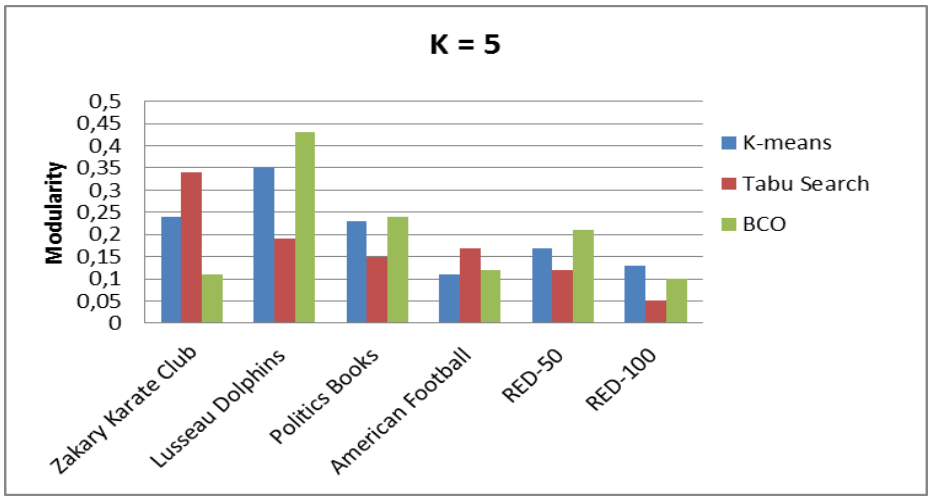


FIG. 3 – *Comparison between the different classification algorithms.*

5 Conclusion

We are interested in this paper to the problem of detection of communities in social networks. Although several works have been proposed in this field, several other tracks remain to be explored. In our approach, we used classification techniques, applying as a first step, the K-means algorithm which constitutes an initial solution. In this algorithm we used the centrality function, one of the functions commonly used in graph-oriented approaches. In order to optimize this classification, two meta-heuristics were used: the Tabu Search and the Bee Colony Optimization algorithm. The results of experiments on well-known datasets in this field show that Bee Colony Optimization has given better performance in terms of modularity.

The perspectives of our work concern mainly, the application of other meta-heuristics including those already used in the literature in order to make further comparisons with the existing related work. Also, it will be necessary to compare our work with some graph-oriented approaches such as the following well-known algorithms: Edge-Betweeness, Label propagation and Fast-Greedy. Finally, we envisage exploring the use of other classification techniques as K-means has the disadvantage of prior identification of the number of clusters

References

Babers, R., Hassanien, A. E. (2017). A Nature-Inspired Metaheuristic Cuckoo Search Algorithm for Community Detection in Social Networks. International Journal of Service Science, Management, Engineering, and Technology (IJSSMET), 8(1): 50-62.

Bedi, P., and Sharma, C. (2016). Community detection in social networks. *WIREs Data Mining and Knowledge Discovery*, 6(3):115–135

Blondel, V., Guillaume, J., Lambiotte, R., Lefebvre, E. (2008). Fast unfolding of communities in larges networks. *Journal of Statically Mechanics: Theory and Experiment.*

Fortunato, S. (2009). Community detection in graphs. Physics Reports, 486(3-5):103

Hopcroft, J., Khan, O., Kulis, B., Selman, B. (2004).Tracking evolving communities in large linked networks. In: the national academy of sciences of the United States of America, 1: 5249–5253.

Newman, M., and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review, 69(2).*

Papadopoulos, S., Kompatsiaris, Y., Vakali, A., and Spyridonos, P. (2011). Community detection in Social Media. *Data Mining and Knowledge Discovery, (June):1–40.*

Plantié, M., and Crampes, M. (2013). Survey on Social Community Detection. *Springer Publishers. Social Media Retrieval, Springer Publishers. Computer Communications and Networks,* 65-85

Sharma, J., Annappa, B. (2016). Community detection using meta-heuristic approach: Bat algorithm variants. *Ninth International Conference on Contemporary Computing (IC3)*

Wang, C., Tang, W., Sun, B., Fang, J., Wang, Y. (2015). Review on community detection algorithms in social networks. *IEEE International Conference on Progress in Informatics and Computing (PIC): 18-20*

Yang, B., Liu, D., Liu, J., and Furht, B. (2010). Discovering communities from Social Networks: *Methodologies and Applications . Springer US, Boston, MA.*

Entrepôt de données NOSQL orienté graphe : Règles de modélisation

Amal Sellami*, Ahlem Nabli**,
Faiez Gargouri***

*MIRACL Laboratory, Faculty of Economics and Management of Sfax, University Sfax, Tunisia,
sellami.amal91@gmail.com
** MIRACL Laboratory, Faculty of computer sciences and information technologies, University Al-Baha, KSA,
ahlem.nabli@fsegs.rnu.tn
***MIRACL Laboratory, Higher Institute of Computer Science and Multimedia of Sfax, University Sfax, Tunisia,
faiez.gargouri@isims.usf.tn

Résumé. Nous assistons aujourd'hui à une croissance importante de données issues des médias sociaux et de l'internet des objets. Face à cette croissance exponentielle de données, les approches d'entrepôt de données (ED) classiques doivent être adaptées et de nombreuses applications web sont orientées vers l'usage des données sous forme de graphes. Une telle stratégie, est sensée fournir plus de solutions et d'outils permettant de gérer avec plus d'efficacité l'énorme volumes de données complexes. Les BD NoSQL offrent des atouts intéressants tels que l'évolutivité et la flexibilité. Ce type de BD constitue une piste intéressante pour la construction des ED capables de supporter des grandes masses de données. Pour pouvoir bénéficier des avantages des systèmes NoSQL, il est nécessaire d'avoir un modèle conceptuel d'ED basée sur le paradigme de graphe. Dans ce contexte, nous proposons de nouvelles règles permettant la modélisation conceptuel d'un schéma d'ED en modèle NoSQL orienté graphe.

Mots clés. Entrepôt de données, règles de modélisation, schéma multidimensionnel, modèle NoSQL orienté-graphe.

1 Introduction

Aujourd'hui, de multiples acteurs de la technologie numérique produisent des quantités infinies de données. Capteurs, réseaux sociaux ou e-commerce, ils génèrent tous de l'information qui s'incrémente en temps réel. En raison de leur importance, les bases de données sont considérées comme le moyen le plus populaire au niveau du stockage, recherche et manipulation de ce type de données. Cette grande masse de données d'information diversifiée appelée Big Data est souvent caractérisée non seulement par la diversité de sa nature (ensemble d'informations hétérogènes) mais aussi apparaissent sous divers formats.

Le Big Data est défini comme étant des collections de données aussi vastes et complexes à gérer. Le stockage et le traitement de ces collections de données à l'aide des outils de gestion

comme les bases de données classiques est devenu de nos jours très difficile. De même, les bases de données relationnelles, qui ont été le support parfait de stockage de données pendant de nombreuses décennies ne sont plus adaptées à ce phénomène. Pour pallier ces difficultés, les bases de données NoSQL (Not Only SQL) sont de plus en plus envisagées. En effet, ce nouveau type de base de données, permet d'exploiter de nouvelles approches pour implanter les bases de données et, en particulier, les entrepôts de données.

Dans ce contexte, les médias sociaux et l'émergence de Facebook, LinkedIn et Twitter ont accéléré l'émergence des bases de données NoSQL et en particulier celles orientée graphe. Ces dernières représentent le format de base avec lequel les données dans les différents médias sont stockées. De nombreuses applications web s'orientent, de nos jours, vers l'usage des données sous forme de graphes. Une telle stratégie, nécessite des outils spécifiques pour la gestion et l'analyse d'une quantité énorme de données hautement connectées (Par exemple, Open Street Map, FlockDB la base de données graphiques interne de Twitter et TAO le projet de Facebook). Vu la complexité de la structure de graphe que peut engendrer ces sources de données ainsi que leur volume (variété et vélocité) l'automatisation du processus de construction d'ED s'impose. De surcroit, si on prend en considération les atouts promis par les BD NOSQL orientée graphe en terme de flexibilité analytique et afin de faciliter la recherche (sans utilisation des jointures), il est important de choisir ce type de base comme une structure de stockage pour les ED. Par conséquent, il est nécessaire aussi d'avoir un modèle conceptuel pour la modélisation des entrepôts de données. Ce modèle facilitera l'implantation d'ED NOSQL orienté graphe.

C'est dans ce cadre que s'inscrivent les travaux présentés dans cet article. En effet, notre premier objectif est de proposer une modélisation conceptuelle du schéma d'ED selon le modèle NoSQL orienté graphe.

La suite de l'article est structurée comme suit. La section 2 expose un état de l'art sur la transformation du schéma d'ED en modèle NoSQL. La section 3 argumente, d'abord, le choix du modèle NoSQL orienté graphe, ensuite, présente la formalisation de ses notions de bases. Dans la section 4, nous présentons une formalisation du schéma multidimensionnel. La section 5, quant à elle, propose des règles de modélisation des concepts multidimensionnels selon le formalisme orienté graphe. Enfin, la section 6 conclut le papier et trace quelques perspectives.

2 Etat de l'art

En 2014, une première approche a été proposée pour coupler le modèle orienté graphe et l'OLAP Castelltort et Laurent (2014). Dans cette approche les auteurs proposent de structurer les données dans le système NoSQL orienté graphes Neo4J. Ils présentent deux formalismes pour représenter le fait et les dimensions au niveau du modèle logique orienté graphes. Le formalisme assure deux types de relations, celles liant le fait aux dimensions, et celles reliant les attributs des dimensions entre eux. Ces dernières relations permettent de préserver la relation hiérarchique. L'approche de Dehdouh et al. (2014) propose des méthodes d'implantation des cubes OLAP dans le modèle NoSQL en colonnes. Les auteurs ont développé un banc d'essai décisionnel en NoSQL colonnes basé sur SSB (Star Schema Benchmark). Pour représenter les tables de faits et des dimensions dans un système NoSQL en colonnes, les auteurs ont proposé, dans Dehdouh et al. (2015), trois approches à savoir : NLA (Normalized Logical

Approach) où les tables de faits et des dimensions sont stockées séparément sur différentes tables. Ainsi, on note, d'abord, l'utilisation du DLA (Denormalized Logical Approach) qui est un processus favorisant la dénormalisation du schéma conceptuel dimensionnel et regroupant les faits et les dimensions dans une table unique appelée BigFact Table, où chaque famille de colonnes est composée d'un seul attribut. Ensuite, le DLA-CF (Denormalized Logical Approach by using Column Family) qui est un processus permettant d'encapsuler les tables de faits et des dimensions dans une même table d'une manière à ce que chacune devient une famille de colonnes.

Dans les travaux de Chevalier et al. (2015), les auteurs ont présenté une approche basée sur des règles de transformation d'un modèle conceptuel multidimensionnel en un modèle logique NoSQL orienté colonnes ou documents. Ils proposent alors trois modèles pour implémenter un entrepôt de donnée dans HBase. Dans Scabora et al. (2016), les auteurs s'orientent vers un modèle de données NoSQL orienté colonnes pour résoudre le problème de distribution des attributs entre les familles de colonnes afin d'optimiser les requêtes et faciliter la gestion des données. Quant aux travaux de Freitas et al. (2016), proposent une approche complète R2NoSQL pour transformer un modèle conceptuel d'un entrepôt de données en étoile dans une base de données NoSQL pouvant être soit orientée colonnes, soit orientée documents.

Récemment les travaux de Yangui et al. (2016) proposent deux approches automatique pour transformer le schéma multidimensionnel en étoile dans deux modèles orientés soit colonnes soit documents. La première approche propose des règles de transformations dites simples où la hiérarchie des attributs des dimensions n'est pas considérée. Dans les secondes approches les auteurs proposent une politique de nommage pour retrouver la relation hiérarchique des attributs. Les deux approches ont été évaluées avec des requêtes simples. Elles nécessitent aussi une expérimentation avec des requêtes OLAP plus complexes (drill down, roll up). Cependant, ces deux approches ne prennent pas en considération la construction du treillis d'agrégats. D'autres travaux étudient les processus de traduction de schémas conceptuels en NoSQL. A titre d'exemples, les travaux de Abdelhédi et al. (2016) définissent des processus de traduction de diagramme de classes UML en NoSQL orienté colonnes avec HBase. Les travaux de Raouf et Anne (2016) proposent une transformation du RDF ou modèle NoSQL orienté graphe en proposant un ensemble de règles de transformation.

D'une manière générale, la majorité de ces travaux s'inscrivent dans le cadre de transformation du schéma d'ED dans le modèle NoSQL orienté colonnes ou orienté documents. Ils proposent tous une dénormalisation complète des données relationnelles dans la base de données NoSQL cible. Il est à noter que les modèles orientés colonnes et documents partagent l'inconvénient majeur des BD relationnels en l'occurrence le problème de jointure. Un tel problème doit être mis en évidence surtout lors de l'interrogation.

Par ailleurs, peu de travaux se sont intéressés au modèle orienté graphe supportant les requêtes complexes sans passer par l'utilisation des jointures.

Partant de ces constats, nous avons opté pour un processus d'implémentation d'un entrepôt de données multidimensionnelles en se basant sur les modèles NoSQL orientés graphe. Nous proposons ainsi dans cet article de définir, des règles permettant la représentation d'un schéma conceptuel d'ED en modèle NoSQL orienté graphe.

3 Modèle NoSQL orienté graphe

L'émergence des systèmes NoSQL permettent d'envisager de nouvelles approches pour implanter un entrepôt de données. Dans cette section, nous argumentons le choix du modèle NoSQL graphe, puis nous proposons une formalisation du modèle choisit.

3.1 Choix du modèle

Les systèmes NoSQL orientés graphes ont montré leurs avantages par rapport aux systèmes relationnels en termes de flexibilité et de gestion de données massives. Particulièrement, les bases de données orientées graphes sont conçues pour résoudre des problèmes très complexes qu'une base de données relationnelle serait incapable de le faire. Les réseaux sociaux (Facebook, Twitter, etc..), caractérisés par des millions d'utilisateurs constituent un bon exemple : amis, fans, famille etc. Le défi ici n'est pas le nombre d'élément à gérer, mais le nombre de relations qu'il peut y avoir entre tous ces éléments. En effet, il y a potentiellement n^2 relations à stocker pour n éléments. Même s'il existe des solutions comme les jointures, qui sont trop couteuses, les bases de données relationnelles se confrontent très vite aussi bien à des problèmes de performances que des problèmes de complexité dans l'élaboration des requêtes.

Les bases de données graphes sont également utilisées pour la gestion d'importantes structures informatiques. Elles permettent également l'élaboration de liens entre les divers intérêts que pourraient avoir un internaute. Le but étant de pouvoir lui proposer des produits susceptibles de l'intéresser. Ainsi, les publicités s'affichant sur Facebook sont très souvent en relation avec les recherches effectuées sur Google. Les propositions d'achats de sites de vente en ligne tels que EBay et Amazon, quant à elles, sont en relation avec des achats déjà effectués. Comme son nom l'indique, ces bases de données reposent sur la théorie des graphes. Avec cette théorie, il serait possible de sauvegarder les données sous forme de graphes. Hyper Graph DB, Neo4j, FlockDB, Big Data sont des exemples de bases de données orientées graphe.

Dans cet article une attention particulière est accordé au système le plus connu et le plus répandu à savoir le Neo4J. Ce dernier est largement étudié par (Holzschuher et Peinl, 2013) et (Vukotic et al., 2015). Neo4j conserve certaines caractéristiques des systèmes relationnels (transactions) mais avec une nouvelle philosophie de structuration des données. Basé sur la théorie des graphes, il bénéficie également d'un espace de stockage extensible et tolérant aux pannes. Neo4j est très populaire grâce à sa structure reposant sur deux objets fondamentaux : les nœuds et les relations. Il ne permet pas de découper les données et en faire des partitions mais en revanche il respecte les propriétés ACID. De ce fait, il permet le respect de l'intégrité des données via des fonctions de verrou lors des modifications des données. Il assure aussi la disponibilité des données grâce à la réplication des données et sa capacité à agir rapidement pour désigner un nouveau maître en cas de panne.
Un point positif et important d'une telle de base est qu'elle soit parfaitement adaptée à la gestion des données relationnelles même dans un contexte de « Big Data ». De plus avec une architecture modelable, elle peut être adaptée selon les besoins rencontrés. Ainsi, elle peut fournir des performances rapides pour les requêtes analytiques Emanuela et Hristo (2016). Toute ces raisons motivent notre choix pour la construction d'un entrepôt de données NoSQL orienté graphe.

3.2 Formalisation du modèle

Le modèle orienté graphes repose sur quatre notions ; nœud, relation label et propriété. Chaque nœud possède des propriétés et des labels. Les relations relient les nœuds et possèdent éventuellement des propriétés et un type. La FIG. 1 montre un exemple d'une BD orientée graphe.

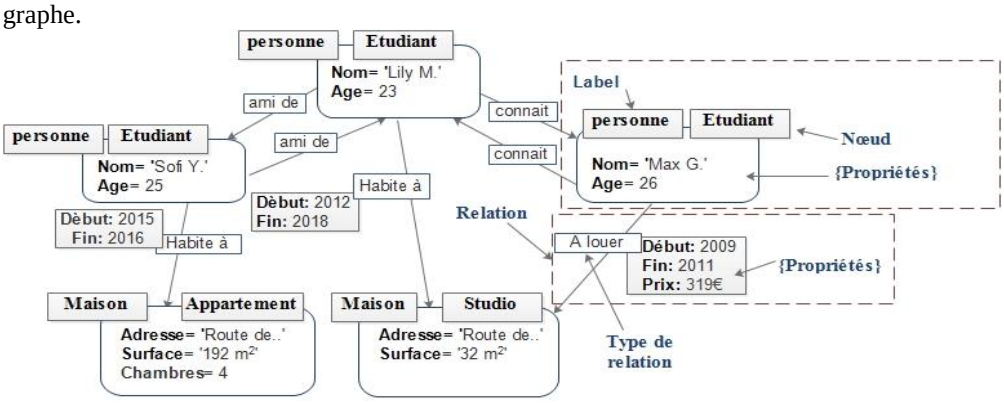


FIG.1 –*Modèle d'une BD NoSQL orientée graphe.*

Une ***base de données orientée graphe***, notée $\mathbf{BD}^{OG}$ est définie par (V^{OG}, L^{OG}, E^{OG}, P^{OG}) où :

- V^{OG}= {V_1,…,V_n} est l'ensemble des nœuds qui représentent les entités,
- L^{OG}= {L_1,…,L_z} est l'ensemble des labels attribués au nœud,
- E^{OG}= {E_1,…,E_m} est l'ensemble des arêtes qui représentent les relations entre les nœuds (i.e. un sous-ensemble de V× V),
- P : l'ensemble des propriétés attribuées à chaque composant de la base de données orientée graphe (nœud/arc).

Un ***nœud***, notée $\mathbf{V}\in V^{OG}$, est défini par (id^V, P^V, L^V) où :

- id^V est l'identifiant du nœud,
- P^V est l'ensemble des propriétés qui décrivent un nœud ($P^V \subset P^{OG}$),
- L^V est l'ensemble des étiquettes ou labels attachés au nœud($L^V \subset L^{OG}$).

Une ***relation***, notée **R,** est définie par (id^R, Vi^R, Vo^R, T^R, P^R) où :

- id^R est un identifiant,
- Ni^R est l'identifiant du nœud entrant,
- No^R est l'identifiant du nœud sortant,
- T^R est le type de relation qui porte le nom de la relation,
- P^R est un ensemble de propriétés d'une relation.

4 Formalisation du schéma multidimensionnel

La modélisation conceptuelle est la phase de fondation nécessaire. Dans cette phase, les entrepôts de données sont modélisés de manière multidimensionnelle. Les concepts de base de

la modélisation multidimensionnelle sont : les faits, les mesures, les dimensions et les hiérarchies (cf. FIG. 2).

- ***Un fait*** : modélise le sujet de l'analyse, et formé de mesures correspondant aux informations de l'activité analysée. Ces mesures sont numériques et généralement valorisées de façon continue. On peut donc les additionner, les dénombrer ou bien calculer le minimum, le maximum ou la moyenne.
- ***Une dimension*** : modélise un axe d'analyse et se compose de paramètres correspondant aux informations faisant varier les mesures de l'activité.
- ***Une hiérarchie de paramètres d'une dimension :*** définie des niveaux de détail de l'analyse sur cette dimension.

Un schéma multidimensionnel peut être décrit comme suit :

Un ***schéma multidimensionnel***, noté **E**, est défini par (F^E, D^E, Fonc) où :

- $F^E=\{F_1,...,F_n\}$ est un ensemble fini de faits,
- $D^E=\{D_1,\ldots,D_m\}$ est un ensemble fini de dimensions,
- Fonc est une fonction qui associe à chaque fait de F^E un ensemble de dimensions qui peuvent être utilisées pour analyser le fait (lien Fait-Dimension).

Un ***fait***, noté $\mathbf{F}\in F^E$, est défini par (N^F, M^F) où :

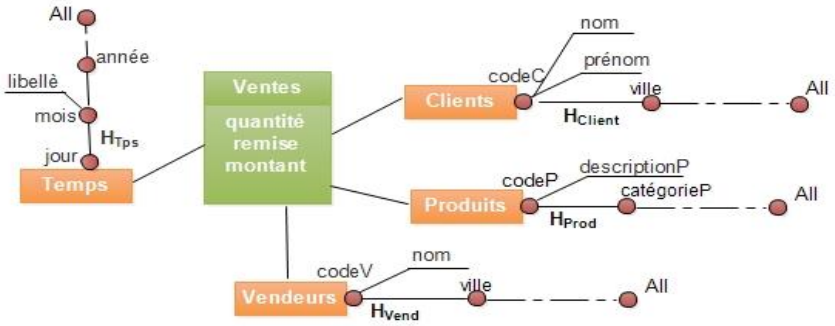
- N^F est le nom du fait,
- M^F est un ensemble de mesures dont chacune est associée avec une fonction d'agrégation.

Une ***dimension***, noté $\mathbf{D}_i\in D^E$, est définie par (N^D, id^D, H^D) où :

- N^D est le nom de la dimension,
- id^D est l'identifiant de la dimension,
- $H^D=\{H_1,...,H_K\}$ est un ensemble de hiérarchies.

Une ***hiérarchie***, noté $\mathbf{H}_i\in H^D$, est définie par (N^{Hi}, $Param^{Hi}$, Pred, $Weak^P$) où :

- N^{Hi} est le nom de la hiérarchie,
- $Param^{Hi}$ est un ensemble de paramètres,
- Pred est une fonction associant à chaque paramètre son prédécesseur,
- $Weak^{Hi}$ est une fonction associant à chaque paramètre d'éventuelles informations complémentaires appelées attributs faibles.



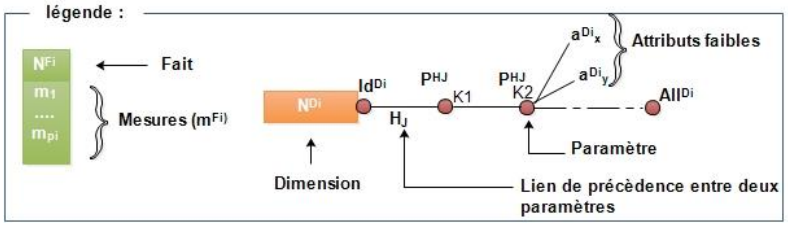


FIG. 2 –*Exemple de modèle conceptuel multidimensionnel.*

5 Règles de modélisation du schéma multidimensionnel orienté graphe

Nous proposons dans cette section des règles de modélisation relatives aux concepts d'un entrepôt de donnée basé sur le paradigme graphique. Ces règles sont en nombre de quatre à savoir *Modélisation de fait et mesures***,** *Modélisation du nom et identifiant d'une dimension, Modélisation de hiérarchies et Modélisation du lien Fait-Dimension.*

Modélisation de fait et mesures. Chaque *fait* est représenté par un nœud. Le label du nœud prend le nom du concept du modèle multidimensionnel qui est le *fait* puis on accorde le nom du fait comme étant un deuxième label au même nœud. Chaque mesure est représentée par une propriété du nœud **(**cf. FIG.3).

Règle 1. Chaque fait $F \in F^E$ est représenté par un nœud V (id^V, P^V, L^V) Où :

- Le label $l_1$est type de concept multidimensionnelle : l_1='Fait' / L^V= $\{l_1\}$
- Le label $l_2$est le nom du fait: l_2=N^F / L^V=$L^V \cup \{l_2\}$
- Chaque mesure $m_i \in M^F$ est représenté par une propriété *p* avec $p \leftarrow m_i$ / P^V= $P^V \cup \{p\}$

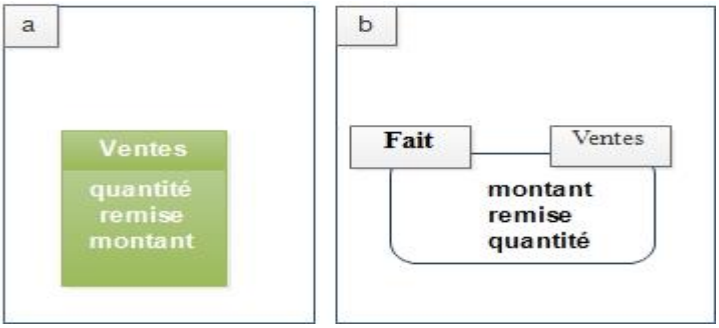


FIG.3 –*a est un exemple du fait du MCM, b la modélisation du concept fait selon le modèle NoSQL orienté graphe.*

Modélisation du nom et identifiant d'une dimension D (N^D, a_i). Chaque dimension est représentée par un nœud. Le label du nœud prend le nom du concept du modèle multidimensionnel qui est *dimension*. Puis on accorde le nom de la dimension comme étant un deuxième label au même nœud. Ensuite, l'identifiant est représenté par une propriété du nœud. En fin, tout attribut faible associé à l'identifiant est représenté par une propriété dans le même nœud **(**cf. FIG.4).

Règle 2. Chaque nom et identifiant d'une dimension sont représentés par un nœud V (id^V, P^V, L^V), avec :

- Le label $l_1$est type de concept multidimensionnelle avecl_1='Dimension' / L^V= {l_1}
- Le label $l_2$est le nom de la dimension avecl_2=N^D / L^V=L^V∪ {l_2}
- L'identifiant a_i, modélisé par une propriété p avec p←a_i / P^V= P^V∪{p}
- Chaque attributs faible a_f associé à a_i est représenté par une propriété p avec p←a_f / P^V= P^V∪{p}

Règle 3. Modélisation de hiérarchies H^D : une hiérarchie se compose d'un ensemble de paramètres et d'un lien de précédence entre ses paramètres. Chaque paramètre est représenté par un nœud. Le label du nœud prend le nom du concept du modèle multidimensionnel qui est *paramètre*. On accorde, ensuite, le nom du paramètre comme étant un deuxième label au même nœud. Puis, chaque attribut faible est représenté dans le nœud sous forme de propriété. En fin, chaque lien de précédence est représenté par une relation **(**cf. FIG.4).

Règle3.1. Modélisation d'un paramètre

Chaque paramètre a_i est représenté par un nœud V (id^V, P^V, L^V). Où :

- Le label $l_1$est type de concept multidimensionnelle : l_1='Paramètre' / L^V= {l_1}
- Le label $l_2$est le nom du paramètre : l_2=a_i / L^V=L^V∪ {l_2}
- Chaque attribut faible a_f associé à a_i, s'il existe, est représenté par une propriété p avec p←a_f / P^V= P^V∪{p}

Règle3.2. Modélisation du lien de précédence entre paramètres a_i→a_{i-1}

Chaque a_i→a_{i-1}⊂ H^D est représenté par un relation **R** définie par (id^R, Vi^R, Vo^R, T^R) Où :

- Vi^R est le nœud représentant a_i
- Vo^R est le nœud représentant a_{i-1}
- Le type$t_1$est le nom de la relation : t_1='Précède' / T^R= {t_1}

La figure 4 montre un exemple de modélisation d'une dimension par l'application des règles 2 et 3.

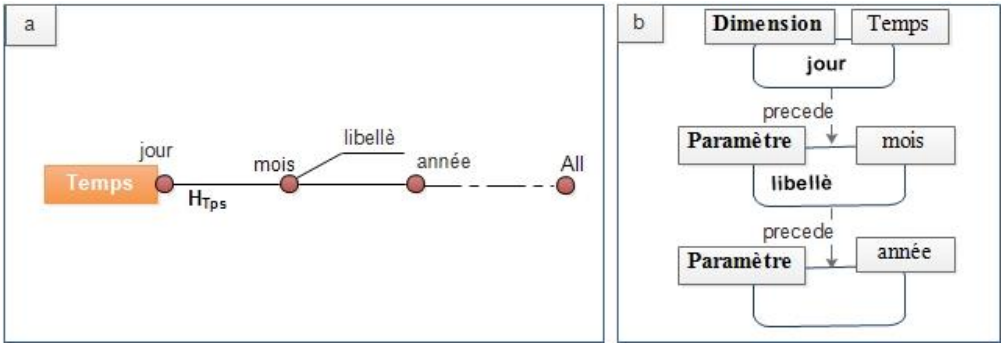


FIG. 4 –*a est un exemple de la dimension du MCM, b la modélisation du concept dimension selon le modèle NoSQL orienté graphe.*

Modélisation du lien Fait-Dimension. Chaque lien entre fait et dimension est représenté sous forme d'une relation ayant comme nœud source un nœud modélisant un fait et comme nœud destination un nœud modélisant une dimension. La relation prend le nom 'Lien Fait-Dimension' (cf. FIG. 5).

Règle 4. Chaque lien *Fait-Dimension* est représenté par un relation R définie par (id^R, Vi^R, Vo^R, T^R) **(**cf. FIG.5) Où :

- Vi^R est le nœud représentant *le fait*
- Vo^R est le nœud représentant *la dimension*
- Le type t_1 est le nom de la relation : t_1='Lien Fait-Dimension' / T^R= {t_1}

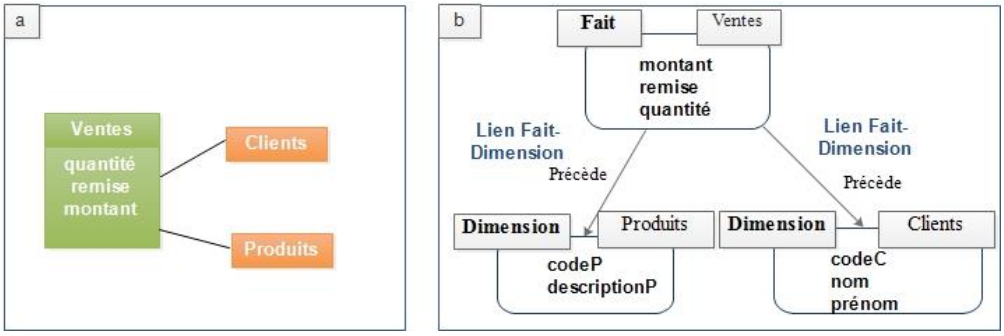


FIG. 5–a *est un exemple de modélisation du lien Fait-Dimension du MCM, b la modélisation du lien fait-dimension selon le modèle NoSQL orienté graphe.*

Les règles de modélisation proposées peuvent être résumées dans le tableau de correspondance suivant :

Schéma multidimensionnel	modèle orienté graphe
Fait	Nœud
Mesure	Propriété dans un nœud
Dimension	Nœud
Paramètre	Nœud
Attribut Faible	Propriété dans un nœud
Lien de précédence $a_i \rightarrow a_{i-1}$	Relation
Lien Fait-Dimension	Relation

TAB. 1 – *Mappage entre les composants du schéma multidimensionnel et le modèle orienté graphe*

L'application des règles de modélisation proposées sur le schéma conceptuel de la figure 2 fournit le modèle conceptuel du schéma d'entrepôts de données selon le formalisme graphique illustré par la FIG.6.

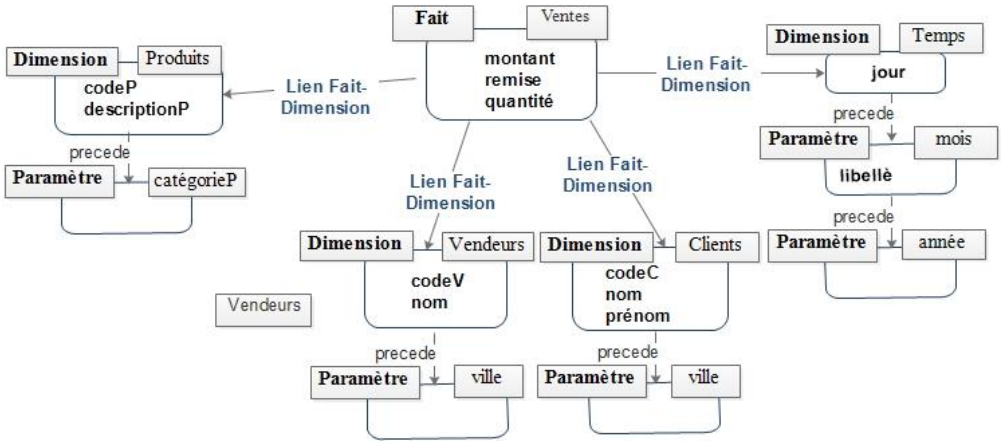


FIG. 6–*Schéma multidimensionnel selon le formalisme orienté graphe.*

6 Conclusion

L'avènement du Big Data a souligné la difficulté du modèle relationnel à traiter de gros volumes de données. Il a permis le développement de nouvelles technologies reposant sur une architecture à la fois décentralisée et extensible à travers l'utilisation des systèmes NoSQL. Dans cet article, nous avons proposé un ensemble de règles permettant la représentation d'un schéma conceptuel d'ED en modèle NoSQL orienté graphe. En termes de perspectives nous visons, en premier lieu de proposer des règles de transformation du modèle conceptuel au modèle logique orienté graphe. Ensuite, nous nous focalisons sur l'automatisation de création d'un entrepôt de données NoSQL orientée graphe.

Références

AleksaVukotic, Nicki Watt, Tareq Abedrabbo, Dominic Fox, and Jonas Partner (2015). Neo4j in Action, Manning.

Arnaud Castelltort et Anne Laurent (2014).NoSQL Graph based OLAP Analysis: In SCITEPRESS - Science and and Technology Publications, 217–224.

Emanuela Mitreva et Hristo Kyurkchiev (2016). Performance Study of SQL and Graph Solutions for Analytical Loads. Conference: Information Systems and Grid Technologies, At Sofia.

Fatma Abdelhédi, Amal AitBrahim, Faten Atigui, and Gilles Zurfluh (2016). Processus de transformation MDA d'un schéma conceptuel de données en un schéma logique NoSQL.

Florian Holzschuher et René Peinl (2013). Performance of graph query languages: comparison of cypher, gremlin and native access in Neo4j. Proceedings of the Joint EDBT/ICDT 2013 Workshops. ACM, 2013. p. 195-204.

K. Dehdouh, O. Boussaid, F. Bentayeb (2014). Columnar NoSQL Star Schema Benchmark. International Conference on Model and Data Engineering. Springer, Cham, 2014. p. 281-288.

K. Dehdouh, O. Boussaid, F. Bentayeb (2015). Using the column oriented NoSQL model for implementing big data warehouses. Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2015. p. 469.

Lucas C. Scabora, Jaqueline J. Brito, Ricardo Rodrigues Ciferri, and Cristina Dutra de AguiarCiferri (2016). Physical Data Warehouse Design on NoSQL Databases - OLAP Query Processing over HBase. International Conference on Enterprise Information Systems, XVIII. Institute for Systems and Technologies of Information, Control and Communication-INSTICC, 2016.

M. Chevalier, M. El Malki, A. Kopliku, O.Teste, T. Tournier(2015). Implementing Multidimensional Data Warehouses into NoSQL. International Conference on Enterprise Information Systems.

M. Chevalier, M. El Malki, A. Kopliku, O. Teste, R. Tournier(2015). Implementation of MultidimensionalDatabases in Column-Oriented NoSQL Systems. In : East European Conference on Advances in Databases and Information Systems. Springer, Cham, 2015. p. 79-91.

M. Chevalier, M. El Malki, A., O. Teste, R. Tournier(2015). Implementation of Multidimensional Data bases with Document-Oriented NoSQL. International Conference on Big Data Analytics and Knowledge Discovery. Springer, Cham, 2015. p. 379-390.

Myller Claudino de Freitas, Damires Yluska Souza, and Ana Carolina Salgado (2016). Conceptual Mappings to Convert Relational into NoSQL Databases: In SCITEPRESS - Science and and Technology Publications.

Rania Yangui, Ahlem Nabli, and Faiez Gargouri (2016). Automatic Transformation of Data Warehouse Schema to NoSQL Data Base: Comparative Study. Procedia Computer Science, 2016, vol. 96, p. 255-264.

Raouf Bouhali, Anne Laurent (2016). Exploiting RDF Open Data Using NoSQL Graph Databases. In : IFIP International Conference on Artificial Intelligence Applications and Innovations. Springer, Cham, 2015. p. 177-190.

Summary

With the increase amount of data in the different social media and the Internet of Things, data warehouse approaches need to be adapted. Recently, many web applications are moving towards the use of data in the form of graphs. Indeed, these databases (DB) provide very effective tools for managing huge volumes of complex data. Generally, classic relational DBs are inadequate when dealing with massive data. As a result, NoSQL DBs have interesting assets such as scalability and flexibility. These type of DBs are an interesting way to build EDs that can support large amounts of data. In order to benefit from the advantages of NoSQL systems, it is necessary to have a conceptual model of ED based on the graph paradigm. In this article, we propose a new rules allowing the modeling of a conceptual schema of the data warehouse in NoSQL graph oriented model.

Vers une architecture intégrée pour la gestion des données spatiales massives en télécommunications

El Hassane Nassif*, Hajji Hicham*,
Reda Yaagoubi*, Hassan Badir**

nassif.hassane@gmail.com, h.hajji@iav.ac.ma, r.yaagoubi@iav.ac.ma
*Ecole ESGIT, IAV H2
Rabat

hbadir@uae.ac.ma
**Ecole ENSAT TANGER
Tanger
hbadir@uae.ac.ma

Résumé. La gestion des données spatiales dans les télécoms est devenue dernièrement un vrai challenge pour les décideurs et les utilisateurs. Bien que des solutions commerciales existent pour la gestion de ces données à travers des architectures basées sur les modèles relationnels, il est à remarquer que ces approches existantes ne répondent pas efficacement aux différentes contraintes que posent les données spatiales en télécoms, notamment en termes de volumétrie, de vitesse et de complexité. En effet, le décideur télécom a actuellement besoin de plateforme technique lui permettant de mieux ingérer, stocker et interroger les données spatiales issues des différentes traces clients résultantes des activités: voix, messages courts et internet. Nous présentons dans notre papier une architecture Big Data intégrée pour la gestion des données spatiales massives en télécom. Également, nous traiterons certains aspects importants comme l'indexation de ces données spatiales et leur traitement en étendant le framework SparkSQL.

1 Introduction

Nous assistons depuis quelques années à une démocratisation des offres télécoms et à une montée en force de l'usage des smartphones et des réseaux sociaux. Ces deux phénomènes ont conduit à des changements profonds dans le comportement des abonnés. Par conséquent, nous observons aujourd'hui une explosion du volume, de la vitesse et de la variété des données clients recueillies par les opérateurs de télécommunications notamment les données spatiales vectorielles issues de la géolocalisation des abonnées. Cette nouvelle tendance met à la disposition de l'opérateur une mine de données à référence spatiale émanant de l'activité de ses abonnés. Elle lui offre de nouvelles opportunités pour se différencier des concurrents. En même temps, elle représente un défi technologique. En effet, l'exploitation des données spatiales massives à travers les architectures classiques basées sur les modèles relationnels se retrouve rapidement dépassée par les contraintes liées aux données massives (volume, vitesse, variété).
Si les plateformes Big Data peuvent prendre en charge de façon efficiente les données massives, arrivant à grande vitesse et sous divers formats, elles restent néanmoins inadaptées pour la gestion des données à référence spatiales. En effet, la prise en compte de l'attribut de

localisation s'avère plus complexe pourvu que l'on veuille l'utiliser dans sa structure métrique. Même si, à titre d'exemple, rien n'empêche d'utiliser des structures classiques pour modéliser des points (x,y) dans une solution Big Data. Mais on sera dans l'incapacité d'effectuer des tests d'appartenance, répondre à une requête sur la distance entre des objets géolocalisés ou simplement reconstituer un contour.

Plusieurs initiatives de recherche (Olasz et al. 2016) ont prouvé qu'il est possible d'étendre les solutions BigData afin qu'elles puissent prendre en charge la composante spatiale et améliorer de façon significative la performance des traitements. Malheureusement, l'amélioration des délais de traitements ne suffit pas si l'on a des latences dans la collecte, le stockage ou la lecture de ce type de données surtout dans des cas d'utilisations nécessitant le temps réel ou presque. Cela signifie qu'il est nécessaire d'optimiser, de bout en bout, la chaine Big Data spatiale et pas uniquement la couche de traitement.

Dans ce travail, nous nous basons sur un cas pratique issu du domaine Télécom et nous présentons une architecture intégrée reposant sur un ensemble de technologies Big Data variées afin de garantir une efficience au niveau des différentes couches applicatives, incluant la collecte des données spatiales vectorielles, leur ingestion, leur stockage et puis leur traitement. Et puisque la modélisation des données est nécessaire au développement des solutions informatiques de toute nature, Big Data spatial y compris, nous avons conçu une entité spatiale de base que nous avons appelée GeoActivity pour nous permettre de manipuler les données spatiales issues des données brutes Télécoms.

2 Travaux liés

Le tableau TAB. 1 liste des travaux de recherches qui se sont penchés sur la résolution des problématiques liées à l'exploitation des données spatiales dans un environnement Big Data.

<table>
<tr><td rowspan="2">Initiatives de recherches</td><td rowspan="2">Solutions associées</td><td colspan="3">Couches applicatives</td></tr>
<tr><td>Collecte et ingestion</td><td>Stockage</td><td>Traitements et requêtes spatiales</td></tr>
<tr><td>Eldawy A. et Mokbel F.2013</td><td>SpatialHadoop</td><td rowspan="6">Non prévues</td><td>HDFS (système de fichiers distribués de hadoop)</td><td>Hadoop/ MapReduce</td></tr>
<tr><td>Nishimura S. et al.2011</td><td>MD-Hbase</td><td>HBASE (base de données orientée colonne)</td><td rowspan="2">Fonctions Hive personnalisées</td></tr>
<tr><td>Aji A. et al.2013</td><td>Hadoop-GIS</td><td>HIVE (entrepôt de donnée intégrée sur Hadoop)</td></tr>
<tr><td>Tang M et al.2016</td><td>LocationSpark</td><td rowspan="3">Non prévu</td><td>Spark /RDD</td></tr>
<tr><td>Huang.Z et al.2017</td><td>GeoSpark</td><td rowspan="2">Spark/RDD, DataFrame et SparkSQL</td></tr>
<tr><td>Xie D. et al.2016</td><td>Simba</td></tr>
</table>

TAB. 1 –*Initiatives de recherche dans le big data spatial.*

Ces travaux traitent uniquement les deux couches : stockage et traitement et ne propose pas une architecture intégrée. En outre, ils se basent principalement sur Hadoop et Spark. Ces deux plateformes implémentent le même paradigme de calcul parallèle MapReduce (Jeffrey et Sanja, 2004) mais de deux manières différentes. La première se base sur les accès au disque à chaque fois qu'il y a besoin de manipuler la donnée, ce qui détériore les performances surtout dans le cas des traitements itératifs. Ceci n'est pas le cas pour Spark, car il privilégie les traitements en mémoire tout en limitant l'accès au disque. D'après ses concepteurs (Matei Z. et al. ,2012), la force de Spark réside dans son abstraction de base, les RDDs (Resilient Distributed Datasets). Ce sont des collections distribuées d'objets avec les caractéristiques suivantes :

- Elles ne se calculent pas de façon automatique, car chaque RDD dispose d'un graphe orienté acyclique qui permet de le générer en cas de besoin.
- Elles sont partitionnées sur les différents nœuds du cluster Spark.
- Elles tolèrent des pannes, car elles disposent de toutes les informations pour recalculer les partitions perdues suite à un problème sur un des nœuds Spark.
- Elles peuvent-être persistées pour une utilisation ultérieure si besoin.

En 2014, Spark avait battu le record précédemment détenu par Hadoop en triant 100 To de données en 23 minutes[1]. Ce qui explique que les travaux cités dans le tableau TAB.1 et ayant utilisé Spark mettent en avant leur performance face à ceux basés sur Hadoop. Toutefois il y a besoin de préciser les RDDs sont complexes à manier et à maintenir. Cela est dû au

[1] http://sortbenchmark.org/

langage bas niveau offert par Spark pour manipuler les RDDs, ce qui conduit généralement à des problèmes de performance. Depuis quelque temps, Spark a évolué pour intégrer deux nouvelles abstractions DataFrame et DataSet, manipulables à travers un langage relationnel de haut niveau permettant à l'utilisateur final de s'affranchir de la complexité et du langage bas niveau de manipulation des RDDs. Ainsi, cet utilisateur pourra se concentrer non pas sur la façon avec laquelle il va dérouler l'enchainement des traitements et calculs sur Spark, mais plutôt sur la logique métier et le résultat qu'il souhaite atteindre. Un des exemples de plate-forme Big Data Spatial qui a utilisé cette abstraction est le projet SIMBA (Xie D. et al.2016). Bien que ce dernier a proposé une couche de traitement spatial en utilisant SparkSql[2] (Armbrust, M. et al.(2015)), nous rappelons qu'il est nécessaire dans notre cas de mettre en place une architecture qui prend en charge la donnée ainsi que les traitements de bout en bout depuis la collecte de la donnée spatiale brute jusqu'à la mise à disposition des résultats à l'utilisateur final.

3 Architecture intégrée pour les données spatiales massives télécoms.

3.1 Provenance des données spatiales massives et leurs caractérisations

Le schéma FIG 1, illustre la façon avec laquelle l'information spatiale massive est produite dans les télécoms.

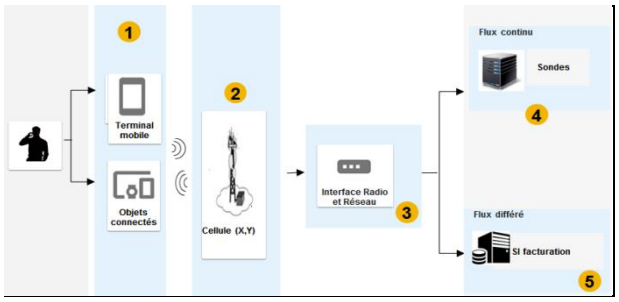


FIG. *1: Échange de signalisation entre le mobile et le réseau télécom.*

La cellule[3] (FIG.1-composante 2) garantit l'accès aux services télécoms pour les clients mobiles à travers des échanges de signalisation avec leurs terminaux mobiles (FIG.1-composante 1). De façon générale, toutes les activités du mobile qu'elles soient de type : appel, message court ou navigation ou juste le fait d'allumer son mobile sont acheminées en temps réels et en différé vers l'infrastructure informatique de l'opérateur (composantes 4 et 5 dans le schéma) à travers des protocoles de signalisation bien définis. De plus, chaque signalisation contient par construction l'identifiant de la cellule ayant servi de relais et le timestamp de génération. Ainsi, il suffit de connaitre la position géographique de cellule pour déduire celle de l'abonné.

[2] https://spark.apache.org/sql/
[3] Il s'agit d'une infrastructure télécom qui rattache le terminal du client au réseau de son opérateur.

3.2 Architecture intégrée et cheminement des données spatiales massives.

La figure 2, illustre le cheminement des données spatiales télécoms issues des activités des abonnés et leur passage par les différentes couches prévues dans notre architecture.

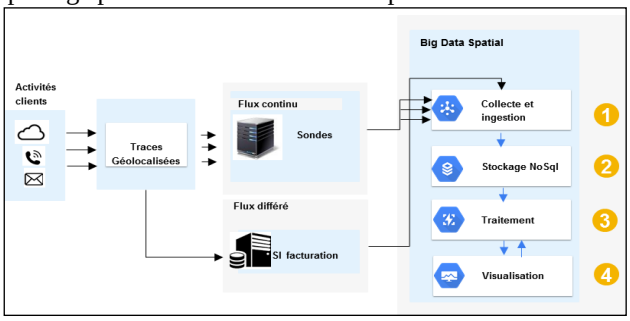


FIG. *2: cheminement des données spatiales dans les télécommunications.*

Ainsi, l'étape 1 est dédiée à la collecte et ingestion des données brutes, l'étape 2 est dédié au stockage dans une base NoSql[4], alors que l'étape 3 est dédiée aux traitements nécessaires avant la mise à disposition des résultats au niveau de l'étape quatre. Dans ce qui suit, nous détaillons les quatre couches en justifiant nos choix.

3.3 Couche de collecte et ingestion des données spatiales

Les procédures d'échanges de signalisation établies en continu entre le terminal mobile et les équipements réseau permettent à l'opérateur de géolocaliser les activités de ses abonnés. Ainsi, pour structurer ces données de signalisation, nous proposons une entité spatiale de référence qu'on a appelée ***GeoActivity***. Cette entité permet de renseigner sur l'événement client, sa date et puis sa localisation ainsi que d'autres caractéristiques.

Cette structure sera détaillée par la suite et sera alimentée par plusieurs données brutes provenant de sources hétérogènes. Ces sources sont caractérisées par leur énorme volume de données générées et la variété de leurs formats. Pour illustrer cette structure, nous avons identifié deux types majeurs de ces sources de données et qui sont les suivants :

- Type 1 : Les données arrivent en mode flux continu et en temps réel depuis les sondes (FIG1- composante 4).
- Type 2 : Les données arrivant en mode batch tel que les fichiers issus des partages FTP et les bases de données relationnelles de l'opérateur (FIG1- composante 5).

Étant donné que la collecte est la première étape dans la chaine de valorisation des données spatiales massives télécoms, il est nécessaire de sélectionner l'outillage qui répond le mieux aux deux types de sources précitées. La figure FIG3 récapitule la méthode adoptée pour la collecte et l'ingestion des données spatiales massives télécoms.

[4] Une nouvelle génération de bases de données qui ne se base pas sur le modèle relationnel.

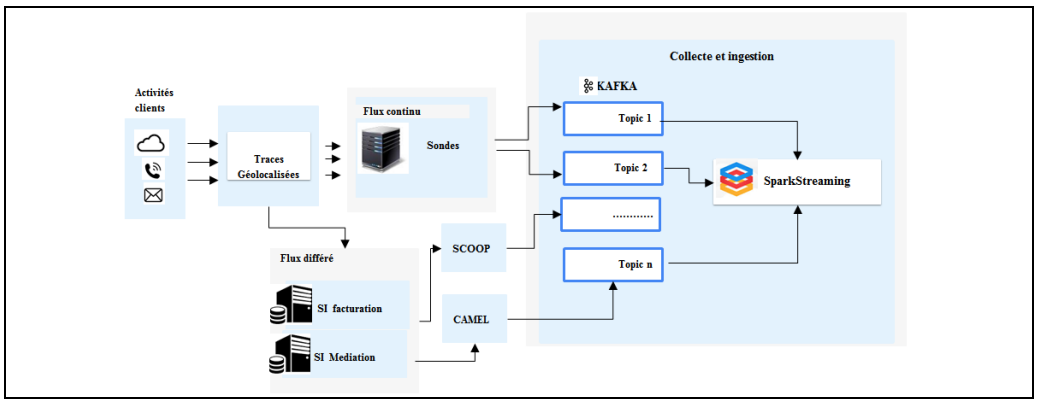


FIG. *3: Collecte et ingestion des données spatiales massives télécoms.*

Bien que SparkStreaming soit capable de collecter les flux de données brutes qui arrivent en continu en les accumulant pendant un laps de temps configurable selon les cas d'utilisations, nous avons opté pour Kafka (https://kafka.apache.org/). Il s'agit d'un système distribué de publication de messages, son choix est motivé par notre volonté de découpler l'ingestion de la collecte. De plus, il permet de faire persister les données de tous types, de cette façon on pourra l'utiliser comme entrepôt intermédiaire afin d'enrichir les données de type 1 provenant des sondes par les données de type 2 (FIG1- composant 4). L'enrichissement consiste en des compléments de données qui renseignent sur le cout de l'évènement client tel que le montant de l'appel par exemple. À noter aussi que Kafka est capable de monter en charge facilement afin de supporter des débits très importants. Aussi, il fonctionne avec une logique de publication-souscription, qui permet d'échanger les données entre les sources de données (émetteurs) et la brique d'ingestion formée par le couple SparkStreaming (https://spark.apache.org/streaming) et SparkSql (consommateurs). Chaque émetteur Type 1 (deux sondes par ville dans notre cas) envoie ses données sous forme de message dans des Topics Kafka spécifiques : voix, sms, data. Une fois reçues, Kafka ajoute le message le plus récent à la fin du Topic choisi par l'émetteur et forme une sorte de séquence ordonnée non modifiable et partitionnée sur plusieurs nœuds. De la même façon, les émetteurs de type 2 utilisent Camel (http://camel.apache.org/) pour collecter les fichiers depuis les partages et Scoop (http://sqoop.apache.org) pour extraire les données depuis les bases de données relationnelles pour ensuite les injecter dans les topics Kafka via des connecteurs spécifiques. Toutes ces données sont ensuite mises à la disposition de l'unique consommateur : Spark Streaming pour pouvoir les ingérer avant le stockage final avec la structure GeoActivity.

3.4 Couche de stockage des données spatiales

Les technologies NoSql ont prouvé leur capacité de traiter des données massives arrivant à haut débit et sous divers formats. Dans notre travail, nous avons opté pour les bases de données NoSql. Ce type de format doit aussi prendre en charge les contraintes de nos données

telles que le volume et la vitesse avec laquelle elles sont générées. D'après les travaux de benchmarking de (EndPoint.2015) portant sur la performance des bases de données NoSql, Cassandra est la meilleure base de données orientée colonne dans les cas d'utilisations qui font le mixe entre le mode transactionnel et le mode analytique, ce qui correspond exactement à notre cas. Le défi avec Cassandra c'est qu'elle ne supporte pas la structure spatiale, donc pour stocker nos données en préservant l'information spatiale il faudra étendre les types de base. En effet, le type de base le plus à même de remplir cette fonction est le type JSON, puisqu'il est déjà supporté au niveau de Cassandra. Nous avons donc opté pour une extension GeoJson en créant deux nouveaux types de données : Geometry (FIG4-etape1) et Simple Feature Properties (FIG4-etape2) qui nous servent dans la définition de notre entité spatiale GeoActivity (FIG4-etape3).

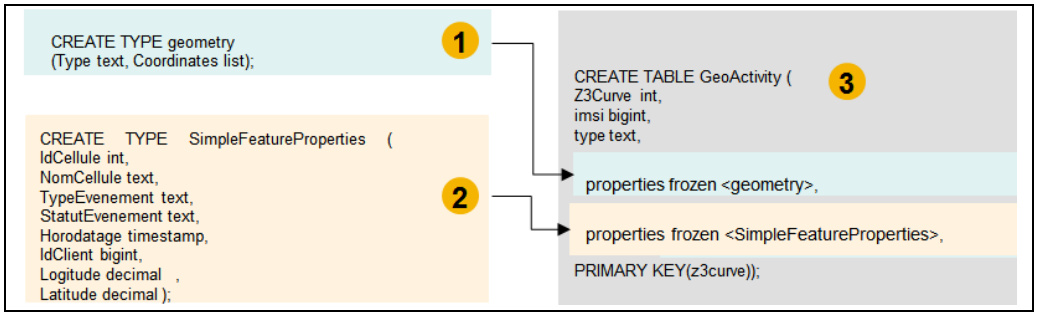


FIG. 4: Modèle logique de la structure spatiale télécom dans Cassandra

Cependant, il ne faut pas oublier que la structure spatiale génère une complexité additionnelle dans les traitements vu qu'elle est multidimensionnelle (au moins deux dimensions). Pour cette raison, nous avons décidé de simplifier le problème en le réduisant à une dimension à travers un procédé d'indexation. Ceci consiste à appliquer une fonction de hachage qui transforme les données multidimensionnelles en des données à dimension unique en préservant la localité spatiale. Parmi les techniques d'indexation qui existent, nous avons choisi les courbes qui remplissent l'espace en l'occurrence la technique Z Order Curve capable de prendre en charge les espaces de deux et trois dimensions. D'après ses auteurs (BÖxhm C. et al. 1999) son principe est de parcourir l'espace en formant une courbe Z et de mettre chaque point sur une ligne tout en préservant l'ordre et la localité. La préservation de la localité est souhaitée dans de nombreuses applications. Dans notre cas, elle nous permettra d'éviter au maximum les latences des accès disques durs et des communications réseau en gardant les données proches géographiquement sur le même serveur. Le figure 5 illustre le model physique de stockage de notre entité GeoActivity au niveau de Cassandra et met en évidence la structure clé/valeur.

KeySpace Cassandra		
GeoActivity (Famille colonne)		
Clé (RowKey)	Nom colonne	Valeur colonne
Z3Curve(x,y,t)	Imsi	Valeur encodée
	Type	
	Geometry	
	Properties	

FIG. 5 – *Modèle physique de l'entité GeoActivity*

L'index Z3curve qui est le résultat de la réduction des trois dimensions : x, y , t (temps) en une seule valeur que nous utilisons comme clé primaire et en même temps comme clé de partitionnement. Quand les données sont lues ou écrites, une fonction appelée « Partitioneur » est utilisée pour calculer la valeur de hachage de la clé de partition. Cette valeur de hachage est utilisée pour déterminer le nœud et la partition qui pourront stocker l'enregistrement en question. Le challenge avec le partitionnement sur Cassandra, et les bases NoSql en général, c'est de faire en sorte à ce que l'attribution des données aux différent serveurs soit exécuter de façon équilibrée pour pouvoir tirer profit du parallélisme. Nous n'avons pas détaillé cet aspect dans cet article, mais il est à noter qu'il existe plusieurs travaux se sont penchés sur le partitionnement dynamique et équilibré dans les bases de données Nosql à travers la mise à disposition d'un mécanisme générique distribué (Konstantinou I. et al.2013). Ce mécanisme peut être installé au-dessus de tout type de base de données NOSQL et permet de réorganiser de façon optimale et dynamique les partitions.

Maintenant que notre structure de stockage GeoActivity est créée sur Cassandra, il ne reste plus qu'à l'alimenter. Comme indiqué dans la partie 3.3, Kafka permet de mettre à disposition les évènements clients qui arrivent en continu pour qu'ils soient transformés et traités par le moteur de traitement Spark. Ce dernier effectue trois opérations principales : d'abord il calcule la clé Z3curve sur la base du triplet (x, y, t) et puis transforme le schéma des données en Geojson pour les injecter à la fin dans Cassandra à travers un connecteur spécifique qui supporte tous les types de celle-ci. L'injection des données se fait à travers un simple Insert Into Json comme décrit dans la figure 6 où nous retrouvons nos deux types de données déjà créés Geometry (1) et Simple Feature Properties (2).

```
INSERT into GeoActivity JSON '{
"z3curve": 37,
"imsi": 60499909808988,
"type": "SimpleFeature",
1 "SimpleFeatureProperties": {
"IdCellule": 12456890,
"NomCellule": "casa123",
"TypeEvenement": "voix",
"StatutEvenement": "Success",
"Horodatage":201801211545663,
"IdClient": 60499909808988,
"Logitude": 3,
"Latitude": 4
},
2 "geometry": {
"type": "Point",
"coordinates": [3, 4]
}
}'
```

FIG. *6: Alimentation de la table GeoActivity*

3.5 Couche de traitement et d'analyses spatiales

Afin de répondre à un maximum de cas d'utilisation des données spatiales dans les télécoms nous proposons une architecture qui permet de concilier entre les deux types de traitements : par lots et en quasi temps réel. Le premier type s'exécute sur des données d'historique et avec des fréquences assez larges : heure, jour, semaine, mois. Le deuxième type s'exécute sur des données récentes au fur et à mesure qu'elles sont acquises par les couches en amont : collecte, ingestion et stockage avec une mise à disposition immédiate à l'utilisateur.

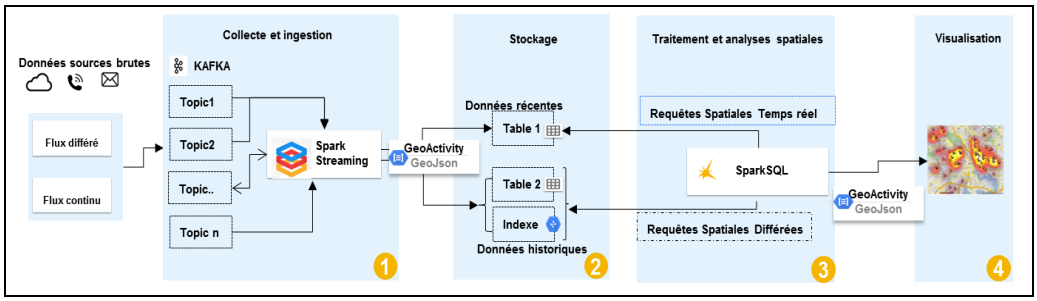


FIG. *7: Flux de données*

Comme l'indique la figure 7, les données sont collectées au fur et à mesure qu'elles arrivent par la couche 1 : collecte et ingestion. Puis, elles sont nettoyées[5], enrichies et transformées par le moteur SparkSteaming. Ce dernier les transforme en un schéma GeoJson pour correspondre à notre entité de stockage GeoActivity. Ensuite, les données sont stockées doublement dans la couche numéro deux, une première fois dans la table 1 pour satisfaire aux requêtes temps réel et une deuxième fois dans la table 2 qui regroupe tout l'historique pour des requêtes spatiales nécessitant plus de temps de calcul.

La couche de traitement (étape 3 FIG.8) est responsable de l'exécution des requêtes spatiales sur les données brutes et leur transformation en information. Ces données sont extraites depuis Cassandra pour répondre aux requêtes et interrogations des utilisateurs tout en assurant une certaine interactivité.

Afin d'implémenter les requêtes spatiales, nous avons utilisé le module SparkSQL (Armbrust, M. et al. (2015)) de Spark. Son avantage est de permettre à l'utilisateur de créer des requêtes analytiques complexes à l'aide du langage SQL et de les exécuter de façon parallélisée sur des données massives partitionnées entre plusieurs serveurs. À noter aussi qu'il a la capacité de faire un maximum de traitements en mémoire, tout en offrant la possibilité d'interroger n'importe quels types de données de la même façon qu'on interroge des tables. SparkSQL offre donc une implémentation distribuée du langage SQL basée sur une connaissance préalable du schéma des données à travers les DataFrames. Contrairement aux RDDs,

5 Le nettoyage que nous avons appliqué dans ce travail consiste à appliquer des règles basiques consistant à rejeter les valeurs manquantes au niveau des champs obligatoires tels que la date de l'événement, la localisation ou le numéro de téléphone par exemple. Une présentation plus détaillée de cette partie est prévue dans le cadre d'un autre article.

les DataFrames sont une collection de données distribuées ayant un schéma et organisées en des colonnes nommées. Quand le type de données est connu, nous pouvons faire évoluer nos DataFrame vers une autre abstraction DataSet. Son principal avantage est d'offrir plus de sécurité aux développeurs à travers un typage statique au moment de la compilation. Enfin, le fait de travailler avec les DataFrames ou les DataSets permet de bénéficier d'un plan d'exécution optimisée. Ceci est possible grâce à la connaissance de la structure de données par Spark, ce qui lui permet de sélectionner le plan d'exécution le plus adapté. Même si SparkSql ne reconnait pas la structure spatiale, il offre néanmoins la possibilité de définir des types de données personnalisés. La figure 8 présente un exemple d'implémentation écrit en scala pour étendre sparkSQL afin de supporter le type Geometry :

```
1 private case class Geometry (Type :String, x: Double, y: Double)

2 private class GeometryUDT extends UserDefinedType[Geometry] {
  override def sqlType: DataType = ArrayType(DoubleType, containsNull = false)
  override def serialize(obj: Any): ArrayData = {
   obj match {
   case features: Geometry => new GenericArrayData(Array(features.Type, features.x, features.y))
   } }  override def deserialize(datum: Any): Geometry = {
   datum match {
   case data: ArrayData if data.numElements() == 3 => {
   val arr = data.toDoubleArray()
   new Geometry(arr(0), arr(1),arr(2))  }  } }
   override def userClass: Class[Geometry] = classOf[Geometry]
   override def asNullable: GeometryUDT = this}
```

FIG. *8: Extension de spark par de nouveaux types spatiales*

De la même façon on a défini le deuxième type SimpleFeatureProperties précédemment présenté sur la figure 4. Par la suite, nous avons créé notre DataSet Geoactivity qui servira à charger les données depuis nos tables Cassandra. La figure 9 illustre la création de DataSet (1), la liaison avec les tables spatiales dans Cassandra pour charger les données (2) et l'interrogation de ces données avec du langage SQL (3):

```
1 case class GeoActivity (
   Z3Curve: Long,
   Imsi: Integer,
   Type: String,
   Properties: GeometryUDT,
   SimpleFeatureProperties: SimpleFeaturePropertiesUDT)

2 val CreateCassandraView = """CREATE TEMPORARY VIEW GeoActivityForSql
          USING org.apache.spark.sql.cassandra
                OPTIONS  (
                  table  "GeoActivity",
                  keyspace  "TelecomDataBase") """
  spark.sql (CreateCassandraView)


3 spark.sql ("select * from GeoActivityForSql").show
```

FIG. *9: Creation du DataSet Geoactivity dans spark et alimentation depuis Cassandra*

Une fois notre structure de données spatiales créée et qu'on a pu la charger avec les données, il reste à l'exploiter à travers la mise en place de fonctions personnalisées. En effet, SparkSql offre la possibilité d'encapsuler toute notre logique spatiale dans des fonctions que nous pouvons appeler dans des requêtes SQL depuis Spark. La figure 10 illustre un exemple où nous créons une nouvelle fonction qui permet de vérifier si l'abonné a fait une activité dans une zone particulière. Pour simplifier cet exemple, nous avons choisi un stade et nous supposons qu'il a une géométrie de type cercle. Dans l'étape 1, nous créons la fonction, ensuite nous la déclarons pour qu'elle soit reconnue par SparkSql. À la fin nous l'utilisons

dans notre requête SparkSql afin d'identifier tous les clients qui ont une activité dans le stade « StadeFoot ».

```
1 val EventDansCercle(A: GeoActivity, C: Cercle): Boolean = {
  LessThanOrEqual((Math.pow((A.getX() - C.getXCentre()), 2) + Math.pow((A.getY() - C.getYCentre()), 2)),
  Math.pow(C.getRayon, 2))
   }
2 spark.udf.register("EventDansCercle", EventDansCercle)

3 spark.sql ("select GeoActivityForSql.imsi from GeoActivityForSql where EventDansCercle
  (GeoActivityForSql.Geometry, StadeFoot)=true").show
```

FIG. *10: Exemple d'implémentation d'une fonction personnalisée sous SparkSql*

4 Conclusion

Les cas d'utilisations des données spatiales massives, notamment en temps réel, nécessitent la mise en place en place d'une architecture intégrée capable de gérer ce type de données de bout en bout. Contrairement aux autres travaux de recherches (TAB.1), nous avons présenté dans cet article un cas pratique issu du domaine Télécom pour expliquer comment on peut profiter des technologies Big Data pour assurer l'efficience des différentes couches applicatives (collecte, ingestion, stockage et puis traitement).

Références

Aji, A. et al. (2013). *Hadoop gis: A high performance spatial data warehousing system over mapreduce*. Proceedings of the VLDB Endowment, 2013, vol. 6, no 11, p. 1009-1020.

Armbrust, M. et al. (2015). *Spark sql: Relational data processing in spark.* In Proceedings of the ACM SIGMOD International Conference on Management of Data (pp. 1383-1394). ACM.

BÖxhm, C. et al. (1999). *Xz-ordering: A space-filling curve for objects with spatial extension.* In International Symposium on Spatial Databases. Springer, Heidelberg. p. 75-90.

Eldawy, A. et Mokbel F. (2013*), A demonstration of spatialhadoop: An efficient mapreduce framework for spatial data*. In Proceedings of the VLDB Endowment, vol. 6, no 12, p. 1230-1233.

EndPoint. (2015). *Benchmarking Top NoSQL Databases*. https://www.endpoint.com

Jeffrey D. et and Sanjay G. (2004). *MapReduce: Simplified Data Processing on Large Clusters*. Google.

Huang, Z. et al. (2017). *GeoSpark SQL: An Effective Framework Enabling Spatial Queries on Spark*. ISPRS International Journal of Geo-Information, vol. 6, no 9, p. 285.

Konstantinou, I. et al. (2013). *DBalancer: Distributed Load Balancing for NoSQL Datastores.* In Proceedings of the ACM SIGMOD International Conference on Management of Data. ACM, p. 1037-1040.

Matei, Z. et al. (2012). *Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing.* In proceedings of the 9th USENIX conference on Networked Systems Design and Implementation. USENIX Association, p. 2-2

Nishimura, S. et al. (2011). *A Scalable Multi-dimensional Data Infrastructure for Location Aware Services.* In Mobile Data Management (MDM), 12th IEEE International Conference on. IEEE, p. 7-16.

Olasz, A et al. (2016). *Geospatial Big Data processing in an open source distributed computing environment.* PeerJ Preprints, vol. 4, p. e2226v1.

Tang, M. et al. (2016) *LocationSpark: A Distributed In-Memory Data Management System for Big Spatial Data.* Proceedings of the VLDB Endowment, vol. 9, no 13, p. 1565-1568.

Xie, D. et al. *(2016) Simba: Efficient In-Memory Spatial Analytics.* In Proceedings of the 2016 International Conference on Management of Data. ACM, p. 1071-1085.

Summary. Spatial data management in Telecom has recently become a real challenge for decision-makers and users. Although commercial solutions exist for managing these data through GIS and Database architectures, it should be noted that these existing approaches do not respond effectively to the various constraints posed by spatial data in telecoms, particularly in terms of volume, speed and complexity. Indeed, the telecom user currently needs a technical platform allowing him to better ingest, store and query the spatial data from the various customer traces resulting from the activities: voice, short messages and internet. We present in our paper an integrated Big Data architecture for spatial data management in telecom. We also covered some important aspects such as indexing this spatial data and processing it by extending the SparkSQL frame-work.

OLAPing Reflexive Multidimensional Fact

Maha BEN KRAIEM*, Jamel FEKI*, **, Ahmed ALGHAMDI**, Franck RAVAT ***

* University of Sfax, MIRACL Laboratory,
Airport Road Km 4, P.O.Box. 1088; 3018 Sfax, Tunisia
Maha.BenKraiem@yahoo.com
** University of Jeddah, FCIT, IS dept
Saudi Arabia
{Jfeki, ahmedg}@uj.edu.sa
*** IRIT, University of Toulouse,
118, Route de Narbonne, 31069 Toulouse Cedex 9, France
Franck.Ravat@irit.fr

Abstract. The multidimensional data model and implementations of social networks come with a set of specific constraints, such as missing data, reflexive relationship on fact instance. However, the conventional OLAP operators and existing models do not provide solutions for handling those specificities. Therefore, further efforts should be invested to extend these operators to take into consideration the specificities of multidimensional modeling of tweets and their manipulation. Face to this issue, we propose, in this paper, two new OLAP operators that enhance existing solutions for OLAP analyses involving a reflexive relationship on the fact instances. For each OLAP operator, we suggest a user-oriented definition as an algebraic formalization, along with an implementation algorithmic.

1 Introduction

The data warehouse has been the backbone of decision support systems for more than two decades and widely accepted and used across the globe in a variety of applications. Contributions of the research community in the data warehousing field, complimented by advancement in the relevant hardware technology, have matured these systems in managing huge volumes of data and providing their access with matchless efficiency to applications and decision-makers. On-Line Analytical Processing (OLAP) is at the core of data warehouse systems enabling multidimensional analysis of warehoused data.

Social media is yet another interesting area producing large data volumes that fascinate the attention of research and business communities. There is growing interests in gaining insights to the way social networks operate, their users behave, engage in conversations, express their opinions and influence others. This involves performing aggregations across conventional and unconventional dimensions in social media data.

Furthermore, businesses can largely benefit from this new resource and market of social media, provided that the underlying technology and systems of data warehousing can deal

with the challenges of heterogeneous data (i.e., semi-structured data) and the speed at which the data originate from social media.

In previous work, we have applied the data warehousing technology to enable comprehensive analysis of massive data volumes generated by the *Twitter* social network. More accurately, we have proposed a multidimensional model dedicated to the OLAP of data exchanged through tweets (Ben Kraiem et al. (2015)). This model takes into account the specificities of data issued from tweets. Among these specificities, we can find links between tweets and their answers' tweets. Regarding this new issue, we have extended the concept of fact by the proposal of a new relationship between fact instances called *reflexive relationship*. This fact-to-fact reflexive relationship allows connecting an instance of the fact to one or several instances of the same fact. Based on this relationship, the fact instances are linked at many successive levels.

Naturally, the concept of levels between fact instances is a novel proposal for which the conventional analysis tools are not designed for. Therefore, we need new OLAP operators to manipulate such a reflexive relationship.

In this paper, we define two new OLAP operators called *FDrilldown* and *FRollup*. They allow navigating down and up through the implicit hierarchical levels of the *fact*; this represents the first step to detect strong connections between fact instances and, therefore discover interesting or amazing topics and then conduct much more deep analysis of such data sets.

We have opted for the following organization of this paper. Section 2 studies representative works related to the OLAP operators that addressed the analysis of facts. Section 3 introduces our motivation example and context. Section 4 proposes two new operators called *FDrilldown* and *FRollup* for fact drilling. For each operator, we formalize it as an algebraic definition and develop an algorithm to implement it. Section 5 provides experimental results and assessments on the efficiency of our proposed OLAP operators.

2 Related works

To the best of our knowledge, no solution for OLAP analysis is proposed for *Drilling down* and *up* on the fact on the multidimensional schema. Only few *querying* operators on fact (Drill-Across, FRotate) are formally proposed in Abelló et al. (2002) and Ravat et al. (2008).

Drill-Across operator relates information contained in two multidimensional facts having the same dimensions. According to Kimball and Ross. (2002), Drill-Across can only be applied when both cubes have the same schema dimensions and the same instances. Other authors relax this restriction. Abelló et al. (2002, 2003) define the Drill-Across as changing a currently analyzed subject F1 (fact) with another fact F2 while keeping the same analysis space (current dimensions). The authors have identified semantic relationships between dimensions and facts: Derivation, Generalization, Association and Flow to extend possibilities to drill across. These relationships between dimensions and facts improve the conformity between attributes and could be used to navigate or Drill-across between Star schemas, even when dimensions are not shared. Cabbibo and Torlone (2004) define drill across as an extension to the natural join where the intersection of the two dimensions is aggregated at the finest grain of the dimensions. Furthermore, Riazati et al. (2008) propose extending the navigation operation drill across to include the non-conformed dimensions.

The FRotate operator in Ravat et al. (2008) consists in using a new fact in the multidimensional table while preserving the characteristics of the current analysis axes. The new fact must share at least the two current (i.e., displayed) dimensions with the current fact. Note that the fact rotation operation, noted FROTATE, is equivalent to the Drill-Across operation Abelló et al. (2003).

According to this study, we may conclude that none of these works offers tools for the decision-maker to navigate (Drilling -down, and -up) through the fact. So far, the OLAP frameworks lack the ability to cope with this problem.

To alleviate this drawback, our proposed OLAP operators namely *FDrilldown* and *FRollup* go further according to a new *Reflexive* relationship on the fact instances. These new operators allow modifying the analysis level in a fact while keeping the same analysis context(i.e., without changing the dimensions for the currently analyzed fact). Hence, data analysts would benefit greatly from the ability to navigate and view combined multidimensional data from multiple levels of fact.

3 Motivation example

Referring to our multidimensional model dedicated to the On-Line Analytical Processing (OLAP) of data exchanged through tweets, our motivation example is built upon the '*Tweet Constellation*' proposed in Ben Kraiem et al. (2014). This model mainly consists of a set of two facts namely *FACTIVITY-TWITTOS* which corresponds to an observation on user accounts and allows the analysis of the user activity over time, and *FACTIVITY-TWEET* fact, which is a reflexive fact. It models links between a tweet and the person concerned by the answer (answered person) and then allows participants and other readers to easily follow the exchange of tweets. Being reflexive, *FACTIVITY-TWEET* allows interconnecting instances of the same fact hierarchically. In practice, if a tweet *tr* is a reply to tweet *t*, *tr* refers *t* (it contains the ID of tweet *t)*. This reflexive relationship between tweets will guarantee that every tweet response inserted to the data warehouse corresponds to an existing tweet so that the analysis of a set of linked tweets becomes possible. Our '*Tweet Constellation*' multidimensional model is composed of five dimensions namely *DTime, DSource, DTweet-Metadata, DPlace* and *DUser*.

Fig. 1 shows a fragment modeling the reflexive fact *FACTIVITY-TWEET* and its dimensions. Further details on this model (i.e. *Tweet Constellation*) are in Ben Kraiem et al. (2014), Ben Kraiem et al. (2015).

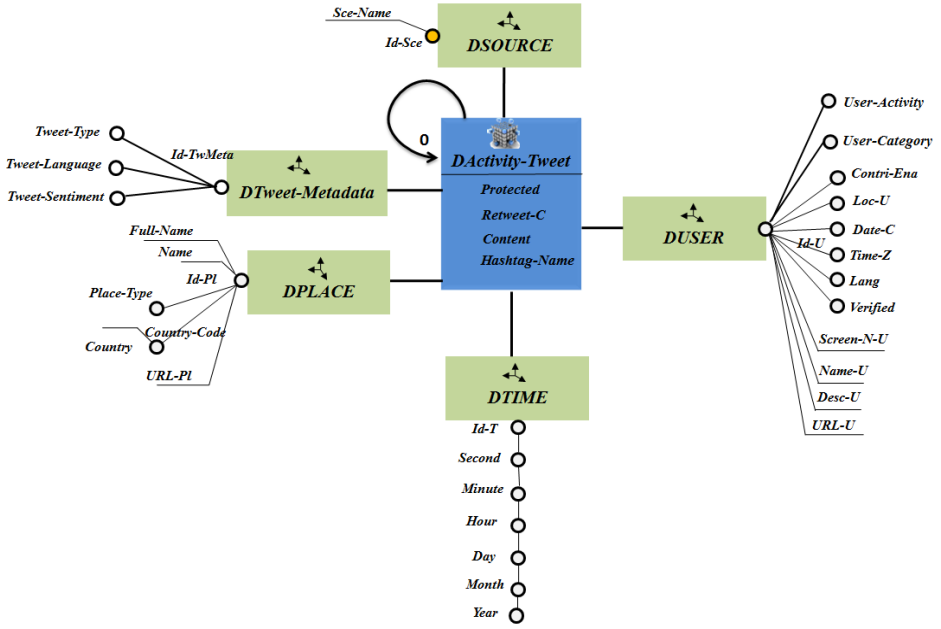


FIG 1. *FACTIVITY-TWEET fact and its dimensions*

Table 1 shows a set of seven reflexive tweets from the fact *FACTIVITY-TWEET*.

N	ID-Twt	Content	Id-Twt-Response	LEVEL
1	946077853262778373	Tu sais que tu n'as rien foutus de ta journée quand ton AppleTV te demande si tu es encore là. #BingeWatching #NoLife	-	**1**
2	946078190027661312	"@lolfr C'est bien aussi :) T'as vu quoi ?"	946077853262778373	**2**
3	946084024375750657	@cegron On my list. J'ai jusqu'au 31 pour rattraper mon retard.	946078190027661312	**2**
4	946078475923935232	@cegron The Punisher. J'étais en retard. ¿	946078190027661312	**3**
5	946078699283206145	"@lolfr Moins que moi, alors. Je n'en suis qu'au 8" .	946078475923935232	**4**
6	946079260711768064	@cegron Ah mais j'ai pas fini. Épisode 6 seulement. ¿	946078699283206145	**5**
7	946080193910853633	"@lolfr I.N.E.X.C.U.S.A.B.L.E ¿ Tu as le spécial noël de Doctor Who aussi"	946079260711768064	**6**

TAB. 1. Sample of seven interconnected instances of tweets from FACTIVITY-TWEET

In example in TAB. 1, we distinguish six hierarchical levels. The first level corresponds to the tweet at line 1. The second level corresponds to tweets at lines 2 and 3, which are

responses to the same tweet in line 1. Finally, tweets from lines 4 to 7 correspond, respectively, to levels 3, 4, 5 and 6 (cf. Fig. 2). Hence, we may notice that due to the reflexive relationship on fact instances, the fact is composed of hierarchical data at multiple levels and allows a decision-maker to navigate between levels. Using levels in OLAP offers further alternatives analyses since it provides users with the flexibility to view data from different perspectives.

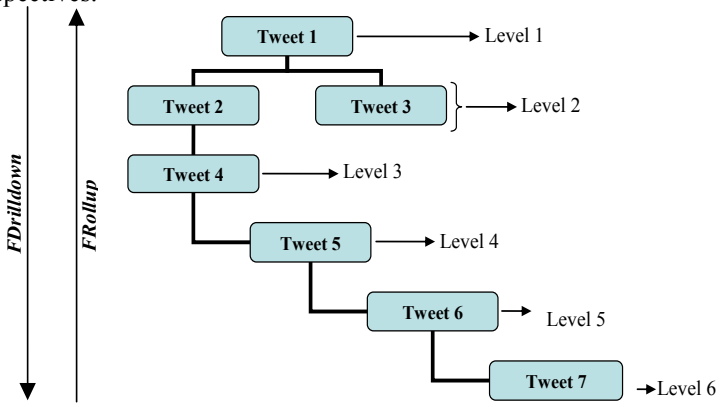


FIG 2. *Hierarchy of levels of tweets listed in Table 1*

However, classical OLAP algebra does not provide solutions for handling navigation between levels of the fact's instances since the multidimensional conventional models do not offer the reflexive relationship. Solving this issue requires appropriate operators. Hereafter, we define two new OLAP operators called *FDrilldown* and *FRollup*. They allow navigating through the hierarchical levels of the *fact*, in order to analyze a measure with more or less precision. The proposed operators are well suited to decision making applications since they can produce an output that leads to many different kinds of analyses. Basically, they allow identifying topics that have elicited a significant number of responses; these topics can be more investigated/explored later using sophisticated techniques as those used in "Text Mining" tools. Thus, we can extract knowledge from tweets and strengthen more semantics.

4 OLAP operators for reflexive fact

The result of an OLAP analysis is usually presented as a *Multidimensional Table* (Gyssens and Lakshmanan, 1997). A multidimensional table is a visualization structure that displays, from a single fact, data calculated according to two of the fact dimensions.

A multidimensional table, denoted *MT*, is defined by *(F, MES, Dim, Hier, Pred)* where:

- *F* is the name of the fact (subject) analyzed,
- *MES*= $\{f_1(m_1),\ldots, f_p(m_p)\}$ is a set of p measures $m_1,\ldots, m_p$, each measure is associated with an aggregation function $f_i,\ldots, f_p, f \subseteq \{SUM, AVG, MAX \ldots\}$,
- *Dim* = $\{D_1, D_2\}$ is the set of two dimensions currently displayed in *MT*,
- *Hier* = $\{H^{D1}, H^{D2}\}$ is a set of two hierarchies currently displayed in *MT* and belonging to dimensions D_1 and D_2 respectively.
- *Pred* = $\{pred_1 \wedge \ldots \wedge pred_s\}$ is a normalized conjunction of predicates (restrictions of dimensions data and fact data).

4.1 FDrilldown Operator

The *FDrilldown* operator applies to a reflexive fact; it consists in moving from coarser-level data to finer level data within the same analyzed fact. This drilling is possible due to the presence of the reflexive relationship on fact instances. Next, we give the *FDrilldown* algebraic formalization.

4.1.1 Conceptual definition

$MT \leftarrow FDrilldown\ (MT_k, F, Lvl_{inf})$	
Input	- MT_k: A multidimensional table currently displayed - F: is the ***reflexive*** fact analyzed in MT_k and on which the drilling operation is applied. - Lvl_{inf} is a level lower than the displayed level in the current fact F.
Output	MT is the result multidimensional table.

TAB 2. Algebraic Formalization of the FDrilldown operator

4.1.2 Logical definition

The algorithm *FDrilldown* develops the logic of the *FDrilldown* operator.
Note that "For each row *r* in the result set, the keyword LEVEL returns the depth in the hierarchy (hierarchical level) of the node represented by row *r*. The LEVEL of the root node is 1, the LEVEL of an immediate child of the root node is 2, and so on"[1].

Algorithm FDrilldown: $MT \leftarrow FDrilldown\ (MT_k, F, Lvl_{inf})$
Input
MT_k: Multidimensional table
F: is the reflexive fact analyzed in MT_k, on which the drilling down operation is applied.
Lvl_{inf} is a level of F, to be reached by the *FDrilldown*
Output
New multidimensional table MT, with the same structure as MT_k
Begin
1. Let Levels = {Lvl_n, Lvl_{n-1},…,Lvl_c} be the set of displayed levels of F with Lvl_c is the finest level, and Lvl_n is the highest level ($c \leq n$)
2. Let *NB* be the number of levels in the reflexive fact F in MT_k
3. Query-Level = **' SELECT MAX (LEVEL) '** || **' FROM '** || F || **' CONNECT BY PRIOR '** || child_expr = parent_expr (i.e., *Id-Twt = Id-Twt-Response*);
4. *NB* = Result of *Query-Level*
5. If $Lvl_c \leq Lvl_{inf}$ OR $Lvl_{inf} \geq NB$ then
6. // Impossible Drilling operation
7. Else

[1]https://www.enterprisedb.com/docs/en/9.5/eeguide/EDB_Postgres_Enterprise_Guide.1.036.html

8. Translate *FDrilldown* (MT_k; *F*, Lvl_{inf}) into query *Q* such as
   Q = ' **SELECT LEVEL** ' || p_n, p_{n-1},…, p_1 || $f_1(m_1)$, $f_2(m_2)$, …. || ' **FROM** ' || D_1, D_2, F || ' **WHERE** ' || Pred || ' **AND** ' || D_1.primary key = F.foreign key-D_1 || ' **AND** ' || D_2.primary key = F.foreign key-D_2 || ' **AND LEVEL =** ' || $\mathbf{Lvl_{inf}}$ || ' **CONNECT BY PRIOR** ' || child_expr = parent_expr (i.e., *Id-Twt = Id-Twt-Response*) || ' **GROUP BY** ' || **LEVEL,** p_n, p_{n-1},…, p_1;
9. MT = Result of query Q.
10. Display MT
11. End if

End

Example 1. To explain how the *FDrilldown* executes, we provide an example of analysis. Suppose that the decision-maker wishes to count the number of tweets (*Count (Id-Twt)*) by *Tweet-Sentiment of the DTWEET-METADATA* dimension and by *Country* of the *PLACE* dimension. As a result for this requirement, (s)he obtains the multidimensional table MT_1 shown in Fig. 3. Each cell in MT_1 gives the number of tweets for each combination of *Country* and *Tweet-Sentiment*.

FACTIVITY-TWEET | Count(Id-Twt)		**DPLACE | H_Geo**				
		Country	Belgium	Canada	France	Spain
DTWEET-METADATA | H_Sen	**Tweet-Sentiment**					
	Positive		167	46	10885	172
	Negative		41	28	6279	97
	Neutral		85	17	7658	43
R = Ø						

FIG 3. MT_1: *Result multidimensional table for Example 1*

To the extent that this sample is representative, most conversations that occur in Twitter appear to be dyadic exchanges of three to six messages. For this reason, based on the results presented in MT_1 (cf. FIG 3), the decision-maker intends to restrict the analysis to tweets that tie in *level 6*. In fact, his aim is to move deeper into a chain of data, from high-level information to more detailed information. Hence, data pertaining to fact can then be pre-summarized and then be available for more analyses (number of intense conversation…). This OLAP analysis is calculated using the following algebraic expression:

$$MT_2 \leftarrow FDrilldown\ (MT_1, FACTIVITY\text{-}TWEET, 6) \qquad (1)$$

After execution, the decision-maker obtains the multidimensional table presented in Fig. 4.

FACTIVITY-TWEET | Count(Id-Twt)		DPLACE | H_Geo		
		Country	Belgium	France
DTWEET-METADATA | H_Sen	Tweet-Sentiment			
	Positive		37	985
	Negative		9	392
R = Ø				

FIG 4. *MT_2: Result of the expression 1*

This result will allow the analyst to get interesting values from topics that have elicited more responses, performing specialization if more details are needed and, finally gleaning valuable insights about the way of propagation of data within each level. It also allows identifying where relevant tweets originate from.

4.2 FRollup Operator

The *FRollup* operator is the reverse of *FDrilldown*, it consists in moving from a finer level to a coarser level on a currently displayed fact based on fact instances linked through the reflexive relationship (Tweet Response – Tweet). Each tweet may be connected to *n (n≥0)* tweets responses within the same fact.

4.2.1 Conceptual definition

$MT \leftarrow FRollup\ (MT_k, F, Lvl_{sup})$	
Input	- MT_k: A multidimensional table currently displayed - F: is a reflexive fact, on which the FRollup operation is applied. Lvl_{sup} is a coarser-graduation level on the current fact.
Output	MT is the resulting multidimensional table.

TAB 3. *Algebraic Formalization of the FRollup operator*

4.2.2 Logical definition

The algorithm *FRollup* develops the logic of the *FRollup* operator.

Algorithm FRollup: $MT \leftarrow FRollup\ ((MT_k, F, Lvl_{sup})$
Input
MT_k: Multidimensional table
F: is the fact actually analyzed in MT_k, on which the rolling up operation will apply. The relationship between the instances of the analyzed fact must be reflexive.
Lvl_{sup} is a level of F, to be reached by the *FRollup*.

Output
Result multidimensional table *MT*, with the same structure as MT_k
Begin
1. Let Levels = {Lvl_n, Lvl_{n-1},…,Lvl_c} be the set of levels displayed for F; *n* and *c* are respectively the lowest and highest levels ($c \geq n$)
2. Let *NB* be the number of levels in the fact F analyzed in MT_k
3. Query-Level = **' SELECT '** || **MAX (LEVEL)** || **' FROM '** || F || **' CONNECT BY PRIOR '** || child_expr = parent_expr (i.e., *Id-Twt = Id-Twt-Response*;
4. *NB* = Result of *Query-Level*
5. If $Lvl_c \geq Lvl_{sup}$ OR $Lvl_{sup} \geq NB$ then
6. // Impossible *FRollup* operation
7. Else
8. Translate *FRollup* (MT_k; *F*, Lvl_{sup}) into query *Q* such as
Q = **' SELECT LEVEL, '** || p_n, p_{n-1},…, p_1 || $f_1(m_1)$, $f_2(m_2)$, …. || **' FROM '** || D_1, D_2, F || **' WHERE '** || Pred || **' AND '** || D_1.primary key = F.foreign key-D_1 || **' AND '** || D_2.primary key = F.foreign key-D_2 || **' AND LEVEL = '** || **Lvl_{sup}** || **' CONNECT BY PRIOR '** || child_expr = parent_expr (i.e., *Id-Twt = Id-Twt-Response* || **' GROUP BY '** || **LEVEL,** p_n, p_{n-1},…, p_1;
9. MT = Result of query Q.
10. Display MT
11. End if
End

Example 2: Assume that the decision-maker starts his analysis by displaying the number of tweets at *level 3* by *Sce-Name (source Name)* on the *DSource* dimension and by *User-Category* on *DUSER* dimension. He obtains a multidimensional table as in Fig. 5. Each cell represents the number of tweets of level 3 for a given *Source Name* and a given *User Category*.

FACTIVITY-TWEET| Count(Id-Twt)		DUSER | H_Cat			
		User-Category	Friendship relationship	Information Seeker	Information Sharing
DSOURCE | H_Sce	Sce-Name				
	Tendances France		-	-	58
	Tweetbot for ioS		2	-	4
	Tweetbot for Mac		-	2	1
	Twitter for Android		37	33	65
	Twitter for iPad		-	-	2
	Twitter for iPhone		-	75	172
	Twitter Web Client		25	13	36
R = Ø					

FIG 5. MT_3. *Result multidimensional table for Example 2*

Suppose the decision-maker carries out the same analysis described in the example 2, but he puts less emphasis on the depth of involved level (*Level 3*) (cf. Fig. 5), the decision-maker continues by rolling up analysis level. This time he expects to get the number of tweets at *level 2*. The corresponding analysis expression is:

$$MT_4 \leftarrow FRollup\ (MT_3,\ FACTIVITY\text{-}TWEET,\ 2) \qquad (2)$$

Fig. 6 shows the obtained result within level 2.

FACTIVITY-TWEET| Count(Id-Twt)		DUSER | H_Cat			
		User-Category	Friendship relationship	Information Seeker	Information Sharing
DSOURCE | H_Sce	Sce-Name				
	Tendances France		3	17	103
	Tweetbot for ioS		4	2	12
	Tweetbot for Mac		5	3	3
	Twitter for Android		157	176	208
	Twitter for iPad		1	4	8
	Twitter for iPhone		256	250	481
	Twitter Web Client		93	66	97
R = Ø					

FIG 6. *MT_4: Result of expression 2*

According to this result, we may conclude that the number of tweets for Information Sharing category and Twitter for iPhone as well as Twitter for Android is important. In fact, information sharing users post news and tend to have a large base of "followers" and answers about that news.

5 Experimental results

In order to evaluate the drilling Up and Down operators using the reflexive relationship between tweets, we have integrated these operators into our software prototype called *OLAP4Tweet* (Ben Kraiem et al., (2015)), developed using JAVA and ORACLE 10g database.

The *OLAP4Tweet* framework is composed of two modules, namely *Analysis Engine* and *Interactive Restitution*. Each module has specific roles and interacts with the other. The *Analysis Engine* module is designed for R-OLAP environment. It is composed of a set of algebraic operators and one parser:

- The set of algebraic operators defines elementary operations that decision-makers can carry out while analyzing. The definition of algebraic operators is independent of tools and implementation languages.
- The operator parser (a) translates algebraic operators into queries, (b) generates corresponding SQL queries and executes them.

The *Interactive Restitution* module contains (a) a graphical implementation of analysis operators in order to facilitate decision-makers' tasks and (b) a graphical interface showing analysis results.

We have loaded a dataset containing 71,739 tweets collected by crawling two hours of public tweets (from Fri Dec 22 10:48:50 UTC 2017 to Fri Dec 22 12:48:50 UTC 2017). These tweets are written in different languages. Once we load the "Tweet Constellation" multidimensional model with data, we can express and execute OLAP queries. For this purpose, we include a user-friendly decision-making process in our analysis framework. A

decision-maker starts an analysis by exploring the proposed model through an interface. (S)he selects measures and attributes related to their analysis needs by clicking and then an SQL code is generated. Queries involved in the experimental assessments aggregate the measure through the *COUNT* aggregation function. Finally, the interface provides the decision-maker with a dashboard interface representing the analysis result in tabular and graphical forms.

To illustrate how the *FDrilldown* and *FRollup* operators perform, we provide an example of analysis. We assume that the decision-maker wants to analyze the *number of tweets* by *Tweet-Sentiment (of the TWEET-METADATA* dimension) and by the parameter *Country* (of the *PLACE* dimension). A bar chart is required by the decision-maker ((s)he just click on Bar chart icon) to display the analysis result (cf. Fig. 7).

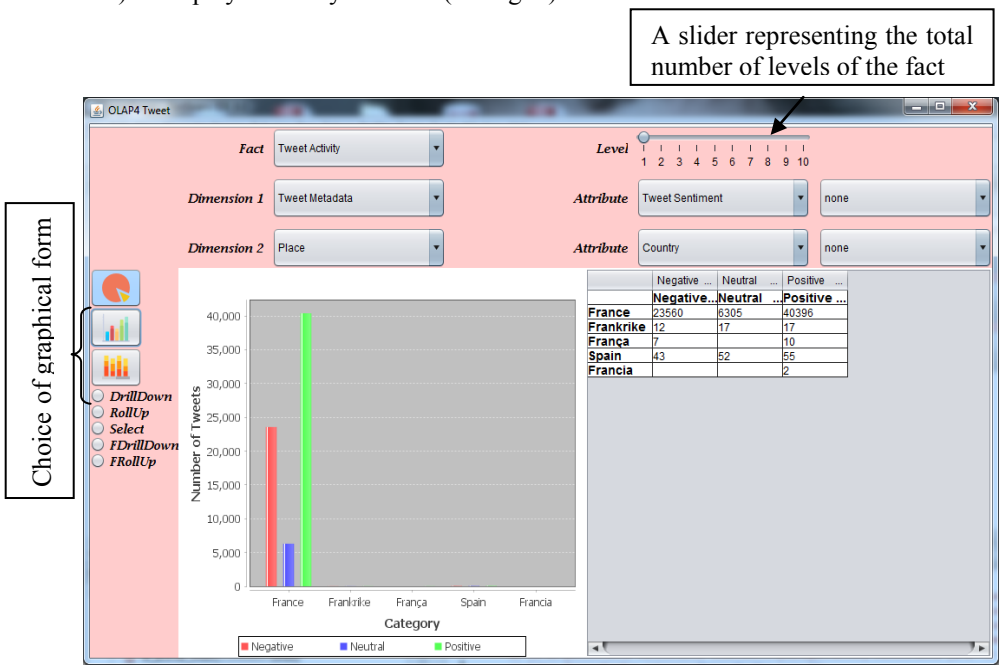


FIG 7. *Number of tweets by Tweet-Sentiment and by Country.*

Based on the analysis result presented in FIG 7 the decision-maker intends to restrict the analysis to tweets that correspond to *level 6*. This level can represent a valuable source of information that could help obtain a full picture of topics. Hence, statistical data pertaining to fact can then be pre-summarized, and then be available for large analyses. A slider (on the right top of the interface in Fig. 7) allows navigating along a set of levels on the fact. Fig. 8 shows the obtained result.

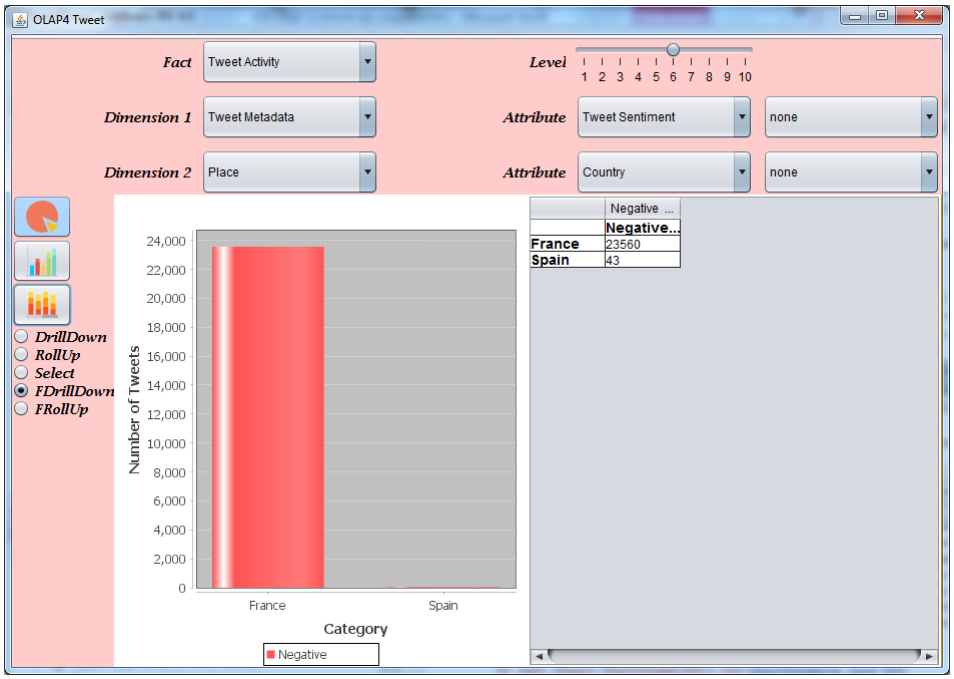


FIG 8. *Interface after the execution of the FDrilldown operator*

6 Conclusion

In order to exploit the reflexive relationship on fact instances, we have proposed two specific OLAP operators namely *FDrilldown* and *FRollup*. They provide solutions for handling an intuitive navigation between different levels within the fact. The proposed operators are well suited to decision making applications since they can produce an output that leads to many different kinds of analyses. They highlight the importance of tweets responses to show how information is propagated through each tweet. Basically, they allow identifying topics that have elicited a significant number of responses; these topics can be more investigated/expolored using sophisticated tools based on "Text Mining" techniques; thus, we can extract knowledge from tweets and strengthen more semantics.

For each of these operators, we have presented an algebraic formalization, and a pseudo code algorithm.

To the best of our knowledge, this is the first initiative that has tackled OLAP operators for drilling down and up within a fact by exploiting the reflexive relationship.

As perspective work, we intend to integrate more analysis operators that take into consideration the specificities of our multidimensional model, as dynamic Data. These operators will help the interpretation of the results of multidimensional analyses on tweets and their metadata. It is also important to use OLAP mining, which integrates on-line analytical processing (OLAP) with data mining so that mining can be performed in different portions of data warehouses and at different levels of abstraction at user's fingertips.

Moreover, we plan to conduct experiments to measure the quality of the result extracted by our OLAP operators.

References

Abelló, A, Samos, J and Saltora, F (2002), *On relationships offering new drill-across possibilities*. In Proceedings of the 5th ACM international Workshop on Data Warehousing and OLAP, McLean, Virginia, USA, ACM Press, 2002, pp 7-13.

Abelló, A, Samos, J and Saltora, F (2003), *Implementing Operations to Navigate Semantic Star Schemas*. In Proceedings of the 6th ACM International Workshop on Data Warehousing and OLAP, New Orleans, Louisiana, USA, ACM Press, 2003, pp 56-62.

Ben Kraiem, M., J. Feki, K. Khrouf, F. Ravat, and O. Teste (2017), *OLAP operators for missing data*. In Revue des Nouvelles Technologies de l'Information (RNTI), Ed. Hermann, vol. B- 13, pages 53-66. 13èmes journées francophones sur les Entrepôts de Données et l'Analyse en Ligne, Business Intelligence & Big Data, 2017.

Ben Kraiem, M., J. Feki, K. Khrouf, F. Ravat, and O. Teste (2015). *Modeling and OLAPing social media: the case of twitter*. In Social Network Analysis and Mining 5(1), 47:1–47:15.

Ben Kraiem, M., J. Feki, K. Khrouf, F. Ravat, and O. Teste (2014). *OLAP of the tweets from modeling toward exploitation*. In 8th International Conference on Research Challenges in Information Science (IEEE RCIS'2014), 45–55.

Cabibbo, L. and Torlone, R (2004). *Dimension compatibility for data mart integration*. In Proceedings of the 12th Italian Symposium on Advanced Database Systems, Cagliari, Italy, 2004, pp. 6-17.

Ravat, F., Teste, O, Tournier, R, and Zurfluh, G. (2007), *Graphical Querying of Multidimensional Databases*. In Advances in Databases and Information Systems, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, Vol. 4690, pp. 298–313.

Ravat, F., Teste, O, Tournier, R, and Zurfluh, G. (2008), *Algebraic and graphic languages for OLAP manipulations*. In International Journal of Data Warehousing and Mining, IJDWM 4(1), 2008, pp.17–46.

Riazati, D, Thom, J.A and Zhang X (2008), *Drill Across & Visualization of Cubes with Non-Conformed Dimensions.* In the Nineteenth Australasian Database Conference (ADC 2008), Wollongong, Australia, January 2008. Conferences in Research and Practice in Information Technology, Vol. 75.

Résumé

Le modèle de données multidimensionnel et les implantations des réseaux sociaux sont assortis d'un ensemble de contraintes, telles que des données manquantes, des relations réflexives sur des instances de fait. Cependant, les opérateurs OLAP classiques et les modèles multidimensionnels ne fournissent pas de solutions pour gérer ces spécificités. Par

conséquent, des efforts méritent d'être déployés pour étendre ces opérateurs afin de prendre en compte la spécificité de la modélisation multidimensionnelle des tweets et de leur manipulation. Face à ce problème, nous proposons de nouveaux opérateurs OLAP qui exploitent l'existence d'une relation réflexive entre les instances d'un fait. Pour chacun de ses opérateurs, nous proposons une définition orientée utilisateur (c'est-à-dire une formalisation algébrique) ainsi qu'une traduction algorithmique pour sa mise en œuvre.

Graph databases and big data technologies in healthcare: A gap analysis

Faiza Deghmani*, Idir Amine Amarouche**
LSI, Department of Computer Science, University of Science and Technology Houari Boumediene (USTHB)
BP 32 El Alia 16111 Bab Azzouar Algiers, Algeria
*deghmani.faiza @gmail.com
**i.a.amarouche@gmail.com

Abstract. Several aspects related to big data technologies in the healthcare area, like architecture and capabilities, have been surveyed. Also, many works propose the use of graph databases in healthcare domain. However, according to the best of our knowledge, there is no work that addresses the challenges related to big data technologies and graph databases in healthcare. For this reason, we address a survey of big data in healthcare based on a graph database. The presented paper exposes a gap analysis based on a set of paper related to the healthcare systems based on graph databases and big data technologies.

1 Introduction

In recent years, digitized data in healthcare are generated at very high speed with the data coming in from internal as well as external sources, such as, mobile devices, wearable sensor devices, Electronic Health Records (EHR), social media and remote health monitoring devices (Mathew et Pillai 2015). These data have a big volume and a variety of formats (images, texts, photos...etc.) and a high level of velocity. As a result, these data meet the main characteristics of big data and motivate the need for required management solutions (Zillner et Neururer 2016; Mathew et Pillai 2015).

Furthermore, rising rates of chronic diseases, increased population, need for evidence-based medicine, inability to process and get insight from ever-increasing heterogeneous health data are also drivers for adopting big data solutions in healthcare field (Mathew et Pillai 2015). Big data technologies should provide capabilities to manage massive data available in the healthcare industry which need to work on prediction, prevention and personalization to improve their outcomes and the quality of patient care (Mathew et Pillai 2015).

This trend can be explained by the limits encountered by relational databases, which are not designed to effectively cope with such large quantities of data and promote the development of NoSQL databases (Fraczek et Plechawska-Wojcik 2017). Several works show that NoSQL databases provide significant advantages, such as, easy and automatic scaling, better performance and high availability which address the limitations of relational databases in distributed healthcare systems (Ercan et Lane 2014).According to (Mathew et Pillai 2015), NoSQL databases expose four data store classes:(1) *Key Value* oriented database which presents data stored as couples of values and their keys; (2) *Column* oriented database which stores data as columns and each column has a key; (3) *Document* oriented database, in which data are stored in documents (JSON format) and (4) *Graph* oriented database in which data are represented as a network and stored as nodes and edges.

Recently, graph databases regained an important interest among the researchers for many reasons. The inherent property of graphs, as a structure, is that represents the strong connectivity within the data. Graph databases are the best for dealing with complex, semi-structure, and mainly densely connected data and it is very fast in terms of queries (kumar Kaliyar 2015). The use of graph databases in healthcare has significant benefits (Park et al. 2014). That's why many researchers proposed the use of graph databases in healthcare systems to offer better analytics either descriptive or predictive(Ling et al. 2014; Khan, Uddin, et Srinivasan 2016; Sen et al. 2017). Also, to understand relationships between entities and to construct efficient data management framework for large scale healthcare system (Park et al. 2014; Khan, Uddin, et Srinivasan 2016). In addition, there are several works that survey many challenges related to big data technologies in medical and health area like architecture and capabilities (Zillner et Neururer 2016; Krishnan 2016; Mathew et Pillai 2015; Asare-Frempong et Jayabalan 2017; Wang et al. 2015; Raghupathi et Raghupathi 2014; Cyganek et al. 2015; Wang et Hajli 2017). The use of NoSQL databases in the health sector was also evaluated (Yaqoob et al. 2016). But to the best of our knowledge, there is no work which has addressed a survey of big data technologies in healthcare based on graph databases. To address this lack, the present paper proposes to review recent papers related to healthcare systems based on graph databases and big data technologies.

This paper is organized as follows. The next section presents background concepts related to graph databases, big data technologies and healthcare data. Section 3 motivates the present paper. Section 4 presents the reviewed work that deal with adoption of graph databases and big data technologies to handle healthcare system requirement. In section 5, we give an analysis of studied systems presented in the previous section and highlight potential challenges. Finally, section 6 concludes the paper and presents our perspectives.

2 Background

In this section, we provide the main definitions related to the core concepts used in the proposed survey.

2.1 Big data

The term big data refers to the huge amount of data that needs new technologies and architectures to find valuable knowledge from it by using new and innovative analysis practices. Various explanations from 3V (i.e. Volume, Variety, and Velocity) to 4V (i.e. Volume, Velocity, Variety and Veracity) have been provided to define big data (Yaqoob et al. 2016). The term "volume" refers to the size of the data, "velocity" refers to the speed of incoming and outgoing data, and "variety" describes the sources and types of data. "Veracity" or "variability" as the fourth V; refers to the messiness and trustworthiness of data (Yaqoob et al. 2016). Big data is also defined by "Value" which refers to the worth of hidden insights inside big data.

To identify the potential benefits offered by big data, it is necessary to understand its architecture and component functionalities. As illustrated by the figure1, big data architecture consists of a data-based logical framework that starts with data capture, proceeds via data transformation, and concludes with data consumption. According to (Wang et al. 2015), this architecture is composed of five major layers:

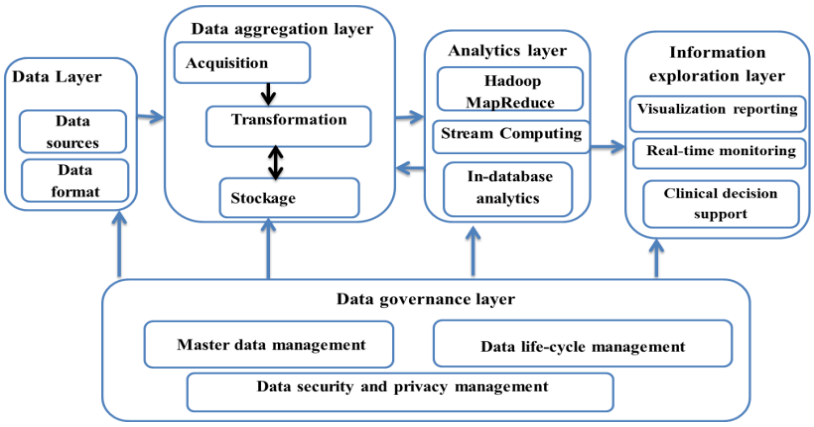


FIG. 1– *Big data Architecture.*

1) **Data layer:** This layer focus on data sources and content format. The data is divided into structured data, semi-structured data and unstructured data. These data are collected from various locations, and will be stored immediately into appropriate databases, depending on the content format.
2) **Data aggregation layer:** It is responsible for handling data. Data will be processed by performing three steps: data acquisition, transformation, and storage. The goal of data acquisition is to read collected data. The transformation engine must be capable of moving, cleaning, splitting, translating, merging, sorting, and validating data. Finally, the data are loaded into the target databases such as Hadoop Distributed File Systems (HDFS) or in a Hadoop cloud for further processing and analysis.
3) **Analytics Layer**: This layer is responsible for processing all kinds of data and performing appropriate analyses. Data analysis can be divided into three components: Hadoop Map/Reduce, stream computing, and in-database analytics, depending on the type of data and the purpose of the analysis.
4) **Information exploration layer:** It generates outputs such as the various visualization reporting options, real-time monitoring of information, and meaningful business insights derived from the analytics platforms to users in the organization.
5) **Big data governance layer:** This layer is composed of Master Data Management (MDM), data life-cycle management, and data security and privacy management that emphasize how to harness data in the organization. The MDM is regarded as the processes, governance, policies, standards, and tools for managing data. Data is properly standardized, removed, and incorporated in order to create the immediacy, completeness, accuracy, and availability of master data for supporting data analysis. The data life-cycle management is the process of managing business information throughout its lifecycle, from archiving data, through maintaining data warehouse, testing and delivering different application systems to deleting and disposing of data. Data security and privacy management is the platform for providing enterprise-level data activities in terms of discovery, configuration assessment, monitoring, auditing, and protection.

2.2 Healthcare data

Healthcare information systems, social media and medical devices are some of the main providers of health data (Zillner et Neururer 2016; Asare-Frempong et Jayabalan 2017). Healthcare social websites, such as "PatientsLikeMe[1]", are generating large sets of health data, by voluntarily sharing data about rare diseases or remarkable experiences with common diseases (Zillner et Neururer 2016). Sensors such as glucose monitors or blood pressure monitors can provide some valuable insights about patients' health conditions (Asare-Frempong et Jayabalan 2017). Data provided by Hospital Information System such as patients' demographics, medical records, lab results and medical images, cost and billing data constitute the Electronic Medical Record (EMR). EMR is defined as the record of the periodic care provided mainly by one institution (Ercan et Lane 2014). EHR is defined as an electronic record that holds a patient's lifetime health-related information or a collection of EMRs for a single individual (Ercan et Lane 2014).

Consequently, health data are distributed, heterogeneous in terms of structure, feature and semantic which makes them particularly challenging to secure, to store, to process, to share and to analyse.

2.3 Big data in healthcare

Several developments in healthcare sector, such as escalating healthcare costs, increased need for healthcare coverage, and shifts in provider reimbursement trends, trigger the demand for big data technologies in order to improve the overall efficiency and quality of care delivery (Zillner et Neururer 2016). Accordingly, taking into consideration its complexity, heterogeneity, fast growing and size we need special tools to analyse it and we should consider healthcare data as big data. Such a situation imparts to healthcare data a 5Vs character (Cyganek et al. 2015), namely:

Volume: When doctor's note stored as a text file is a few kilo bytes; a raw image requires a few megabytes and sophisticated diagnostic tools such as MRI[2] requires Giga bytes. If such a volume size is multiplied by the number of test carried out in the hospital, we should be ready to deal with Tera and Peta bytes. According to (Asare-Frempong et Jayabalan 2017), the data amassed in the healthcare industry is about 500 Petabytes by estimate in 2012.

Velocity: Health data is in motion; new information about patient is added, and some medical records are updated. Therefore, some smart analytic tools applied to EHR analysis require that the models will not be rebuilt from scratch when new data come, but it will be improved. The need to process data in real time coming from streaming data like Remote Patient Monitoring, data from sensor devices and Telemedicine (Mathew et Pillai 2015).

Variety: The EHR include data with heterogeneous structures, i.e., on the one hand structured data in the form of standardized medical information, such as DICOM[3], or using the ICD[4] codes, but on the other hand the most valuable data could be found in unstructured data such as doctor's notes written in natural language (Cyganek et al. 2015).

[1]http://www.patientslikeme.com/

[2]Magnetic Resonance Imaging

[3]Digital imaging and communications in medicine

[4]International Classification of Diseases

Veracity: Data in the healthcare may be noisy and biased, and we may find abnormality as outliers. Also, human errors are as important issue as well, due to the clinical mistakes and it consequences.

Value: Big data analytics tool deployment makes sense if it leads to healthcare improvement. Big data adoption in healthcare was not only to manage the massive health data, analytics of big data can be also applied in the diagnosis of diseases and in the treatment of illnesses, which makes the applications of big data analytics in healthcare a solution to improve the quality of care and allow advancing research in healthcare (Zillner et Neururer 2016; Krishnan 2016).

Big Data technologies will definitely open new opportunities and enable breakthroughs related to, among the others healthcare data analytics addressing different perspectives: (i) descriptive to answer what happened, (ii) diagnostic to answer the reason why it happened, (iii) predictive to understand what will happen and (iv) prescriptive to detect how we can make it happen (Heinrich et al. 2016). The prescriptive analysis aims mainly to offer the optimal solutions or possible courses of action to help users understand what to do in the future while predicting risk of developing a disease(Wang et Hajli 2017).

Big data analytics capabilities: healthcare context

The logical layers presented above would enable healthcare managers to understand how to transform the healthcare data from various sources into meaningful clinical information through big data implementations. In this context, big data analytics capability is defined as the ability to acquire, store, process and analyse large amounts of health data in various forms, and deliver meaningful information to users (Wang et Hajli 2017). These capabilities are derived from the various design principles and functionalities of big data and are confirmed by the real-world use of big data in healthcare contexts. According to (Wang et al. 2015), the main capabilities are described as follows:

- **Traceability**: Traceability is the ability to track output data from all the system's IT components throughout the healthcare's setting units. Thus, big data can track information that is created by the medical devices in real time. This makes it possible to gather location, event and physiological information from each patient wearing the device. This information is deposited in NoSQL databases, for review by medical staff when needed.
- **Unstructured data analytical capability:** An analytical process in a big data management system starts by acquiring data provided from both inside and outside the healthcare sectors. After unstructured data has been gathered across multiple healthcare units, it is stored in a HDFS and NoSQL database that maintain it until it is called up in response to users' requests. The ability to analyse unstructured data plays a pivotal role in the success of big data in healthcare settings since 80% of health data is unstructured.
- **Analytical capability for patterns of care:** Analytical capabilities in healthcare can be used to identify patterns of care and discover associations from massive healthcare records, thus providing a broader view for evidence-based clinical practice.
- **Decision support capability:** Decision support capability emphasizes the ability to produce reports about daily healthcare services to aid managers' decisions. In general, this capability yields shareable information and knowledge such as historical

reporting, executive summaries, drill-down queries and time series comparisons. Such information can be utilized to provide a comprehensive view to detect advanced warnings for disease surveillance and to develop personalized patient care.

- **Predictive capability:** Predictive capability is the ability to apply diverse methods from statistical analysis, modelling, machine learning, and data mining to both structured and unstructured data to determine future outcomes. Predictive capabilities can reduce the degree of uncertainty and enable managers to support preventive care. The Texas Health Harris Methodist Hospital Alliance, for example, analyses information from medical sensors to predict patients' movements and thus provide needed services more efficiently (Wang et al. 2015).

2.4 Big data based on graph databases

Despite the fact that Relational Database Management System (RDBMS) is the most popular and used in academic research, as well as industrial setup, graph databases regained interest among the researchers (kumar Kaliyar 2015). Indeed, there has been developed a large number of systems for handling graph-like data like; social, biological, and other network. Graph databases are the best for dealing with complex, semi-structure, and densely connected data. Generally, graph databases are useful when we are more interested in relationships between data than in the data itself (Angles 2012). Graph database can traverse any number of relationships between entities, and they are efficient in retrieving relevant information after scouring several entities and relationship[5]. In the last time, there has been an increasing work on graph databases; from the current implementations, there are Neo4j, AllegroGraph, DEX, HyperGraphDB, InfiniteGraph and Sones.

According to (Angles 2012), a set of features is proposed in the literature in order to evaluate the data model provided by each graph database, that can be summarized as follows:

Graph data structures: The data structure of graph databases is defined around the notions of graphs, nodes and edges. There are four graph data structures; simple graphs, hyper graphs, nested graphs and attributed graphs (Angles 2012).The basic structure is a simple flat graph defined as a set of nodes (or vertices) connected by edges. Nodes in graph may represent heterogeneous entities. A Hypergraph extends this notion by allowing an edge to relate an arbitrary set of nodes (called a hyperedge). A nested graph is a graph whose nodes can be themselves graphs (called hypernodes). Attributed graphs are graphs where nodes and edges can contain attributes for describing their properties. Simple graph and attributed graph are the most supported by graph databases. Other features are considered which are directed or undirected edges, labelled or unlabelled nodes/edges, and attributed nodes/edges (i.e., edges between edges are possible).Graph data modelling can represent entities, properties and relations at both instance and schema levels.

Query languages: There is not proposal for a standard query language for graph databases. Some of the graph databases support predefined languages, such as AllegroGraph that supports SPARQL the standard query language for RDF. In contrary, Neo4j is developing Cypher, a query language for property graphs.

Integrity constraints: Integrity constraints are general statements and rules that define the set of consistent database states, or changes of state, or both. But integrity constraints are poorly studied in graph databases.

5 https://www.techopedia.com/2/31969/trends/big-data/graph-databases-a-new-way-of-thinking-about-data

3 Motivation of graph databases in healthcare

According to (Kaur et Rani 2013), NoSQL databases enable programmers to model the data closer to the format used in their application domain. Graph databases which stores data as nodes and edges are the most appropriate NoSQL databases to model data in health care because these data contain as many relations between them as the amount of data themselves.

Healthcare systems based on graph databases provide a holistic and unified view of health data which helps doctors to diagnose and to predict diseases more quickly; consequently graph databases face an important big data challenge which is the data representation (Kaur et Rani 2015; Ling et al. 2014).

Graph databases serve as multiple de-normalized tables which can avoid generating and replicating a large number of tables, thus reducing complexity in a database and enhancing data accessibility (Park et al. 2014). In addition, they have an intuitive query structure which facilitates the management, user validation, and exploration of the analytic intent of the query (Park et al. 2014). Directly , they can handle a wide range of queries that would otherwise require deep join operations in normalized relational tables, and performs well for some queries such as those supporting relationship mining (Park et al. 2014).

Graph databases are also recommended to represent the temporal ordering of events while this type of queries is complex in relational database; for example a trajectory of patient's diseases, which is a set of records put together in chronological sequence (Sen et al. 2017).

4 Surveyed papers

Among the research work on big data, we have selected those exploiting graph databases in order to ensure above mentioned big data capabilities by storing, visualizing and analysing massive healthcare data. The first published works date from 2014, that's why the adoption of graph databases in healthcare sector is fairly recent. These works can be classified into three groups according to their objectives, namely, descriptive analytics, predictive analytics, preventive medicine and Large Scale Healthcare System. Descriptive analytics provides the ability to describe the data in summary from for exploratory insights and to answer *"what has happened in the past?"* questions. Predictive analytics allow users to predict or forecast the future for a specific variable, based on the estimation of probability (Wang et Hajli 2017). Preventive medicine is to take appropriate measures when identifying individuals having risk of developing a disease (Khan, Uddin, et Srinivasan 2016) and Large Scale Healthcare System that refers to a system ensuring both of efficient data management and data services (Park et al. 2014).

4.1 Descriptive and predictive analytics: GEMINI (Ling et al. 2014)

Objective: The objective of GEMINI is to respond to predicative tasks such identifying patients at high risk of developing heart disease in the near future, or predicting the probability that patients would re-admit into hospital within 30 days.

Data and structuring: GEMINI extracts clinical data from the CCDR[6] of the national university hospital which has structured sources containing EMR. To interpret the unstructured data, GEMINI uses a well-known medical knowledge base UMLS[7] and NLP engines.

Architecture of framework: The system consists of two components: PROFILING and ANALYTICS. The **PROFILING** component extracts data of each patient from various sources and stores them as information in a patient profile graph. The patient profile graph provides a holistic and unified view of a patient's clinical data. The **ANALYTICS** component analyses the patient profile graphs to infer implicit information and extract relevant features for the prediction tasks.

4.2 Descriptive analytics and preventive medicine

4.2.1 Framework to understand chronic disease progression (Khan, Uddin, et Srinivasan 2016)

Objective: The aim of the framework is to understand chronic disease progression in order to enable stakeholders making preventive measures.

Data and structuring: Hospitals, during the course of patient's admission and upon discharge, report the detailed information in standard format to government departments and respective private health funds. In order to find the health trajectory of chronic disease patients and assess potential risk of developing disease for new patients, patients' admission records are analysed including length of stay; diagnose information, item-wise billing codes and patient's Diagnosis Related Group for each admission episode. The next step is to understand the semantics of the admission data and utilize them to develop a network that can represent the trajectory of chronic disease patients.

Architecture of framework: The framework is divided in 3 parts:

a. Part "a" collects and analyses patients' admission records and then understands the semantic of the admission data;
b. Part "b" creates baseline network to find patients' trajectory. This network represents typical sequence of diagnoses and comorbidity of chronic disease patients. The first phase of creating this network is Patient filtering which is essentially identifying chronic disease patients. And then statistical aggregation generates a graph of chronic disease patient's typical health trajectory based on admission history of chronic disease patients;
c. Part "c" finds the similarity between the Baseline Network of patients with chronic conditions and medical history of a new patient (not diagnosed with chronic disease). The method is named as Longitudinal Distance Matching, which uses sequential phases of rule-based, graph theory and social network analysis methods.

4.2.2 Portinari (Sen et al. 2017):

Objective: Portinari has as aim to explore and visualize future trajectories of patients who have undergone a specific sequence of screening exams in order to personalize cervical cancer screening and consequently reducing the number of cancer cases.

[6]Computerized Clinical Data Repository
[7]Unified Medical Language System

Data and structuring: Data is extracted from a socio-technical system for cervical cancer which is a system that identifies automatically individuals at risk of developing the disease and invites them for a screening exam; and available in the form of events in the life of patients stored in transaction records. In cervical cancer screening an event corresponds to attendance to exam, such as a Human Papilloma Virus, test along with a date and type of diagnosis.

Architecture of framework: The framework is divided in two components, namely: 1) Graph database of screening events. Events are transformed from transaction records into sequences of connected events for individual patients in a graph database implemented in Neo4J. 2) Portinari: It is a web-based data exploration tool, to explore and visualize individual trajectories by querying the graph database. Portinari automatically generates future trajectories of patients who underwent the input sequence of exams and diagnosis by matching similar patients in the graph database. Portinari visualizes the outcome as a Sankey Diagram.

4.3 Large Scale Healthcare System: Framework for efficient data management and data services (Park et al. 2014)

Objective: Authors aimed to construct healthcare graph database; from a normalized relational database using the proposed "3NF Equivalent Graph" (3EG) transformation and to evaluate the performance of queries that require deep join operations over a relational database and its equivalent graph representation.

Data and structuring: Data are stored in 3NF relational healthcare database.

Architecture of framework: it consists of3EG transform which is a graph database design rationale that constructs a graph database from an existing normalized database based on a group of conversion rules.

5 Analysis and discussion:

Table1 presents a comparison between healthcare's systems presented in section 4, based first on the objective of each work. Then, they are analysed according to the proposed architecture of big data; taking into consideration all the layers presented in section 2 except big data governance layer. At last, the systems were compared according to the structure of the graph.

		(Ling et al. 2014)	(Khan, Uddin, et Srinivasan 2016)	(Sen et al. 2017)	(Park et al. 2014)
Objective		Predicative analytics	Prevention of chronic diseases	Prevention of cancer	Large scale healthcare systems
Data structure	Structured data	×	×	×	×
	Unstructured data	×			
Data sources	EMR	×	×	×	×
	Social networks				

	Captured data				
Data transformation	NLP	×			
Data storage	Graph data base management system			Neo4j	Neo4j
Data analysis	Hadoop Map Reduce	×			
	Stream processing				
	In-database analytics				
Information exploration layer	Visualization reporting	×	×	×	
	Real-time monitoring				
	Clinical decision support				
Graph Data Structure	Simple graph	×	×	×	×
	Hyper graph				
	Nested graph				
	Attributed graph	×	×	×	×
	Node Labelled				
	Node attribution		×	×	
	Edge Directed	×	×	×	
	Edge Labelled	×		×	
	Edges attribution		×	×	

TAB. 1 – Comparison between healthcare systems based on graph database and big data technologies.

By analysing Table 1, we draw the following conclusions:

– All the papers exploit structured data provided by EMR systems because it contains efficient and reliable data from which valuable information can be extracted.

– (Ling et al. 2014) exploited unstructured data from doctor's notes. Due to the complexity of this data, authors used several NLP techniques and UMLS dictionary to understand it and to construct the patient's profile graph from this data and had to manage some limitations such as ambiguous mappings, missing mappings and relationships.

– Despite the existence of so much graph database management system, Neo4j is the most used because of many reasons. Neo4j is open source; it has an API and a query language (Cypher) so it is easy to handle and to query the database.

– Most of the works presented a visualization reporting as an output which is very helpful for doctors to draw conclusions about their patients and to predict future events. But the only visualisation was network graph for individuals or events when other dashboards may help a lot in statistic mostly for predictive tasks.

– None of the papers presented a real-time monitoring or clinical decision support which may be beneficial and helpful for descriptive analytics, predicative analytics and

preventive medicine by sending notifications for patients having risk for developing a disease for example or determining the appropriate exam or medication.

- All of the papers opt for simple and attributed graph because those two types of graph are supported by most of graph databases management systems.
- Relationships in health data is as important as nodes, it contains relevant information that's why many papers presented directed edges with attributes and labels.

6 Conclusion

The main contribution of this paper is to initiate a gap analysis related to graph databases and big data technologies in the healthcare area. In this sense, the first published works date from 2014, that's why adoption of graph databases in the healthcare sector is fairly recent, despite the benefits it gives, the research is not completed.

The analysed papers raised some big data solutions in healthcare like descriptive and predictive analytics, preventive medicine and implementing large-scale system. However, there are other aspects, either in big data or healthcare area, are not yet tackled in the context of graph databases which may be the topic of future work, such as prescriptive analytics. This analysis requires the combination of optimization, machine learning, simulation and heuristics-based predictive modelling technique.

References:

Angles, Renzo. 2012. « A Comparison of Current Graph Database Models ». In , 171‑77. IEEE. https://doi.org/10.1109/ICDEW.2012.31.

Asare-Frempong, Justice, et Manoj Jayabalan. 2017. « Exploring the Impact of Big Data in Healthcare and Techniques in Preserving Patients' Privacy », 8 août 2017. http://paper.ijcsns.org/07_book/201708/20170819.pdf.

Cyganek, Boguslaw, Manuel Grana, Andrzej Kasprzak, Krzysztof Walkowiak, et Michal Wozniak. 2015. « Selected aspects of electronic health record analysis from the big data perspective ». In , 1391‑96. IEEE. https://doi.org/10.1109/BIBM.2015.7359881.

Ercan, Mehmet Zahid, et Michael Lane. 2014. « Evaluation of NoSQL databases for EHR systems », décembre 2014.

Fraczek, Konrad, et Malgorzata Plechawska-Wojcik. 2017. « Comparative Analysis of Relational and Non-relational Databases in the Context of Performance in Web Applications ». In *Beyond Databases, Architectures and Structures. Towards Efficient Solutions for Data Analysis and Knowledge Representation*, édité par Stanisław Kozielski, Dariusz Mrozek, Paweł Kasprowski, Bożena Małysiak-Mrozek, et Daniel Kostrzewa, 716:153‑64. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-58274-0_13.

Heinrich, Adrienne, Aizea Lojo, Ernestina menasalvas, et Wilfried Verachtert. 2016. « Big Data technologies in healthcare: Needs, opportunities and challenges ».

Kaliyar, Rohit kumar. 2015. « Graph databases: A survey ». In , 785‑90. IEEE. https://doi.org/10.1109/CCAA.2015.7148480.

Kaur, Karamjit, et Rinkle Rani. 2013. « Modeling and querying data in NoSQL databases ». In , 1‑7. IEEE. https://doi.org/10.1109/BigData.2013.6691765.

Kaur, Karamjit, etRinkle Rani. 2015. « Managing Data in Healthcare Information Systems: Many Models, One Solution ». *Computer* 48 (3): 52‑59. https://doi.org/10.1109/MC.2015.77.

Khan, Arif, Shahadat Uddin, et Uma Srinivasan. 2016. « Adapting Graph Theory and Social Network Measures on Healthcare Data: A New Framework to Understand Chronic Disease Progression ». In , 1‑7. ACM Press. https://doi.org/10.1145/2843043.2843380.

Krishnan, Shankar M. 2016. « Application of Analytics to Big Data in Healthcare ». In , 156‑57. IEEE. https://doi.org/10.1109/SBEC.2016.88.

Ling, Zheng Jye, Quoc Trung Tran, Ju Fan, Gerald C. H. Koh, Thi Nguyen, Chuen Seng Tan, James W. L. Yip, et Meihui Zhang. 2014. « GEMINI: An Integrative Healthcare Analytics System ». *Proceedings of the VLDB Endowment* 7 (13): 1766‑71. https://doi.org/10.14778/2733004.2733081.

Mathew, Prabha Susy, et Anitha S. Pillai. 2015. « Big Data solutions in Healthcare: Problems and perspectives ». In , 1‑6. IEEE. https://doi.org/10.1109/ICIIECS.2015.7193211.

Park, Yubin, Mallikarjun Shankar, Byung-Hoon Park, et Joydeep Ghosh. 2014. « Graph databases for large-scale healthcare systems: A framework for efficient data management and data services ». In , 12‑19. IEEE. https://doi.org/10.1109/ICDEW.2014.6818295.

Raghupathi, Wullianallur, et Viju Raghupathi. 2014. « Big Data Analytics in Healthcare: Promise and Potential ». *Health Information Science and Systems* 2 (1). https://doi.org/10.1186/2047-2501-2-3.

Sen, Sagar, Manoel Horta Ribeiro, Racquel C. De Melo Minardi, Wagner Meira, et Mari Nygard. 2017. « Portinari: A Data Exploration Tool to Personalize Cervical Cancer Screening ». In , 37‑46. IEEE. https://doi.org/10.1109/ICSE-SEIS.2017.6.

Wang, Yichuan, et Nick Hajli. 2017. « Exploring the Path to Big Data Analytics Success in Healthcare ». *Journal of Business Research* 70 (janvier): 287‑99. https://doi.org/10.1016/j.jbusres.2016.08.002.

Wang, Yichuan, Leeann Kung, Chaochi Ting, et Terry Anthony Byrd. 2015. « Beyond a Technical Perspective: Understanding Big Data Capabilities in Health Care ». In , 3044‑53. IEEE. https://doi.org/10.1109/HICSS.2015.368.

Yaqoob, Ibrar, Ibrahim Abaker Targio Hashem, Abdullah Gani, Salimah Mokhtar, Ejaz Ahmed, Nor Badrul Anuar, et Athanasios V. Vasilakos. 2016. « Big Data: From Beginning to Future ». *International Journal of Information Management* 36 (6): 1231‑47. https://doi.org/10.1016/j.ijinfomgt.2016.07.009.

Zillner, Sonja, et Sabrina Neururer. 2016. « Big Data in the Health Sector ». In *New Horizons for a Data-Driven Economy*, édité par José María Cavanillas, Edward Curry, et Wolfgang Wahlster, 179‑94. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-21569-3_10.

Résumé

Plusieurs aspects liés aux technologies de big data dans le domaine médical, tels que l'architecture et les fonctionnalités, ont été étudiés. Aussi, beaucoup de travaux proposent l'usage de bases de données à base de graphes dans le domaine médical ou de santé. Cependant, selon nos connaissances, rares sont les travaux qui abordent le défi lié aux technologies de big data et des bases de données à base de graphes et leur emploi dans le domaine médical. A cet effet, le présent papier tente d'analyser les lacunes et de dresser l'état de l'art de cette question, et ce, à travers l'étude d'un ensemble de papiers qui proposent des solutions de big data basées sur les bases de données à base de graphes qui sont destinées au domaine médical.

Disambiguation Solution for Complex Questions Answering System over Linked Data

Wafa Nouar*, Zizette Boufaida**

*LIRE Laboratory, Abdelhamid Mehri Constantine 2 University, Nouvelle ville Ali Mendjeli, Constantine, Algeria
wafa.nouar@univ-constantine2.dz
**LIRE Laboratory, Abdelhamid Mehri Constantine 2 University, Nouvelle ville Ali Mendjeli, Constantine, Algeria
zizette.boufaida@univ-constantine2.dz

Abstract. Thanks to the development of the Semantic Web, a lot of new structured data has become available on the Web in the form of knowledge Bases (KBs). Semantic Question Answering System (SQAS) provide intuitive access to structured data via natural language and shield end users from technical aspects related to data modelling, vocabularies and query languages. Question answering systems involve natural language. Thus, the former inherits the challenges involved in processing natural language. One of these challenges is dealing with ambiguities. The ambiguity problem can be classified into four types, lexical ambiguity, structural ambiguity, semantic ambiguity, and pragmatic ambiguity. In this paper we propose a prototype of an open-domain question answering system over Linked Data that is able to treat complex natural language questions, equipped with a disambiguation solution for the lexical ambiguity resulting from the phase mapping in order to select the correct meaning.

1 INTRODUCTION

As a part of the Semantic Web, Linked Data means using the Web to connect related data. A large amount of data from various domains such as government, education, life sciences, art and others were made available in the context of the Linked Open Data (LOD) initiative built around DBpedia Auer et al. (2007). The latter is one of the central linked data datasets in LOD, one of the greatest challenges of this new big set of data is querying it.

Nowadays, with the growing amount of knowledge in LOD, interest in question answering over structured data is quickly regaining interest. SQAS was the study concerning Information Retrieval (IR), Information Extraction (IE), and Natural Language Processing (NLP) with the purpose of helping the user to access the information through the natural language and to obtain the concise, meaningful, and needed information from linked resources Hakimov et al. (2013). SQAS aim to bridge the gap between the user and the data, by translat-

ing between information need expressed in natural language on the one hand and structured queries and answers on the other hand Unger et al. (2014).

Such SQAS involve natural language. Thus, the former inherits the challenges involved in processing natural language. One of these challenges is dealing with ambiguities. Ambiguity covers all cases in which a natural language expression can have more than one meaning, in our case can map to more than one vocabulary element in the target dataset. Only one mapping is appropriate, often depending on the context Unger et al. (2014).

Ambiguity is a pervasive phenomenon in natural language, in the context of SQAS it affects the precision. The ambiguity problem can be classified into four types, lexical ambiguity, structural ambiguity, semantic ambiguity, and pragmatic ambiguity. One form of ambiguity that can arise during the mapping phase is the **lexical ambiguity**. This kind of ambiguity results from the interpretation of single words and not from their structure. For example, the word "bank" has ten different senses as a noun alone inWordNet (a lexical database structured as a semantic network).

In the LOD initiative the information comes from different ontologies, lacking a semantic mapping among them and many ontologies describe similar domains with different terminologies, making the resource ambiguity problem of increasing importance when heterogeneous resources or Linked Data are queried. In other words, the systems now need to deal not only with how to map certain terms to the ontology concepts but it also needs to disambiguate and decide which ontology should provide the best answer. So, without a **disambiguation solution**, the SQAS may not return an answer due to a failure of mapping.

Thereby we find that the lexical ambiguity arises during the phase mapping is the most addressed ambiguity by SQAS. Furthermore, in the literature we find that most SQAS deal only on simple questions and discards complex one. However, by exploiting the structure provided by the knowledge graph (KG) and extracting relationship between entities, we can also answer complex questions that require multiple joins, corresponding to paths in the KG. In this paper we present an approach to overcome this type of ambiguity in complex questions using knowledge-based Word Sense Disambiguation (WSD) and relying on NLP techniques.

The rest of this paper is organized as follows: we introduce some useful background notions in Section II. In Section III, we discuss related work in disambiguation in SQAS. Section IV presents our proposition, which is a SQAS that transforms complex naturel questions into formal representation that would be easy interpreted as SPARQL query. Finally, section V concludes the paper and suggests some future work.

2 BACKGROUND

This section presents an overview of some relevant concepts that are word sense disambiguation, and complex queries, which will be used in the rest of the paper, before ex- posing some details concerning SQAS.

2.1 Complex Queries

Simple questions can most often be answered by translating into a set of simple triple pat- tern. Problems arise when several facts have to be found out, connected and then

combined respectively the resulting query has to obey certain restrictions or modalities like a result order, aggregated or filtered results Höffner et al. (2016). In this SQAS we target complex queries, particularly relationship queries that are questions involving multiple relations between queried entities, which in the classical IR setting would require combining cues from multiple documents to obtain an answer. As argued by Yin et al. (2016), question answering over KB falls into two types, namely single-relation question answering and multi-relation question answering. Single-relation questions, such as "How old is Obama?", can be answered by finding one fact triple in KB, and this task has been widely studied. In comparison, reasoning over multiple fact triples is required to answer multi-relation questions where more than one entity and relation are mentioned and the answer can be obtained by the intersection of results from multiple path queries. For instance, the question "*Name a soccer player who plays at forward position at the club Borussia Dortmund*." has a possible answer as the intersection of results from two path queries FORWARD→ plays position → Marco Reus and Borussia-Dortmund → plays-in-club → Marco Reus. Compared with single-relation question answering, multi-relation question answering is yet to be addressed.

The system also answers more complex questions where the queried entities are not directly related to the entities given in the question but are linked to them through a chain of relations. For instance, rather than asking "Who invented x-bar theory?" (Answer: "Noam Chomsky") and then using the result of that question in a follow-up question "Where does Noam Chomsky work?" ("MIT"), we can directly find the result with the question: ""Where does the person who invented x-bar theory work?".

2.2 Word Sense Disambiguation

WSD is a particular case of disambiguation. WSD is the process of identifying the senses of word in textual context, when word has multiple meanings. The most appropriate meaning for a word is selected from a predefined set of possibilities, usually known as a sense inventory.WSD techniques use the notion of context in order to decide a particular word sense, a context could differ widely across WSD methods. One may consider a whole text, a word window, a sentence or some specific words.

Navigli (2009) classified methods applied to this research work into three main approaches: Supervised WSD which uses machine learning techniques to learn a classifier from labeled training sets. Unsupervised WSD which rely on unlabeled corpora, and do not exploit any manually sense-tagged corpus to provide a sense choice for a word in context and Knowledge-based approach, this type of approaches depends on external knowledge sources that provide necessary information to associate senses with words. Knowledge sources can vary from corpora of texts, either unlabeled or annotated with word senses, to machine readable dictionaries, thesauri, glossaries, ontologies, etc.

3 RELATED WORK

In the past years, several question answering systems were published working on-top of linked data. In this section we will describe some SQAS equipped with a disambiguation solution.

Damljanovic et al. (2011) in FREyA (Feedback, Refinement and Extended VocabularY Aggregation) which is a SQAS, aims to investigate whether user interaction coupled with deeper syntactic analysis and usability methods such as feedback and clarification dialogs can be used in combination to improve the performance of Natural Language Interfaces to ontologies. The system attempt to improve recall by enriching the domain lexicon from the user's vocabulary and improves precision by resolving ambiguities more effectively through the dialog. In a first step, it generates a syntactic parse tree in order to identify the answer type. The processing then starts with a lookup, annotating query terms with ontology concepts using an ontology-based gazetteer. In case of ambiguities, FREYA resolves it by assigning a confidence score for each resource and requires users to manually select the resource. The suggestions shown to the user which are the disambiguation candidates are found through ontology reasoning and are initially ranked using the combination of string similarity. The user's selections are saved and used for training the system in order to improve its performance over time. However, FREyA relies heavily on user feedback for the disambiguation and, that interaction requires all users to be familiar with KB structures.

Moreover, in PowerAqua, the component PowerMap provides automatic mapping for inter-ontology concepts and semantic relevance analysis, it calculates semantic relatedness as the distance between corresponding senses in Wordnet's graph in order to determine the similarity between elements in the query and predicates, subjects or objects in the knowledge base, more or less aiming at computing a bijective mapping between the elements of the query and resources or predicates. The researcher in PowerAqua uses word sense disambiguation techniques to disambiguate various interpretations across heterogeneous resources Lopez et al. (2012). PowerAqua's main strength is that it locates and integrates information from different, heterogeneous semantic resources, relying on iterative algorithms, query disambiguation, and ranking and fusion of answers. Its main weakaness, on the other hand that is lacks a deep linguistic analysis and cannot handle complex queries because it cannot capture its structure. Due to limitations in the linguistic tools GATE it cannot cope with aggregation, i.e. questions involving counting, comparisons, and superlatives.

In the system DEANNA, Yahya et al. (2012) have managed phrase detection, entity recognition and entity disambiguation by formulating the SQAS task as an Integer Linear Programming (ILP) problem which is an optimization tool. The authors have used ILP to addresses address the ambiguity of the phase mapping and some ambiguity that arises during the segmentation. It employs semantic coherence which measures co-occurrence of resources in the same context. DEANNA constructs a disambiguation graph which encodes the selection of candidates for resources and properties. The optimization function includes three terms. The first increases if the label of a resource is similar to the corresponding segment. The second increases if two selected resources often occur in the same context. The third tries to maximize the number of selected segments. The follow-up approach Yahya et al. (2013) uses DBpedia and Yago with a mapping of input queries to semantic relations based on text search. At QALD-2, it outperformed almost every other system on factoid questions and every other system on list questions. However, the approach requires detailed textual descriptions of entities and only creates basic graph pattern queries. The main disadvantage is that some dependencies between the segments have to be computed in the question analysis phase.

Furthermore Shekarpour et al. (2013, 2014) have developed the system called SINA, which is one of the initial studies to automate resource disambiguation. The researchers studied entity disambiguation by keyword segmentation in detecting the compatible ontology.

The aim is to maximize the high textual similarity of keywords to resources along with relatedness between the resources. The problem is cast as a Hidden Markov Model (HMM) with the states representing the set of candidate resources extended by OWL reasoning. The transition probabilities are based on the shortest path between the resources. The Viterbi algorithm generates an optimal path though the HMM that is used for disambiguation. The system also presents a novel method for constructing formal queries using disambiguated resources and leveraging the interlinking structure of the underlying datasets. An advantage of this technique is that it is not necessary to know the dependency between the different resources. However, the major drawback is that keyword does not always allow precise specification of the user's intent and lacks a clear specification of the relations among the different entities, so the result sets that are returned may be unmanageably large and of limited relevance so is not adapted for the context of the linked data.

The authors in LiQuate presented a tool to assess the quality related to both incompleteness of links, and ambiguities among labels and links. This quality evaluation is based on queries to a Bayesian Network that models RDF data and dependencies among properties, this probabilistic model is used in LiQuate to reduce the occurrence of ambiguity by performing quality validation on different resources. The different resources may have redundant labels or missing links. The returned probabilities can suggest ambiguities or possible incompleteness in the data or links Ruckhaus et al. (2014).

Table 1 show existing disambiguation solutions for SQAS and indicates our technique.

SQAS	**Disambiguation technique**
Damljanov, et al. (2011)	Assigning confidence score for each resource
Lopez, et al. (2012)	Word sense and triple similarity service
Yahya et al. (2013)	Integer linear program
Shekarpour et al. (2013,2014)	Hidden Markov model and federated queries
Ruckhaus et al. (2014)	Analyse quality of linked data using Bayesian Network
Our approach (2018)	WSD for complex questions using Lesk algorithm and DBpedia abstarct

TAB. 1–*Existing disambiguation solutions and our proposed solution.*

In the literature, we can find other systems that resolve the lexical ambiguity, but they don't deal with complex questions. Other SQAS involve the users to interact with the system to resolve ambiguities. Our proposed SQAS focuses on automating the disambiguation processes and does not need any user feedback. Up to now, WSD approaches have been mostly developed and tested using WordNet. Our disambiguation solution uses the DBpedia datasets, and its abstracts (dbo:abstract) rather than the semantic labels (rdfs:label). This has the potential advantage of providing richer contextual representation as DBpedia abstracts normally contain additional implicit semantics that are not available in the rdfs:label.

4 OUR PROPOSITION

The arrival of huge structured knowledge repositories has opened up opportunities for answering complex questions, involving multiple relations between queried entities, by operating directly on the semantic representations rather than finding answers in natural language texts.

In this section, we present the proposed SQAS. Figure 1 illustrates the proposed SQAS architecture. Our goal is to translate the complex questions into a formal query. The following sections explain in more detail the system component.

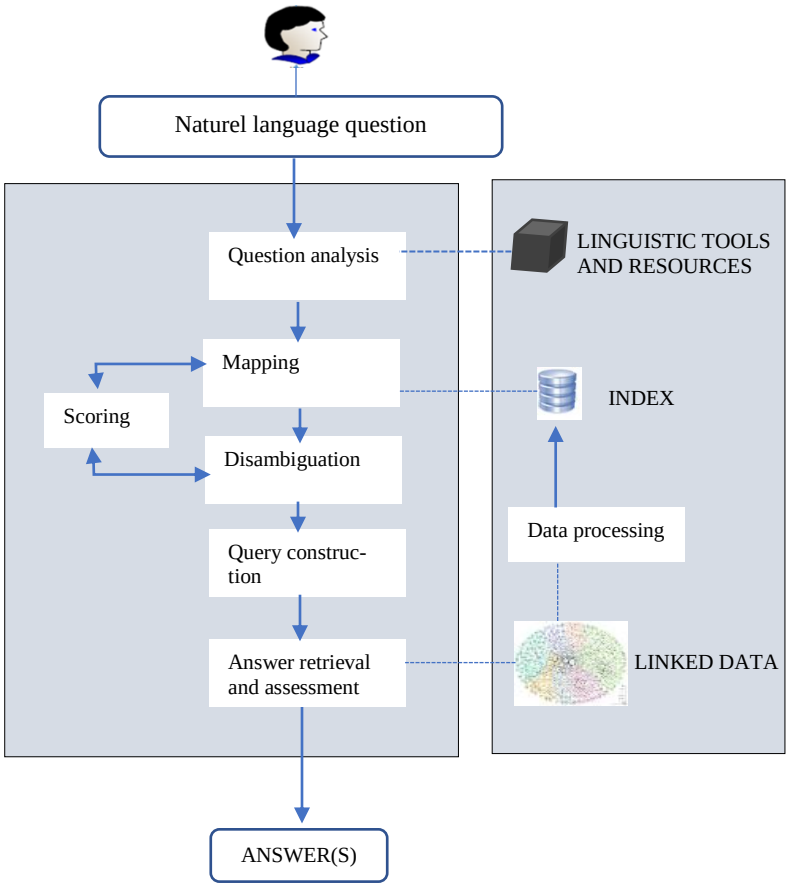


FIG. 1 – *Architecture of the proposed SQAS.*

The scope of the SQAS is open domain, addressing general knowledge. The KB used by the system is DBpedia. The DBpedia dataset is based on cross-domain ontology with most concepts representing places, persons, work, species, and organizations. The ontology was mostly extracted from infoboxes in Wikipedia. Each DBpedia resource is described by a

label, a short and long English abstract, and a link to corresponding Wikipedia page and a link to the image representation of the resource, when available.

The SQAS relies on an index of the dataset in order to match natural language expressions with labels of vocabulary elements. Also reprocessing the information present in a dataset helps to reduce the runtime of a system. The pipeline of the proposed SAQS consists of four phases:

4.1 Question analysis

Consists on the linguistic analysis of the question. A syntactic analysis is executed to determine its structure. In this SQAS, we initiate the syntactic analysis using an existing Naturel Language (NL) analysis tool Stanford CoreNLP (SCNLP) to determine the part of speech (POS) of each word, word dependencies. It contains:

- **Step 1**: Lemmatisation is the process by which a word form is returned to its basic or canonical form, in order to overcome the problems of morphological variations.
- **Step 2:** Detecting the question type, the focus and the expected answer type based on rules over part-of speech tags.
- **Step 3:** POS tags are used mainly to identify which phrases correspond to instances (subjects or objects), to properties, to classes and which phrases do not contain relevant information. To do this, we use Stanford CoreNLP to determine the part-of-speech of each word.
- **Step 4:** Dependency parsing, consists on the parse of the question using a parser based on dependency grammars, the idea behind dependency grammars is that the words in a sentence depend on each other, the parser detect relations between words.

To better understand the sequence of previous steps, let be show the following query: what are associate genes of diseases treated with Cetuximab?

POS Tagging: what/WP are/VBP associate/JJ genes/NNS of/IN diseases/NNS treated/VBN with/IN Cetuximab/NNP ?/.

Table 2 shows some abbreviations used by the Stanford Parser to denote POS tags

Abreviation	Explication
WP	Determiner
DT	Wh-pronoun
VBP	Verb, non3rd person singular present
IN	Preposition or subordinating conjunction
JJ	adjective
NNS	Noun, plural
VBN	Verb, past participle
NNP	Proper noun, singular

TAB. 2– *Some abbreviations used by the Stanford Parser to denote POS tags.*

Dependency parsing of the query:

- dobj(treated-7, what-1)
- auxpass(treated-7, are-2)
- amod(genes-4, associate-3)
- nsubjpass(treated-7, genes-4)
- case(diseases-6, of-5)
- nmod(genes-4, diseases-6)
- root(ROOT-0, treated-7)
- case(Cetuximab-9, with-8)
- nmod(treated-7, Cetuximab-9)

Table 3 shows some abbreviations used by the Stanford Parser to denote grammatical relationships.

Grammatical relations	Explication
dobj	direct object
auxpass	passive auxiliary
amod	adjectival modifier
root	root
nsubjpass	passive nominal subject
nsubj	nominal subject

TAB. 3– *Some abbreviations used by the Stanford Parser to denote grammatical relationships.*

- **Named Entity Recognition (NER):** refers to the task of correctly identifying atomic entities in text that fall into predefined categories such as person, location, organization, etc. To avoid missing useful phrases, we retain all n-grams (groups of n words) as candidate phrases. N-gram strategy is common strategy is to try to map n-grams in the question to entities in the underlying KB. If at least one resource is found, then the corresponding n-gram is considered as a possible named entity. We use *DBpedia* due to their wide coverage of entities. In this example entities are ***associate genes, diseases, Cetuximab.***

- **Linguistic triples generation**: in this step we need to translate natural language words into triples elements, and sentences into triples that semantically and syntactically comply with the question posed by the user. The corresponding query triples generated are:
  - ***1/ associate genes / of ?/ diseases***
  - **2/** ***diseases / treated with / Cetuximab***

We note that in this example the property in 1 is unknown as it stems from the semantically light preposition ***of.***

4.2 Mapping

In this phase the generated linguistic triples are matched against all available LOD datasets to find resources that have complete or partial answers to the question. In our approach, two types of mapping methods can be used:

- **Syntactic matching:** measures the distance between the phrase and the labels of the different resources of the KB. It uses the techniques of comparing strings, such as stemming, and by using string similarity that refers to the Levenshtein distance between the text string (such as Paris), and the labels describing the entity URIs (for example, Paris Hilton, Paris and Paris, Ontario).
- **Semantic matching:** to deal with problem when the question element and label of the resources of the KB have only a semantic relation. The SAQS use Redirects which is a way to collect new labels of a resource which is to follow the ***owl:sameAs*** links. The labels in the connected KB can be used then in the same way as in the original KB. In our case the DBpedia dataset has multiple URIs within the dataset and from other datasets connected with ***owl:sameAs*** relations and thus referring to the same concepts.

In this phase, the ambiguity occurs via two primary scenarios:

- **The first scenario** is the lexical ambiguity of the matches, it happens when there are multiple KB triples matched against the linguistic triples, the disambiguation phase must determine which of the resources identified during the phase mapping are the right ones.
- **The second scenario** of ambiguity is when there are no matches found because the linguistic triples do not map directly with any of the KB triples. In this case the SQAS enumerates the synonym of the user's terminologies to generate more relevant matches with KB concepts in the KBs.

4.3 Disambiguation

It is the core contribution of this paper, which performs disambiguation i.e. identifying the right entity among a number of entities with the same names.

The application of WSD to IR presents both computational, and effectiveness limitations. Furthermore, a typical query context, as in Web searches, may be too short for sense disambiguation. So, it became necessary to profit from the long context as feature of complex questions for the disambiguation phase. The class of complex questions we target are those composed of multiple clauses, each centered around a relationship, which collectively describe a single variable (with possibly multiple bindings in the KG). For example, the question "actors starring in The Departed who were born in Cambridge, MA" is composed of two clauses: "actors starring in The Departed" (e.g., MattDamon, MarkWahlberg, JackNicholson) and the relative clause "actors who were born in Cambridge, MA" (e.g., MattDamon, UmaThurman, MarkWahlberg). MattDamon and MarkWahlberg are answers to the complete question. The challenges include little context for mentions in questions. To overcome this problem, we exploit the context of complex question.

Many of the LOD datasets have textual definitions attached to resources like DBpedia dataset. Based on this remark, we adopt the basic ideas from typical Lesk algorithm and Bag-of-Words (BOW) approach.

4.3.1 Lesk algorithm

Lesk (1986) is a classical knowledge-based WSD algorithm which disambiguates a word by selecting a sense whose definition overlaps the most with the words in its context. The Lesk algorithm is based on the assumption that words in a given neighborhood (section of text) will tend to share a common topic. A simplified version of the Lesk algorithm is to compare the dictionary definition of an ambiguous word with the terms contained in its neighborhood. Versions have been adapted to use WordNet. An implementation might look like this:

- For every sense of the word being disambiguated one should count the amount of words that are in both neighborhood of that word and in the dictionary definition of that sense.
- The sense that is to be chosen is the sense which has the biggest number of this count.

Simplified LESK Algorithm:

```
Function SIMPLIFIED LESK (word, sentence) returns best sense of word
best-sense <- most frequent sense for word
max-overlap <- 0
context <- set of words in sentence
for each sense in senses of word do
signature <- set of words in the gloss and examples of sense
overlap <- COMPUTEOVERLAP (signature, context)
    if overlap > max-overlap then
        max-overlap <- overlap
        best-sense <- sense
end return (best-sense)
```

FIG. 2 – *Simplified LESK Algorithm* Vasilescu et al. (2004).

The COMPUTEOVERLAP function returns the number of words in common between two sets, ignoring function words or other words on a stop list.

Unfortunately, Lesk's approach is very sensitive to the exact wording of definitions, so the absence of a certain word can radically change the results. Further, the algorithm determines overlaps only among the glosses of the senses being considered. This is a significant limitation in that dictionary glosses tend to be fairly short and do not provide sufficient vocabulary to relate fine-grained sense distinctions. So, in our solution, we suggest using DBpedia abstract.

4.3.2 Bag-of-Words

The Bag-of-Words approach is a model, used in NLP, to find out the actual meaning of a word having different meaning due to different contexts. In this approach, there is a bag for each sense of a keyword (disambiguated word) and all the bags are manually populated. When the meaning of a keyword would be disambiguated, the sentence (containing the keyword) is picked up and the entire sentence would be broken into separate words. Then, each word of the sentence (except stop words) would be compared with each word of each "sense" bags searching for the maximum frequency of words in common.

4.3.3 Creation of glosses from DBpedia

In our disambiguation solution, Lesk algorithm takes advantage of resource definitions provided by DBpedia. The glosses of entities are taken from DBpedia abstracts. In DBpedia short abstracts (frst paragraph) are represented using **rdfs:comment** and a long abstract (text before a table of contents, at most 500 words) are represented using the property **dbpedia:abstract** from each article.

Property	Value
rdfs:comment	• Cetuximab (INN) is an epidermal growth factor receptor (EGFR) inhibitor used for the treatment of metastatic colorectal cancer, metastatic non-small cell lung cancer and head and neck cancer…
dbo:abstract	• Cetuximab (INN) is an epidermal growth factor receptor (EGFR) inhibitor used for the treatment of metastatic colorectal cancer, metastatic non-small cell lung cancer and head and neck cancer. Cetuximab is a chimeric (mouse/human) monoclonal antibody given by intravenous infusion that is distributed under the trade name Erbitux in the U.S. and Canada by the drug company Bristol-Myers Squibb and outside the U.S. and Canada by the drug company Merck KGaA. In Japan, Merck KGaA, Bristol-Myers Squibb and Eli Lilly have a co-distribution…

TAB. 4– *Example DBpedia representation of the concept Cetuximab.*

4.4 Query construction

To construct the query, we make the assumption that it is possible to deduce the SPARQL query from the structure of the question, using candidate RDF triples extracted from natural language questions and the dependency graph and POS tags of the question. The SPARQL query generated will be send to an endpoint. Figure 3 show a diagram that model the behavior of this SQAS.

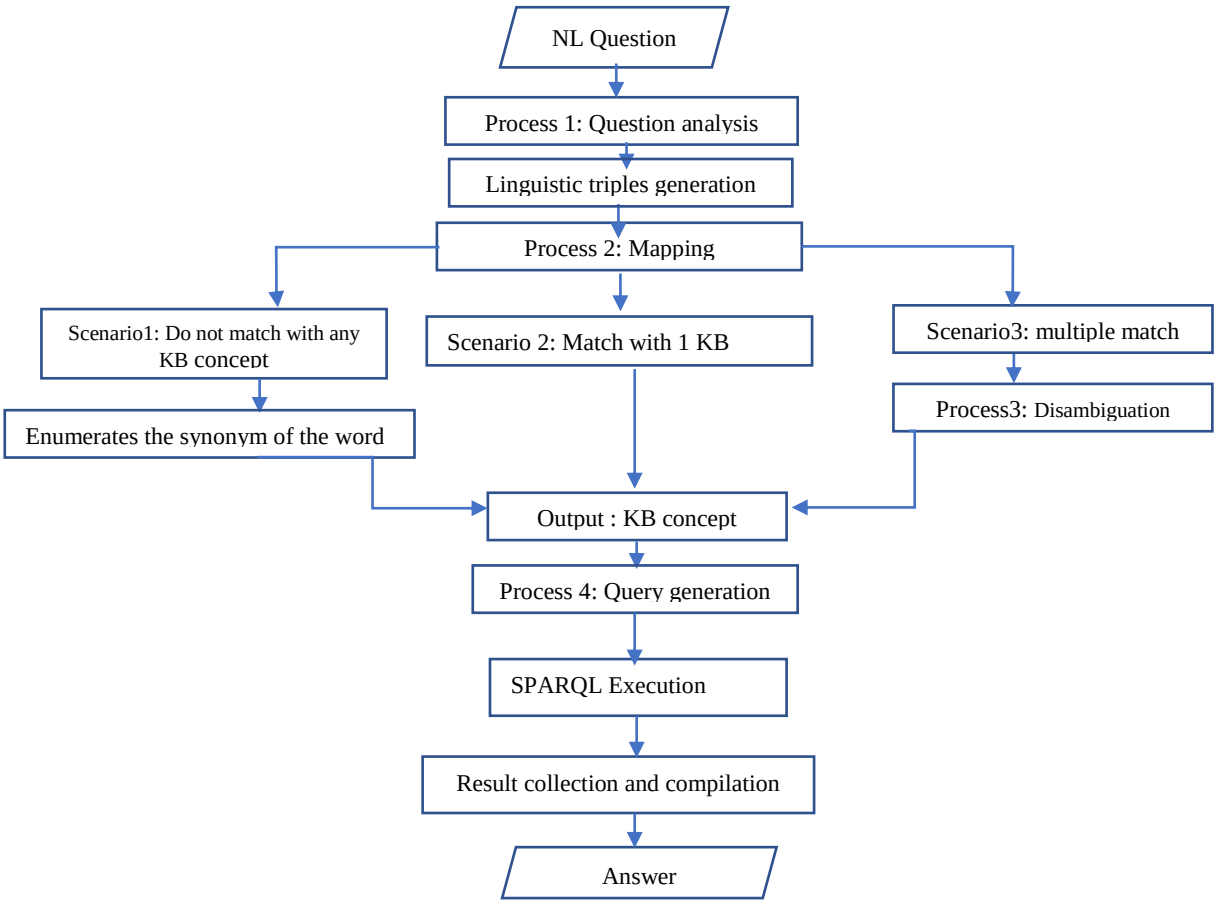


FIG. 3 – *Ambiguity scenarios in a SQAS.*

5 Case study

We use DBpedia to verify if a named entity is ambiguous using the dbo:wikiPageDisambiguates property, and to extract all possible senses of an ambiguous entity, then select one of them as the correct one.

Consider for example the question **Where is born the person Who directed the Lord of the Rings**?. In this example the named entity "The Lord of the Rings" could refer to different entities: The Lord of the Rings (book), The_Lord_of_the_Rings (film). In order to select the correct meaning we take these two glosses from DBpedia abstract.

dbo:abstract (book)	The Lord of the Rings is an epic high-fantasy novel written by English author J. R. R. Tolkien. The story began as a sequel to Tolkien's 1937 fantasy novel The Hobbit, but eventually developed into a much larger work. Written in

	stages between 1937 and 1949, The Lord of the Rings is one of the best-selling novels ever written, with over 150 million copies sold…
dbo:abstract (film)	The Lord of the Rings is a 1978 American high fantasy animated film directed by Ralph Bakshi. It is an adaptation of J. R. R. Tolkien's high fantasy epic The Lord of the Rings, comprising The Fellowship of the Ring and the first half of The Two Towers. Set in Middle-earth, the film follows a group of hobbits, elves, men, dwarves, and wizards who form a fellowship…

TAB 5*: The glosses taken from DBpedia abstract.*

After matching all word (and not only the named entity) of the query to DBpedia abstract, the correct meaning of the named entity The Lord of the Rings in the query is determined by locating the sense that overlaps the most between the glosses of the named entity and the given context. So, in this case "The Lord of the Rings" is clearly referring to the film and not to the book because there is a word in common with DBpedia abstract extracted from the page of The_Lord_of_the_Rings (film), which is the property "directed ". After disambiguation phase the SPARQL query is generated:

```
SELECT DISTINCT  ?c WHERE {
?c   rdf : type   dbo:location.
?p   rdf : type   dbo:person.
?p   dbp:directed   res: The_Lord_of_the_Rings.
?c   dbp:birthPlace   ?p.
}
```

We use the following prefixes: dbo for http://dbpedia.org/ontology/, dbp for http://dbpedia.org/property/, and res for http://dbpedia.org/resource/.

6 CONCLUSION

Users pose NL questions based on their writing skills, and a SQAS must have the capability to process a variety of NL questions that are unknown during the SQAS design and development stage. Interestingly, it would seem very easy to answer complex question by executing a structured query over available knowledge bases or the existing and rapidly increasing set of Linked Data sources.

In this work we opted for the disambiguation of complex questions, and we have introduced a methodology for translating natural language questions into structured queries over Linked Data. The proposed SQAS combines multiple techniques of NLP. So as future work we intend to develop a LOD based semantic relatedness measure to perform disambiguation by picking a candidate solution that maximizes the total relatedness.

REFERENCES

Auer, S., C. Bizer, G. Kobilarov, J. Lehmann, and Z. Ives (2007). DBpedia: A Nucleus for a Web of Open Data *in: Proceedings of the 6th International Semantic Web Conference.*

Damljanovic, D., M .Agatonovic, H .Cunningham (2011). FREyA: an interactive way of querying linked data using natural language. *In the Semantic Web: ESWC 2011 Workshops, Springer.*

Hakimov, S., H. Tunc, M. Akimaliev, and E. Dogdu (2013). Semantic question answering system over linked data using relational patterns. *In: Proceedings of EDBT/ICDT, Genoa.*

Höffner, K., S. Walter, E. Marx, J. Lehmann, A. Ngonga, and R. Usbeck(2016) . Overcoming challenges of semantic question answering in the semantic web. *Semantic Web Journal.*

Lesk, M-E (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. *In Proceedings of the SIGDOC Conference, Toronto.*

Lopez, V., M. Fernández, E. Motta, N. Stieler (2012). PowerAqua: supporting users in querying and exploring the semantic web. *Semantic Web,3: 249–265*.

Navigli, R (2009). Word sense disambiguation: A survey. *ACM Computational Surveys, 41(2).*

Ruckhaus, E., M.-E. Vidal, S. Castillo, O. Burguillos and O. Baldizán (2013). Analyzing Linked Data Quality with LiQuate. *In: OTM Workshops*.

Shekarpour, S., A-C. Ngonga Ngomo, and S. Auer (2013). Question answering on interlinked data. *In Proceedings of the 22nd International World Wide Web Conference (WWW), 1145_1156*.

Shekarpour, S., E. Marx, A-C. Ngonga Ngomo, and S. Auer (2014). SINA: semantic interpretation of user queries for question answering on interlinked data. *Web Semantics: Science, Services and Agents on the world wide web 39–51.*

Unger, C., A. Freitas, and P. Cimiano (2014). An introduction to question answering over linked data. *In Reasoning Web International Summer School. Springer, 100–140.*

Vasilescu, F., P. Langlais, G. Lapalme (2004). Evaluating Variants of the Lesk Approach for Disambiguating Words. *LREC, Portugal.*

Wenpeng,Y., M. Yu, B. Xiang, B. Zhou, and H. Sch¨utze (2016). Simple question answering by attentive convolutional neural network. *International Conference on Computational Linguistics, pages 1746–1756.*

Yahya, M., K. Berberich, S. Elbassuoni, and G. Weikum (2013). Robust question answering over the web of linked data. *In Proceedings of CIKM, 1107–1116*.

Yahya, M., K. Berberich, S. Elbassuoni, M. Ramanath, V. Tresp, and G. Weikum (2012). Natural language questions for the web of data. *In Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL), 379–390.*

Détection des intrusions et aide à la décision

David PIERROT*, Nouria HARBI**, Jérôme DARMONT**

*Université de Lyon, ERIC EA 3083
5 avenue Pierre Mendès France F69676 Bron Cedex France
david.pierrot1@univ-lyon2.fr
**{nouria.harbi,jerome.darmont}@univ-lyon2.fr

Résumé. Les conséquences d'une intrusion dans un système d'information peuvent s'avéver problématiques pour l'existence d'une entreprise ou d'une organisation. Les impacts sont synonymes d'une perte financière, d'image de marque et de sérieux. La détection d'une intrusion n'est pas une finalité en soit, la réduction du delta détection-réaction est devenue prioritaire. Nous proposons une méthode prenant en compte les aspects techniques par l'utilisation d'une méthode hybride de Data mining mais aussi les aspects fonctionnels. L'addition de ces deux aspects permet d'obtenir une vision générale sur l'hygiène du système d'information mais aussi une orientation sur la surveillance et les corrections à apporter.

1 Introduction

La détection d'intrusions est devenue une priorité pour garantir le maintien opérationnel d'un système d'information. Au cours des dernières décennies, de nombreuses approches ont été créées pour détecter les intrusions. Il existe deux catégories principales de détection d'intrusions : les abus et les anomalies. La détection des abus est basée sur des modèles spécifiques identifiés par des signatures. Cette méthode souffre de signatures manquantes pour les attaques inconnues. L'inconvénient principal qui est déjà largement démontré repose sur les faux positifs. Cette problématique était déjà évoquée lors de la conférence Black Hat en 2006 (Zambon et Bolzoni, 2006). Il n'est donc pas étonnant de retrouver cette difficulté en 2018 comme le soulignent parfaitement H.Rais et T.Mehmood. (2018)

La seconde méthode de détection basée sur les anomalies est obtenue à partir d'un modèle de comportement normal : une dérivation significative générera une alerte.

De nombreuses approches ont été développées pour améliorer la détection des anomalies à l'aide de méthodes de statistiques ou d'exploration de données (data mining). Cependant, la plupart des approches nécessitent de capturer le trafic réseau avec des outils spécifiques. Ceci peut avoir un impact sur la performance globale du réseau. La détection d'une intrusion est certainement possible, mais sa prise en compte et son traitement restent nébuleux. Les aspects de sécurité fonctionnelle tels que la gestion des risques ne sont jamais exploités. Ce point peut avoir un impact sur une prise de décision rapide faute de connaissance de l'actif visé. Par conséquent, les IDS existantes et les travaux de recherche sont assez difficiles à mettre en

œuvre sans un effort significatif et soutenu. L'objectif de cet article est : 1) d'analyser et d'expliquer l'état actuel des pratiques de détection d'intrusions ; 2) de discuter du travail que nous avons effectué pour faciliter la visualisation du flux de données du système d'information et la détection d'intrusions / d'attaques. Dans la continuité du document de David Pierrot (2016), nous proposons d'ajouter la mise en œuvre de la gestion des risques et l'utilisation des résultats obtenus lors d'un audit de sécurité. Notre principale contribution est l'extraction et l'analyse de journaux de type Firewall en combinant des méthodes de data mining pour détecter automatiquement les dérivations de comportements et les abus. Nous intégrerons le résultat d'une analyse de risque afin de cibler les actifs les plus sensibles visés par un comportement anormal. Notre travail tente de prouver qu'il est possible sans sonde de détection ou base de signatures de détecter des intrusions et d'appliquer une méthode de prise en charge.

2 Étude de l'existant

Le Data mining offre diverses solutions avantageuses pour réaliser la détection et l'analyse des intrusions. Trois types d'approches d'exploration de données sont possibles comme le précisent Deepa et Kavitha (2012). L'apprentissage supervisé utilise un ensemble de données labellisées pour effectuer des prédictions (classification ou régression). La labellisation des données peut être assistée d'un expert. L'apprentissage non supervisé est appliqué sur des ensembles de données non étiquetés pour découvrir des similitudes. Enfin, l'apprentissage semi-supervisé est une combinaison de techniques d'apprentissage sur les données étiquetées et non étiquetées.

Il est également possible d'utiliser des méthodes hybrides combinant un apprentissage supervisé et non supervisé.

Pour tester les différentes méthodes d'apprentissage, des "benchmarks" (jeux de données réels ou synthétiques) ont été proposés.

KDD99 est sans doute le plus cité et utilisé (University of California, Irvine, 1999) et ceci avec une existence de 20 années. Il est construit à partir de sept semaines de trafic réseau capturé avec TCPdump (Garcia, 2017). Il fournit des données étiquetées pour quatre types d'attaques :

— DOS : déni de service ;
— R2L : accès non autorisé depuis une machine distante ;
— U2R : accès non autorisé aux privilèges du super-utilisateur local (root) ;
— Probe (sonde) : surveillance.

NSL-KDD (`http://www.unb.ca/cic/research/datasets/nsl.html`) est similaire à KDD99. Il est dénué de doublons et avec un nombre d'enregistrements limité. Ainsi, les méthodes d'apprentissage peuvent être entièrement utilisées dans des délais raisonnables (Elkhadir et al., 2016). Enfin, le jeu de données ORNL (`https://www.ornl.gov`) propose une capture de réseau à partir de l'IDS open-source Snort qui présente les mêmes attaques que KDD99 (Cisco, 2017). Ces trois ensembles de données sont basés sur la bibliothèque libpcap généralement utilisée pour les captures du trafic réseau (Garcia, 2017). L'acquisition de données reste donc similaire pour ces derniers. Nous pouvons définir cette acquisition comme symétrique, c'est à dire que la volumétrie de capture et de sauvegarde correspond à la quantité des flux entrants et sortants.

2.1 Détection des intrusions par l'apprentissage supervisé

Les réseaux de neurones artificiels (ANN : Analysis Neural Networks) peuvent être utilisés pour détecter les cyber-menaces (Bognar, 2016). Le thème central de cette approche est d'apprendre et de modéliser le comportement en imitant le cerveau humain. Ces expériences ont été menées sur KDD99. Les 41 premiers attributs sont considérés comme des valeurs d'entrée exogènes. Le 49 $^{\text{ième}}$ attribut est la valeur labellisée, simplifiée avec la valeur 1 en cas d'attaque, et 0 pour les activités normales. L'algorithme de rétropropagation (backpropagation) de Levenberg-Marquardt (Levenberg, 1944; Marquardt, 1963) est utilisé comme méthode d'apprentissage supervisé avec dix "layer" cachés. Il peut prédire les attaques et les comportements normaux avec seulement 3% de faux positifs. Cependant, les ANN présentent des inconvénients intrinsèques tels qu'une lenteur de convergence et la nécessité d'une importante quantité de données d'apprentissage. De plus, des problèmes de sur-apprentissage doivent être traités.

Les arbres de décision (DT) peuvent également être utilisés pour la détection d'intrusions. L'algorithme J48 (une extension de C4.5) a été appliqué sur le jeu de données ORNL. Bien que les arbres de décision obtiennent une meilleure précision que les autres méthodes d'apprentissage supervisées telles que Naive Bayes ou Support Vector Machines (Gupta et al., 2016), les principaux inconvénients sont la consommation de puissance de calcul et de mémoire.

2.2 Détection des intrusions par l'apprentissage non-supervisé

La détection d'intrusions basée sur des méthodes de data mining non-supervisées permettent une automatisation complète. Ces dernières n'exigent pas d'intervention d'expert. Lee et al. (1999) ont travaillé sur la classification, du méta-apprentissage et des règles d'association à partir de données d'audit. L'objectif est de calculer des modèles qui capturent avec précision le comportement des intrusions et des activités normales selon une fréquence temporelle. De telles méthodes ont tendance à générer beaucoup de règles d'association et donc a augmenter exponentiellement le niveau de complexité du système. Un modèle améliorant la précision de détection des intrusions a été obtenu en utilisant l'algorithme DBSCAN (Ajboye et al., 2015). Il s'agit d'identifier des points au sein d'une classe (cluster). DBSCAN nécessite deux paramètres, le rayon et les points requis minimum à l'intérieur d'une classe, et ceci pour déterminer sa taille. Le résultat fournit des régions denses. Si certaines instances n'appartiennent à aucune classe, elles sont considérées comme aberrantes. La détermination du rayon ainsi que le nombre de points reste difficile à estimer Cet algorithme nécessite de fortes ressources de mémoire et de calcul.

2.3 Détection des intrusions par une méthode d'apprentissage hybride

L'utilisation des méthodes data mining uniques donne des résultats limités (Tanpure et al., 2016). Ainsi, des approches hybrides ont été proposées ces dernières années. Il s'agit en fait de mixer les deux méthodes d'apprentissage (supervisée et non supervisée). Une IDS a été développée sur un ensemble de méthodes utilisant l'algorithme K-means, les réseaux de neurones flous (FNN) et les arbre de décisions (C4.5). K-means attribue une valeur (38 type d'attaques de KDD99) au cluster avec les centroïdes les plus proches. FNN apprend les paramètres avec une rétropropagation plus rapide que ANN pour obtenir la classification dans les catégories d'attaques KDD99. Enfin, C4.5 utilise la nouvelle sortie d'étiquette de FNN pour classer les

données entrantes et détecter les attaques réseau (Meghana et Dhamdhere, 2015). Les principaux inconvénients de cette approche résident à nouveau dans le besoin d'une grande puissance de calcul et dans la détermination du paramètre k classe de K-means. L'IDS APMINING repose sur un algorithme de règles d'association appliqué à une collection de connexions sans attaque de KDD99 vers des données de trafic entrant (Emna Bahri, 2013). Le résultat fournit deux profils : normal et anormal. Ensuite, la labellisation est affinée avec une méthode supervisée pour obtenir une classification finale (normale ou attaque). Cependant, le processus de génération des règles d'association est relativement long avec une consommation mémoire importante. Pour renforcer l'exploration de données sur KDD99, un moteur d'analyse basé sur quatre parties utilise une IDS, le jeu de données KDD99, la classification Naive Bayes et le K-means (Tanpure et al., 2016). Les données des sondes aident à améliorer la base de données KDD99. Si des données (flux réseau) sont déjà présentes dans la base de données, K-means est utilisé pour déterminer s')il s'agit d'une activité malveillante ou normale. En cas d'activité malveillante, un message d'alerte est envoyé. Si les données sont nouvelles, K-means est également sollicité. Le résultat est envoyé à un classificateur Naive Bayes pour analyser les relations potentielles. Ce principe de fonctionnement est intéressant car il tient compte à la fois du comportement et des relations. Cependant, il est difficile d'ajouter une nouvelle attaque dans la base de données KDD99 sans expertise. Chaque paquet réseau doit être analysé et comparé avec les contenus existants.

Une autre méthode hybride mise au point par Sunita et al. (2016) consiste à utiliser l'ANN et le Fuzzy C-Means (FCM). L'ANN et la rétropropagation sont utilisés pour l'apprentissage supervisé et FCM pour la formation des classes. Le trafic réseau entrant est classé par l'AAN avec six neurones et quatre "layer" cachés. Le résultat est envoyé à la méthode FCM qui est configurée avec cinq classes correspondant à KDD99 (quatre types d'attaques et une de trafic normal). Ce système résout les problèmes relatifs à la classification, mais il n'intègre pas de nouvelles attaques. Ce point est considéré comme fastidieux et chronophage. De plus, il requiert également de l'expertise.

L'utilisation de l'analyse des composants principaux du noyau (KPCA) avec le K-plus proche voisin (KNN) pour la détection d'intrusion a été étudiée par Elkhadir et al. (2016). Cette approche démontre la supériorité du KPCA sur l'analyse du composant principal (ACP). KPCA est une forme non linéaire de l'ACP et effectue une réduction des dimensions. Le résultat du KPCA (taux de détection) est ensuite envoyé à KNN. Enfin, les données sont classées en attaque ou en connexion normale. Le choix de KNN est discutable, puisque cette méthode est coûteuse en calcul et que les données volumineuses sont souvent la norme dans les captures de paquets réseau.

Nous pouvons noter que toutes les études existantes sont basées sur des outils de capture de paquets. Comme nous l'avons déjà souligné, les outils de capture de paquets ont besoin d'expertise et d'une capacité d'absorption de trafic réseau selon la quantité de flux. Afin de reproduire les différents travaux existants, il est obligatoire d'utiliser les outils de capture de paquets et l'ensemble de données comme KDD99 ou NSL-KKD. Dans ces deux ensembles de données, R2L (accès non autorisé) et U2R (accès non autorisé aux privilèges du super utilisateur local) peuvent être considérés comme incomplets ou même hors de propos. Les outils de capture de paquets sont généralement aveugles devant une connexion cryptographiquement sécurisée entre deux hôtes. Les connexions sécurisées rendent difficile l'analyse de l'escalade des privilèges, comme des bugs ou des shellshock (exécuter des commandes arbitraires pour

obtenir un accès non autorisé à un serveur).

3 Approche combinée pour la détection des intrusions

3.1 Motivations et propositions

La gestion de sécurité étant relativement coûteuse, il convient que cette dernière soit abordable par tous et de limiter l'intervention des experts. La détection des anomalies ou des intrusions doit être suffisamment compréhensible afin d'automatiser au maximum les actions en découlant.

Notre étude portera sur quatre phases qui couvrent un spectre relativement large. Ces phases se décomposent de la façon suivante :

— Phase 1 : "Monitoring et visualisation" des données réseau.
— Phase 2 : Analyse des comportements et alertes, phase qui s'appuiera sur des méthodes de Data Mining.
— Phase 3 : "Scoring" des risques et phase d'évaluation.
— Phase 4 : Détermination d'un plan d'actions.

Nous avons opté pour une phase monitoring/visualisation, qui est une approche conventionnelle et une évaluation du risque déterminée par l'analyse du comportement. Enfin, le plan d'action permet d'arrêter toute action anormale. La nouveauté réside principalement dans la petite quantité d'informations nécessaires et la définition des classes de comportement sans une détermination en apriori.

Les travaux présentés dans le précédent chapitre sont basés sur des flux provenant de jeux de données. De ce fait, nous ne pouvons reproduire des résultats similaires car, par défaut, les variables étudiées (durant la phase 1) sont inférieures au nombre de variables utilisées par KDD99 (41 variables). Il est donc judicieux d'utiliser certaines méthodes de Data Mining sur des événements de type "Firewall" pour la détection des anomalies. Le challenge repose sur la possibilité, à partir d'un équipement de filtrage qui par sa nature ne peut délivrer autant d'information qu'une sonde, d'identifier les intrusions.

3.2 Les séquences d'une intrusion

Avant de présenter nos travaux, nous souhaitons décrire brièvement l'organisation d'une attaque. La méthode présentée ci-dessous est dite "éthique", elle est basée sur les cinq étapes suivantes.

1. La reconnaissance, basée sur une recherche d'informations à partir d'Internet.
2. Balayage réseau (inventaire des services et ports, des systèmes d'exploitation et des versions logiciels serveurs utilisés).
3. Obtenir l'accès (exploitation des vulnérabilités et obtenir un accès).
4. Maintenir l'accès (rendre l'accès permanent).
5. Couvrir les traces (effacer et réduire les traces).

La première étape est difficilement détectable car dépendante de la vie numérique d'une entreprise ou de son personnel. Nos travaux seront axés sur les étapes deux et trois. Toute attaque informatique commence généralement par une prise de renseignements (phase 2).

Il existe des méthodes plus détaillées comme la Cyber Kill Chain (Yadav et Mallari, 2016) qui aborde la notion de pivot et implicitement d'APT [1]. Dans les faits, les actions importantes restent similaires.

3.3 Phase 1 : Monitoring et visualisation

Cette phase est utilisée pour fournir une visualisation complète des données aux utilisateurs concernés (ingénieurs de sécurité, analystes de réseau, responsables de la sécurité de l'information). Les captures ou sondes étant relativement lourdes en déploiement, nous avons opté pour l'utilisation des journaux issus d'un Firewall. La mission d'un Firewall est de filtrer le trafic réseau selon une politique basée sur le flux autorisé dans un réseau sur son origine, sa destination et les services souhaités (Al-Shaer et Hamed, 2003).

Grâce à sa position, un Firewall offre une visibilité complète et selon les points suivants :

— Adresse IP source, adresse IP de destination.
— Port de destination et protocole.
— Date à laquelle le Firewall a appliqué une règle de filtrage.
— Numéro de règle de filtrage (ID) et actions du Firewall (acceptées ou rejetées)

Les journaux d'accès sont envoyés à un serveur syslog-ng (Balabit, 2016). Par la suite, un traitement de découpage des différents champs est réalisé via des expressions régulières (PCRE). Le script Afterglow (Marty, 2013) et la librairie Graphviz (ATT, 2016) sont utilisés pour créer des graphiques.

3.4 Phase 2 : Analyse des comportements et alertes

La phase 2 permet l'analyse des données et la détection d'un comportement anormal. Nous exploitons les données décrites dans la section 3.3. Nous effectuons une transformation afin de réduire les modalités. A titre d'exemple, la variable de port de destination dispose de 65535 modalités. Nous choisissons de regrouper les variables dans les trois catégories suivantes :

— les ports inférieurs à 1024 acceptés et refusés ;
— ports allant de 1024 à 65535 acceptés et refusés ;
— ports d'administration (portadm), activité sur les ports d'administration d'un actif (acceptés et refusés).

Nous réalisons un agrégat des actions réalisées par les adresses IP source. Ceci permet de considérer le nombre total de transactions effectuées par la même adresse IP source et le nombre de flux rejetés (action refusée) et autorisés (action autorisée) par le Firewall.

Nous avons demandé l'avis de cinq experts pour déterminer si le comportement observé pouvait être défini comme à risque (étiquetage des agrégats proposés dans le tableau 1).

L'apprentissage supervisé nous donne la possibilité de construire un estimateur qui peut prédire le risque à partir des adresses IP source. L'analyse des experts nous donne une image de la politique de sécurité sous la forme d'un ensemble de données d'apprentissage. Pour obtenir la meilleure précision de détection, nous avons testé différents algorithmes d'apprentissage supervisé en utilisant la validation croisée (10 fold cross validation). Le meilleur résultat a été obtenu en utilisant l'algorithme Random Forest (Table 2).

1. Advanced Persistent Threat : Menace Permanente Avancé

TAB. 1 – *Données agrégées selon l'IP source*

	Variable name	Description
1	ipsrc	Adresse IP source
2	nbr	Nombre d'occurence de l'adresse IP source est presente
	cnbripdst	Nombre de d'adresse IP différentes contactées
	cnportdst	Nombre de ports de destination contactés
	permit	Nombre de d'occurrence autorisées par le Firewall
	deny	Nombre de d'occurrence rejetées par le Firewall
3	inf1024permit	Nombre de port <1024 autorisés par le Firewall
	sup1024permit	Nombre de port $\geq$ 1024 autorisés par le Firewall
	adminpermit	Nombre de ports d'administration (21, 22, 23, 3389, 3306) accepté par le Firewall
	inf1024deny	Nombre de ports <1024 rejetés par le Firewall
	sup1024deny	Nombre de ports $\geq$ 1024 rejetés par le Firewall
	admindeny	Nombre de ports d'administration (21, 22, 23, 3389, 3306) rejetés par le Firewall
4	risk	Variable à prédire et étiquetée par l'expert

TAB. 2 – *Comparaison des taux d'erreur en apprentissage supervisé*

Algorithme	Taux d'erreurs (%)
CART	9.09
C4.5	11.3
C5.0	8.5
Random forest	8.2
Boosting (Adaboost)	14.3

Durant cette phase et comme dans la plupart des travaux, nous nous concentrons uniquement sur l'adresse IP source. Or, il existe des systèmes d'anonymisation comme "Tor" ou l'utilisation de VPN anonymes. Malgré cela, notre premier cycle d'analyse nous informe que le système d'information (IP de destination) est analysé en vue d'une potentielle intrusion.

En utilisant deux méthodes d'apprentissage différentes, il devient possible d'avoir une vue d'ensemble sur les dérivations de comportements ainsi que sur les actifs visés (Meesala et Xavier, 2015). La première étape consiste à créer un jeu de données et d'extraire les flux reçus sur chaque serveur. L'étape suivante repose sur l'identification des comportements différents. Nous voulons regrouper le comportement d'IP source en fonction de l'adresse IP de destination. Cette analyse de classe s'articule autour du concept de placement d'un ensemble d'objets dans le même groupe ou classe. En suivant cette méthodologie, il est possible d'identifier et de vérifier tout écart de comportement.

Le principal problème avec les méthodes de clustering est de déterminer le nombre de classes à utiliser. Pour trouver la meilleure méthode déterminant le bon nombre de classes, nous utilisons la validation interne. Nous avons testé différents algorithmes et après avoir analysé les résultats, nous avons opté pour le Partitioning Around Medoids (PAM) décrit par Lamiaa et Manal (2013). Cette méthode a fourni le meilleur résultat par sa rapidité et l'ensemble des définitions de k classe. La figure 1 montre les résultats obtenus par rapport aux méthodes utilisant k classe en apriori. Le temps de calcul pour la détermination des classes pour 53 serveurs a été de 0.353 secondes. Ce résultat est utilisé en tant que référentiel. En d'autres termes,

chaque serveur (actif) dispose d'un nombre de classes de comportement et toute dérivation sera considérée comme anormale. Nous avons réalisé différents tests d'activité malveillante et nous avons pu constater une diminution des classes sur les actifs visés. Ceci confirme le principe d'attaque. Pour confirmer le résultat obtenu, nous avons utilisé le coefficient de variation (CV). Il permet d'obtenir un score Breunig (2001) et il est possible de se focaliser sur l'actif subissant le comportement le plus déviant.

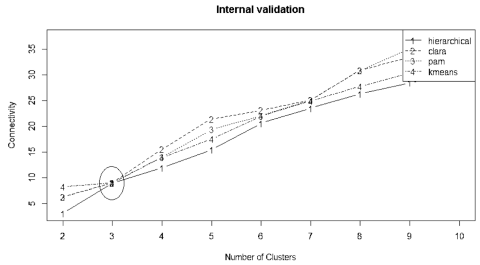


FIG. 1 – *Validation interne pour un serveur*

3.5 Phase 3 : Scoring du risque et évaluation

Il existe plusieurs méthodes d'analyse de risques. Dans les faits, peu importe la méthode comme le précise la norme ISO 27001 (ISO 27001 :2013 chapitre 9.1 point b) du moment que cette dernière soit documentée et reproductible. Nous utilisons la méthode EBIOS qui traite de l'analyse contextuelle en fonction de la dépendance du système d'information (DCSSI, 2003). Il ne s'agit pas de décrire une méthode d'évaluation des risques, mais de savoir comment utiliser le résultat obtenu et de l'inclure dans notre cadre de recherche. La figure 2 montre un graphique réalisé à partir d'une analyse de risques EBIOS. Il a été décidé que l'acceptation du risque résiduel est fixée à un niveau inférieure à 6.

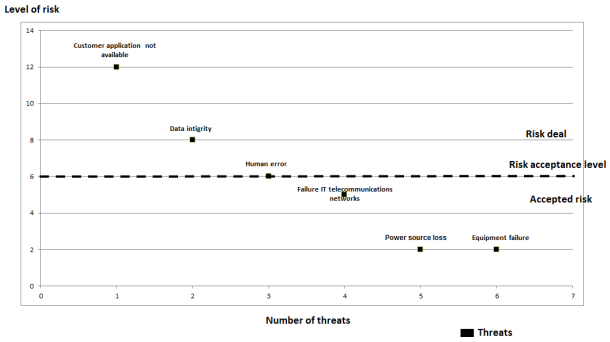


FIG. 2 – *Exemple graphique d'une analyse de risques*

Nous pouvons utiliser le niveau d'acceptation du risque de chaque actif comme indicateur de gravité. La détection du nouveau comportement à partir d'une adresse IP source identifiée comme à risque (phase 2 : Random Forest et PAM) pourra être priorisée (score du risque phase 3). Avec ces trois différentes informations, nous construisons un indicateur d'attaque (IOA : violation de la politique de sécurité phase 2, dérivation du comportement phase 2).

3.6 Phase 4 : Détermination du plan d'actions

A l'aide de la phase 3, nous savons exactement quels actifs protéger selon le "scoring" du risque. Il est souvent recommandé de réaliser une évaluation des vulnérabilités et des tests de pénétration.

Les audits visent à indiquer la gravité des vulnérabilités. Une vulnérabilité est qualifiée de faible, moyenne, élevée ou critique selon un catalogue de menaces de sécurité connues. Au cours de cette phase, nous utilisons le référentiel construit dans les phases 2 (classe IP de destination) et 3 (évaluation des risques). Nous implémentons les quatre qualifications d'audit qui correspondent à la gravité de la vulnérabilité sur chaque actif (IP de destination). Pour ce faire, nous utilisons le scanner de sécurité OpenVas largement utilisé par la communauté de la sécurité. Le rapport d'audit est exporté au format XML et intégré dans notre référentiel. De plus, à la suite du rapport, les CVE (vulnérabilités et expositions communes) sont listées. Ces dernières sont une liste de vulnérabilités avec les actions à effectuer pour corriger les faiblesses de sécurité établies. A l'issue de cette phase, nous obtenons le résultat suivant :

— serveur (IP de destination) nombre de classes ;
— numéro de niveau de risque ;
— nombre de vulnérabilités élevées trouvées ;
— nombre de vulnérabilités moyennes trouvées ;
— nombre de vulnérabilités faibles trouvées.

La notation des risques et le plan d'actions doivent permettre d'identifier et de prioriser une série d'actions prédéterminées.

4 Preuve de concept (POC)

4.1 Cas d'utilisation

Nous avons travaillé sur l'architecture d'une entreprise du domaine de la santé constituée de plusieurs dizaines d'employés. Notre analyse se concentre sur trois réseaux interconnectés au sein d'un réseau étendu (WAN), et protégé par un équipement de filtrage. Ainsi, nous avons pu créer plusieurs jeux de données incluant des attaques dans une échelle temporelle relativement longue (d'une à quarante huit heures).

4.2 Résultats

L'exécution de la Phase 1 permet de visualiser l'activité du réseau avec un outil graphique, avec une meilleure compréhension que l'examen des données brutes. Nous estimons que les résultats peuvent être considérés comme satisfaisants. Nous proposons une interface simple dans laquelle les flux de données réseau peuvent être évalués (diagnostic ou pertinence de la politique de filtrage). Nous avons intégré les journaux Firewall vers un conteneur Syslog-Ng et nous avons appliqué l'apprentissage supervisé de Random Forest. Le résultat fournit une variable "risque" avec 2 modalités (Oui et Non). Nous pouvons détecter une activité malveillante (Table 3, ligne 3) : le flux autorisé de 0,7 % pourrait être défini comme un marqueur d'une intrusion. Lors des tests de méthodes non supervisées, l'ACP et le regroupement agglomératif hiérarchique ont fourni une bonne visualisation graphique. Par conséquent, nous avons décidé d'intégrer ces fonctions dans notre outil D113. De plus, l'évaluation visuelle de la technique

TAB. 3 – *Aggregate flow with risk analysis result*

Sum	Action denied	Action allowed	Inf 1024	Sup 1024	Adm ports	Risk
16	0.0	100.0	0.0	0.0	0.0	No
12	0.0	100.0	0.0	0.0	0.0	No
3296	99.3	**0.7**	61.9	38	0.1	**Yes**
36	100.0	0.0	0.0	100	0.0	Yes

de la tendance (VAT), pour identifier visuellement le nombre de classes (Bezdek et Hathaway, 2002) nous donne une confirmation graphique du nombre de classes. Les trois graphiques (PCA, VAT, Dendrogram) montrent l'existence de 3 classes. PAMK donne les mêmes résultats sans aucune interprétation humaine.

4.3 Retour d'experts

Le tableau 4 montre le retour des experts sur l'utilité de notre système, suivant l'approche de Ghoniem et al. (2014). Nous avons demandé à cinq experts (deux ingénieurs de sécurité E1 et E2, un chef de la sécurité des informations E3, un consultant en sécurité E4 et un analyste de réseau E5) d'évaluer sur une note comprise entre 0 et 5 les quatre phases de notre approche. De leur point de vue, l'outil D113 offre la possibilité d'accéder à tous les flux rejetés, y compris le balayage des ports et les attaques par force brute. La représentation des flux acceptés donne un aperçu global intéressant.

TAB. 4 – *Overview of expert feedback*

Question about the usefulness of	E1	E2	E3	E4	E5
Visualization (Phase 1)	5	4	3	4	4
Policy derivation (Phase 2)	5	4	4	4	5
Behavior derivation (Phase 2)	5	5	5	4	5
Risk management (Phase 3)	3	4	4	2	3
Action plan (Phase 4)	3	4	3	4	3

La combinaison de l'apprentissage supervisé et non supervisé ainsi que la notation des risques permet d'identifier les serveurs affectés et leurs importances dans le système d'information. L'utilité des Phases 3 et 4 n'a pas été remise en question, mais nécessite un suivi régulier et un travail supplémentaire.

5 Conclusion et perspectives

Nous avons proposé IDS mixant les apprentissages supervisés et non supervisés. En utilisant ceci, nous sommes capables de détecter les violations de la politique de sécurité et les dérivations de comportement. L'ajout de l'analyse des risques et de l'audit de vulnérabilité permettent de nous concentrer sur les serveurs les plus sensibles avec les vulnérabilités clairement identifiées. Par conséquent, la connaissance complète du système d'information est moindre. Selon les experts, la manière de présenter l'information (IP source à risque, modification du comportement sur l'IP de destination et notation des risques) doit être plus facile à comprendre. De plus, les attaques orientées WEB doivent être prises en compte. Les algorithmes Bagging, K-mean++ doivent être testés pour améliorer les méthodes d'apprentissage.

Références

Ajboye, A. et al. (2015). Anomaly Detection in Dataset for Improved Model Accuracy Using DBSCAN Clustering Algorithm. pp. 39–46.

Al-Shaer, E. et H. Hamed (2003). Firewall policy advisor for anomaly detection and rules. In *International Symposium on Integrated Network Management*, Volume 118, pp. 17–30.

ATT (Last accessed September 10, 2016). Graphviz – Graph Visualization Software. http://www.graphviz.org.

Balabit (Last accessed September 03, 2016). Reliable, scalable, secure central log management. https://www.balabit.com/log-management.

Bezdek, J. C. et R. J. Hathaway (2002). VAT: A Tool for Visual Assessment of (Cluster) Tendency. In *International Joint Conference on Neural Networks*, pp. 2225–2230.

Bognar, E. (2016). Data mining in cyber threat analysis neural networks for intrusion detection. Volume 15, pp. 187–197.

Breunig, R. (2001). An almost unbiased estimator of the coefficient of variation. *Economics Letters*, 15–19.

Cisco (2017). Snort. https://www.snort.org.

David Pierrot, Nouria Harbi, J. D. (2016). Hybrid intrusion detection in information systems. In *3rd International Conference on Information Science and Security (ICISS 16), Pattaya, Thailand*, pp. 27–32.

DCSSI (2003). The EBIOS method – Expression of Needs and Identification of Security Objectives. https://www.ssi.gouv.fr/archive/en/confidence/documents/methods/ebiosv2-methode-plaquette-2003-09-01_en.pdf.

Deepa, A. J. et V. Kavitha (2012). A comprehenseive survey on approaches to intrusion detection system. *International Conference on Modelling Optimization and Computing 38*, 2063 – 2069.

Elkhadir, Z. et al. (2016). Intrusion Detection System Using PCA and Kernel PCA Methods. *International Journal of Computer Science 43*, 72–79.

Emna Bahri, N. H. (2013). Real detection intrusion using supervised and unsupervised learning. pp. 321–326.

Garcia, L. M. (2017). TCPdump and Libpcap. http://www.tcpdump.org.

Ghoniem, M. et al. (2014). VAFLE: Visual Analytics of Firewall Log Events. In *Visualization and Data Analysis*.

Gupta, M. et al. (2016). Intrusion Detection Using Decision Tree Based Data Mining Technique. *International Journal for Research in Applied Science and Engineering Technology 4*, 24–28.

H.Rais et T.Mehmood. (2018). Dynamic ant colony system with three level update feature selection for intrusion detection, injs. *nternational Journal of Network Security, 20*, 184–192.

Lamiaa, F. et H. Manal (2013). Using Modified Partitioning Around Medoids Clustering Technique in Mobile Network Planning. *International Journal of Computer Science Issues 9*,

10–25.

Lee, W. et al. (1999). A data mining framework for building intrusion detection model. In *IEEE Symposium on Security and Privacy*, pp. 120–132.

Levenberg, K. (1944). *A method for the solution of certain problems in least squares*, Volume 2.

Marquardt, D. (1963). *An algorithm for least-squares estimation of nonlinear parameters*, Volume 11.

Marty, R. (2013). AfterGlow. http://afterglow.sourceforge.net.

Meesala, S. et B. Xavier (2015). A Hybrid Intrusion Detection System Based on C5.0 Decision Tree and One-Class SVM. *International Journal of Current Engineering and Technology 5*, 59–70.

Meghana, S. et V. Dhamdhere (2015). Hybrid approach for Intrusion Detection Using Data Mining. *International Journal of Innovative Research in Science, Engineering and Technology 4*, 5588–5595.

Nguyen, H. et al. (2011). An efficient Local Region and Clustering-Based Ensemble System for Intrusion Detection. In *15th International Database Engineering and Applications Symposium*, Volume 185–191.

Sunita, S. et al. (2016). A Hybrid approach of Intrusion Detection using ANN and FCM. *European Journal Advances in Engineering and Technology 3*, 6–14.

Tanpure, S. et al. (2016). Intrusion detection system in data mining using hybrid approach. *National Conference on Advances in Computing, Communication and Networking 5*, 18–21.

University of California, Irvine (1999). KDD Cup 1999 Data. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

Yadav, S. et D. Mallari (2016). Technical Aspects of Cyber Kill Chain. *Communications in Computer and Information Science 536*, 438–452.

Zambon, E. et D. Bolzoni (2006). Network intrusion detection systems.

Summary

The consequences of an intrusion into an information system can be problematic for the existence of a company or an organization. The impacts are synonymous with financial loss, brand loss and seriousness. The detection of an intrusion is not an end in itself, the reduction of the delta detection-reaction has become a priority. We propose a method dealing with the technical aspects by the use of a hybrid method of data mining but also the functional aspects. The addition of these two aspects makes it possible to obtain a general vision on the hygiene of the information system but also an orientation on the monitoring and the corrections to be made.

Chapter

6

Enterprise Data-driven: Big data and Digital transformation

ASD'2018

Content

Strategic analytics for the agile and smart enterprise

Jabir Brahim*
Noureddine Falih*
Khalid Rahmani*
* LIMATI Laboratory, Polydisciplinary Faculty
Mghila, BP 592
Beni Mellal, Morocco
ibra.jabir@gmail.com
nourfald@yahoo.fr
kh.rahmani@hotmail.fr
http://fp.usms.ac.ma/

Abstract. Helping enterprise take a rational and quick decision and make consumer achieve their business goals are the real trends. For this purpose, business analytics has been emerged and exploit data through various tools to extract a 'value'.
Indeed, no one can deny that the business analytics can help enterprises to stay competitive and efficient, but the existed analytic solutions still experiencing problems of speed and accuracy and do not have reached the desired objectives. Therefore, it is the time to think about new strategic analytics which can be faster than existed traditional ones. In this paper, we present a literature review of this strategic analytics concept to help the enterprise to succeed in its processes with more agility and intelligence. In other future works, we will present our contribution to concretize this notion of strategic analytics by a specific and original approach.

1 Introduction

Enterprises collect billions of data and try to exploit it to improve and enhance their competitiveness. That brings us to the 1970s (Power, 2005) when the first application was designed to support decision-making based on data collected. Over the years, various applications and approaches of decision-making (Saaty, 2008) appeared and expanded this domain of business intelligence, but collecting and analyzing the right data to make the right decisions with the existed analytic solutions and strategies often need costly statistical and complicated analysis. So the traditional analytic solutions are neither so simple nor so fast. That is the reason why we should look towards new strategy analytics delivering informed, accurate answers to strategic business questions (Steads, 2000) in the right time, and simplify the complex tasks of prediction; many research focused on this domain of business analytics for enterprises. We can legitimately present the different approaches always built about strategic analytics for agile and smart enterprise, and propose an instead analytic strategy that can reduce the time needed for the overall cycle of collecting, analyzing, and acting on enterprise data.

2 From a traditional enterprise to a smart and agile one

In the socio-economic world, the transition from traditional enterprise to smart and agile one (Larson and Chang, 2016) requires some methods and approaches defined under a solution which is business analytics; this science boosts traditional enterprise to come smart and agile using data analytics to achieve their objectives with fast and intelligent manner.

As presented in figure 1, another group of researchers (Evans and Lindner, 2012) claims that conjunction of three main approaches used for a long time is the primary key that has been contributed to appear the smart and agile enterprises. These approaches provide insights and predictions and help to accord profit, revenue, and shareholder return, and these three approaches are: statistical methods, business intelligence, Modeling/optimization, these three approaches are the perspective of advanced business analytics for a smarter company.

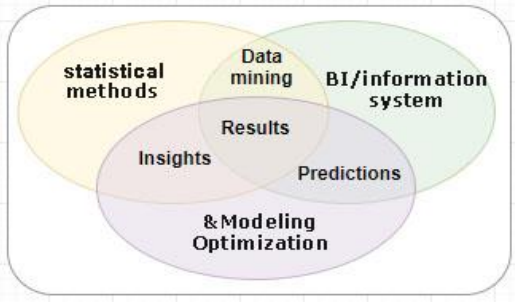


FIG. 1 – *Agile and smart enterprise perspectives*

3 Business analytics

In this section, we define business analytics (BA) as a critical part of the strategic analytics for smart and agile enterprises.

3.1 Overview

We can define Business Analytics as an advanced concept from business intelligence. It is also an approach uniting various disciplines to get value from data to make better decisions (Kohavi and al., 2002). The delivering of this value attained via systematic analysis with different strategies to develop existing processes, identify new opportunities, discovering more product features, changing and evolving new services and systems, better understand the behavior of customers, and expect problems before they happen. These disciplines are computer science, statistics, data management, decision science, and scientific research methods ...

Other researchers affirm that business analytics is the point where advanced analytics technique has been emerged and operates to collect data, discover predictions, and provide insights to take a fast, easy and better decision (Hota, 2011). The principal purpose of business analytics is helping enterprises to build an appropriate analytic strategy with an approach making the enterprise more agile and smart (Acito, 2014).

The figure 2 below shows that the principal function of BA is to extract value from different formats of data (structured unstructured and semi-structured) collected in several data sources and distribute it to the appropriate people, in the correct formats at the right time.

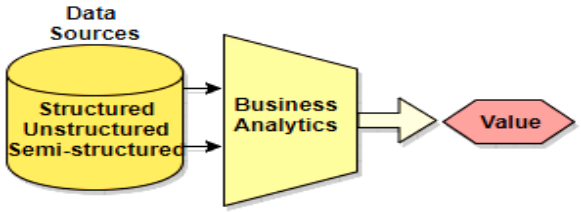


FIG. 2 – *Extracting value from several forms of Data.*

Business analytics is regularly inspected from three critical perspectives: descriptive, predictive, and prescriptive (Evans and Lindner, 2012).

3.1.1 Descriptive analytics

Also called business reporting, it is the most used and most well-understood type of analytic (Banerjee and al., 2013), business start with it to answer the question of ' what happened and/or what is happening? It is a unique tool contribute to identify, collect, filter and classify data to convert it into value for understanding and analyzing business performance, the principal output of this type is the determination of business problems opportunities.

Example of enterprise questions answered by this analytics:

- What is the percentage of sales for each month?
- What was our revenue last year?
- How many and what types of complaints did we resolve?
- What are the less-selling products?

3.1.2 Predictive analytics

This kind of analytics come to answer the question of "what will happen and/or why will it happen?", it is an accurate projection of the future used to discover predictions and identify the reasoning as to why, by the analyzing of the past performance, examine historical data and detect patterns in the vast amount of data (Delen and Demirkan, 2013).

Predictive analytics help enterprise to predict behavior, detect trends, and expect problems before they happen, using data and mathematical methods like data mining, text mining, statistical time-series, forecasting ...

Example of the enterprise questions that predictive analytics can answer:

- What will happen if selling decrease by 5 percent?
- What will happen if supplier prices increase by 5 percent?
- What do we expect to pay for water and electricity over the next year?
- What is the risk of losing money on new business investment?

3.1.3 Prescriptive analytics

Prescriptive modeling uses data and mathematical methods to identify a set of high-value alternatives to minimize or maximize some objectives, the major outcome for this type of analytics is to answer the question of "What should I do and/or Why should I do it?" (Sun, 2015), It is the key that can lead to the best possible course of action, enhance enterprise performance through its several business areas, operations, finance, and marketing ... popular methods used include optimization modeling, simulation modeling, multi-criteria decision modeling, expert systems and group support systems (Figure 3).

Enterprise uses prescriptive analytics to answer questions such as:

- How many products do we need to maximize revenue?
- What is the best way to minimize costs and fees?
- What is the alternative plan to maintain max of profit if supplier prices increase?

The figure 3 shows the analysis levels of the business analytics which clarifies that Descriptive Analytics provide insights into the past, predictive Analytics help understanding the future and prescriptive analytics to advise on the possible outcome.

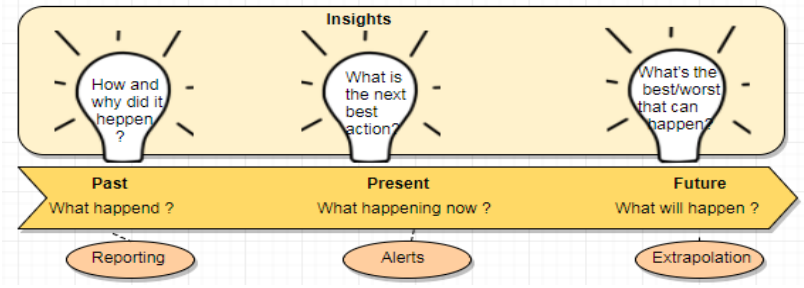


FIG. 3 – *Business analytics perspectives*

4 Strategic Analytics

Strategic analytics is an integrated analytics approach, rather than strategic planning (Klatt and al., 2011) considered as a comprehensive vision and road-map for an enterprise, helps to exploit data-dependent capabilities and provide insight and direction. Strategic analytics facilitates the realization of business objectives through reporting of data to analyze trends, creating predictive models to foresee future problems and opportunities and analyzing/optimizing business processes to enhance organizational performance faster, simple and more objective.

4.1 Strategy Analytics for smart and agile enterprises

As it mentioned in the figure 4 strategy Analytics for smart and agile enterprises, combines business planning experience with advanced analytical tools, technology businesses and add end-user research to deliver actionable information and insights that support strategic planners in dynamic high technology markets.

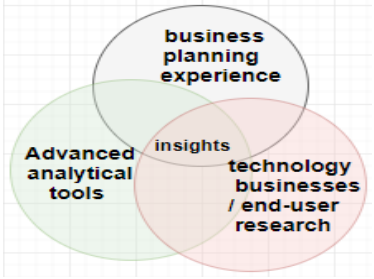


FIG. 4 – *Strategy analytic for smart and agile enterprise perspective*

In short, strategic analytics for smart and agile enterprise apply analytics in a manner consistent with the large volume of data, its format, and the speed of its processing. This concept will offer a new solution for the known problems of the current methods, and the most of them turn around: simple integrating various format of data from several sources, fast processing and providing insights, offers text and graphics visualization, and supports streaming requests (Zaharia and al., 2016). In the next chapter, we will describe technological side and the solutions existed.

5 Current Solutions

In this section, we identify and describe the existed technology solutions and other development tools that can contribute to creating and elaborating efficient strategic analytics. This strategic analytic offers analytics options to produce predictions and insights and help to make better decisions for smart and agile enterprises, we have focused on open source technologies, many of them involve the predictive analytics, descriptive analytics and prescriptive analytics which makes from it a potential strategic solution:

5.1 Statistical analysis solution

Statistical analysis methods are the tools offering the possibility of the integration organization, analysis, interpretation and presentation of data used to solve problems of economic importance and technological innovation since a long time (Miner and al., 2012), here is two popular statistical analysis solutions :

IBM SPSS. is a platform that offers advanced analytics, it is among the first analytic open source solutions introduced based on machine learning algorithms (Field, 2013), designed to provide insights and predict what will happen, in order to support enterprise finding new opportunities and minimize risk.

The figure 5 presents the distributed architecture of SPSS IBM, and it illustrates the SPSS server side where the most operations processed. The SPSS client part where the results passed once the processing is complete, it also contains a Data Base server it could be a data-warehouse, and service repository service as manage the life cycle of data mining models and related predictive objects and other operations.

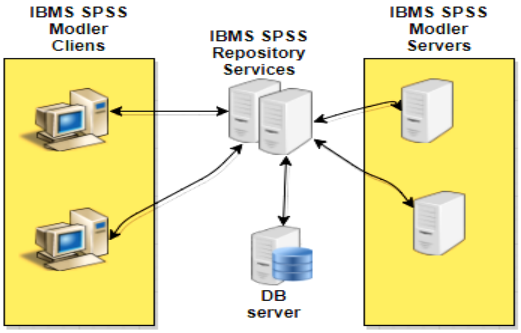


FIG. 5 – *IBM SPSS Architecture*

R. is one from the most statistical open source solutions known today, it is created as a language and environment for statistical computing. This technique inspired from S language by Ross Ihaka and Robert Gentleman (Patil, 2007), R offers a broad diversity of statistical methods as linear and nonlinear modeling, classical statistical tests, time-series analysis, clustering, also of a graphical presentation.

The figure 6 indicates that the executable Rscript of an application R, is used to execute a script "R": it takes as input a script format Core_Library_Resource_R, as well as a data source, which can be an instance of type xml. The module will attempt to display the generated files in the working directory according to their names.

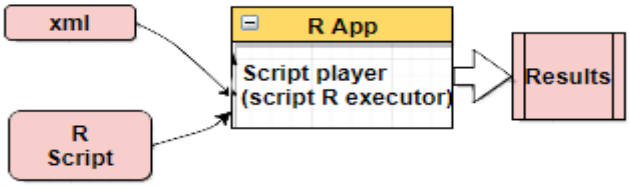


FIG. 6 – *R Statistical Architecture*

5.2 Data mining

Data mining is a potential process of discovering patterns and behaviors in amount data sets using several methods like machine learning, database systems. It is an essential process where creative methods are applied to extract value and provide advanced analytics (Hand, 2007).

We can also talk over the text analytics tools which is the special process that convert unstructured text data into data with meaning, used for analysis to measure customer behaviors, products, feedback, reviews, sentimental analysis, in fact, to support decision making. Below we present an open source solution that analyzes, evaluate and interpret data for extracting value (Tan, 1999).

Weka. is an open source technique of data machine learning includes several algorithms for data mining operations (Markov and Russel, 2006). It is an analytics solution allows enterprises to discover structure in data store systems, provide insights and predictions, and almost enhance their performance through interaction with data.

Apache mahout. is a special distributed framework created to help the implementation of the statistic and the mathematic algorithms and gain value from data (Ngersoll, 2009). Mahout can support distributed process streaming of a large amount of data, runs on systems using Hadoop core (Meng and al., 2016).

As it marked on figure 7; after data mining tool collects data from several sources, it executes it with the different techniques: regression (predictive) association Rule Discovery (descriptive), classification (predictive), clustering (descriptive) and there are many more, the results are then graphically displayed.

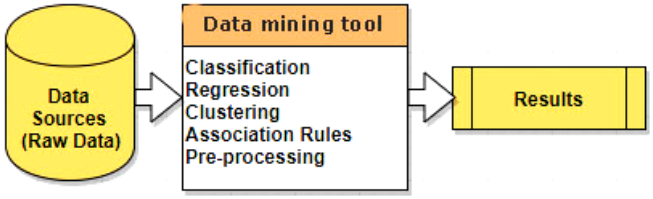


FIG. 7 – *Data mining Architecture*

5.3 Predictive data analytics

Predictive analytics is a powerful advanced analytics approach used to provide insights and make predictions about unknown future (Kelleher and al., 2015). This kind of solutions is an analytics solution that enterprise uses to predict some customer's behavior and other opportunities, many tools have appeared, for this reason, we going to shed light on:

KNIME. This analytics solution brings advanced analytics solution created in January 2004 at University of Konstanz (Berthold and al., 2009) to help enterprise discover potential hidden on data and predict the future behavior of their customers to take a rational decision.

As we shall see in figure 8, the predictive analytics system is capable of handling a large quantity of structured and unstructured data from several sources, to integrate it using a particular database connection rather than concatenation technique. After the filtering level, the data analyzed with different methods (decision tree, statistics, text analytics …). Then various forms of display can be used to visualize the results (interface table, tag cloud, scatter bar …) the final step is the deployment where the predictions and insights generated by different techniques (PMML writer…).

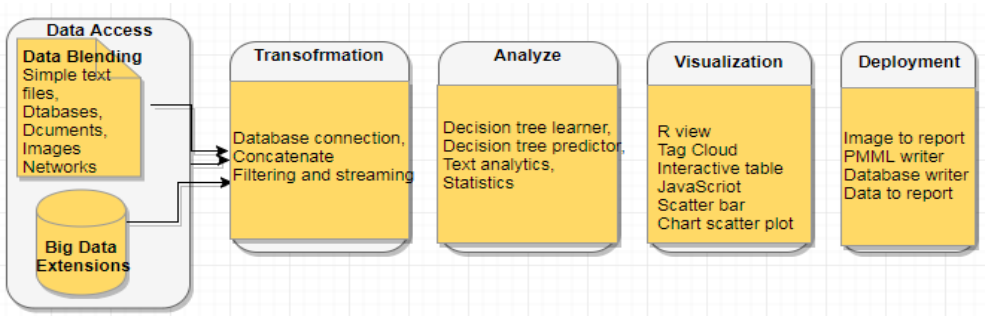


FIG. 8 – *Predictive analytics system Architecture*

5.4 Quantitative analytics

The quantitative data analysis used to convert data to numerical forms and subject them to statistical analyses. This approach can be divided into two categories; the first is descriptive statistics exploratory data analysis (EDA), second is confirmatory data analysis (CDA). The quantitative analytics provide advanced analytics tools to extract knowledge that are exploited in the corporate domain to produce value for decision-making about risk management, investments, and pricing (Julio and Ikenberry, 2004).

Quantitative analytic provides quantifiable and easy to understand results may include the calculation of frequencies of variables and differences between variables.

ELKI. Environment for Loping KDD-Applications Index-Structures is an open source of quantitative analysis written in Java; it uses to provide high performance (Schubert and al., 2015). Allows easy and fair evaluation and benchmarking of algorithms, it handles an efficient data management tools which are index-structures.

The figure 9 shed light on the quantitative analytic system that collects data from different type of sources (applications, flat files …). Later analyzed it with quantitative analytics algorithms belong to clustering based on such queries (k-means, density-based, hierarchical clustering …), it allows algorithms to be in interaction with database index structures to improve the speed of data retrieval operations.

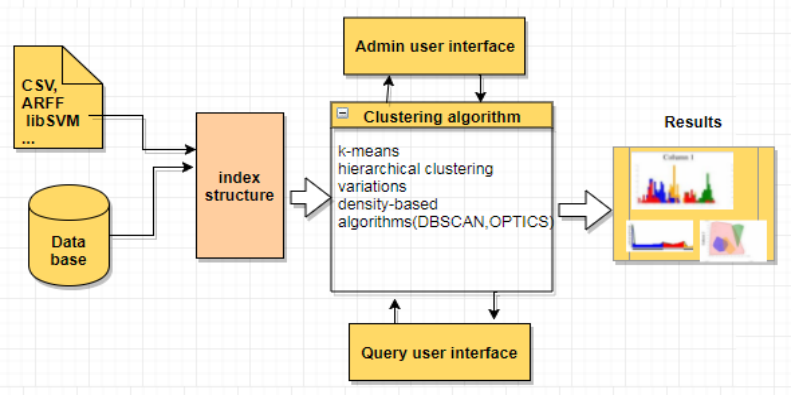


FIG. 9 – *Quantitative Data analysis system architecture*

5.5 Qualitative analytics

Qualitative data is the data that not easily reduced to numbers; it is related to concepts, opinions, values, and behaviors of people in context social, it may be a structured or unstructured text, rather than audio and video.

Qualitative data analysis is the range of the processes and procedures assist with qualitative research (discourse analysis, transcription analysis, content analysis, recursive abstraction...). This analysis gives meaning to the qualitative data that have been collected and provide such explanation, understanding or interpretation to extract value from (Lewis, 2007), the principal idea is to process and examine the meaningful and symbolic content of qualitative data QDA miner is a power technology support qualitative analytics.

QDA miner. is a special qualitative analysis environment developed by Normand Peladeau in 2004 (Péladeau, 2004). It used for coding, annotating, retrieving and analyzing data whatever its size (Lewis, 2007), it is an analytic solution adopted by enterprises to get information from data, and presents the results as a table, graphs and schema for easy comprehension to be used later in decision making.

As it clarified in figure 10, the qualitative data analytics system imports unstructured data (emails, providers, social media, RSS, files...) analyze it with different approaches (clustering, crosstabulation, frequency analysis, proximity plots...) then presenting predictions graphically.

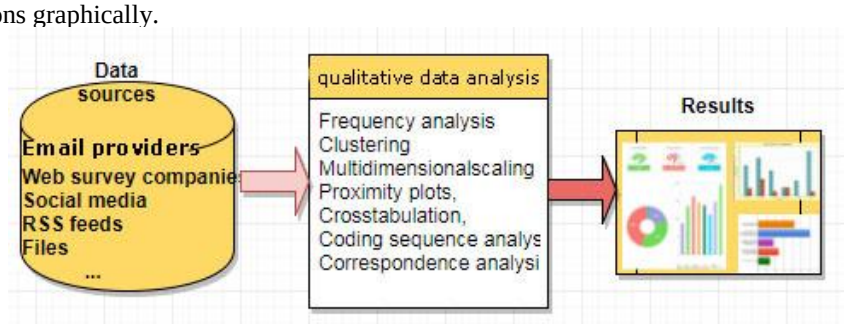


FIG. 10 – *Qualitative Data analysis system architecture*

6 Discussion

Before comparing between a several business analytics tools we will define the list criteria used to evaluate and select a BI analytics tool, below are the critical capabilities considered in the comparing process:

- Security
- Data Source
- Visualizations
- Metadata Management
- Ease of Use
- Scalability
- Speed
- Dependency
- Real-time operations
- Time of latency and Memory
- Advanced Analytics
- High fault tolerance

The table 1 below illustrates benchmarking between the different current solutions and presents the strong points and the gaps for each one based on the list criteria above.

Solutions	pros	cons
IBM SPSS	Handling huge datasets Visualize results, Good ETL capabilities Multiple data sources	Uploading /downloading datasets difficult sometimes. The graphics are not easy to manipulate. The graphic is not always accurate. Slow execution of multiple models together. Clunky with complex statistics
R Solution	Data handling and storage facility. Graphical facilities for data analysis. Free and open source	Not easy to use Less packages quality Memory management problems High consumption of memory. Poor parallelization Design flaws Difficulty handling big data Requires compilers Minimal GUI
WEKA	Open source (GNU licence) Easy to use Easy to modify Provide access to SQL databases. Provides various machine	Relatively slow (JAVA) Some features available only from the command line Curse of dimensionality Limit memory and less performance

	learning Algorithms for data mining tasks	Lacking in the representation of the results Data format constraints Weak in interfacing with other software
APACHE mahout	Real-time recommendations Large data process Distributed Support text processing High scalability	Poor visualization Less support scientific libraries Legacy dependencies to other techniques Slow Complex algorithms
KNIME	Open source platform Fast to deploy, Easy to scale	No intermediate results No interactive execution Not all nodes can be streamed
ELKI	Construct more complex statistical models Provide reliable results High scalability open source Simple Derives multiple databases	Visualization is less rich Loss of data richness Time spent on maintenance A lot of drawbacks Can't be in accordance with possible semantics. POI retrieved by the user is less than the allowed value Accuracy decrease
QDA Miner	Open Source Achieve deeper insights Support several data formats Can Deal with Large Data Sets Improves Validity/Auditability	Not easy to use Not quick Poor in analytical capabilities. Can Impose Deterministic understandings Produce Nonsensical Findings Imperfections in results Pressure to engage large data sets

TAB. 1– Discussion of the advantages and disadvantages of the analytics solutions

From the comparing table above we can undoubtedly notice that all analytic solutions existed have some limitations sometimes are obvious, and none of them offerings yet full strategic analytics for the socio-economic world. For this purpose, our contribution is realizing and establishing a new approach aims developing new strategic analytics for the smart and agile enterprise can respond to the gap of the existing solution and allows an easy and fast making decision.

7 Recommandation :

To establish successful strategic analytics solutions, the technology side is the most important; it is the main key to success the BI projects. So it is necessary to select the business intelligence analytics tool which is the best fit for the smart and agile enterprise, the table below determines our recommendations for the optimal features to evaluate, select and deploy the BA tools based on the above discussion:

Criteria	Recommendations
Security	User and user role-based security. Integrate with other pre-existing security applications. Detect and prioritize threats
Data Source	Blend data from various data sources Support multiple data sources at the same time. Support several data formats
Visualization	Rich analytic Dashboards Graphical facilities Allow export graphic results Easy to understand and manipulate Allows multiple visualizations
Metadata Management	Handle huge datasets Process a massive amount of data quickly Large data process
Ease of Use	Easy to deploy Less time to set-up Easy to manipulate Easy to scale Easy to modify Graphical facilities for data analysis Understood results Simple algorithms
Scalability	Insights accessible to users Dashboards and reports are available Easily deployed on the web
Speed	Good ETL capabilities Handling and processing data quickly Streaming access
	Powerful in interfacing with other software.

Dependency	Self-hosting compilers. Don't require others technologies
Real time operations	Real-time access Real-time processing Real-time recommendations
Time of latency and Memory	Lowest latency Efficient memory management. High performance memory. Low consumption of memory

TAB. 2– Recommendations for the optimal features to evaluate and select a BA tool to deploy a strategic analytics solution.

8 Conclusion

In this paper, we described the approach of strategic analytics for the smart and agile enterprise, as well as some, existed technique solutions and our recommendations to select business analytics tools to deploy strategic analytics for enterprises.So we release that there is always a need to realize and create a clear approach and propose an analytics solution for the smart and agile enterprise that respond to the gap of the existing ones.

Finally, new instead strategic analytics will be used to produce insights and predictions, to create a rational decision and allows efficient advanced analytics. It will be an excellent solution to help the enterprise to identify new opportunities, discover more product features, change and evolve new services and systems, better understand customers, analyze and predict each behavior and expect problems before they happen, will be our future contribution.

References

Acito, F., & Khatri, V. (2014). *Business analytics: Why now and what next*?. Elsevier, 565-570

Baysal, O., Holmes, R., & Godfrey, M. W. (2013). *Developer dashboards: The need for qualitative analytics*. IEEE software, 30(4), 46-52.

Banerjee, A., Bandyopadhyay, T., & Acharya, P. (2013). Data analytics: Hyped up aspirations or true potential*?*. Vikalpa, 38(4), 1-12.

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., ... & Wiswedel, B. (2009). *KNIME-the Konstanz information miner: version 2.0 and beyond*. AcM SIGKDD explorations Newsletter, 11(1), 26-31.

Delen, D., & Demirkan, H. (2013). *Data, information and analytics as services*. Elsevier ,359-363

Evans, J. R., & Lindner, C. H. (2012). *Business analytics: the next frontier for decision sciences*. Decision Line, 43(2), 4-6.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage.

Hand, D. J. (2007). *Principles of data mining.* Drug safety, 30(7), 621-622.

Hota, Jyotiranjan. (2011). *Business Analytics :A Tool for Organizational Transformation*. CSI Communications. 35. 21-22.

Ingersoll, G. (2009). *Introducing apache mahout. Scalable, commercialfriendly machine learning for building intelligent applications.* IBM.

Julio, B., & Ikenberry, D. L. (2004). *Reappearing dividends*. *Journal of applied corporate finance*, 16(4), 89-100.

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT Press.

Klatt, T., Schlaefke, M., & Moeller, K. (2011). *Integrating business analytics into strategic planning for better performance*. Journal of business strategy, 32(6), 30-39.

Kohavi, R., Rothleder, N. J., & Simoudis, E. (2002). *Emerging trends in business analytics*. Communications of the ACM, 45(8), 45-48.

Larson, D., & Chang, V. (2016). *A review and future direction of agile, business intelligence, analytics and data science*. International Journal of Information Management, 36(5), 700-710.

Lewis, R. B., & Maas, S. M. (2007). *QDA Miner 2.0: Mixed-model qualitative data analysis software*. Field methods, 19(1), 87-108.

Markov, Z., & Russell, I. (2006). *An introduction to the WEKA data mining system*. ACM SIGCSE Bulletin, 38(3), 367-368.

Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., ... & Xin, D. (2016). *Mllib: Machine learning in apache spark*. *The Journal of Machine Learning Research*, 17(1), 1235-1241.

Miner, G., Elder IV, J., & Hill, T. (2012). *Practical text mining and statistical analysis for non-structured text data applications.* Academic Press.

Patil, S. (2016). *Big Data Analytics Using R*. International Research Journal of Engineering and Technology (IRJET), 4.

Péladeau, N. (2004). *QDA miner qualitative data analysis software*, user's guide. Montreal: Provalis Research.

Power, D. J. (2007). *A brief history of decision support systems*. DSSResources. COM, World Wide Web, http://DSSResources. COM/history/dsshistory. html, version, 4.

Saaty, T. L. (2008). *Decision making with the analytic hierarchy process*. International journal of services sciences, 1(1), 83-98.

Schubert, E., Koos, A., Emrich, T., Züfle, A., Schmid, K. A., & Zimek, A. (2015). *A framework for clustering uncertain data*. Proceedings of the VLDB Endowment, 8(12), 1976-1979.

Stead, J. G., & Stead, E. (2000). *Eco-enterprise strategy: standing for sustainability*. Journal of Business Ethics, 24(4), 313-329.

Sun, Z., Zou, H., & Strang, K. (2015, October). *Big data analytics as a service for business intelligence*. In Conference on e-Business, e-Services and e-Society (pp. 200-211). Springer, Cham.

Tan, A. H. (1999, April). *Text mining: The state of the art and the challenges*. In Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases (Vol. 8, pp. 65-70). sn.

Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., ... & Ghodsi, A. (2016). *Apache spark: a unified engine for big data processing*. Communications of the ACM, 59(11), 56-65.

Résumé

Aujourd'hui les entreprises utilisent des outils analytiques pour l'extraction de connaissances à partir de données afin de prédire l'impact d'une décision et donc en prendre de meilleures, et tout ça pour atteindre leurs objectifs rapidement.

En effet, personne ne peut nier que « Business Analytics » aide les entreprises à rester compétitives et performante, mais les solutions analytiques qui existent pour le moment connaissent toujours des problèmes au niveau de la rapidité et de la précision et n'ont pas atteint les objectifs souhaités. Par conséquent, c'est le temps de penser à une nouvelle « Strategic analytics » qui peut être plus rapide et qui va apporter des remèdes aux lacunes connus par les solutions existantes. Dans cet article, nous présentons une revue de la littérature de ce concept « Strategic Analytics » pour aider l'entreprise à réussir dans ses processus économiques avec plus d'agilité et d'intelligence. Dans d'autres travaux futurs, nous allons présenter notre contribution pour concrétiser cette notion de « Strategic Analytics » par une approche spécifique et originale.

Structural analysis for IS performance measuring Case study

Noureddine Falih*, Azedine Boulmakoul**
*Polydisciplinary Faculty, LIMATI , B.P 592 Mghila - Benimellal – Morocco
nourfald@yahoo.fr
http://fp.usms.ac.ma/
**FST, Computer Science Department, B.P. 146 - Mohammedia - Maroc
azedine.boulmakoul@yahoo.fr
http://www.fstm.ac.ma/

Abstract. Contemporary corporate performance is measured by some key performance indicators (KPIs) that need to be carefully monitored to ensure that processes are being executed efficiently to achieve their intended objectives. In this work, we are particularly interested in measuring performance related to Information Systems (IS), which is another pillar of IS governance in the company. For this purpose, we propose a framework with structural analysis that can provide a multiview meta-knowledge library, enriching all the dashboards common to the company for decision making. In practice, we propose a case study based on Galois lattices, in order to evaluate the synchronization level between some processes and key performance indicators used in the company.

1 Introduction

For several decades, the information technologies have become the most valuable actors in any institution. The modern enterprise is strongly structured by IT processes responding to several business processes Hartono et al. (2003). According to Gerard Balantzian, the Information System is a package of four essential components: IT (Software and Hardware), information, processes and resources Balantzian (2006). The information system guarantees communication between the operating system and the decision-making system as well as the exchange with the environment. The information system is strongly sensitive to strategic evolutions of the enterprise: organizational change, change of objectives, modified variety, new objects and business processes. Otherwise, in the early 1990s, the emergence of the corporate governance concept as a mean of creating value served as an important model for Information Systems management. A few years later, with the birth of the notion of information systems governance, this discipline became an indispensable component of corporate governance. According to the Information Systems Audit and Control Association « ISACA», Measuring performance is one of the main pillars of IS governance. The organizational performance from Information System is considered in many works as the result of some determinants which is presented as a coherent arrangement Grembergen et al. (2003). In this article, we propose a structural analysis integrated formally in the extended enterprise ISO 19440 meta-model Boulmakoul et al. (2009). This extension integrates the necessary structures for developing systemic tools, in order to measuring the IS performance as a contribution of IS governance. This paper is structured as follows: We present, in Section 2, the state of the art of the IS performance measurement. In section3, we remember the extended enterprise meta-modeling approach for measuring IS performance. Then, we propose, in

section 4, a structural framework to technically operationalize this approach before we deploy, in section 5, the proposed approach in a Moroccan Telecom company. We particularly study the structural matrix "Process, KPI" in order to integrate it into a specific platform for viewing the lattice generated. This architecture enriched by a specific analysis methodology helps to evaluate the level of performance of the IT processes studied by their contribution or not to the improvement of the associated KPIs. The conclusion of this work defines the strengths of this approach and other further developments.

2 IS performance measurement : State of the art

2.1 The objective and subjective measurements of information systems

According to R. Reix, the performance measurement of Information System is based on a synergy of objective and subjective measures reflecting information systems user perceptions Reix (2004).

2.1.1 The objective measurements

This concerns essentially the Information Systems efficiency (ratio of the results to the resources used). However, it is often reduced to cost tracking Bodhuin et al. (2004).

2.1.2 The subjective measurements

The subjective measures reflect the perceptions of the IS user and thus make it possible to consider the IS effectiveness according to the following indicators: User satisfaction and Usability. Also, all studies found that regardless of the type of measure adopted, the performance measurement of Information System is done in relation to the following determinants Basu et al. (2002):

- System quality ;
- Information quality ;
- System use ;
- Service quality.

However, Wim Van Grembergen has shown that the Balanced Scorecard BSC can ideally be applied to Information System to measure its performance Grembergen and Saull (2003).

2.2 Balanced Scorecard

Between 1992 and 1996, Kaplan and Norton introduced the "balanced scorecard" into the company. The founding concept of this approach was that the evaluation of a company cannot be reduced to a simple financial estimate but must be supplemented by the results of measurement indicators relating to customer satisfaction, efficiency of internal processes and Capacity for innovation. The two authors proposed a three-level structure (missions, objectives and evaluation) for each of the four perspectives: finance, customer relations, processes and innovation Kaplan and Norton (1992). The implementation of a balanced scorecard involves the following steps:

Step 1: Identification of strategies for achieving financial objectives (causal relationships between organizational / innovation axes, internal processes, customers, finance);
Step 2: Selection of the best performance indicators and objectives to drive the strategy;
Step 3: Implementation of the BSC.

Nevertheless, to apply the balanced scorecard concept to the IS function, the four dimensions cited above have been redefined by the ITGI (Information Technology Governance Institute). According to this organization, the balanced scorecard of information systems (IS BSC) can be developed by considering the following questions:

- The added value for the company: How is the information system seen by The business departments of the company ?
- User satisfaction: How is the IS seen by its users ?
- Operational excellence: how effective are the IS function processes ?
- Future Directions: How well is the IS positioned to identify future needs and requirements ?

Otherwise, Wim Van Grembergen has demonstrated that the Balanced Scorecard BSC could ideally be applied to the measurement of Information Systems performance Grembergen and Saull (2003).

2.3 Strategic maps

The strategic map allows companies to describe the links between intangible assets and value creation. It allows managers to align investments in people, technology and capital of the organization for maximum impact. The strategy map is a visual framework used to integrate the four perspectives of a Balance Scorecard (financial, client, internal, learning and growth). It illustrates the time-based dynamics of a strategy and the relationships that link desired outcomes in the customer and financial perspectives to outstanding performance in critical internal processes (Figure 1) Kaplan and Norton (2004). The strategy map is based on several principles, including the following:

- Strategy balances the contradictory forces of short-term financial objectives for cost reduction and increased productivity;
- Strategy is based on a differentiated customer value proposition ensuring customer satisfaction and subsequently sustainable value creation ;
- Value is created through internal business processes Strategy maps and BSCs describe what the organization hopes to achieve ;
- Strategy consists of simultaneous, complementary themes or clusters of internal processes that provide strategic and competitive benefits ;
- Strategic alignment determines the value of intangible assets such as human capital, information and organization that constitute the three components in the learning and growth perspective.

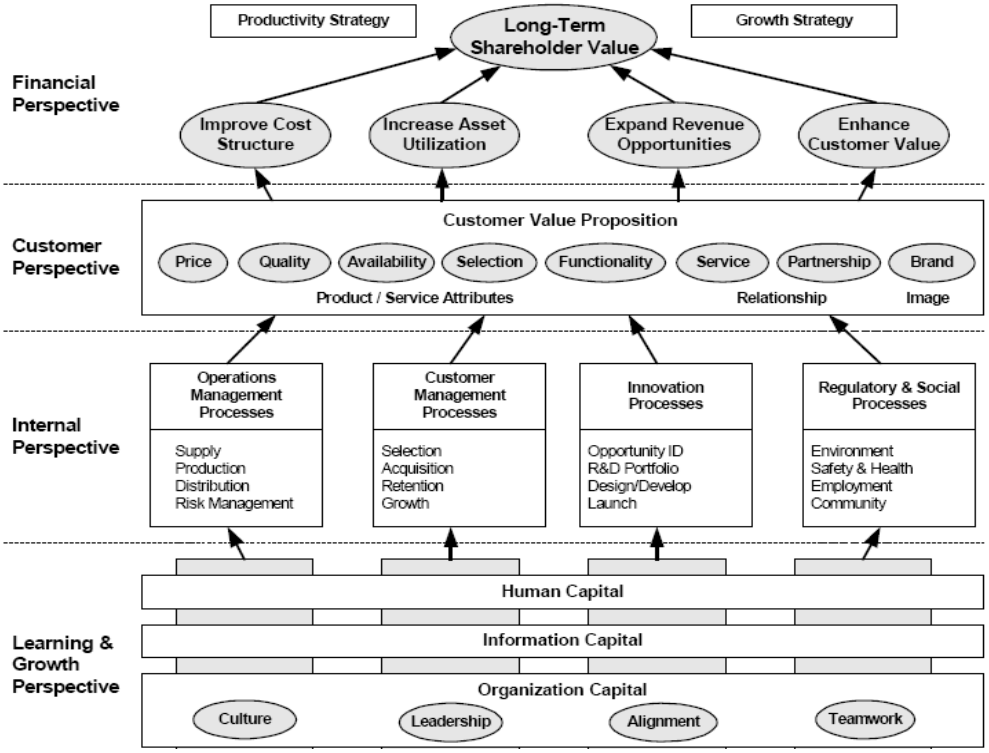


FIG. 1 – *A Strategy Map Represents How the Organization Creates Value*

3 Integrated structural analysis in a holistic meta-modeling for IS performance measurement

In this section, we propose another form of performance measurement of information systems named holistic meta-modeling resulting from a formal extension of the ISO/DSI 19440 Meta-model Boulmakoul et al. (2009). This meta-modeling highlight the alignment between the functional, informational, organizational and resource views in a systemic context and supports some constructs allowing a structural analysis for better synchronization of IS and Strategy. Indeed, the structure of the basic Meta-model allows the expression of the alignment in the above-mentioned forms. However, the formulation of the alignment is not explicit in the modeling of the four views. The proposed formal extension of the ISO 19440 Meta-model has been discussed and argued in our previous works Boulmakoul et al. (2009) and Boulmakoul et al. (2012). The proposed holistic meta-modeling integrates structural analysis and allows us to establish alignment between different views of the company. The proposed meta-modeling provides a formal framework that gives the company global representativeness with a holistic view. In this article, our analysis goes beyond the explicit definitions of functional, organizational, informational and resource entities to also mingle the relationships and associations linking these entities in order to better situate this notion of multiview coupling (Figure 2).

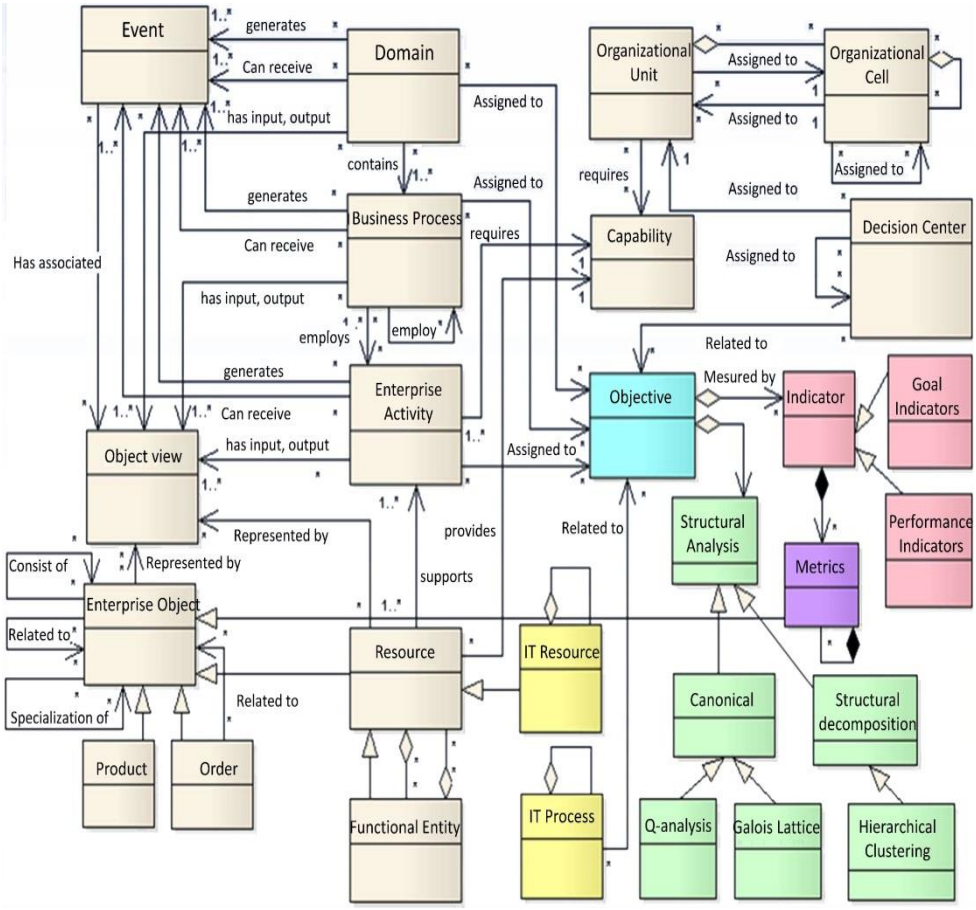


FIG. 2 – *Holistic meta-Modeling for IS strategic alignment Boulmakoul et al. (2009)*

3.1 Structural Analysis

The structural analysis (SA) is based on the modeling of the existing relationships between the different entities of the company and the analysis of their impacts on the overall performance Leung and Bockstedt (2009). We model these relationships using a simple 1/0 association (yes / no). For example, if a given process measured by some key performance indicators (KPI), this KPI will be associated particularly with this process and not necessarily with the others. Other associations could be defined between activity, resources, objective, information object or events. In this simple way of modeling, the basic relationships can be established between the different entities of an enterprise structure. Such an association makes it possible to model hierarchical and non-hierarchical relationships. As we will see in our case study, the interpretation of relations is always clear by the semantic meanings of the entities concerned. To model complex relationships within an organizational structure, simple associations can be extended to include inheritance, aggregation and common characteristics of multiplicity (Roy 2005).

The entities and their associated relations in structural analysis can be implemented using the well-known relational model. In particular, the Entity-Relation model can be implemented in some tables of a relational database. By highlighting associations between specific entities, the analysis can define different views of a company and, therefore, have in-depth insights into performance from several perspectives. Structural analysis (SA) can be used to produce a functional, organizational, informational or resource perspective. This form of analysis is useful in a fairly broad range of business scenarios, from business design to real-time performance monitoring (Figure 3).

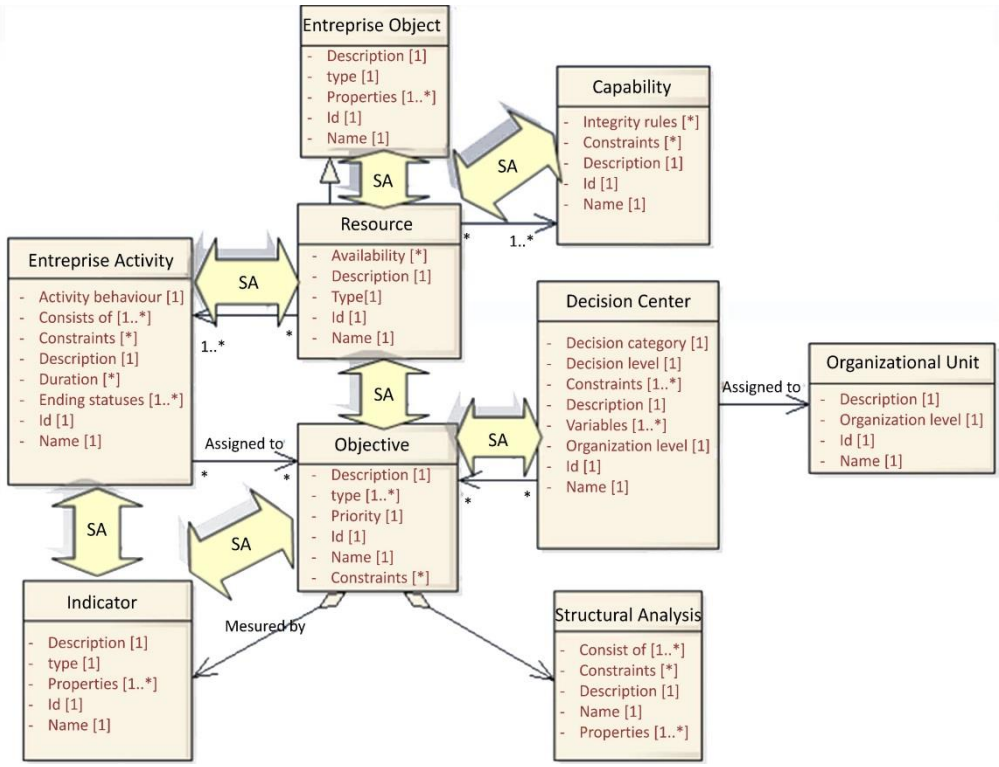


FIG. 3 – *Multiview structural analysis (Boulmakoul et al. 2012)*

4 Structural Framework

In this section, we present architecture for the implementation of the proposed Structural Analysis (SA) approach for the IS performance measurement. The aim of this architecture is to provide a generic tool to be modeled by a computer system in order to undertake a practical and realistic structural analysis of the different components of the company. This architecture is carried out in three essential phases: a configuration phase, an analysis and evaluation phase and a phase dedicated to the exploitation of the results for structural analysis purposes (Figure 4). In the following, we describe each of these phases.

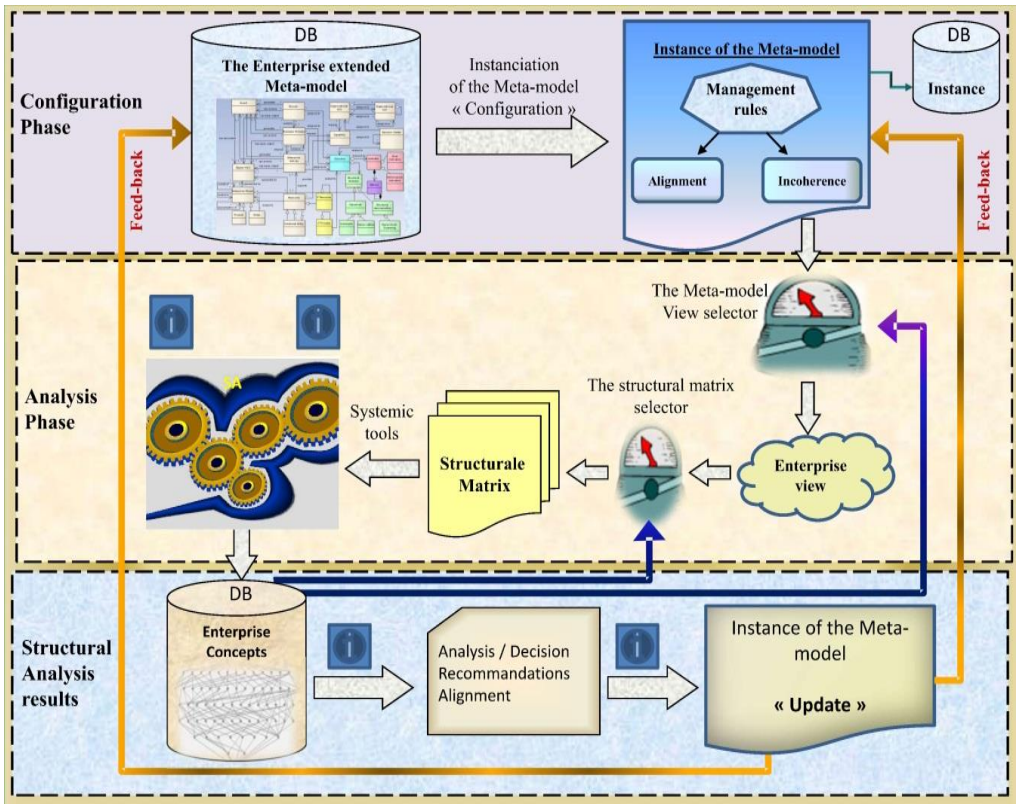


FIG. 4 – *Implementation architecture for structural analysis*

4.1 Configuration Phase

The extended meta-modeling of ISO/DIS 19440 incorporating the proposed structural analysis allows for a more complete representation of the company in its systemic capsule. Such a structure allows a holistic multiview vision based on the structural paradigm aimed at modeling any kind of company, institution or organization. The first step that characterizes the implementation architecture of the structural analysis aims at constituting a database regrouping all the tables coming from the Meta-model (process, activity, objective, resource, indicator, etc.). According to the context chosen, the Meta-model instantiation provides a specific model that reflects the business in a particular area. This means a projection of the Meta-model on one or more views characterizing a predetermined aspect. Then, this instance is used for a configuration of the Meta-model to define explicitly management rules governing the level of performance of several components in the company as process, activity, resource decision-making etc. This approach is a colossal project that requires both human and technical intervention. So, the identification of the parameters of congruence or inconsistency between the company constituents is essential for any structural analysis and alignment evaluation. This allows a response to the fundamental question "Are the X and Y components synchronized, yes or no?".

We present, for example, some management rules characterizing the bilateral relations linking two or more components of the model studied:

- Completion of the Activity « A » requires, at most, 2 human resources;
- The achievement of the objective « O » should not use more than 3 processes;
- The budget allocated to resources for carrying out the process « Pi » is limited to 1M $;
- The indicator of customer satisfaction should not be degraded due to a reduction in resources;
- A business process can deploy up to 3 resources
- the performance of a process can be evaluated by KPIs
- An objective can be measured by one or more KPI…

The result of the structural analysis of the different matrices that can be generated refers to the explicit definition of the predefined management rules in order to identify possible incoherencies. Thus, we can establish a model configuration of the firm studied, in order to begin the analysis phase and to identify any inconsistencies requiring urgent actions.

4.2 Analysis phase

This step is purely technical and tactical evaluating overall the alignment between different constituents of the Meta-model. So, we can realize the projection of the Meta-model of the company on all views in order to reduce the complexity and overlap of the company's activities. Here, we use a computer tool that plays the role of a views selector able to identify appropriate components of the Meta-model in a given company context. This solution makes it possible to generate company views according to the target domain. Once we have chosen a particular view, we also use another structural matrix selector that generates a structural matrix from combinations of two components of the Meta-model (object, attribute). Each structural matrix is dedicated to an analysis based on systemic tools such as Galois lattices. The visualization of the binary relations linking these matrix components makes it possible to develop Galois lattices providing pertinent information based on the study of the closed thus generated. We use the same technique for all possible combinations of each view. All generated lattices are stored in a specific database for any purpose of analysis. This phase is essentially based on a structural analysis to identify all the knowledge needed for better performance measurement and decision making support.

4.3 Exploitation of results

The database with all of the lattices generated from the structural analysis is a robust warehouse to evaluate the synchronization between the various components of the company. Based on the management rules defined in the initial model, we can thus observe all the inconsistencies between the different views of the Meta-model. The results obtained offer a good opportunity to understand the different facets of alignment and performance level. This technique ultimately leads to the creation of notes, suggestions and recommendations that can help the company's senior managers to review the structure of the company by means of multidimensional actions in order to improve the overall performance of the company.

5 Case study

This case study concerns a business entity that centralizes the processing of all demands resulting from the indirect sale of a telecom company in Morocco. This entity was created to expand the sale of telecom products and services beyond urban areas, with a very dense and scattered sales network throughout the country.
In the studied entity, there are some processes, activities and resources used to achieve prefixed objectives and measured by some key performance indicators (KPI) or metrics. In this study, we apply a structural analysis on (Process / KPI) matrix in order to evaluate the performance of all processes depending on their goals (objectives) achieved (Table 1).
By analogy, the same scenario could be projected on the other organizational components of the company, in order to have a global and holistic view of the information systems performance.

5.1 Choice of the structural matrix to be studied

In this work, we are particularly interested at the structural matrix (Process / Indicator): The elements of this matrix are essentially constituted of processes measured by some KPIs and generate intersections allowing better analysis of the performance measurement for the entity studied (Table 1). Of course, in other contexts, we can build several structural matrices by applying couplings like (Process / Resource), (Process / Objective), (Process / Process), (Resource / Objective), (Resource / KPI), (Activity / Resource), (Activity / KPI), (Activity / Objective), (Product / resource), (Product / activity) etc.

	KPI1	KPI2	KPI3	KPI4	KPI5	KPI6	KPI7	KPI8	KPI9
P1	1	1	1	1	0	1	1	1	1
P2	1	1	1	1	0	1	1	1	1
P3	1	1	1	1	0	1	1	1	1
P4	0	0	0	0	0	1	0	1	1
P5	0	0	0	0	1	1	0	1	1
P6	0	0	0	0	1	1	1	1	1

Table 1. (Process / KPI) structural matrix

5.2 Implementation

For the implementation of this architecture, we use a MySQL database with tables fed by data representing the different classes defined in the Meta-model. A web interface allows selecting through a DragAndDrop objects set and attributes set to constitute structural matrices appropriate to the chosen context. Each matrix thus generated is integrated into an open source solution called Galicia to visualize the associated trellis thus generated. The lattices obtained are stored in xml format in a database which will then serve as a relevant analysis for all structural matrixes (Figure 5).

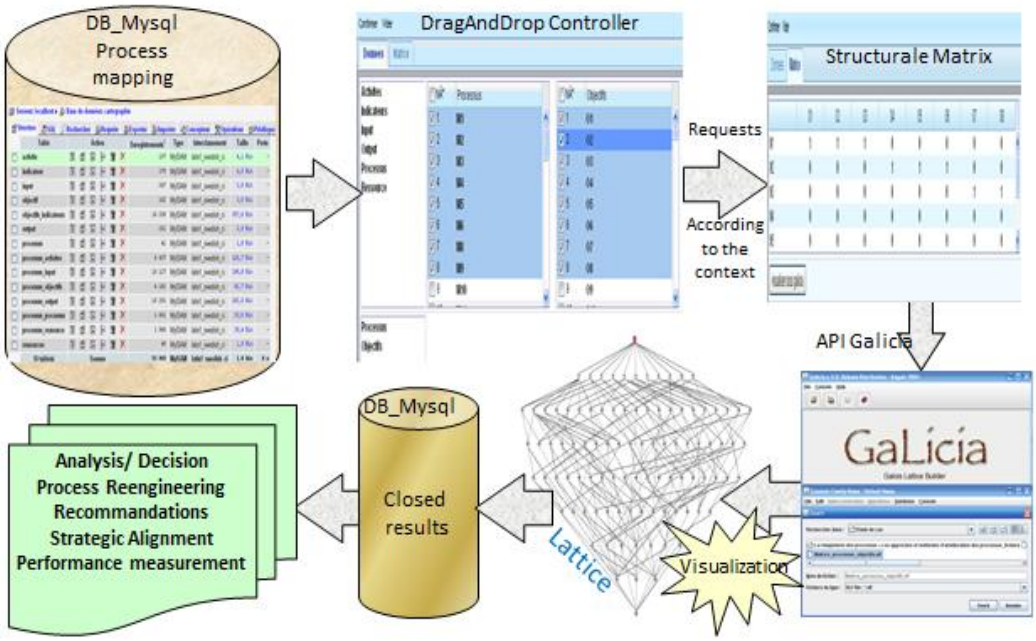


FIG. 5 – *Prototype for the structural analysis implementation*

5.3 Analysis

The visualization of the generated Galois lattice gives a hierarchical representation allowing to answer questions like "what are the processes measured by this or that KPI?" and detailing the relations that exist between the process objects and their predicates as well as the related relationships (Figure 6). The analysis of the closed generated in the Galois lattice contributes in particular to the Process Reengineering. The main objective of this technique is to improve service quality, to ensure customer satisfaction and reduce costs by improving profitability and productivity in the company. To meet these objectives, two approaches are proposed: the first one (vertical) is characterized by an improvement of the performance of a process without worrying about the other processes, which can have an impact on the performance. The second approach (transversal) is characterized by a service offered to a customer which is often the outcome of a more or less complex circuit and by a customer who is concerned solely with the quality of the end product and not with the Company structure.

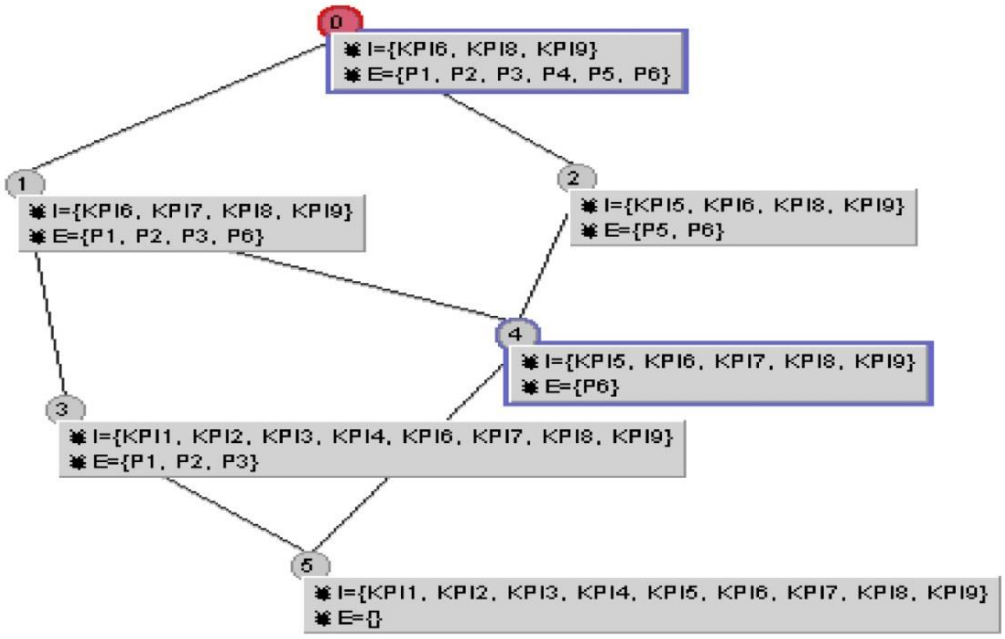


FIG. 6 – *Galois lattice generated in the Galicia platform*

As methodology of analysis, we propose a method contributing to the process reengineering by identifying non-value added processes to streamline the process model Boulmakoul et al. (2009, 2012).

Notation

Π : Set of processes, Θ : Set of strategic axes, λ application embodying the impact force of a process on a target, $\lambda : \Pi \times \Theta \rightarrow R+$, ϖ aggregate function, $\varpi : R+ \times R+ \times \ldots \times R+ \rightarrow R+$.
For each process Pi we associate the δ aggregate measure, impact on the overall axes.
$\delta(Pi) = (\varpi((\lambda(Pi,A1), \ldots, \lambda(Pi,Aj), \ldots, (\lambda(Pi,An))$.
The standard measure μ is given by: $\mu(Pi) = \delta(Pi) / \Sigma(\delta(Pi))$.

Analysis methodology

- Calculate $\mu(Pi)\ \forall\ Pi \in \Pi$.
- Establish a descending sort of process, as the μ measure.
- Δ is the list of processes impacting the goals according to the Pareto rules Rizzo (2009).
- $\Lambda = \Pi - \Delta$ processes impacting low objectives.
- For each process Pi of Λ follow the closed $\Phi i,j$ containing Pi according to a Guttman scale Boulmakoul et al. (2012).
- For each closed $\Phi i,j$ analyze expenditures related processes objectives.
- Audit responsibility centers that deploy $\Phi i,j$ processes.

This analysis methodology allows reviewing the evaluation of some processes measured by KPIs that are not very beneficial to the fixed objectives. So, when one or more KPIs are deemed unsatisfactory, corrective actions are initiated. In the case of large firms, a first step is to identify potential problems that contribute to an insufficient KPI. The structural analysis provides an overview of the processes, resources, activities or products directly related to the suspect KPIs. This technique indicates which processes are measured by a particular KPI. For example, if the time to serve a client is too long, one or more activation processes might be responsible. By analogy, we can generalize to take up all the components responsible for the degradation of the KPIs, and take the necessary measures. This multiview analysis provides insight into the extent of a particular Process-KPI combination, depending to the number of products affected. An unsatisfactory KPI associated with many processes (or resources) impacting many products is naturally at the heart of a reflection of recovery and improvement. We limit ourselves to this simple analysis, which is the first step in a wider study that can identify all the possible inconsistencies that characterize the other constituents of the enterprise meta-model. Other forms of structural analysis will be at the heart of the case studies of other work such as risk analysis, outsourcing analysis, reuse analysis, information analysis, etc. Indeed, we can dig deeper to analyze the closed through other purely computer techniques such as analytics tools or mathematical methodologies, including the Guttman Scalograms from the lattices generated. These different trends of analysis can be discussed in other works.

6 Conclusion

In this paper we propose another view of performance evaluation based on a structural analysis integrated formally in the ISO 19440 Meta-model. This approach constitutes an instrumental causality of the IS performance evaluation. The structural analysis of the couplings linking the various components of the company is likely to provide the top management with pertinent information allowing a multiview evaluation of the performance of information system as contribution of IS governance. This technique could be used in conjunction with analytical tools for organizing, designing, process reengineering or decision-making.

References

Balantzian G. (2006), Le plan de gouvernance du Système d'Information. Etat de l'art, méthodes et cas concrets: Dunod.

Basu, V, Hartono, E, Lederer, AL & Sethi V. (2002). The impact of organizational commitment, senior management involvement, and team involvement on strategic information systems planning, Information & Management, 39: 513–524.

Bodhuin, T. Esposito, R. Pacelli, C. and Tortorella, M. (2004). Impact Analysis for Supporting the Co- Evolution of Business Processes and Supporting Software Systems. In CAiSE Workshops (2). Riga, Latvia, 46–150.

Boulmakoul, A. Falih, N. and Marghoubi, R. (2009). Meta-Modelling and Structural Paradigm for Strategic Alignment of Information Systems. Proceedings of the 4th Mediterranean Conference on Information Systems-MCIS. Athens, Greece, 1070–1081.

Boulmakoul, A. Falih, N. and Marghoubi, R. (2012). Deploying Holistic Meta-modeling for Strategic Information System Alignment. Information Technology Journal, 8: 946–958.

Grembergen, V. and SAULL, R. (2003). Linking the IT balanced scorecard to the business objectives at a major Canadian financial groups. Journal of Information Technology Cases and Applications, 5: 23–50.

Hartono, E. Lederer, A. Sethi, V. and Zhuang, Y. (2003) Key predictors of the implementation of strategic information systems plans. Data base for advances in information systems, 34: 41–53.

Kaplan, R.S. and Norton, D.P. (1992). The Balanced Scorecard: Measures that Drive Performance. Harvard Business Review, 70-79.

Kaplan, R.S. and Norton, D.P. (2004). Strategy Maps: Converting Intangible Assets Into Tangible Out comes. Harvard Business School Press. ISBN 1-59139-134-2.

Leung, Y. and Bockstedt, J. (2009). Structural Analysis of a Business Enterprise. Service Science, 1: 169–188.

Reix, R. (2004). Système d'information et management des organisations. Vuibert, Paris, France.

Rizzo, M. (2009). Goodness-of-fit tests for Pareto distributions, Astin Bull Journal, 39: 691–715.

Roy, B. (2005). Multiple criteria decision analysis: State of the Art Surveys Figueira. Greco and Ehrgott éditeurs. Springer's international, Switzerland.

Summary

Les entreprises contemporaines sont mesurées à l'aide d'indicateurs clés de performance (KPI) qui doivent être surveillés attentivement pour s'assurer que les processus sont exécutés efficacement en vue d'atteindre les objectifs prévus. Dans ce travail, nous nous intéressons particulièrement à la mesure de la performance liée aux systèmes d'information (SI), qui constitue un autre axe pilier de la gouvernance des SI. Pour cela, nous proposons un framework d'analyse structurelle pouvant fournir une bibliothèque de méta-connaissances multi-vues, enrichissant tous les tableaux de bord communs à l'entreprise pour l'aide à la prise de décision. En pratique, nous proposons une étude de cas basée sur le concept de Treillis de Galois, afin d'évaluer la synchronisation entre certains processus métier et les indicateurs clés de performance utilisés dans l'entreprise.

Lean To Identification and Categorization Of Wastes In IT Service
Focusing on IT Operation Processes

[1] **W. Berrahal, R. Marghoubi and Z. Elakkaoui**

[1] NATIONAL POSTAL AND TELECOMMUNICATION INSTITUTE
CEDOC "2TI", LABORATORY, "SEEDS"
av ALLal EL Fassi - Madinat AL Irfane - Rabat, Morocco

Abstract. *This paper discusses the application Lean thinking as a process of continuous improvement for a great achievement of ITSM. Especially, in IT operation based on ITIL frameworks. Our Lean approach aims to minimize wastes throughout the service operation cycle in order to maximize IT service efficiency and customer satisfaction. The biggest challenge is the characteristics of IT service environment, such as inseparability, perishability, variability and intangibility. That can represent a difference between manufacturing wastes and IT service one. Facing this special challenge, a conceptual lean IT operation model is developed. The model intercepts external and internal wastes during IT service delivery and offers a way to project lean on IT operation while respecting continuous improvement processes of the service.*

Keywords : Lean Service, Wastes, IT Operation, IT Service Wastes

1. General Introduction

Successful implementation of an Information Technology Service Management system (ITSMS), for a better delivering of IT services remains a real challenge. Even Though it focuses on the frameworks of best practices, such as Information Technology Infrastructure Library (ITIL). Our approach to apply the Lean to IT has a goal to identify and eliminate sources of productivity loss and wastes in the implementation process. It also aims also to provide a pragmatic and efficient model and to continuously promote the human factor and reduce the resistance to change, therefore, provide at a specific time of the customer or end-user needs. It means to be Just-In-Time (JIT). Moreover, the current study intends to reduce the gap between the research committee and the professional world, and initiates a critical reflection in relation to the implementation of IT Service Management standard.

Companies often have the constraint of innovation and change, especially in the service industry. Facing the arduous competition, the massive flow of information, the technological complexity and organizational, namely the approaches (Agile or V-cycle). The fight against waste becomes a major problem for the companies. Lean management has been particularly assessed in the automotive industry. And this was confirmed by (Wolmak and Jones 1990) in Massachusetts Institute of Technology study (MIT) occur Toyota company In order to examine the Toyota production system (TPS) and which gave rise to the famous book "the machine that will change the world". Furthermore, in 2007 Toyota exceeded "General Motors" that occupied the first place of automotive producer

		Manufacturing	IT Services
Seven Wastes	Waiting	Workers having to stand waiting for the next processing step or supply [17]	– Delay on completing service output [16] – Delays in terms of employees or customers waiting for information or service delivery Sometimes referred to as queuing and occurs when there are periods of inactivity in a downstream process because an upstream activity has not delivered on time [12].
	Transport	Carrying work in process long distances, creating inefficient transport, or moving materials, parts, or finished goods into or out of storage or between processes [17]	– Unnecessary movement of material and information [16] – Needless, non-adding-value movement of resources Unnecessary motion or movement of materials, such as work in progress (WIP) being transported from one operation to another [12]
	Extra processing	Taking unneeded steps due to poor tool and product design or providing higher-quality products than is necessary [17]	– Lack of standardization in the offer or processes, procedures, formats, including expired or outdated with no standard time defined Extra operations such as rework, reprocessing, handling or storage that occur because of defects, overproduction or excess inventory [12]
	Inventory	Excess raw material, or finished goods [17]	– Excess work-in-process such as queues and pending request [16].All inventory that is not directly required to fulfil current customer orders. Inventory includes raw materials, work-in-progress and finished goods. – Inventory all requires additional handling and space. Its presence can also significantly increase extra processing [12]
	Motion	Any wasted mouvement employeeshave to perform during the course of their work [17]	– Unnecessary movement of people in serviceareas with a poor Layout [16] – Non-adding-value movement of resources Refers to the extra steps taken by employees and equipment to accommodate inefficient layout, defects, reprocessing, overproduction or excess inventory.Motion takes time and adds no value to the product or service [12]
	Defect	Production of defective parts or correction. Repair or rework, scrap [17]	– Mistake in any service processes such as error in data entry [16] – Finished goods or services that do not conform to the specification or customer's expectation, thus causing customer dissatisfaction [12]
	Overproduction	Producing items for which there are no orders [17].	– Occurs when operations continue after they should have ceased. This results in an excess of products, products being made too early and increased inventory [12].

Table 1. Comparative of the Specificities of Lean Wastes : Between Manufacturing and IT Service

of the world. Since then, efforts had been made to expose examples of lean method used in service and the term ”lean service” started to be used in literature [16].

The choice of ITIL as a Lean application environment comes from the fact that has proved best adherence to Information Technology Service Management [5]. However, operation Management of IT service processes as it is described in best practices literature, not assure the best productivity. Our proposed model aims a continuous improvement approach for increase efficiency in daily operation without waste.

The list of standard wastes are defining in Table 1. This table shows a difference between the first definition of the Wastes by ”Taiich Ohno” and recent years definition which is in the research articles dealing with the subject. The ultimate goal of this comparison is to show that the waste in the IT service has its specificities compare to manufacturing. For example, the defects in the manufacturing are instead local, furthermore, in IT service they can be related to external dimension as as error in data entry or uniformity of specification or customer's expectation.

2. Lean IT To Service Operation

Lean service is the application of lean thinking in the service industry. It's involved to eliminate service process waste, consequently the cost can be reduced and better service can be provided in accordance with customer requirements [4]. which can also be seen as a managerial philosophy that increases the value perceived by users, by adding service features and continuously removing wastes (i.e. non value added activities), which are concealed in hidden of process [5]. there are also some researchers who centered it around creating more value with less work. Moreover, there is much differences and specificities between goods production and those of the services. This is mainly due to service activities and its interactions with customers or consumers of the services (end-users). All these points mentioned above, are inherent of the service characteristics and its nature, which represent an environment area to respect. which we also list in Table 1. It presents the main differences in the basic principles of lean between manufacturing and the IT service.
Lean IT is the expansion of lean service principles to management of Information Technology (IT). In this paper, we aim to understand how ”Lean IT” help improve ”IT operations” in order to explore option opportunities to merge Its two components into a single IT operation. We identify set of capabilities, people, process and technology. It Specially includes operational work, through standard processes areas, namely event, incident, problem, request and access management [10] [14].

In the following section, we will be identifying the fundamental elements and specificities of Waste IT service show Table 2, in order to develop a conceptual Model which we use for identifying wastes in the context of IT operation processes systems.

The root causes are mentioning in Table2 related of waste generate in IT Service delivery, in order to provide a correlation between characteristics of IT Service and classic categories of waste. We should mention that one must not have confusion between the variability, that is a service characteristic and Unevenness or ”Mura” a Japanese word meaning, irregularity, lack of uniformity, inequality. The table of wastes classification for IT Services aims to provide a new vision for waste determination. Moreover, we consider that undapted flux are part of dealing daily works and consequently cause more waste.

Characteristics of IT service	LEAN classic wastes							Examples of wastes root
	Waiting	Extra Processing	Over Production	Defect	Transport	Motion	Inventory	
Variability	✓	✓	✓		✓			– **Variability :** Change of information, concepts, ideas and heterogeneity of platforms.
Inseparabality	✓			✓				– **Inseparability :** The service's generation and consumption occur simultaneously.
Intangibility				✓		✓		– **Intangibility :** Feeling and perception of,the customer in relation to service delivery. The quality of a service is based on customer's feelings and expectations.
Perishability		✓					✓	– **Perishability :** Services cannot be produced and stored to be sold at a later stage.
Unadapted Flow	✓	✓	✓	✓	✓		✓	– **Unadapted Flow :** Failure demand, flow demand, flow excess, flawed flow.

Table 2. Wastes Classification for IT Services

2.1. Failure of IT operation processes

Moreover, IT operations is characterized over specialization and concentration of shared resources that could cause severe consequences [11]. Therefore flow interruptions, the excessive dependence on overburdened individuals with special skills and Communication breakdowns.

Lean approach would seem necessary to redress the balance and streamline the work. This fact lies that there is all aspect of IT operational work that can benefit from Lean. Lean operations reflect performance improvements in the areas of cost efficiency, conformance quality, and delivery speed and reliability [18]

The daily operations management from IT infrastructure, including five relevant basic management processes: Event management, Incident management, Problem management, Request management and Access management[4]. All this processes receive the input flows and accomplish specified activities which can be root of wastes. Our package diagram fig 1 established the interface between IT operation processes and wastes component. The list below describe the service operation processes and their potential wastes in more details :

• Event Management : This process represents the continuous flux of activity that underlies delivery of IT services . It includes the flow in network traffic and fluctuates changes in customer requests.The volume of data handled through batch routines. Consequently, alerts set at the wrong thresholds will trigger actions that represent waste [11], this phenomenon is related to demand variability customer and excessive use of services in specific periods. Over time, this can result an unnecessary work, misdirected efforts, or confused communication, all evidence of waste.

• Incident Management : It's the process of restoring normal operations when an incident defined as an interruption to normal operation or a reduction in service quality, occurs [10]. Namely restoring the service as quickly as possible. Incident Management uses standardized procedures for resolution. However, the incident is not run as a linear flow. Several exceptions go against standardized practice and generate waste. This waste is the result of external or internal flows, such as the lack of understanding of

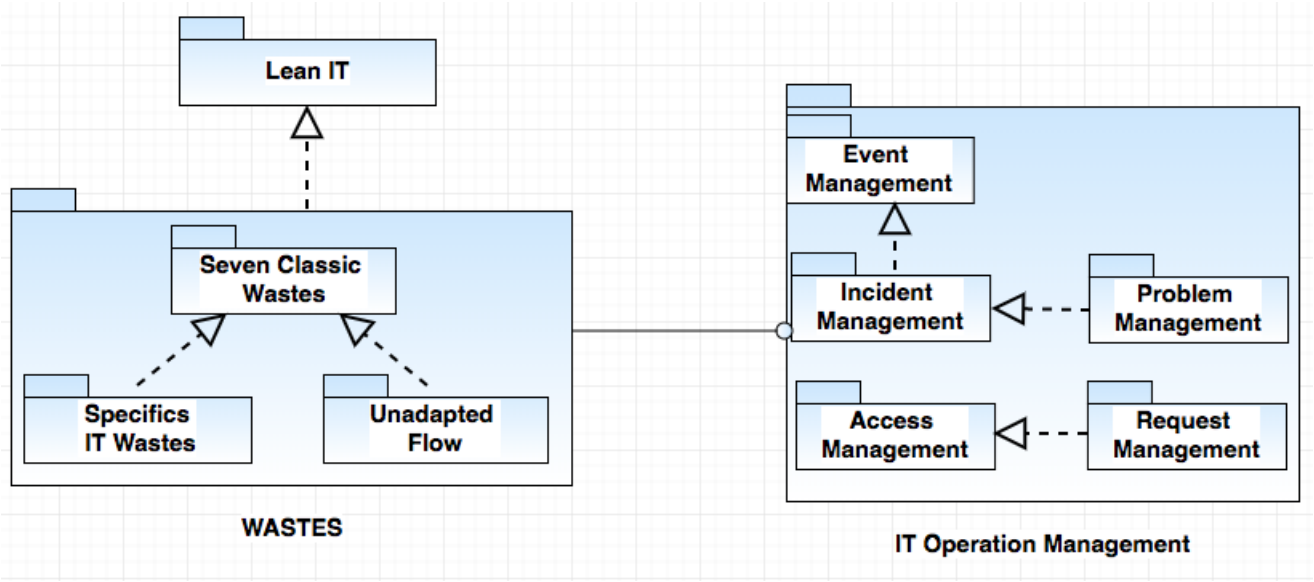


Figure 1. A package diagram for Lean IT Operation system

user/customer requirements and expectations, thus generating ineffective solutions. In addition, the escalation procedure is also a serious problem related to delays.

• Problem Management :It's the process that identifies root cause, resolves problems, and thereby helps resolve their associated incidents. [4]. Problem Management can be viewed as the process of managing problems. it provides the top-down model, in exactly the same way the incident. Accordingly, it will generate a critical gap regard to customer expected. This result is a consequence of the root causes waste, such as Inadequately tested changes to service, systems Defining service strategy without the stakeholders opinions, and taking technical decision without the presence of technicians.

• Request Management : In the ITIL framework, a Service Request is a trigger for a Request, such as password resets, granting of privileges. This process can easily increase inventories, which represents a type of the waste. Particularly, because of rigidity of fulfillment process that delivers services, and excesses information demands.

• Access Management : The access processes aim to authorize users the right to use a service. It's less critical process but represents also a generator of wastes, because it delays between realization and verification.

During our diagnostic of IT operation processes, we have been noting that these processes represent a rigidity and a lack of flexibility. That means, they cannot adapt customer's request variability and technological change. Consequently, it increases backlog of incident (inventory), increase delays (waiting) and promote of the reworks (defect). Against this phenomenon, which increase the volume of the wastes. Therefore, degrades quality level and efficiency in operations. A continual improvement approach Lean, may provide a major contribution to eliminate the wastes and customer satisfaction. The proposed model fig 2 show a map of components of IT service operations and manner where the wastes generated. In order to apply Lean tools improvement in the right zone and proactively. Our global map of IT operation service flow fig 2 aims to identify the sources of the wastes and differentiate between external waste occurred from higher-level and user

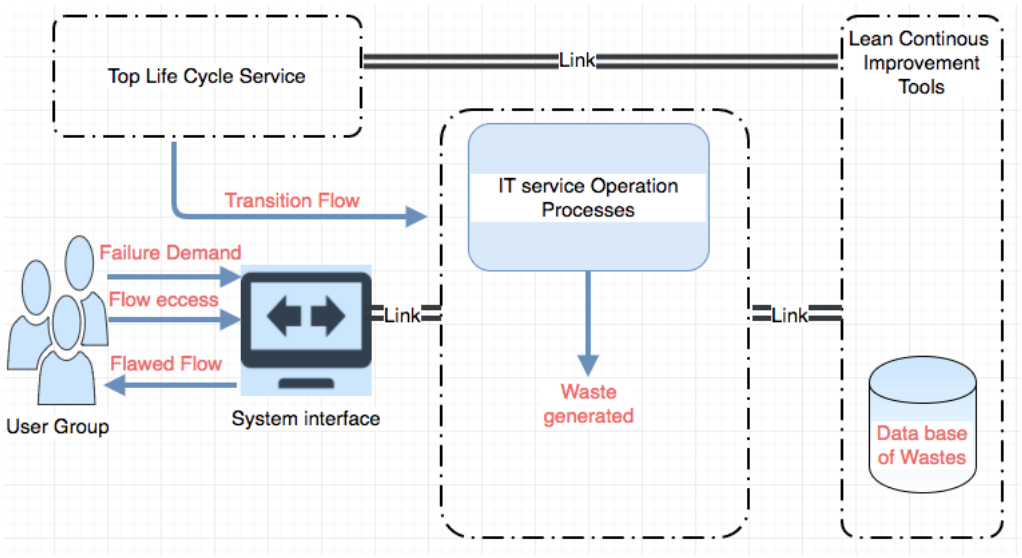


Figure 2. Global IT operation Service Flow

Group and information technology. The system shows the important component required for operation in IT service. And internal wastes generate towards operation processes running.

3. Conceptual Model of waste identification for IT operation processes

In information management area, particularly in the IT operation, the systems are considered as various dimensions of waste that do not occupy an equivalent space, the effects are intangible and the value, or lack of, are far less clear and arguably highly subjective [12]. In this context, our research articulates on the way in which to provide perfect value to the customer through an excellent work-flow process that has zero waste.

Focusing on the frameworks of best practices for delivering and governing IT services, such as Information Technology Infrastructure Library, we have observed that application of continuous improvement processes is not able to eliminate the IT generated wastes [13]. This derives from quality system which is inappropriate to provide tools adapted of waste phenomenon. Consequently, in this paper we will be studying the possibility of eliminating waste through the implementation of an efficient IT operation management, supported by the best practice standard. The following work-flow fig 3 was given as a suggested new vision for IT operational processes treatments. Particularly, the incident and problem processes. The ultimate purpose of the activity diagram shown in fig 3 is identifying all susceptible zone, and representing a potential waste, namely decision and escalation steps.

3.1. New Top model of IT operation

Many researchers have attempted to classify service features as a set of the following key service functions : First, the feelings and perception of the customers in relation to service delivery. Second, the simultaneous production and consumption of services. Third, the lack of consistent, homogeneous and repetitive quality, Fourth, produces and stores services at a later stage. And finally the lack of property [9]. Moreover, information

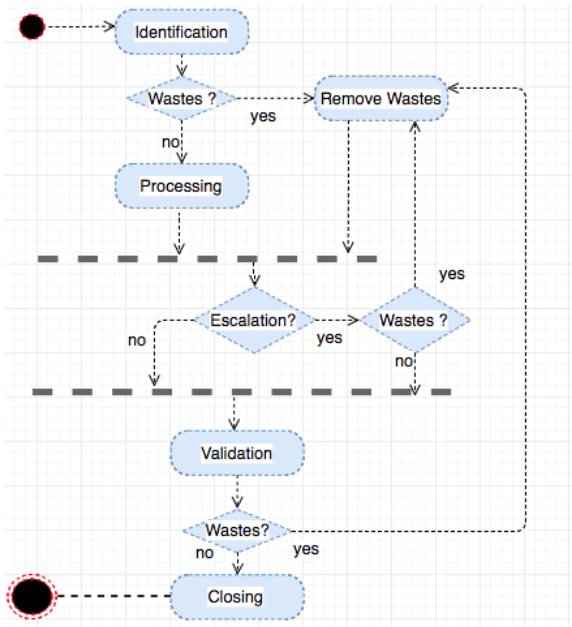


Figure 3. New Top-model for incident work-flow

flows in IT operation system which group of following categories: failure demand, flow demand, flow excess, flawed flow where the different exchanges are represented on fig 2.

Eliminate the wastes in the daily operation is paramount to acquire an excellence IT service operation. One of the big challenges is the waste in IT is intangible. That mean that service is an intangible process, while the a good one is the physical output of a process [7]. In addition, services are activities or a series of activities rather than things. This is why services are at least to some extent produced and consumed simultaneously. Likewise the customer participates in the production process at least to some extent. In order to raise this challenge we propose a top model of incident treatment. which aims to redefine the classic process of ITIL frameworks fig 3.

4. Conclusion

All production chains are confronted with the phenomenon of waste (Porter 1985), this waste impacts the quality of services provided, delays in production, increases costs and degrades customer satisfaction.
A typical application for strategy instruction has been described in the standards references and does not guarantee for success, but rather innovate in order to conceive a specific approach and pragmatic for root out waste and inefficiency. The proposed Top-model for incident work-flow attempts, is institute a best practices standards, and provides a viewing perspective for successful implementation of IT operation management without waste throughout a LEAN IT Management.

References

[1] N. Ahmada, Zulkifli M. Shamsudinb, *Systematic Approach to Successful Implementation of ITIL*, Procedia Computer Science, Elsevier pp. 237-244, 2013.

[2] R. Estevesa, P. Paulo Alves, *Implementation of an Information Technology Infrastructure Library Process – The Resistance to Change*, Procedia Technology, Elsevier pp. 505-510, 2013.

[3] G. Ananda, Peter T. Wardb, Mohan V. Tatikondac, David A. Schillingb, *Continuous Improvement Beyond The Lean Understanding*, Forty Sixth CIRP Conference on Manufacturing Systems 2013: Elsevier Vol.7, pp. 575-579, 2013.

[4] Li. Qu, Man. Ma, Guannan. Zhang *Waste Analysis of Lean Service*, International Conference on Management and Service Science (2011).

[5] Ahmed. Adel *Lean Systems Adaption in Services Industry*, Master of International Business Administration, ESLSCA, 2012.

[6] V.K. Bakliwal *Production and Operation Management*, Mark publishers 1st edition, 2011.

[7] Richard B. Chase, F. Robert Jacobs, Nicholas, J. Aquilano, *Operations Management for Competitive Advantage*, Irwin/McGraw-Hill 11th editioned, 2005

[8] Gronroos. Christia *1947- Service management and marketing.*, Lexington, Mass. : Lexington Books, ©1990

[9] E. Andrés-Lópeza,Is. González-Requenaa,, A. Sanz-Loberaa *Lean Service: Reassessment of Lean Manufacturing for Service Activities*, The Manufacturing Engineering Society International Conference, MESIC Procedia Engineering 132 23 – 30, 2015

[10] Howard. Williams, Rebecca. Duray *Making IT Lean Applying Lean Practices to the Work of IT*, Taylor Francis Group, LLC International Standard Book Number-13: 978-1-4398-7603-9 (eBook) 2013

[11] Steven C. Bell, Michael A. Orzen *Lean IT:Enabling and Sustaining Your Lean Transformation*, Taylor Francis Group, an Informa business International Standard Book Number-13: 978-1-4398-1757-5 (Ebook)

[12] B. J. Hicks *Lean Information Management: Understanding And Eliminating Waste*, International Journal of Information Management Elsevier Vol.27, Issue 4, Pages 233-249: August 2007

[13] W. berrahal,R. Marghoubi *Lean continuous Improvement To Information technology Service Management Implementation*, The 2nd International Conference on Information Technology for Organizations Development March pp. 63-64 : 2016

[14] L. bZhu1, M. Songi *ITIL-based IT Service Management Applied in Telecom Business Operation and Maintenance System*, 2 Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT 09) Huangshan, P. R. China, 26-28,Dec. 2009, pp. 243-246

[15] D. Cannon, D. Wheeldon *Service Operation*, The 2nd 1re ed. Office Of Government Commerce (OGC), V.3 : The Stationery Office : 2007

[16] R. Asnan, N. Nordine, S. T. Othman *Managing change on Lean Implementation in Service Sector*, The 2nd Global Conference on business and Social Science GCBSC : 2015

[17] Taiich. Ohno, *The bible of the Toyota Production System* , publication bureau Diamond Company Book ©1973

[18] H. Soltana, S. Mostafab *Lean and agile performance framework for manufacturing enterprises* , 2nd International Materials, Industrial, and Manufacturing Engineering Conference, MIMEC : 2015

Vers l'évolution des bases de données orientées graphes : opérations d'évolution

Soumaya Boukettaya[*,**] Ahlem Nabli[*,***] Faïez Gargouri[*,****]

[*]MIRACL Laboratory, Sfax University, Tunisia
[**]Faculty of Economics and Management of Sfax, Sfax University, Tunisia
soumayaboukettaya@gmail.com,
[***]Faculty of computer sciences and information technology, Al-Baha University, KSA
ahlem.nabli@fsegs.rnu.tn
[****]Institute of Computer Science and Multimedia, Sfax University, Tunisia,
faiez.gargouri@isims.usf.tn

Résumé. Les bases de données NoSQL deviennent populaires en tant que back-end aux applications web d'aujourd'hui. En fait, le modèle orienté graphe peut gérer efficacement des données fortement connectées. En raison de leur flexibilité, les bases de données orientées graphes ne nécessitent pas la définition de schéma global, mais le schéma est souvent maintenu dans le code source de l'application. Avec cette flexibilité, les développeurs peuvent gérer la diversité des données, mais ils peuvent avoir du mal avec la croissance de la structure des données et la gestion correcte des données persistantes. En effet, le problème d'évolution n'est pas bien traité surtout lorsqu'il s'agit d'analyser le code source de l'application et son historique. Dans cet article, nous proposons une approche pour contrôler l'évolution des schémas des bases de données orientées graphes en analysant le code source des applications. Nous nous intéressons principalement à détailler quelques opérations de migration, ainsi que les contraintes inhérentes.

1 Introduction

Avec l'essor de l'Internet et l'apparition de l'ère du Big Data, les données deviennent très variées en termes de structure et de volume (ex. : données issues des médias sociaux). Le modèle NoSQL orienté graphe est l'un des modèles NOSQL capables de gérer efficacement des données hautement connectées (Réseaux sociaux, graphiques de trafic, lieux géographiques, etc.). Ainsi, beaucoup d'applications Web s'orientent vers l'utilisation des bases de données orientées graphes comme back-end à leurs codes.

Les bases de données orientées graphes, ont un niveau de flexibilité élevé et ne nécessitent pas la déclaration d'un schéma global. En fait, le modèle de données est maintenu dans le code source de l'application. Cependant, lorsqu'il s'agit de manipuler des données dans des applications déployées continuellement sur la même base de données, cette flexibilité s'accompagne de difficultés pour gérer l'entropie croissante de la structure de données et assurer une bonne

manipulation des données existantes. Ce problème peut entraîner de graves pertes de données et des erreurs d'exécution.

Les entités stockées par les différentes versions de l'application peuvent se distinguer dans leur structure. En effet, les bases de données orientées graphes sont capables d'évaluer les requêtes sur les entités structurellement hétérogènes. Essentiellement, cela nécessite des méthodes pratiques pour contrôler l'évolution des données dans les bases de données orientées graphes (BDoG). L'objectif de nos travaux est de proposer une approche pour contrôler l'évolution de schéma dans les (BDoG), dont le but d'éviter toute erreur ou perte de données. Dans le contexte de l'orienté graphe, élaguer la nécessité de déclarer des schémas explicites ne doit pas être confondu avec l'absence totale d'un schéma. Dans la plupart des cas, le schéma est implicite dans les données et les applications (Klettke et al., 2015). Cela rend les développeurs confrontés aux changements de la structure de données et à la gestion des données persistantes.

Le problème d'évolution des entités reste récent et n'est pas bien abordé. Dans ce papier, nous nous intéressons à proposer une approche pour contrôler l'évolution des schémas dans les (BDoG) tout en offrant des opérateurs de migration des données adaptées à la structure d'une (BDoG. Ce papier est structuré comme suit : à la section 2, nous présentons un exemple motivant, la section 3 est conservée pour un résumé des travaux connexes. À la section 4, nous proposons notre approche et à la section 5 nous détaillons nos opérations de migration simples, et nous clôturons ce papier par une conclusion à la section 6.

2 Exemple motivant

Étant utilisée comme back-end, toute nouvelle version du code d'application peut entraîner tant de changements, dans la structure de la (BDoG), qui conduisent à l'évolution de son schéma. En général, tout type de manipulation de données a des conséquences sur les différentes entités de la base de données et sur le schéma global.

En s'inspirant du travail de (Saur et al., 2016), nous considérons le cas d'un magasin en ligne qui garde la trace des bons de commande. L'application stocke les bons de commande dans une (BDoG). Supposons que nous voulons mettre à jour les commandes pour supporter la tarification différenciée, qui nécessite de changer le format de données dans le noeud «PRODUCT».Une instance de la base dans ses deux versions est montrée dans la figure 1.

A l'instant t_0, la version V_1 de la base de données contient un noeud «CLIENT», un noeud «ORDRES» et un noeud «PRODUCT», initialement tout produit est composé des attributs «product» et «price». Nous supposons qu'à l'instant t_6 l'application vise à mettre à jour l'un des produits pour supporter une tarification différenciée. Dans ce cas, un nouveau noeud de produit avec de nouveaux attributs sera créé automatiquement au lieu de mettre à jour le noeud existant (dans notre exemple : le noeud avec produit «Cookies»).

3 État de l'art

L'évolution des bases de données, qui est au coeur d'un système d'information, représente un problème de maintenance difficile. Un très large corpus de littérature existe aujourd'hui reflétant le vaste travail sur l'évolution du schéma et sa gestion (par exemple, le modèle relationnel (Curino et al., 2013), (Qiu et al., 2013), (Manousis et al., 2015), (Cleve et al., 2015),

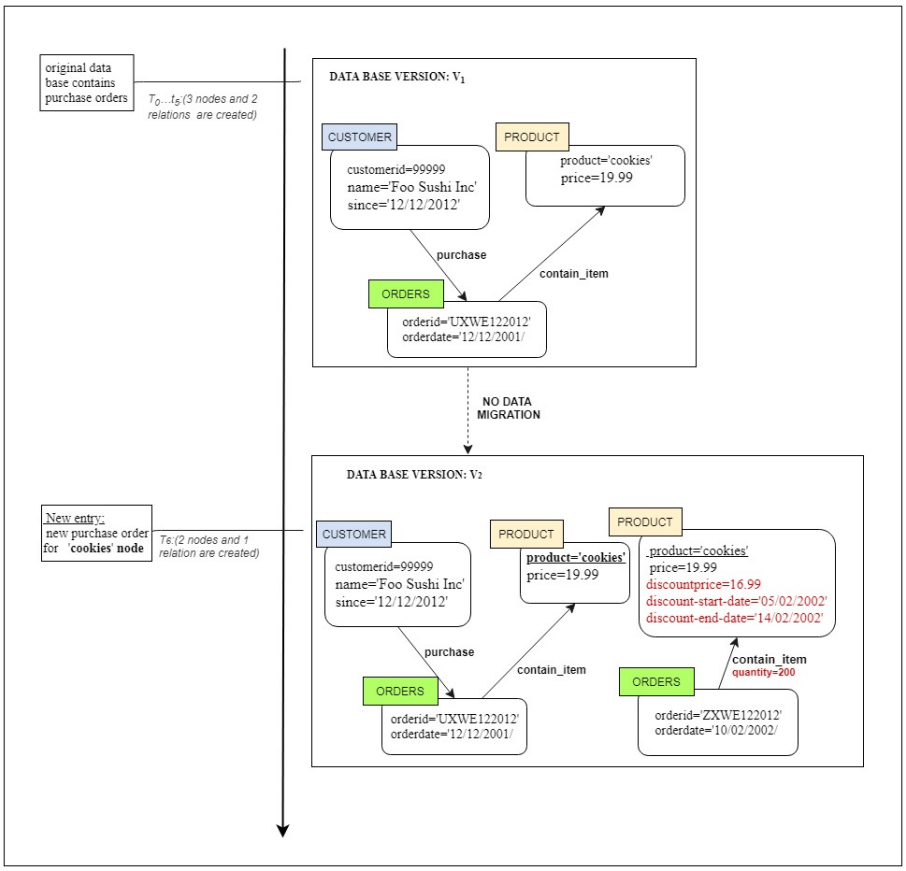


FIG. 1 – *Évolution de schéma entre les versions de l'application.*

le modèle orienté objet (Li, 1999), les ontologies (Mahfoudh, 2015), les entrepôts de données (Subotic et al., 2014). Étant un domaine de recherche récent, l'évolution des données et des schémas dans les bases de données NoSQL n'est pas très répandue. En effet, il existe peu de travaux sur les modèles NoSQL (Scherzinger et al., 2013) (Scherzinger et al., 2015b) (Scherzinger et al., 2015) (Scherzinger et al., 2015a) (Scherzinger et al., 2016), même dans le modèle orienté graphe, le problème d'évolution n'est pas bien traité surtout lorsqu'il s'agit d'analyser le code source de l'application et son historique.

D'une façon générale, il existe deux stratégies d'évolution de schémas dans les bases de données NoSQL. Une stratégie « a priori » consiste à suivre la migration des données au fil du temps afin de pouvoir contrôler l'évolution implicite des schémas. La deuxième stratégie est « a posteriori », qui consiste à détecter les entités qui ont subi des changements à partir des différentes versions de la base de données.

Dans le cadre de l'évolution des données, certains travaux adoptent l'approche à priori et

traitent la migration des données et ses stratégies. L'un des travaux dans ce domaine (Ringlstetter et al., 2016) utilise des mappeurs objets NoSQL pour présenter une analyse sur l'évolution des données dans les bases de données NoSQL (Ringlstetter et al., 2016). Également, le travail de (Scherzinger et al., 2016) présente ControVol, un plug-in IDE qui suit les changements évolutifs dans les mappages objet-NoSQL en se concentrant principalement sur la détection des changements de types d'attributs qui ne sont pas compatibles avec les anciennes entités de la base de données. Les autres dans (Scherzinger et al., 2015), et (Florian Haubold, 2017) présentent d'autres extensions pour le plug-in ControVol. La première extension dans (Scherzinger et al., 2015b), présente un Framework qui vérifie le type des classes de déclarations objet-mappeur et permet de corriger les incompatibilités de type immédiatement, déjà pendant le processus de développement. La deuxième extension est ControVolFlex dans (Florian Haubold, 2017), apporte, la possibilité pour les développeurs de choisir leur propre stratégie de migration. En fait, toutes les données héritées peuvent être migrées au moyen de scripts de transformation NotaQL ou peuvent être migrées, comme le déclarent les annotations d'objet-mappeur.

Cependant, dans (Scherzinger et al., 2013), (Scherzinger et al., 2015a) et (Klettke et al., 2016) les auteurs présentent comme solutions des langages et des stratégies pour la migration des données.Quoique, les solutions présentées permettent de résoudre le problème des données persistantes, aucun de ces travaux ne décrit une solution claire pour suivre l'historique de l'évolution des entités. (Scherzinger et al., 2015a), présente un langage d'évolution de schéma NoSQL, tandis que (Scherzinger et al., 2013) décrit principalement un langage de programmation de base de données NoSQL. (Klettke et al., 2016) présente des opérations de migrations composites comme solution d'évolution de schéma, tout en prenant en considération les sauts de versions.

Dans le cadre de l'évolution des schémas (Klettke et al., 2015), (Störl et al., 2017) exposent des solutions pour l'extraction des schémas et la gestion de son évolution. (Klettke et al., 2015), décrit un processus pour l'extraction de schémas à partir des documents JSON, et (Störl et al., 2017) présente un Framework pour la gestion automatique de l'évolution des schémas. Bien que ces travaux traitent la gestion dynamique de l'évolution des schémas, ils ne traitent pas de la propagation des changements de schéma.

Contrairement aux travaux précédents (Meurice et Cleve, 2017) dévoile une solution pour l'extraction des schémas à partir des codes fournis par des applications et de présenter une analyse sur l'historique de l'évolution des entités.

Bien que les travaux présentés traitent l'évolution des données et des schémas dans les bases de données NoSQL, la plupart ne supportent pas les bases de données orientées graphes. (Castelltort et Laurent, 2013) est le seul travail étudié qui offre une solution pour suivre l'historique de l'évolution des entités dans une base de données orientée graphes. Toutefois, la solution proposée risque non seulement d'avoir des problèmes de consistances, mais aussi d'augmenter la complexité de système.

Dans le cadre de cet article, nous proposons une approche pour contrôler l'évolution des schémas des bases de données orientées graphes en mettant en évidence quelques opérations de migration.

4 Approche d'évolution de BD orientée graphe

Le fait que les bases de données orientées graphes soient fondées sur la théorie des graphes le rend différent pour gérer l'évolution des schémas et la migration des données sur ces bases de données.

Cette section présente notre approche permettant aux développeurs de comprendre et d'analyser l'évolution des schémas dans des bases de données orientées graphes. Notre approche, résumée dans la figure 2, est composée de trois phases à savoir :*extraction du schéma courant*, *processus de migration des données* et enfin *la réalisation de la migration des données*.

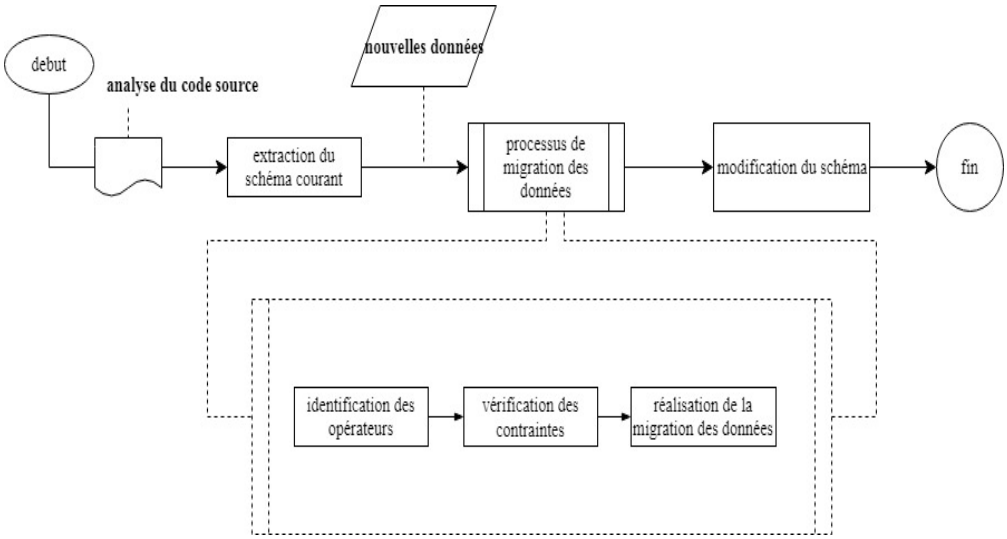


FIG. 2 – *Aperçu de notre approche.*

Extraction de schéma. La première étape de notre approche vise à déduire le schéma de la base de données en analysant de manière statique les accès à la base de données à partir du code source. La détection du schéma global de la base de données peut aider à contrôler la migration des données et à maintenir leur compatibilité avec les données persistantes.

Certains SGBD offrent des requêtes pour extraire le schéma global d'une (BDoG), mais nous avons noté qu'ils ne couvrent pas tous les scénarios des entités qui existent dans la base de données.En effet, les (BDoG) sont souvent énormes, ce qui rend la complexité de leur gestion très grande et difficile à maintenir. Travailler sur toute la base de données peut entraîner une interruption de l'application pendant une longue période. Ainsi, cette étape consiste à charger le schéma de la base de données (cette fonction souvent offerte par les SGBDs) et à analyser chaque nouvel accès à la base de données dans la version actuelle du code afin d'extraire la partie du schéma à laquelle l'application accède actuellement et potentiellement les parties qui peuvent être affectées par la modification de l'entité à migrer.

Processus de migration des données. Ce processus contient trois phases principales qui sont l'identification des opérations, la vérification des contraintes et la réalisation de la migration.

— *Identification des opérations.*Cette étape consiste à identifier les opérations de migration qui seront effectuées sur les différentes composantes de la BDoG à partir des applications. Généralement, ces opérations peuvent être les opérations d'ajout, de mise à jour et de suppression, mais elles peuvent aussi être des opérations de copie et de déplacement.

— *Vérification des contraintes.* Afin de maintenir une évolution de schéma fiable, nous devons vérifier certaines conditions (en fonction de l'opération déterminée) avant de migrer réellement les données. Cette étape est considérée comme cruciale, afin d'éviter de graves pertes de données et des erreurs d'exécution pouvant entraîner un plantage de l'application.

— *Réalisation de la migration.* Après spécification de l'opération et la vérification de ses conditions, les données peuvent être migrées en toute sécurité.

Modification de schéma. Toute migration de données conduit à un changement de schéma et donc à son évolution d'une version à l'autre. En conséquence, une opération d'évolution de schéma peut être écrite comme suit :

$$\forall e \in E : e_{v+1} \longleftarrow Migration_{v+1}\ (e_v)$$

$$\text{évolution(e)} \longleftarrow \sum_{i=1}^{v} Migration_{v+1}\ (e_v)$$

E Décrit l'ensemble d'entités à évoluer (E peut être l'ensemble des noeuds, l'ensemble des relations, l'ensemble des propriétés ou même l'ensemble des labels d'une BDoG). v représente le nombre d'éléments à évoluer.

Dans le cadre de ce papier, nous allons nous concentrer sur l'étape de de la vérification des contraintes. nous allons détailler pour chaque opérateur de migration ces différentes contraintes à vérifier, afin d'éviter les pertes de données et les erreurs d'exécution.

5 Opérations de migration simples

Principalement, il existe deux types d'opération : à savoir les opérations de type simple et les opérations de type multiple. Les opérations de type simple incluent l'ajout, la suppression et la mise à jour, tant dit que celles de type multiple incluent les opérations de déplacement des entités et l'opération d'avoir une copie de la même entité. Dans le cadre de ce papier, nous allons présenter les opérations de type simple.

Avant de détailler les opérations de migration, nous présentons, dans ce qui suit, une formalisation d'une (BDoG) qui facilite la présentation des opérations de migration.

5.1 Formalisation du modèle orienté graphe

Les bases de données orientées graphes sont considérées comme étant des modèles NoSQL les plus puissants, qui se basent sur la théorie des graphes pour manipuler et stocker les données. Conçu pour explorer les données fortement connectées, la structure des bases de données

orientée graphes permet la modélisation des données complexes d'une façon simple et intuitive où il n'y a pas une différence entre les données et les relations.

Une base de données orientée graphe est composée principalement des noeuds qui représentent les différentes entités et des arcs représentent les différentes relations qui organisent les noeuds. Les noeuds et les relations peuvent avoir des propriétés formées par une paire clé / valeur. une base de données orientée graphe est un couple (N, R) avec :

— N : l'ensemble total des noeuds n_i qui forment les entités de la BDoG.
— R : l'ensemble total des relations qui joignent les différents noeuds.

Noeuds. Chaque noeud n est composé par son identifiant id_n et éventuellement un ensemble de propriétés P_n. Il convient de noter que l'identifiant ne contient aucune information sémantique exprimée à travers un ou plusieurs labels. Un noeud peut être présenté comme suit :

$$\forall n \in N : \text{n=}(id_n, P_n, L_n)$$

— Id_n : est l'identifiant unique de chaque noeud n.
— P_n : est l'ensemble des propriétés (p_1, ..., p_n) liées au noeud. Il est à noter qu'une propriété est formée par une paire clé / valeur.
— L_n : est l'ensemble des labels (l_1, ..., l_n) attachés au noeud.

Relations. Une relation r est définie comme (id_r, Ne_r, Ns_r, T_r, P_r) qui contient l'ensemble des noeuds entrants/sortants, le type de relation et son ensemble des propriétés.

$$\forall r \in R : \text{n=}(id_r, N_{er}, N_{sr}, T_r, P_r)$$

.

— Id_r : un identifiant attribuer automatiquement et qui ne contient pas de valeurs sémantiques.
— N_{er} : l'identifiant du noeud entrant,
— N_{sr} : l'identifiant du noeud sortant,
— T_r : le type de relation, il est une chaine de caractère qui porte le nom de la relation.
— P_r : un ensemble de propriétés (p_1, ..., p_r) spécifique à une relation particulière.

Il convient de noter que le noeud de début (entrant) N_{er} et le noeud de fin (sortant) N_{sr} peuvent être identiques (un noeud peut avoir relations à lui-même) (Castelltort et Laurent, 2013). Dans ce cas on parle de boucle dans la théorie des graphes et une relation réflexive dans les bases de données.

5.2 Opération d'ajout

Dans le contexte des (BDoG), l'opération d'ajout concerne l'ajout d'un noeud et l'ajout d'une relation.

5.2.1 Opération d'ajout d'un noeud

Pour éviter la redondance ou la perte de données le temps où l'application tente d'ajouter un noeud n(id_n, P_n, L_n), la vérification de l'existence de celui-ci doit être faite. Dans ce cas, il faut vérifier l'ensemble de ses labels L_n et son ensemble de propriétés P_n. En effet, le nouveau

noeud à ajouter ne doit pas avoir le même ensemble de labels et le même ensemble de propriétés qu'un noeud existant. L'opération d'ajout d'un noeud s'écrit comme suit :

```
Ajout-Noeud
Entrées : BDoGv(Nv, Rv) :BDoG dans sa version v
          n(idn,Pn,Ln) :le noeud à ajouter
Résultat : BDoGv+1 (Nv+1, Rv) BDoG après évolution
Pré-conditions : ∄ m (idm,Pm,Lm)∈ N\ Lm-Ln=∅ ou Pm-Pn=∅
Traitement :
          Nv+1=N∪n
          Pv+1=P∪Pn,
          Lv+1=L∪Ln
POST-conditions : Le noeud ajouté ne doit pas être
obsolète.
```

5.2.2 Opération d'ajout d'une relation

Une application en cours d'utilisation, peut souhaiter l'ajout d'une nouvelle relation entre les noeuds. Dans ce cas, il faut d'abord vérifier l'existence des deux noeuds qui sont la source et la destination de relation. De plus, on doit vérifier l'existence de la relation elle-même, et s'il s'agit d'une relation inversée d'une relation existante ou non. L'opération d'ajout d'une relation s'écrit comme suit :

```
Ajout-Relation
Entrées : BDoGv(Nv, Rv) :BDoG dans sa version v
          r (idr, Ner, Nsr, Tr, Pr) :la relation à ajouter
Résultat : BDoGv+1 (Nv+1, Rv) BDoG après évolution
Pré-conditions : (Ner∈ N et Nsr∈ N) et
          ∄ w(idw, New,Nsw , Tw, Pw)\(New=Ner et Nsw=Nsr et
Tw=Tr)
          ou (New=Nsr et Nsw=Ner et Tw=Tr)
Traitement :
          Rv+1=R∪r
          Pv+1=P∪Pr
```

5.2.3 Opération d'ajout d'une label ou une propriété

L'ajout d'un label ou d'une propriété ne peut être fait sans spécifier le noeud ou la relation qui lui est attachée. Cela entre dans le cadre d'ajout d'un nouveau noeud ou une relation ou mettre à jour un noeud ou une relation.

5.3 Opération de modification

L'opération de modification indique tout type de changement au niveau des noeuds ou des relations.

5.3.1 Opération de modification sur les noeuds

La mise à jour des noeuds peut être effectuée par de nombreuses opérations telles que :
— ajout/suppression d'une propriété.
— ajout/suppression d'un label.
— mise à jour d'une propriété : renommage, retypage.

Modification d'un noeud par ajout de propriété. À chaque fois où l'application essaye de mettre à jour un noeud en y ajoutant une nouvelle propriété à un noeud, on doit vérifier l'existence du noeud en premier lieu, puisque cette propriété ne figure pas dans le noeud spécifié. L'opération d'ajout d'une propriété à un noeud s'écrit de la manière suivante :

```
Modification-noeud
Entrées : BDoGv(Nv, Rv) :BDoG dans sa version v
          n(idn,Pn,Ln) :le noeud à modifier
          Pa : la propriété à ajouter
Résultat : BDoGv+1 (Nv+1, Rv) BDoG après évolution
Pré-conditions : n∈N et Pa∉Pn
Traitement :
          Pn=Pn∪{Pa}
          Pv+1=Pn∪{Pa}
```

Modification d'un noeud par suppression de propriété. Dans le cas où l'application essayerait de mettre à jour un noeud en y supprimant une propriété, on doit vérifier l'existence du noeud, puisque cette propriété existe dans le noeud spécifié. L'opération de suppression d'une propriété d'un noeud s'écrit de la manière suivante :

```
Modification-noeud
Entrées : BDoGv(Nv, Rv) :BDoG dans sa version v
          n(idn,Pn,Ln) :le noeud à modifier
          Ps : la propriété à supprimer
Résultat : BDoGv+1 (Nv+1, Rv) BDoG après évolution
Pré-conditions : n∈N et Ps∈Pn
Traitement :
          Pn=Pn\{Ps}
          Pv+1=Pn\{Pa}
```

Modification d'un noeud par renommage de propriété. Quand l'application tente de renommer une propriété d'un noeud, on doit vérifier l'existence du noeud, et aussi que cette propriété doit figurer dans le noeud spécifié. L'opération de renommage s'écrit comme suit :

```
Modification-noeud
Entrées : BDoGv(Nv, Rv) :BDoG dans sa version v
          n(idn,Pn,Ln) :le noeud à modifier
          Pold : la propriété avec l'encien nom
          Pnew : la propriété avec le nouveau nom
Résultat : BDoGv+1 (Nv+1, Rv) BDoG après évolution
Pré-conditions : n∈N et Pold∈Pn
Traitement :
          renommer pold par pnew
          Pv+1=Pv
```

Modification d'un noeud par ajout d'un label. Lorsque l'opération déclenchée par l'application consiste à ajouter un label à un noeud, on doit vérifier l'existence du noeud, et que ce label n'est pas attribué au noeud spécifié. En plus, il est à noter que les SGBD sont sensibles à la casse (Ex. "ACTOR" et "actor" sont considérés comme deux labels différents). L'opération d'ajout d'un label à un noeud s'écrit de la manière suivante :

```
Modification-noeud
Entrées : BDoGv(Nv, Rv) :BDoG dans sa version v
          n(idn,Pn,Ln) :le noeud à modifier
          La : le label à ajouter
Résultat : BDoGv+1 (Nv+1, Rv) BDoG après évolution
Pré-conditions : n∈N et La∈Ln
Traitement :
          Ln=Ln∪{La}
          Lv+1=Ln∪{La}
```

Modification d'un noeud par suppression de label. Si l'application vise à supprimer un label d'un noeud, après avoir vérifié l'existence du noeud, on doit vérifier que le label existe parmi les labels du noeud spécifié et qu'il n'est pas le seul label dans la BDG. L'opération de suppression de label d'un noeud s'écrit de la manière suivante :

```
Modification-noeud
Entrées : BDoGv(Nv, Rv) :BDoG dans sa version v
          n(idn,Pn,Ln) :le noeud à modifier
          La : le label à supprimer
Résultat : BDoGv+1 (Nv+1, Rv) BDoG après évolution
Pré-conditions : n∈N et Ls∈Ls
Traitement :
          Ln=Ln\{Ls}
          Lv+1=Ln\{Ls}
```

5.3.2 Opération de modification sur les relations

La mise à jour des relations peut être effectuée par de nombreuses opérations telles que :

— changer le noeud de départ et/ou le noeud d'arrivée :Cette opération revient à ajouter une nouvelle relation avec un noeud de départ ou un noeud d'arrivés différents.
— modifier le type de la relation : Changer le type d'une relation est considéré comme une opération composée.
— ajout/suppression/modification d'une propriété : Ce cas est similaire au cas de l'ajout /modification/ ou suppression des propriétés pour les noeuds.

5.4 Opération de suppression

Le dernier cas est lorsque l'application vise à supprimer un noeud ou une relation. L'opération de suppression d'un noeud nécessite la vérification de l'existence du noeud à supprimer, de plus elle nécessite la vérification d'existence des relations reliées au noeud que l'on veut supprimer.

L'opération de suppression d'un noeud ou une relation se rassemble à supprimer une propriété. Dans ce papier on va présenter uniquement celle de la suppression d'un noeud qui s'écrit de la manière suivante :

```
Suppression-noeud
Entrées : BDoGv(Nv, Rv) :BDoG dans sa version v
          n(idn,Pn,Ln) :le noeud à supprimer
Résultat : BDoGv+1 (Nv+1, Rv) BDoG après évolution
Pré-conditions : ∄ r(idr, Ner,Nsr , Tr, Pr)\(Ner=n ou Nsr=n)
Traitement :
          R=R\{r}; Rv+1=R\{r}
          N=N\{n}; Nv+1=N\{n}
          Ln=Ln\{Ls}; Lv+1=Ln\{Ls}
          Pn=Pn\{Ps}; Pv+1=Pn\{Ps}
```

6 Conclusion

Ce travail étudie l'évolution des schémas et la migration des données dans des bases de données orientées graphes. Comme une base de données évolue, son schéma le fait aussi. Toutefois, les bases de données orientées graphes ne sont pas encore équipées d'outils de gestion de schéma pratiques.

Dans cet article, nous mettons en évidence les bases pour la gestion systématique de l'évolution des schémas dans le cadre de bases de données orientées graphes. Nous avons présenté la formalisation de la base de données de graphes et nous avons défini des opérations de migration simples en se focalisant sur les différentes conditions pour chaque opération. Dans des futurs travaux, nous allons proposer des processus pour la migration des entités d'une base de données orientée graphe en mettant en évidence les opérations de migration proposées. nous visons aussi à programmer les différentes étapes de notre approche et de définir des opérations de migration complexes.

Références

Castelltort, A. et A. Laurent (2013). Representing history in graph-oriented nosql databases : A versioning system. In *Digital Information Management (ICDIM), 2013 Eighth International Conference on*, pp. 228–234. IEEE.

Cleve, A., M. Gobert, L. Meurice, J. Maes, et J. Weber (2015). Understanding database schema evolution: A case study. *Science of Computer Programming 97*, 113–121.

Curino, C., H. J. Moon, A. Deutsch, et C. Zaniolo (2013). Automating the database schema evolution process. *The VLDB Journal 22*(1), 73–98.

Florian Haubold, Johannes Schildgen, S. S. S. D. (2017). Controvol flex: Flexible schema evolution for nosql application development.

Klettke, M., U. Störl, S. Scherzinger, et O. Regensburg (2015). Schema extraction and structural outlier detection for json-based nosql data stores. In *BTW*, Volume 2105, pp. 425–444.

Klettke, M., U. Störl, M. Shenavai, et S. Scherzinger (2016). Nosql schema evolution and big data migration at scale. In *Big Data (Big Data), 2016 IEEE International Conference on*, pp. 2764–2774. IEEE.

Li, X. (1999). A survey of schema evolution in object-oriented databases. In *Technology of Object-Oriented Languages and Systems, 1999. TOOLS 31. Proceedings*, pp. 362–371. IEEE.

Mahfoudh, M. (2015). *Adaptation d'ontologies avec les grammaires de graphes typés : évolution et fusion*. Thèse de doctorat, Université de Haute Alsace-Mulhouse.

Manousis, P., P. Vassiliadis, A. Zarras, et G. Papastefanatos (2015). Schema evolution for databases and data warehouses. In *European Business Intelligence Summer School*, pp. 1–31. Springer.

Meurice, L. et A. Cleve (2017). Supporting schema evolution in schema-less nosql data stores. In *Software Analysis, Evolution and Reengineering (SANER), 2017 IEEE 24th International Conference on*, pp. 457–461. IEEE.

Qiu, D., B. Li, et Z. Su (2013). An empirical analysis of the co-evolution of schema and code in database applications. In *Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering*, pp. 125–135. ACM.

Ringlstetter, A., S. Scherzinger, et T. F. Bissyandé (2016). Data model evolution using object-nosql mappers : Folklore or state-of-the-art ? In *Proceedings of the 2nd International Workshop on BIG Data Software Engineering*, pp. 33–36. ACM.

Saur, K., T. Dumitraş, et M. Hicks (2016). Evolving nosql databases without downtime. In *Software Maintenance and Evolution (ICSME), 2016 IEEE International Conference on*, pp. 166–176. IEEE.

Scherzinger, S., T. Cerqueus, et E. C. de Almeida (2015). Controvol : A framework for controlled schema evolution in nosql application development. In *2015 IEEE 31st International Conference on Data Engineering*, pp. 1464–1467. IEEE.

Scherzinger, S., E. C. de Almeida, T. Cerqueus, L. B. de Almeida, et P. Holanda (2016). Finding and fixing type mismatches in the evolution of object-nosql mappings. In *EDBT/ICDT Workshops*.

Scherzinger, S., M. Klettke, et U. Störl (2013). Managing schema evolution in nosql data stores. *arXiv preprint arXiv:1308.0514*.

Scherzinger, S., M. Klettke, et U. Störl (2015a). Cleager: Eager schema evolution in nosql document stores. In *BTW*, pp. 659–662.

Scherzinger, S., U. Störl, et M. Klettke (2015b). A datalog-based protocol for lazy data migration in agile nosql application development. In *Proceedings of the 15th Symposium on Database Programming Languages*, pp. 41–44. ACM.

Störl, U., D. Müller, M. Klettke, et S. Scherzinger (2017). Enabling efficient agile software development of nosql-backed applications. In *BTW*, pp. 611–614.

Subotic, D., P. Poscic, et V. Jovanovic (2014). Data warehouse schema evolution: state of the art. In *Central European Conference on Information and Intelligent Systems*, pp. 18. Faculty of Organization and Informatics Varazdin.

Summary

NoSQL databases are becoming popular as a backend to today's web applications. In fact, the graph-oriented model can effectively manage highly interconnected data. Because of their flexibility, graph-oriented databases do not require global schema definition, but the schema is often maintained in the source code of the application. With this flexibility, devel-opers can manage the diversity of data, but they can struggle with the growth of the data structure and the correct management of persistent data. Indeed, the problem of evolution is not well treated, especially when it comes to analyzing the source code of the application and its history. In this article, we propose an approach to control the evolution of schemas of graph-oriented data bases by analyzing the source code of applications and we are mainly interested in detailing some migration operations.

Specific criteria to measure the strategic alignment in the informatics system

Khalid EL KHOURASSANI*, Rabia MARGOUBI**, Abdeslam ENNOUAARY***

*Rues Al Jouaze & Al Joumaize, Rabat, Maroc
khourassani@mhupv.gov.ma
**Avenue Allal El Fassi، Rabat, Maroc
m.rabia@inpt.ac.ma
***Avenue Allal El Fassi، Rabat, Maroc
abdeslam@inpt.ac.ma

Abstract. Firms in general aim to meet their goals and improve both their image in the market and their turnover through sticking to good strategies, having a clear vision and abide by the criteria of IT-governance which emphasizes on: the aligning strategies, managing risks, managing resource, measuring performance and creating value.
This study focuses only on one pillar of It-governance: Strategic alignment in the informatics systems and offers different methods to measure it. Each method uses some attributes. But in fact, also after measuring it, it is interesting to know the relationship of the process of COBIT and the Strategic Alignment. That is why, this study is important to any manger or it-governance professional whose aim is to improve the strategic alignment in its informatics system, especially, thanks to the Meta Model ISO 19440 , today companies can choose their own attributes to measure the Strategic alignment in their informatics system and pick up the processes to use[1][9].

1 Introduction

Each company aims to enhance its image through an efficient policy, bring about new ideas and new breakthrough so that it can attract more customers and guarantee a respectful share in a certain market. Especially, we live this competitive era which characterized by the speed of information and technology. In the same context, it is undeniable the crucial role of It-governance to reach the above mentioned goals and many more.

My purpose in the following lines is to shed light on the methods to measure the Strategic Alignment in the informatics System, and to notice the relationships between this pillar of IT- governance and the process of COBIT.

According to Weil and Ross, the IT governance specifies the right decision and the ac-countable standard to encourage the desirable behavior in using IT [2], which consists in setting a good management and ambitious leadership to meet the planned objectives. Anyway, there are five axes of IT governance, but we are going to focus only on the criteria to measure one pillar: the Strategic Alignment in Informatics system, which have been carefully chosen to help any IT firm, which wants to have better Strategic Alignment in its informatics system.

This paper is organized as follows: the abstract of the topic, an introduction. the second section describes the State of the Art , after that ,in the third section the five pillars of IT Governance are represented , and next , the fourth section , investigates the question of defining criteria to measure the Strategic Alignment in the informatics Systems , at that time we are going to represent some different cases , and give the criteria to measure it in each case , afterwards we will represent the methodology used to choose each criterion , also the paper , shed light about the relationship between the different PO of Cobit and the Strategic Alignment in the IS , and finally we will talk about the limits and the perspectives on this work in the conclusion.

2 State of the Art

This section presents the state of the Art related to our work, we will give the definitions of the terms: Strategic Alignment, and then we will talk about the former works which have been done to know the relation between it and IT governance, and also, the works related to measure it in the IS, afterwards we will talk also about a former work related to The Metamodel ISO 19440, and give the definition of it.

- **Strategic Alignment** : Henderson and Venkatraman define the Strategic Alignment in the informatics systems as an organizational process where the mission , goals , objectives and Activities of the IS change over time with changes in the organization [1] , there are other definitions of the term "Strategic Alignment" but this one is appropriate to the context of our study .

- **Relation between The Strategic Alignment and IT governance:** A lot of studies focused on this topic, and especially the correlation between the process of Cobit and the Strategic Alignment, [4] [5] [6] [8], suggested a method to know the relationship between the PO of Cobit and the Strategic Alignment, this method will be explained later, also other studies [12] [13] demonstrated how a high Strategic Alignment can have a very positive effect on the firm's productivity.

- **Measuring the Strategic Alignment in the IS**: our state of the Art demonstrated the measuring the Strategic Alignment in the IS is not an easy task , it's different from one case to another , but there are some mathematical approaches of this study which use some methods and give a percentage K% about how the Strategic Alignment is high [12] . In addition to that, there are some managerial approaches which define some criteria to measure the Strategic Alignment in the IS, which are going to present it later. Few researchers has addressed the question of measuring the Strategic Alignment in the IS , and previous works have been limited to measure it , but each one in a specific case , Despite this interest as far as we know no one defined criteria of measuring the Strategic Alignment in the IS.

- **The Metamodel ISO 19440:** it is a process oriented standard model, which offers four views of the firm: the organizational view, the functional view, the informational view and the view of the resources. The functional view is related to the process, and the view of the resources is related to the resources used by the process [1]. a former study [1] , suggested a method to raise the strategic alignment in the IS , by using this Metamodel , it's possible to get a matrix which indicates , which resource used by which process , and then by using the Gallois Lattices , it's possible to know the different consistency's level between the firms parts , in a multi-views approach. This corresponds in fact with the aim of the Strategic Alignment.

The state of the art showed also, a serious need, to know the level of the Strategic Alignment in the IS of a certain firm and the process which raise it .A former study [14], suggested a mathematical approach which uses a function and then, gives an idea about the best standard of it governance to use according to a specific axe, but to make this mathematical calculation, it was suggested to

know the relation about the Cobit process and the different axes of it governance. More over this methodology has been applied to cases without focusing on the relation between the PO of Cobit and the axes of IT governance.

3 The five pillars of IT Governance

According to the IT Governance Institute, IT governance is the responsibility of the board of the directors and executive management. It is an integral part of enterprise governance and consists in the leadership and organizational structures and processes that ensure that the organization's IT sustains and extends the organization strategies and objectives. ISACA organization (Information Systems Audit Control and Association) considers that it should be done by 5 pillars:

- Strategic Alignement ;
- Value Delivery ;
- Risk management ;
- Resource management ;
- Performance measurement ;

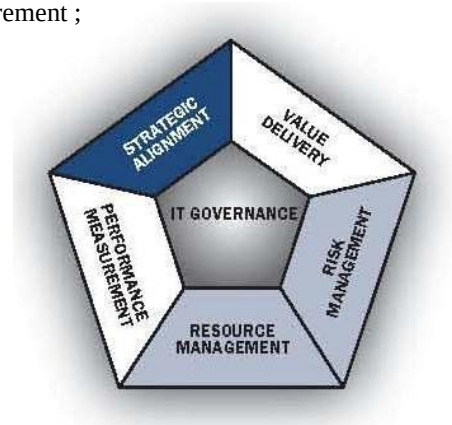


FIG. 1 –5 pillars of IT governance [2].

Strategic Alignment is applied in each department in the firm to be aligned with the other departments objectives that includes the major objective. Actually, each company can be leveled according to what extent its Strategic Alignment in Informatics Systems is mature which is valued by the Luftman model. But; it is possible to be done by other methods.

Well, Measuring the Strategic Alignment is not as easy as it could be imagined because the criteria used for the mentioned objective can be not the same from one case to another and we will talk about that in a detailed way after.

Therefore, if firm want to have the better It-governance possible , they must focus on measuring the Strategic Alignment ,and after focusing on the process of COBIT that raise it, and this paper describes what each Process can do to raise the mentioned pillar.

4 Measuring the Strategic Alignment in Informatics Systems

This section talks about analyses the criteria to measure the Strategic Alignment in the IS , it gives an overview about the luftmodel which used to do this task, but we will see , how many this model tends to focus on the measuring in general case, in fact , this raises many questions about the validity of this model in other cases such as : Clients/Vendors firms , or an inter-enterprise.

4.1 Methods to measure the Strategic Alignment in Informatics Systems

As mentioned before, Strategic Alignment consists of that the objectives of each department in the firm must be aligned with the other departments objectives, and included in its major objective [15]. What's more, it is not easy, to measure it, Most of the studies which were done at this level, concerned only managerial approaches. In fact, there is a model to measure the mentioned pillar: it is the Luftman Model which measures the Maturity of the Strategic Alignment in Informatics Systems (5 levels: 1. Initial process, 2.committed process, 3.Established Process, 4.Improved Process, 5.Optimized process).

Moreover, the Luftman Model verifies the characteristics of these Attributes [7]:

- **Communications**: The effectiveness of exchange of information.
- **Competency / value measurements**: The metrics used to measure the competency.
- **Governance**: Defines the persons who have the authority to make IT decisions, management and leadership.
- **Partnership**: The style of relationship between the business and IT.
- **Scope and Architecture**: The Architecture of the business.
- **Skills**: Measures the human resources skills.

Nevertheless, those attributes cannot be the same in other cases, in fact, for a business Clients / Vendors , it is important to care about the Strategic Alignment between the IT and Vendors. On one hand, a vendor deploys more services, generates more revenue, and invests into Research and development. On the other hand, the Customer implements new services, generates more revenue, and invests into more Services. In this case, it would be wise to add a new Attribute to measure the trust of the Customer to the Vendor [3]. Thus, the Attributes of the Luftman model are not the same from one case to another.

Communication	Value	Governance	Trust
Understanding Client Business by Vendor. Understanding Client Business by Client. Inter/intra Organization. Protocol Rigidity. Knowledge Sharing. Liaison Effectiveness.	Clients Metrics. Vendor Metrics. Balanced Metrics. Service Level Agreements. Benchmarking. Formal Assessment. Continuous Improvement.	Client Strategic Planning. Vendor Strategic Planning. Organization Structure. Budgetary Control. It investment management. Steering Committee. Prioritization Process.	Trusting Deliverables. Trusting Timelines and Schedules. Trusting Competencies. Trusting after-sale support (SLA). Trusting Business processes and procedures. Communication Transparency and incident Reporting.
Partnership	Scope and architecture	Skills	
Business perception of other business value. Role of Vendor in Client's Strategic Planning. Shared Goals and Risks. Vendor Account Management. Business Champion.	Traditional Enabler/Driver. Standards Articulation. Architectural Integration. Architectural Transparency, Agility, Flexibility. Manage Emerging Technology.	Innovation Entrepreneurship. Cultural locus of power. Management Style. Change Readiness. Career Training. Hiring and Retaining.	

FIG. 2 – the attributes used to measure Strategic Alignment In informatics Systems of Clients/Vendors firm [3].

What's more, in an inter-enterprise, the maturity model to measure the Strategic Alignment in Informatics Systems can be suggested and done in a focus group session; which can be inspired from the method of the maturity of software's production which uses SW-CMM proposed by Carnegie Mellon University's Software Engineering Institute: in this case, Any mature model which allow inter-enterprise collaborations to evaluate their Strategic Align-ment and provides a good-business plan transition to make improvements is accepted.

It was proposed that the criteria to evaluate Strategic Alignment in Informatics Systems would be as follows [11]:

- **Company's Architecture:** the interconnections relationships, the technology infra-structure.
- **Business /IT process:** the process to meet the common goals of the companies.
- **Workflow Structures:** responsibilities and roles specifications
- **IT Governance:** the leadership and management that consists of setting strategies and meeting IT enterprises goals.
- **Coordination:** the mechanisms to manage interactions and to share common resources.

In addition to that, the criteria can be chosen as any company's professionals want; a focus group session can be organized to choose criteria. Therefore measuring the Strategic Alignment in Informatics Systems depends on the case, as we have mentioned. The attributes of measuring its maturity are not the same. And can also, be proposed in focus group sessions.

4.2 The need to compare the Standards according to those criteria to create a good IT-governance

A business, that aims to improve its turnover, will have to set a good IT-governance, and especially to have a good Strategic Alignment in Informatics Systems. Depending on the case , it would be measured by some criteria as we have seen in the last chapter, neverthe-less, in a good Strategic Alignment , the characteristics in the criteria seen , would be all very good ,to have the best possible level of maturity of the Strategic Alignment .

To summarize, the attributes used in measuring the Strategic Alignment in Informatics Systems, are:

- **In general kind of firm:** Communications, Competency / value measurements, Governance, Partnership, Scope and Architecture, Skills.

- **In Clients/Vendors firm:** Communications, Competency / value measurements, Governance, Partnership, Scope and Architecture, Skills, trust.

- **In an inter-enterprise collaboration:** Company's Architecture, Business /IT pro-cess, Workflow Structures, IT governance Coordination.

A focus group session used can suggest other criteria, at the moment; this paper considers just those criteria.

What's more, one of the main problems is to know the relationships between the process of Cobit and the level of the Strategic Alignment in the informatics Systems. Generally speaking, this is a serious problematic and hard to identify directly. In fact let's suppose that there is a company A which uses PO1, PO2, PO3 of Cobit , after measuring the Strategic Alignment in its informatics Systems we have discovered that is equal to 4/5 for example , and after we have done the same experience of the company B which uses PO1, PO2, PO3, PO4 , we have discovered that is equal to 3/5 , it has become lower , because there is no alignment between the Process 3 and the Process 4 , So the most remarkable limitation is it's impossible directly to predict the relationship between PO3 and the Strategic Alignment in the Informatics system, but further analysis showed that there is still one method to know it is to do it by statistics methods, which will be explained after. The correlation between the Strategic Alignment and the PO of Cobit is a result which will interest any manager.

Thanks to the Meta model ISO 19440, it is possible to make model to the firm, and in that model, the business can include the criteria chosen. Because this Model offers a framework and constructs for Enterprise Modeling. And in the model proposed, it is possible to choose the attributes of measuring Strategic Alignment in the informatics systems. And also pick up the process of COBIT that was identified as important to raise it [1] [9].

5 The methodology used to choose the criteria

- The limit of this study is that it focuses only the Strategic Alignment In informatics Systems, as one pillar of IT governance; nevertheless, it is possible to make other studies by focusing on other pillars. It helps any firm which wants to improve this pillar, and from this angle the comparison is made.

- Looking the attributes used to measure the Strategic Alignment In informatics Systems.

- Identifying the relationship between the process of COBIT and the Strategic Alignment in the informatics system maturity.

To measure the Strategic Alignment we need to define criteria that put these attributes together into a table that make sense of the level of this pillar according to the type of the business.

12 criteria to measure the strategic alignment in the informatics system
Communications
Competency / value measurements,
Governance
Partnership
Scope and Architecture
Skills
Trust – if the company is Clients/Vendors Company
Company's Architecture in an inter-enterprise collaboration.
Business /IT process in an inter-enterprise collaboration
Workflow Structures in an inter-enterprise collaboration
It governance in an inter-enterprise collaboration
Coordination in an inter-enterprise collaboration

TAB. 1 –*Criteria used measure the Strategic Alignment*

6 The relationship between the process of COBIT and the 12 criteria :

To know the relationship between the process of COBIT and the 12 criteria, we need to make a survey [4] [5] [6] [8]. This would be the best method to know which process must be used to guarantee a high level of maturity to each criterion, the methodology of the survey is explained below:

- Specifying a context of the study and making a questionnaire to different kind of firm, with different sizes, by using the Delphi method to represent firm in the con-text.
- The questionnaire must ask questions about, the level of maturity of each criterion, and the governance practice process adopted in the business.

The results of the survey can be generalized for the firm in the context, because it's based on the Delphi method, and this is how, we can know, the processes used to make each criterion in a high maturity level.

So , we have seen that the fact of measuring the Strategic Alignment in the Informatics System is not easy as we could predict , the situation is different from one case to another , and in addition to that , it's hard to perceive directly the relationship between the Process of Cobit and this pillar of IT Governance , but for a manager who aims to raise it in its company he will be interested to know which Process he can focus on, in fact , if they were identified ,he can focus on them to raise the Strategic Alignment .

We have identified in this study 12 Criteria which are used to measure the measure the Strategic Alignment, depending of the type of its business, as were mentioned. What's more, As Cobit is a standard of IT Governance, if used; the manager must focus on the Process which raise it. But, the fact of identifying those processes is a serious Problematic. As it was mentioned in an example before, the relationship between them and the Strategic Alignment is hard to perceive. It's pretty confusing to tell what a Process can do for the Strategic Alignment. This Study proposes a survey to identify what each Process can do, which depends of the context of the study. After identifying them, any manager can focuses on or include them in the Meta Model ISO 19440, to design its own firm.

7 Conclusion

This study shed light on the methods to measure to Strategic alignment in informatics systems, and after proposes a survey to know the relationship between the process of Cobit and the Strategic alignment.

This study has addressed only one pillar of IT governance; nevertheless, it would be interesting to know the other methods used to measure other it governance pillars, like for example the created value, in fact if we have the criteria to measure the created value, we can do such surveys to know the process that raise it.

As a perspective to this study, further works need to be done : it is necessary to do the suggested survey to know the results of the comparison. In fact that would be an interesting result for any it-governance professional to make his own design to his firm since it is allowed and possible by using the Meta Model ISO 19440.

References

[1] Boulmakoul.A, Falih.N et Marghoubi.R (2013), *Deploying Holistic Meta-modeling for Strategic Information System Alignment.*

[2] COBIT 4.1; IT Governance Institute®, *http://www.isaca.org/*

[3] Dana.S.A, et Zualkerman I.A. (2012) *Measuring Strategic Alignment between Information Technology Clients and Vendors.* 2012 IEEE International Conference on Management of Innovation& Technology (ICMIT)

[4] De Haes.S, Greenberger W.V. (2011) *Analyzing the Relationship between IT Governance and Business/IT Alignment Maturity.* Proceedings of the 41st Hawaii International Conference on System Sciences

[5] Hosseinbeig.S, Karimzadgan M.D, Vahdat.D, et Askari .R. (2011) *Combination of IT Strategic Alignment and IT Governance to Evaluate Strategic Alignment Maturity.* 2011 5th International Conference on Application of Information and Communication Technologies (AICT)

[6] Hosseinbeig.S, Karimzadgan M.D, Vahdat.D, et Askari .R. (2011) *IT Strategic Alignment Maturity and IT Governance*.The 4th International Conference on Interaction Sciences

[7] Gosh.H., (1994). *A Comparison Of ISO 9000 And SEI/CMM For Software Engineering Organizations.* Software Testing, Reliability and Quality Assurance, 1994. Conference Proceedings.

[8] Maidin S.S, Arshad .N.H. (2010) *Information Technology Governance Practices in Malaysian Public Sector*.2010 International Conference on Financial Theory and Engineering

[9] Martin R.A. *ISO/DIS 19439 & 19440 Framework and Constructs for Enterprise Modeling.*

[10] Weil,P. et Ross,J. (2004). *IT Governance*, Boston Massachusetts, Harvard Business School Press.

[11] Tapia .R.S,Daneva.M, Van Eck.P (2008), *Validating Adequacy and Suitability of Business-IT Alignment Criteria in an inter-enterprise Maturity Model.*11th IEEE International Enterprise Distributed Object Computing Conference

[12] Elmanouar.A , Sadok.H, Benkhayat.A (2015) , *Firm Business Strategy And IT Strategy Alignment: A Proposal Of A New Model* , Computer science & Information Technologies" (CSIT'2015), 14-17 September 2015, Lviv , Ukranie

[13] Denford.J.S, Schobel.K.B (2012), *The Chief Information Officer and Chief Financial Officer Dyad – How an Effective Relationship Impacts Individual Effectiveness and Strategic Alignment,* 2012 45th Hawaii International Conference on System Sciences

[14] Chakir.A, Chergui.M , Medromi.H, Sayouti.A (2015) , *An approach to select effectively the best framework IT according to the axes of the governance IT, to handle and to set up an objective IT* ,Third World Conference on Complex Systems (WCCS).

[15] *www.piloter.org*

Résumé

Généralement , les entreprises visent à atteindre leurs objectifs et à améliorer leur image sur le marché et augmenter leur chiffre d'affaires en adoptant de bonnes stratégies tout en ayant une vision claire en respectant les critères de la gouvernance des technologies de l'information qui mettent l'accent sur : l'alignement stratégique, la gestion des risques, la gestion des ressources humaines, la mesure de la performance, et la création de la valeur. Cette étude se concentre seulement sur un seul axe de la gouvernance des technologies de l'information : l'alignement stratégique, et propose de différentes méthodes pour le mesurer, chaque méthode utilise des attributs. Mais en fait, après l'avoir mesuré, il est intéressant de connaître la relation entre le processus de COBIT et l'alignement stratégique. C'est la raison de laquelle cette étude est importante pour tout manager ou professionnel de la gouvernance qui a comme but d'améliorer l'alignement stratégique dans son système d'information. Surtout, que grâce au Meta model ISO 19440, les entreprises peuvent choisir leurs propres attributs pour mesurer l'alignement stratégique dans leur système d'informations et choisir les processus à utiliser.

Chapter

7

The Internet of Things: components challenges and opportunities

ASD'2018

Content

Distributed industrial communication based on MQTT and Modbus in the context of future industry

Mohamed Tabaa*,Safa Saadaoui*, Fabrice Monteiro**, Aamre Khalil**,
Abbas Dandache**, Abdellah Daissaoui*, Karim Alami*

* Pluridisciplinary Laboratory of Research and Innovation (LPRI), EMSI Casablanca, Morocco
med.tabaa@gmail.com

** Industrial Engineering, Production and Maintenance Laboratory (LGIPM), Lorraine University, Metz France

Abstract. Over the next few years, the Industrial Wireless Sensor Network (IWSN), which is a major part of the Industrial Object Internet (IIoT), will play a crucial role in transforming the industrial world by opening a new era of economic and competitive growth in the Industrial Revolution known as "Industry 4.0". IIoT is able to help organizations achieve better benefits in industrial manufacturing markets by increasing productivity, reducing costs and developing new services and products. In this paper we present a wireless industrial communication system based on Node-RED platform using Modbus protocol for smart factories.

1 Introduction

In 1980, the 3rd Industrial Revolution was the cause of a huge evolution of the industrial world. This evolution has been accompanied by network complexity that exceeds the reliability and strength of automation systems. Thus, the need to find a more comprehensive means of information exchange, which can ensure interconnection on a wider and more refined scale (A.Ajithkumar and al, 2017) (M.Ehrlich,L and al, 2016) (U.Gungor and al, 2009). Today, with the new technology, the gateway to the 4th Industrial Revolution, known as Industry 4.0, makes it possible to add intelligence to industrial systems based on interconnected physical or virtual objects, capable of communicating and transmitting information in a less-complex way. This, with less error, based on two main strategies, the Internet of Things (The Internet Of Things, IoT) and Cyber Physical Systems (CPS) (M.Wollshlager and al, 2017) (Jiafu wan and al, 2016).

Industrial wireless sensor networks (IWSN) have several advantages, including reducing the cost of deploying and building a controlled workspace. By installing IWSNs on workstations and attaching labels to current products, information about production operations can be collected efficiently and flexibly, and cyber-physical decisions can be made instantly with great accuracy (X.Shen and al, 2004).

To make manufacturing operations more agile, more flexible, and more responsive to customer needs and to promote competitive advantages, industrial companies intend to rely on the fourth industrial revolution for more automation and flexibility. With industry 4.0, it is now possible to create an intelligent factory where wireless sensors and many other advanced technologies are used. These tools enable the company to react more quickly to market changes and thus optimize production and improve customer satisfaction (M.Ehrlich,L and al, 2016) (Jiafu wan and al, 2016) (Federico Tramaris and al 2016).

With this paper we are particularly interested in the application of the Internet of things in the industrial field especially for smart factories wireless communication. We aim to replace all the heavy wiring in the factories with an intelligent and agile system based on smart things and smart communication for control and monitoring. We present in this paper an industrial communication based on Node-RED Framework using Modbus protocol and MQTT protocol.

This article is organized as follow: in section 2 we present a state-of-art of industry 4.0. the strategy of cyper-physical system in industry will be presented in section 3. In section 4, we present the main objective of IoT in industrial application. Implementation with Node-RED and all discussion will be presented in section4. Finally, conclusion and perspectives.

2 Industry 4.0

In the context of industry 4.0 and smart manufacturing, it is essential to support factory automation as well as flexibility in industrial environments that are considered difficult environments for wireless communication due to high noise, physical barriers, multi-path and interference from co-existing wireless devices (ShiyoungWang and al, 2016). The industry 4.0 concept designates the use of digital technologies and consists of building a controlled workspace using a large-scale deployment of wireless sensors. Introducing digital technologies into a manufacturing company requires building a digital factory to create digital products and provide a digital customer experience.

The strength of Industry 4.0 lies in the integration of human, machine and systems at the same time. It is not yet clear how future developments will actually progress; the results of research in this direction still cannot tell when the era of automation will end (Figure 1).

Today, it is the largest projects that drive the debate on industry 4.0, its future projections and the different long-term visions. According to the researchers, we will have to wait at least ten years to see the deployment of this revolution's technologies. The journey from the era of automation to the era of industry 4.0 is not so obvious, training needs to be put in place, new standards, costly installations and high investment requirements, basically a new ground with new ground rules. Overall, it can also be assumed that the individual elements of industry 4.0 will be carried out on a larger scale in subsequent stages. Automated systems will continue to play a central role in production control over the next five years. However, they will have to meet additional requirements such as providing data for new business models and exchanging information online with other operational systems. The merging of virtual and physical worlds with cyber-physical systems and the resulting fusion of technical and business processes paves the way for a new industrial era better defined by the concept of the intelligent factory (M.Ehrlich,L and al, 2016) (U.Gungor and al, 2009).

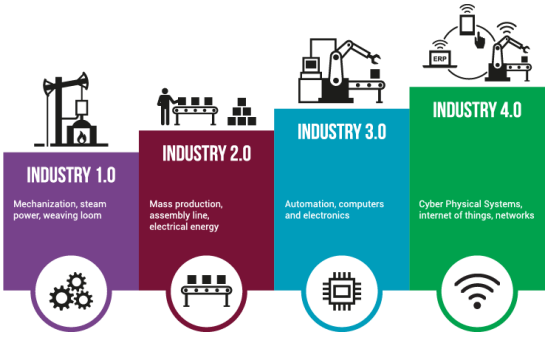


FIG 1 : Industry 4.0

3 CPS for industrial communication

Industrial sectors, the automotive industry, energy saving and, in particular, production technology will be transformed by new value chain models. Globalization, urbanization, demographic change and energy transformation are the transformative forces that stimulate the technological impulse to identify solutions for a world in flux. In the future, industry 4.0 will make contributions to human safety, efficiency, comfort and health in a way that is not imagined before. In doing so, they will play an important role in tackling the fundamental challenges posed by demographic change, scarcity of natural resources, sustainable mobility and energy change (JayLee and al, 2015).

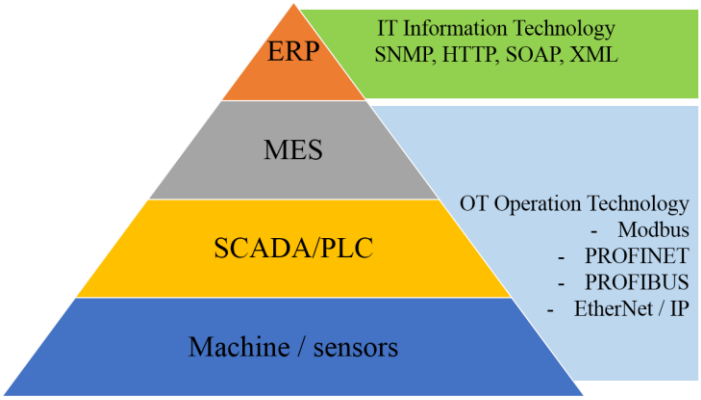


FIG 2 : CPS in Industry 4.0

The deployment of cyber-physical systems in production systems gives birth to the intelligent factory (Fig.2.). Intelligent plant products, resources and processes are characterized by cyber-physical systems. Offering significant real-time benefits in terms of quality, time, resources and costs compared to conventional production systems. The Smart Factory is designed according to sustainable, service-oriented business practices. These insist on adaptability, flexibility, self-adaptability and learning characteristics, fault tolerance and risk man-

agement, which shows that the application of the Industry 4.0 concept necessarily implies optimizing the cost and time of production, and thus a large margin of gain that could contribute to industrial development.

4 Industrial Internet of Thing

The Internet of Things (IoT) is a system of interdependent computer peripherals, machines, sensors, objects, animals or people with unique identifiers that can transmit and receive data over a network, without human intervention or computer interaction. It is the network interconnection of objects equipped with ubiquitous intelligence that has evolved from the convergence of wireless technologies, micro-electromechanical systems (MEMS), micro-services and the Internet. It also analyzed the data generated by unstructured machines to provide information (KunWang and al, 2016).

Technology	Standard	Frequency	Range	Transmission Speed
Bluetooth	Bluetooth 4.2	2,4 GHz (ISM)	50-150 m (Smart/BLE)	Mbit/s (Smart/BLE)
Zigbee	IEEE802.15.4	2,4 GHz	10-100 m	250 Kbit/s
Z-wave	Z-Wave Alliance ZAD12837/ITU-T G.9959	900MHz (ISM)	30 m	9,6 / 40 / 100 Kbit/s
6lowPan	RFC6282	2,4 GHz	--	--
Wi-fi	802.11n	2,4 GHz and 5 GHz	50m	600 Mbit/s max
Sigfox	Sigfox	900 MHz	30-50 km (E ruraux), 3-10 km (E urbains)	10-1 000 bit/s
LoRa	LoRa	3 frequencies	15 km	0,3-50 Kbit/s

Table.1. Communication technology

Practical applications of IoT technology are now found in many industries, including agriculture, chemicals, pharmaceuticals and petroleum, health, energy and transportation. The Industrial Internet of Things (IoT), which is the industrial application of IoT in industry, opens huge opportunities for a large number of new applications that promise to improve productivity in factories, and ensure a better allocation of resources. This revolutionary technology is attracting increasing attention from researchers and practitioners around the world. The set of protocols in the Internet of Things represents a language common to all connected systems, whatever their brand, operating system or software tools used. Table 1 shows the famous technologies used for IoT (JulienMinerauda and al,2016) (MakkelIglessias and al, 2017).

5 Distributed Industrial Communication

Industry 4.0 draws the intention on an application scenario that consists in building or forming a network of plants distributed geographically by using a flexible adaptation of the production and resource sharing capacities that ensures a wide and secure communication.

In this session we are developing our distributed architecture of industrial systems based on Node-Red and Modbus protocol.

5.1 Context

The implementation of industry 4.0 requires the adaptation of new methodologies and technologies. The industrial ecosystem is limited to installations wired between all sensors and actuators, based on a set of industrial protocols: Modbus, profibus, profinet and others. In our architecture, we opted for the Modbus protocol. This is a dialogue protocol based on a hierarchical structure between a master and several slaves, as shown in the figure 3.

The integration of IoT in the industrial field requires the use of wireless sensor-actuator networks that operate in real time for cyber-physics industrial systems, and when talking about cyber-physics systems we are faced with a wireless control system that includes several control loops that connect sensors, controllers and actuators via a wireless mesh network..

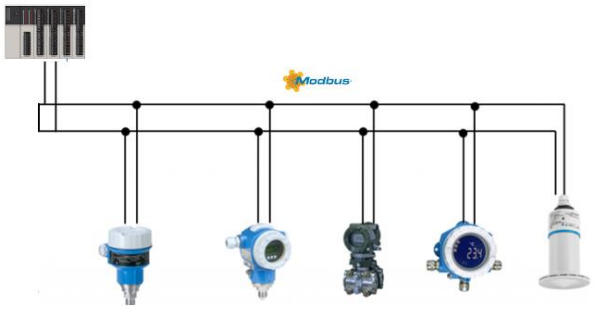


FIG 3 : Industrial Modbus communication

5.2 Discussions

To test this industrial wireless communication in the form of a network of connected objects, the IoT device integration layer has been added to the software infrastructure. This layer consisted of virtual nodes (simulated MQTT clients) and industrial communication architecture based on Node-RED and Modbus protocol. To add a node to the system, its identifiers must be registered with the broker MQTT. Virtual nodes, nodes for the sole purpose of simulating communication flows (Master-Slave) in the system, were simulated using client-mqtt libraries. Figure 4 shows an a virtual node created for sending information in the Modbus frame-responsive form to a Windows machine.

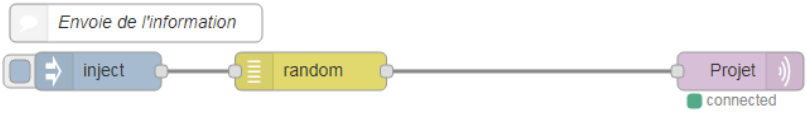


FIG 4 : Sending information

Network connectivity was provided by a single on-board computer via MODBUS frame exchange. A gateway has been set up to read/write MODBUS registers and transmit/receive messages via MQTT broker. The figure 5 shows the reception of the data via the gateway as well as the nature of the frame used and the figure shows the implementation of a node flow on the gateway. The gateway accessed the MODBUS over a wireless network to read its registers, then reformulate and send data to the external message broker via MQTT managed by AMQP Cloud. The exchange of information between the master and all slaves via MQTT and modbus is shown in Figure 6.

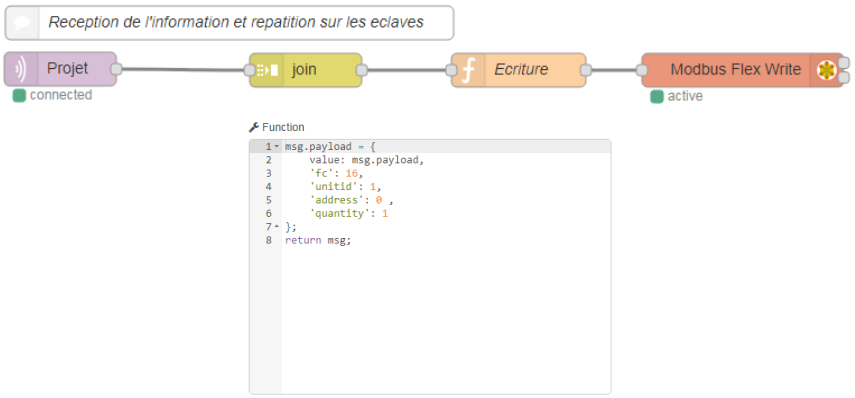


FIG.5 : reception of information and distribution among slaves

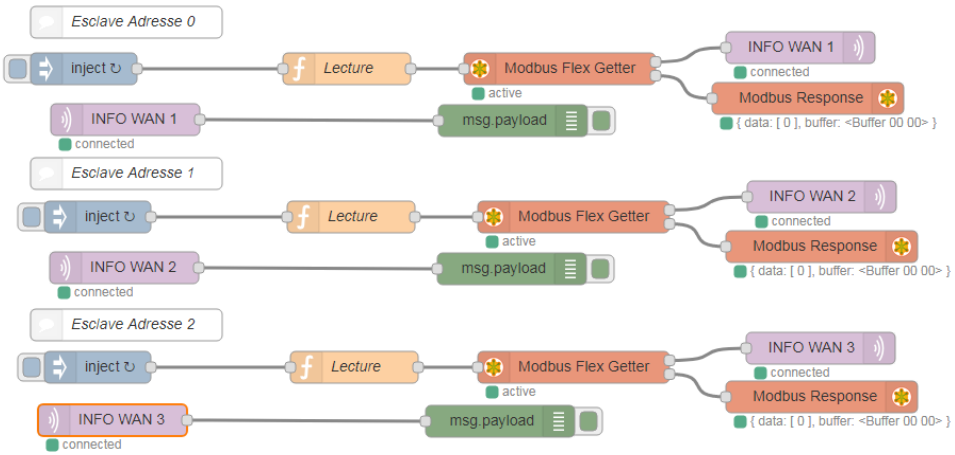


FIG.6 : Master/slave communication via MQTT protocol.

Industrial connected objects have several advantages, including reducing the cost of deploying and building a controlled workspace. By installing IoT systems on workstations and attaching labels to current products, information on production operations can be collected efficiently and flexibly, and cyber-physical decisions can be made instantly with great accu-

racy. Figure 7 shows a proposal for plant digitization, which provides intelligent and real-time access to all plant sensors.

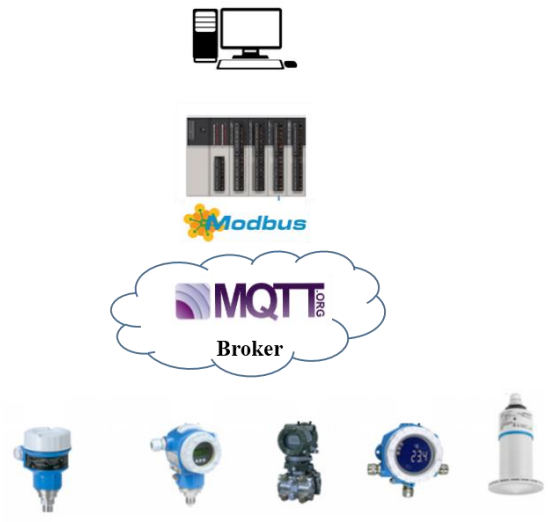


FIG.7 : .Industrial wireless communication

6 Conclusion et perspectives

In automation technology, the introduction of the Internet of Things (IoT) and Cyber Physical Systems (CPS) have revolutionized the industrial world with the introduction of Industry 4.0 technology to create smart factories. In this paper, we presented an industrial communication strategy based on data flow "Node-RED" using the Modbus industrial protocol and the MQTT communication protocol for M2M industrial communication. This new concept uses wireless communication networks to connect industrial machinery and equipment without the use of binding cables. Thus, the use of wireless sensor networks in industrial environments is undoubtedly beneficial to the development of automation technologies.

As perspectives, we want to broaden communication by adding more slaves by ensuring a distributed, test other industrial protocols (Profibus, Profinet) and reliable architecture for industrial communications.

Références

X. Shen, Z. Wang, and Y. Sun (2004), "Wireless sensor networks for industrial applications," in *Proceedings of the Fifth World Congress on Intelligent Control and Automation*, vol. 4, pp. 3636-3640.

A. Ajith Kumar, Knut Ovsthus, Lars Kristensen (2014), "An industrial perspective on wireless sensor network – A survey of requirements, protocols and chalenges", IEEE Communications survey and tutorials.

M. Wollschlaeger, T. Sauter, and J. Jasperneite (2017), "The future of industrial communication: Automation networks in the era of IoT and Industry 4.0," *IEEE Industrial Electronics. Mag.*, vol. 11, no. 1, pp. 17–27.

M. Ehrlich, L. Wisniewski, and J. Jasperneite (2016) "State of the art and future applications of industrial wireles sensor networks", in Proc. Kommunickation in der Automation (KommA), pp. 80-87, Nov.

V. Gungor and G. Hanke (2009) "Industrial wireless sensor networks: Challenges, design principales, and technical approaches," IEEE Transactions on Industrial Electronics, vol. 56, pp. 4258-4265.

Shiyoung wang, Jiafu wan, Di Li, Chunhua Zhang (2016), " Implementing Smart Factory of Industrie 4.0 : An Outlook", International Journal of Distributed Sensor Networks, Article ID 3159805, 10 pages.

Jiafu Wan, Shenglong Tang, Zhaogang Shu, Di Li, Shiyong Wang, Muhammad Imran, Athanasios V.Vasilakos (2016) « Software-Defined Industrial Internet of Things in the context of industry 4.0 », IEEE Sensors Journal, Vol. 16, NO. 20.

Jyotirmoy Banik, Ricardo Arjona, Kumaran Vijayasankar, Arvind Kandhalu (2017) « Improving performance in industrial internet of things using multi-radio nodes and multiple gatewayes », international conference on computing, networking and communications, January 26-29.

Federico Tramarin, Stefano Vitturi, Michele Luvisotto, Andrea Zanella (2016) " on the use of IEEE 802.11n for industrial communications", IEEE Transactions on industrial informatics, vol.12, NO.5.

Kun Wang, Yihui Wang, Yanfei Sun, Song Guo, Jinsong Wu (2016) "Green Industrial internet of things architecture : an energy-efficient perspective", IEEE Cpmmunications Magazine – Communications Standards Supplement.

Jay Lee, Behrad Bagheri , Hung-An Kao, (2015) « A Cyber-Physical Systems architecture for Industry 4.0-based manufacturing systems », Manufacturing Letters 3 18–23.

Julien Minerauda,, Oleksiy Mazhelisb, Xiang Su, Sasu Tarkoma (2016) « A gap analysis of Internet-of-Things platforms », Computer Communications, Special issue on the Internet of Things: Research challenges and SolutionsMarch 11, 2016.

Markel Iglesias-Urkia, Adri´an Orive, Aitor Urbieta (2017) , Analysis of CoAP Implementations for Industrial Internet of Things: A Survey, The 8th International Conference on Ambient Systems, Networks and Technologies, Procedia Computer Science 109C 188–19.

Résumé :

Au cours des prochaines années, le réseau de capteurs sans fils industriels (IWSN) qui constitue une partie principale de l'internet industriel des objets (IIoT), joue un rôle crucial dans la transformation du monde industriel en ouvrant une nouvelle ère de croissance économique et compétitive de la révolution industrielle dite « *Industrie 4.0* ». L'IIoT est en mesure d'aider les organisations à obtenir de meilleurs bénéfices sur les marchés de la fabrication industrielle en augmentant la productivité, en réduisant les coûts et en développant de nouveaux services et produits. Dans ce papier nous développons un système de communication distribué Maitre/esclave basé deux protocoles de communication à savoir : MQTT et Modbus. L'implémentation est basée sur Node-RED dont nous souhaitons participer aux développements des usines futures

Multicast routing in wireless sensor networks with neural networks in fixed time

Nadia SABER*, Mohammed Mestari**

*LPRI, EMSI Casablanca
nadiasaber4@gmail.com,
**SSDIA, ENSET Mohammedia
mestari@enset-media.ac.ma

Abstract. Multicast routing, in wireless sensor networks (WSNs), that respects several Quality of Service constraints, is important for supporting data communication under the banner of Internet of Things (IoT). In this paper, we propose a new neural networks architecture in order to deal with the challenging problem of multicast routing in IoT, by constructing multicast routing tree in fixed time. In our method, this problem is considered as a multiobjective optimization problem (MOP), which will be converted into a single optimization problem (SOP), by using an entropy-based process. The architecture herein designed reduces substantially the complexity especially in large networks, and uses only two kinds of neurons.

1 Introduction

Multicast routing, in wireless sensor networks (WSNs), that respects several Quality of Service constraints, consists of transmitting simultaneously a message from a single source to a set of destinations; It is important for supporting data communication under the banner of Internet of Things (IoT)

In last decades, multicast routing has become more and more popular, and has motivated several researchers, because of the progress in the area of multimedia communications, file sharing, interactive games, videoconferencing, on-demand video, radio and TV transmission... etc

The multicast routing problem, in graph theory, is known as the Steiner tree problem, and has been shown to be NP-complete (non deterministic polynomial-time complete) Wang and Crowcroft (1996) and Chang and Wang (1999).

Several methods are proposed in the literature to solve the Steiner tree problem. Chow (1991) and Salama et al. (1997) proposed two exact algorithms to solve this problem, but they are not viable in very large networks, because of their high degree of computational complexity. Heuristic proposed by Kompella et al. (1993) is one of the famous methods used to solve this problem, because they construct a feasible solution within reasonable time. This method assumes that the source node can obtain topology information about the communication network through the routing protocol,

but it suffers from some drawbacks such as failure on the central node, or high communication cost in keeping network information up-to-date, especially in large networks. To overcome this, Bauer and Varma (1996), and Jia (1998) proposed distributed algorithms for the routing problem, where each node operates based on its local routing information and the coordination with other nodes is done via network message passing.

The QoS-constrained multicast routing problem include routing that guarantees the required quality of service (QoS), such as bandwidth requirement, delay constraint on transmitting information between a source node and each destination. Many metaheuristics are proposed to solve the QoS-constrained multicast routing problem: genetic algorithms (Zhengying et al. (2001), Tseng et al. (2006), Lu and Zhu (2013)), tabu search (Youssef et al. (2002), Wang et al. (2004), Yang (2002)), ant colony optimization (Tseng et al. (2008), Yin et al. (2014), Wang et al. (2009)), fuzzy-based algorithms (Nie et al. (2006), Su et al. (2008)).

To solve the multicast routing problem, neural networks were first proposed by Rauch and Winarske (1988), by defining proper energy functions and deriving associated weights between neurons. In 1982, Hopfield presented the Hopfield neural network Hopfield (1982), since then, many researchers have been exploring HHNs and improving their performance on different real time applications. Pornavalai et al. (1995) proposed a modification of HNNs to solve constrained multicast routing, but Gee and Prager (1995) demonstrated that they are not efficient in large networks.

The multicast routing problem can be formulated as a single-objective problem (SOP) where only one generic cost function is considered (Kompella et al. (1993), Bauer and Varma (1996), Jia (1998), Zhengying et al. (2001), Saber et al. (2016)), or as a multi-objective problem (MOP), where several objective functions may be optimized (minimized or maximized) in conflicting situations (Crichigno and Barán (2004b), Crichigno and Barán (2004c), Crichigno and Barán (2004a), Roy and Das (2004)). The proposed method to solve this problem, construct an optimal multicast tree under several constraints, and do not need any central node to keep information about of the whole network.

The construction and the calculation of the weight functions, associated to each link, and to the whole network, necessitate the use of weighting network (WN) and the FCN network. These two networks have a very simple configuration, and are construed by combining several linear neurons. To sort the weight functions associated to each link, we use the sorting network, that will be implemented on the basis of adjustable order statistic filters AOSF (Mestari (2004)).

The neural network architecture herein proposed, solve this problem in fixed time, regardless of the network size, and use only two kinds of neurons :linear and threshold-logic neurons. These neurons have been used in constructing various kinds of neural networks: Mestari et al. (2015), Mestari (2004), Khouil et al. (2014b,a, 2016), Saber et al. (2014a,b, 2016). Among all the neurons proposed in the literature, they are probably the easiest to implement in hardware.

This paper is organized as follows: section 2 will give the mathematical notation, section 3 will give a description of our method to construct the multicast tree, the neural

networks architecture for the routing problem will be presented in section 4, and in section 5 we will give illustrative examples, and section 6 will present conclusion.

2 Mathematical notation

2.1 Physical network

A communication network in an IoT environnement can be represented by a connected non-directed graph $G = (V, E)$, where $V = \{v_1, v_2, \ldots, v_N\}$ is a set of vertices, and E is a set of edges: $E \subseteq \{(v_i, v_j)/v_i \in V, v_j \in V$, where $v_i \neq v_j, 1 \leq i \leq N$ and $1 \leq j \leq N\}$.

N and M are used to represent the number of vertices and edges in the graphe and i.e. $|V| = N$ and $|E| = M$.

We denote by s the source node, u a destination node, and $U = \{u_1, u_2, \ldots, u_m\}$ is a set of destination nodes. We denote by $T = (V_T, E_T)$ a multicast tree, where $V_T \subseteq V, E_T \subseteq E$, and by $\omega_k(T)$ the sum of k^{th} weights on edges of T.Let $P_T(s, u)$ be a unique path in the tree T from the source node s to a destination node $u \in U$ and $\omega_k(P_T(s, u))$ is the sum of k^{th} weights of $P_T(s, u)$.

Edges are numbred from 1 to M and each edge has K independant additive weights $\omega_1, ..., \omega_K$ where $\omega_k(e)$ is the k^{th} weight of edge e.

Let $W_1, ..., W_K$ be the K QoS constraints. The problem of Multi-constrained tree is to find a multicast tree T such that $\omega_k(T) \leq W_k$, $1 \leq k \leq K$. A multicast tree that satisfies $\omega_k(T) \leq W_k$, $1 \leq k \leq K$ is said to be a feasible tree. We use $\{T^f\}$ to denote all the feasible trees in $G(V, E)$. Foe each feasible tree $T_i \in \{T^f\}$, there exists a smallest $\theta_i \in]0, 1]$ such that $\omega_k(T_{opt}) \leq \theta_i.W_k$, $1 \leq k \leq K$.

The graph G is represented by a matrix $X = (x_{ik})_{\substack{1 \leq i \leq M \\ 1 \leq k \leq K}}$ where the variable x_{ik} represents the k^{th} of the i^{th} edge.

2.2 Weight functions

In this contribution, we adopt an entropy-based weight agregation algorithm, to reduce the multi-constrained multicast routing problem to a single optimisation problem.

To define the weight function W_{ij} associated to link (v_i, v_j), we must first define two parameters f_k^U and f_k^L:

$$f_k^U = \max_{1 \leq i \leq M} \{x_{ik}\} \tag{1}$$

$$f_k^L = \min_{1 \leq i \leq M} \{x_{ik}\} \tag{2}$$

These parameters are used in the next step to calculate the matrix $R = [r_{ik}]_{m \times K}$ where:

$$r_{ik} = \begin{cases} \dfrac{f_k^L}{x_{ik}} & if\ x_{ik} < 0 \\ \dfrac{x_{ik}}{f_k^U} & otherwise \end{cases} \tag{3}$$

Obviously $0 \leq r_{ik} \leq 1$, the normalized matrix $P = [p_{ik}]_{m \times K}$ can be computed by:

$$p_{ik} = \frac{r_{ik}}{\sum_{i=1}^{m} r_{ik}} \tag{4}$$

Then δ_k, d_k and α_k are calculated by :

$$\delta_k = -\frac{1}{\ln m} \sum_{i=1}^{m} p_{ik}. \ln p_{ik} \tag{5}$$

$$d_k = 1 - \delta_k \tag{6}$$

$$\alpha_k = \frac{d_k}{\sum_{i=1}^{K} d_k} \tag{7}$$

The weight function W_{ij} associated to link (v_i, v_j) is defined as follows:

$$W_{ij} = \sum_{k=1}^{K} \alpha_k . \omega_k(v_i, v_j) \tag{8}$$

The weight function W_T associated to a multicast tree T is defined as follows:

$$W_T = \sum_{(v_i, v_j) \in T} W_{ij} \tag{9}$$

3 Our algorithm

The majority of the proposed heuristics for the routing problem are centralized, that require a central node to be responsible for computing the entire routing tree, and this central node must have the full knowledge about the global network, and this is not efficient in very large network. Furthermore, the proposed heuristics find an optimal tree under a delay bound, and consider only the cost of tree, without taking into account other QoS requirements, such as bandwidth, packet loss or required power in transmission.

3.1 Assumptions

We assume that:

1) All parameters associated to each link are independent ;
2) The network is connected ;
3) Every node has a unique index ;
4) The routing protocol will collect state information of the communication network (e.g. the group membership, available resources and application requirements) and deliver this information throughout the communication network ;
5) The network nodes are robust and can include a system that stores statistical information of the sent and received packets in order to calculate packet loss rate between nodes ;
6) Each node have all information about its adjacent nodes, and all the paths to other nodes in the network.

3.2 Basic idea of our heuristic

This algorithm mimics Kruskal's MST algorithm Kruskal (1956). In our method, the routing trees are constructed gradually. In a first time, all nodes adjacent to a given node are sorted in ascending order of their weight function defined by equation (8), and then stocked in a table associated to each node v_i. The construction starts with a tree T containing only the source node s, and then all links connecting s to other nodes will be deleted from all tables. In each step, if there are multicast nodes adjacent to the constructed tree T, we select the closest multicast node to join the tree, if not, the closet node will join the tree.

3.3 Pseudo code of the proposed method

input:

— The network topology $G = (V, E)$;

— The matrix $X = (x_{ik})_{\substack{1 \leq i \leq M \\ 1 \leq k \leq K}}$;

— s: the source node ;

— U: the set of destination nodes;

— $W_1, ..., W_K$: the K QoS constraints

output:

— The multicast tree T

procedure:

1) Calculate the weight function W_{ij} associated to each link (v_i, v_j): $W_{ij} = \sum_{k=1}^{K} \alpha_k . \omega_k(v_i, v_j)$;
2) For each node $v_i \in V$, sorting all link adjacent to it in ascending order for their weight function ;
3) Initialize the multicast tree with the source node $T \leftarrow \{s\}$;
4) Where U is not entirely included in T
   — if there are multicast nodes adjacent to T
   select the closet multicast node adjacent to T by adding the edge connecting this node to the multicast tree
   — else select the closet node adjacent to T by adding the edge connecting this node to the multicast tree
   — remove the selected node from all tables.

The proposed method minimizes the weight function W_T (equation (9)) associated to a multicast tree T, by selecting, in each step, the link with minimal weight function W_{ij} (equation (8)) to join the tree. But this method not enforces fulfilling constraints.

This algorithm generate a single solution. To obtain a set containing a large number of solutions, we can execute this algorithm to generate successive approximations of the multicast tree.

4 Neural networks architecture for the routing problem

In this section, we study how to develop an appropriate neural network architecture for implementing the proposed algorithm for the routing problem. First, we describe sorting network, which sorts all links adjacent to a giving node. The weighting network which calculates the weight function W_{ij}.

4.1 neurons used

All neural networks proposed in this paper use only two kinds of neurons: the linear and the threshold-logic neuron. The only difference between these neurons is in their activation functions.

4.2 Neural Networks for calculation of f_k^U and f_k^L

In this subsection we develop two neural networks for calculation of f_k^U and f_k^L given in (1) and (2).

The calculation of f_k^U and f_k^L requires the use of AOSF network developed by Mestari (2004).

AOSF network accepts an array of real numbers as input and outputs in fixed time the k^{th} largest element of the array.

The task of finding the k^{th} largest element of an input array can be done in two phases as follows:

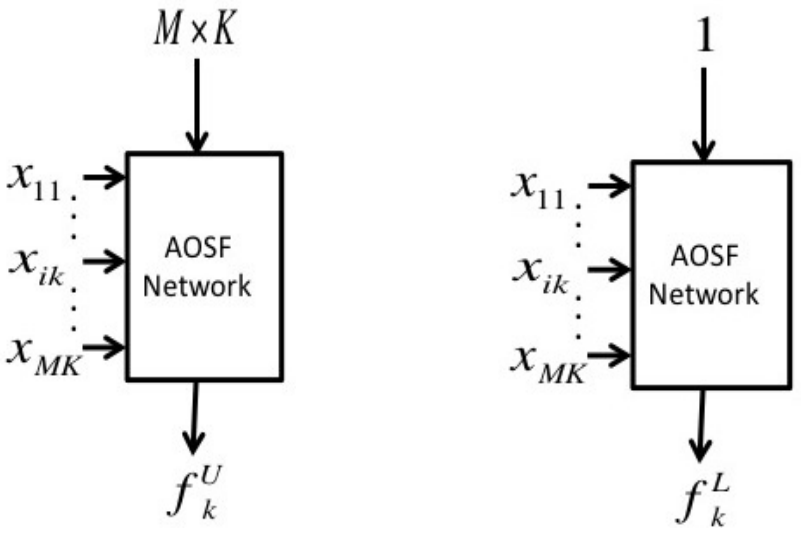
1. Compute the order in the input array of any element ;
2. Select and transfer to the output the element corresponding to order k, chosen by the decision maker.

The AOSF network consist of two kinds of neurons, arranged in 11 layers, thus the total processing time is 11 times the processing time of a single neuron. As the number of elements in the input array increases, only the number of neurons in each layer increases, not the number of layers themselves.

The implementation of networks for calculating f_k^U and f_k^L are given respectively by Fig. 1 and Fig. 2.

4.3 Sorting network

The function of sorting network for weight functions (Fig. 2) is to giving the order of the weight function W_{ij} associated to any link (v_i, v_j) adjacent to a given node v_i, and arranging them in ascending order. $W_{(ik)}$ represents the k^{th} largest weight function. Sorting network (Fig. 2) consists of n_i adjustable order statistic filter AOSF developed by Mestari (2004), set up in parallel, where n_i is the number of link adjacent to node i. Sorting network for our algorithm (Fig. 2), accepts as input an array of real numbers, gives all orders statistics of the input array, and outputs the input array sorted in ascending order. The k^{th} order statistic of the input array is defined as being



(a) Neural network for calculation of f_k^U (b) Neural network for calculation of f_k^L

FIG. 1: Neural networks for calculation of f_k^U and f_k^L

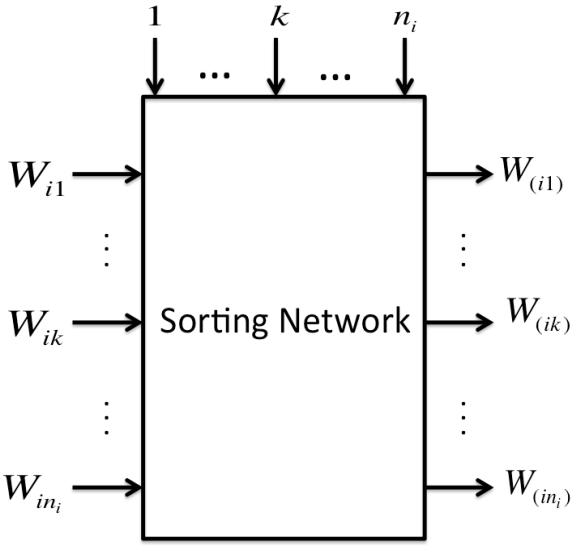


FIG. 2: Sorting network for weight functions

the k^{th} largest element of the array. Sorting time is constant, and is equal to the processing time of a single AOSF.

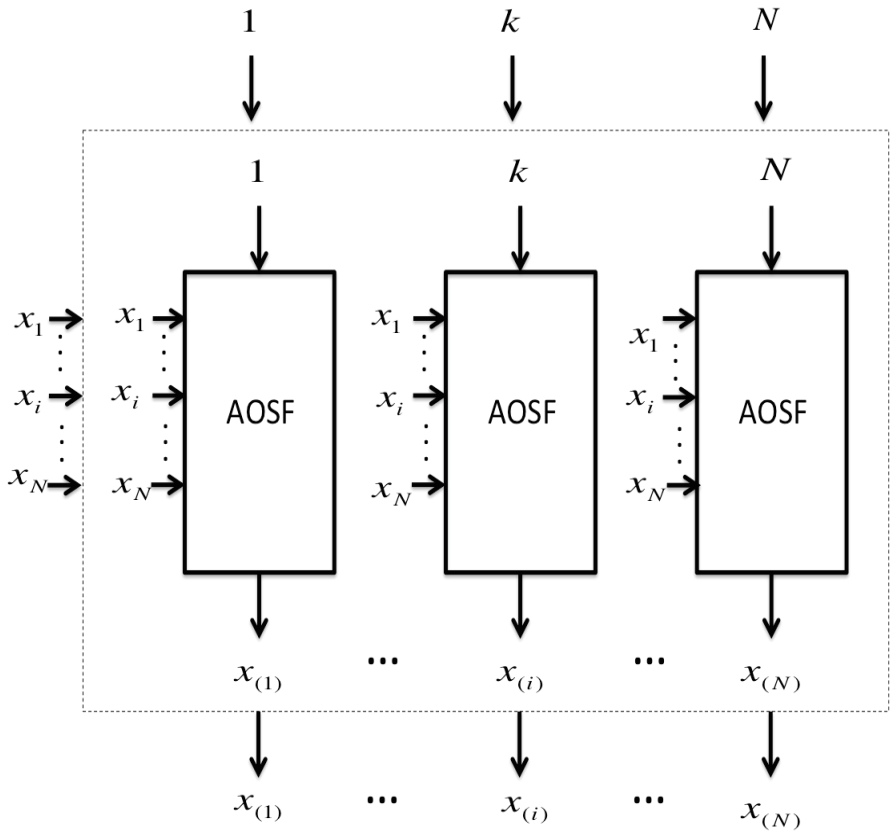


FIG. 3: Sorting network

The sorting network have a very simple configuration, and is composed only of two kinds of neurons: linear and threshold-logic neurons. The synaptic weights are all fixed , most of them being just +1 or -1, which makes hardware implementation easy.

To sort an input array of n_i elements, n_i AOSF set up in parallel, each AOSF outputs the k^{th} largest element of the input array, where $k \in \{1, ..., n_i\}$.

4.4 Weighting network

The function of the weighting network WN is to calculate the weight function W_{ij} associated to each link (v_i, v_j): $W_{ij} = \sum_{k=1}^{K} \alpha_k.\omega_k(v_i, v_j)$. The weighting network is shown by Fig. 4.

The function computed by WN is defined as:

$$W_{ij} = \sum_{k=1}^{K} \alpha_k.\omega_k(v_i, v_j) \tag{10}$$

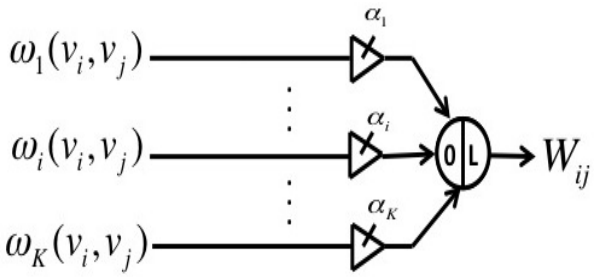


FIG. 4: Weighting network

5 Conclusion

In this stydy, we propose an heuristic to find multicast tree by minimizing a weight function calculated by an antropy-based weight-aggregation method. This algorithm does not require a central node to be responsible for computing the tree. We have also proposed some neural networks to implement this method, by using only two kinds of neurons: linear and threshold logic neurons. Neural networks proposed for our method have a simple configuration, and their processing time remains constant and independent of the network size, due to the massive parallelism property offered by neural networks.

References

Bauer, F. and A. Varma (1996). Distributed algorithms for multicast path setup in data networks. *IEEE/ACM Transactions on Networking (TON) 4*(2), 181–191.

Chang, R.-S. and C.-D. Wang (1999). Improved www multimedia transmission performance in http/tcp over atm networks. *IEEE transactions on multimedia 1*(3), 278–290.

Chow, C.-H. (1991). On multicast path finding algorithms. In *INFOCOM'91. Proceedings. Tenth Annual Joint Conference of the IEEE Computer and Communications Societies. Networking in the 90s., IEEE*, pp. 1274–1283. IEEE.

Crichigno, J. and B. Barán (2004a). A multicast routing algorithm using multiobjective optimization. In *International Conference on Telecommunications*, pp. 1107–1113. Springer.

Crichigno, J. and B. Barán (2004b). Multiobjective multicast routing algorithm. In *International Conference on Telecommunications*, pp. 1029–1034. Springer.

Crichigno, J. and B. Barán (2004c). Multiobjective multicast routing algorithm for traffic engineering. In *Computer Communications and Networks, 2004. ICCCN 2004. Proceedings. 13th International Conference on*, pp. 301–306. IEEE.

Gee, A. H. and R. W. Prager (1995). Limitations of neural networks for solving traveling salesman problems. *IEEE Transactions on Neural Networks 6*(1), 280–282.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences 79*(8), 2554–2558.

Jia, X. (1998). A distributed algorithm of delay-bounded multicast routing for multimedia applications in wide area networks. *IEEE/ACM Transactions on Networking (TON) 6*(6), 828–837.

Khouil, M., H. I. ENSET, M. I. Sanou, M. M. Mestari, M. A. Aitelmahjoub, and E. Casablanca (2016). Planification of an optimal path for a mobile robot using neural networks. *Applied Mathematical Sciences 10*(13), 637–652.

Khouil, M., N. Saber, and M. Mestari (2014a). Collision detection for three dimension objects in a fixed time. In *Information Science and Technology (CIST), 2014 Third IEEE International Colloquium in*, pp. 235–240. IEEE.

Khouil, M., N. Saber, and M. Mestari (2014b). Neural network in fixed time for collision detection between two convex polyhedra. *network 1*, 6.

Kompella, V. P., J. C. Pasquale, and G. C. Polyzos (1993). Multicast routing for multimedia communication. *IEEE/ACM transactions on networking 1*(3), 286–292.

Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society 7*(1), 48–50.

Lu, T. and J. Zhu (2013). A genetic algorithm for finding a path subject to two constraints. *Applied Soft Computing 13*(2), 891–898.

Mestari, M. (2004). An analog neural network implementation in fixed time of adjustable-order statistic filters and applications. *IEEE transactions on Neural Networks 15*(3), 766–785.

Mestari, M., M. Benzirar, N. Saber, and M. Khouil (2015). Solving nonlinear equality constrained multiobjective optimization problems using neural networks. *IEEE transactions on neural networks and learning systems 26*(10), 2500–2520.

Nie, J., J. Wen, J. Luo, X. He, and Z. Zhou (2006). An adaptive fuzzy logic based secure routing protocol in mobile ad hoc networks. *Fuzzy sets and systems 157*(12), 1704–1712.

Pornavalai, C., G. Chakraborty, and N. Shiratori (1995). A neural network approach to multicast routing in real-time communication networks. In *Network Protocols, 1995. Proceedings., 1995 International Conference on*, pp. 332–339. IEEE.

Rauch, H. E. and T. Winarske (1988). Neural networks for routing communication traffic. *IEEE Control Systems Magazine 8*(2), 26–31.

Roy, A. and S. K. Das (2004). Qm 2 rp: a qos-based mobile multicast routing protocol using multi-objective genetic algorithm. *Wireless Networks 10*(3), 271–286.

Saber, N., H. I. ENSET, C. B. Hassan II, M. Mestari, and A. A. El Mahjoub (2016). The multi-constrained least-cost multicast problem with neural networks in fixed time. *Applied Mathematical Sciences 10*(19), 931–945.

Saber, N., M. Khouil, and M. Mestari (2014a). Delay constrained multicast tree with neural networks in fixed time. *INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY 3*, 11.

Saber, N., M. Khouil, and M. Mestari (2014b). Neural networks based on adjustable-order statistic filters for multimedia multicast routing. In *Information Science and Technology (CIST), 2014 Third IEEE International Colloquium in*, pp. 435–439. IEEE.

Salama, H. F., D. S. Reeves, and Y. Viniotis (1997). Evaluation of multicast routing algorithms for real-time communication on high-speed networks. *IEEE Journal on Selected Areas in Communications 15*(3), 332–345.

Su, B.-L., M.-S. Wang, and Y.-M. Huang (2008). Fuzzy logic weighted multi-criteria of dynamic route lifetime for reliable multicast routing in ad hoc networks. *Expert Systems with Applications 35*(1-2), 476–484.

Tseng, S.-Y., Y.-M. Huang, and C.-C. Lin (2006). Genetic algorithm for delay-and degree-constrained multimedia broadcasting on overlay networks. *Computer Communications 29*(17), 3625–3632.

Tseng, S.-Y., C.-C. Lin, and Y.-M. Huang (2008). Ant colony-based algorithm for constructing broadcasting tree with degree and delay constraints. *Expert Systems with Applications 35*(3), 1473–1481.

Wang, H., J. Fang, H. Wang, and Y.-M. Sun (2004). Tsdlmra: an efficient multicast routing algorithm based on tabu search. *Journal of Network and Computer Applications 27*(2), 77–90.

Wang, H., Z. Shi, and S. Li (2009). Multicast routing for delay variation bound using a modified ant colony algorithm. *Journal of Network and Computer Applications 32*(1), 258–272.

Wang, Z. and J. Crowcroft (1996). Quality-of-service routing for supporting multimedia applications. *IEEE Journal on selected areas in communications 14*(7), 1228–1234.

Yang, W.-L. (2002). A tabu-search based algorithm for the multicast-streams distribution problem. *Computer networks 39*(6), 729–747.

Yin, P.-Y., R.-I. Chang, C.-C. Chao, and Y.-T. Chu (2014). Niched ant colony optimization with colony guides for qos multicast routing. *Journal of Network and Computer Applications 40*, 61–72.

Youssef, H., A. Al-Mulhem, S. M. Sait, and M. A. Tahir (2002). Qos-driven multicast tree generation using tabu search. *Computer Communications 25*(11-12), 1140–1149.

Zhengying, W., S. Bingxin, and Z. Erdun (2001). Bandwidth-delay-constrained least-cost multicast routing based on heuristic genetic algorithm. *Computer communications 24*(7), 685–692.

Multicast routing with neural networks in fixed time

Résumé

Le routage multicast, dans les réseaux de capteurs sans fil, qui répond à plusieurs contraintes de qualité de service, est important pour la communication de données dans le contexte d'Internet des Objets. Dans cet article, nous proposons une nouvelle architecture de réseaux de neurones artificiels afin de résoudre le problème complexe du routage multicast, en construisant un arbre de routage multicast à temps fixe. Dans notre méthode, ce problème est considéré comme un problème d'optimisation multiobjectif, qui sera converti par la suite en un problème d'optimisation mono-objectif, en utilisant un processus basé sur l'entropie. L'architecture conçue ici réduit considérablement la complexité notamment dans les grands réseaux et n'utilise que deux types de neurones.

Vers un nouveau modèle pour l'équilibrage de charge dans le CloudIoT

Sofiane Benabbes *, Abderrahim Necib **
Sofiane Mounine Hemam***

*Université Abbès Laghrour de Khenchela-Algérie
benabbes.sofiane@gmail.com
** Université Abbès Laghrour de Khenchela-Algérie
miharredbanecib@gmail.com
***Laboratoire ICOSI, Université Abbès Laghrour de Khenchela-Algérie
Sofiane.hemam@gmail.com

Résumé. CloudIoT est un nouveau paradigme qui, a émergé suite à l'intégration du Cloud Computing et l'Internet des Objets. Dans le CloudIoT, les capteurs physiques sont chargés de détecter et de transmettre les données vers le Cloud afin qu'elles soient traitées et stockées. La quantité de données à traiter augmente de jour en jour ce qui nécessite un mécanisme d'équilibrage de charge afin de répartir les données capturées entre les différentes ressources du CloudIoT.
Dans ce papier, nous proposons une architecture pour équilibrer la charge entre les machines virtuelles (MVs) d'un CloudIot. Cette architecture se compose de quatre composants essentiels: le Classiffieur, le routeur, l'équilibreur local et l'équilibreur général. Ce modèle se base sur le pire temps d'exécution et la taille d'une tâche, et sur le centre de gravité des classes. Les résultats obtenus à travers l'étude de cas montrent que notre modèle permet de réguler la charge entre les MVs.

1 Introduction

Le Cloud Computing est un modèle informatique basé sur l'internet. Selon NIST« National Institute of Standards and Technology », le Cloud Computing est un modèle permettant d'établir un accès à la demande en réseau vers un bassin partagé de ressources informatiques configurable. Ces ressources peuvent être: des réseaux, des serveurs, de l'espace de stockage, des applications et des services. Elles peuvent être approvisionnées rapidement avec un minimum d'effort de gestion et d'interaction avec le fournisseur de services (Mell et Grance. 2009). D'un autre côté, l'Internet des objets (IdO) repose sur des nœuds (objets) intelligents et auto-configurés, interconnectés dans une infrastructure de réseau dynamique et globale (Botta, et al. 2015). Il est généralement caractérisé par de petits objets du monde réel, largement distribuées, avec une capacité de stockage et de traitement limitée, ce qui implique des préoccupations concernant la fiabilité, la performance, la sécurité et la confidentialité (Botta, et al. 2015). A l'opposé de l'Internet des objets, le Cloud Computing a des capacités prati-

quement illimitées en termes de puissance de stockage et de traitement, c'est une technologie beaucoup plus mature.

A partir de ces deux technologies (Cloud-Computing et Internet des Objets) est né un nouveau paradigme dans lequel le Cloud Computing et l'Internet des Objets sont fusionnés. Ce nouveau paradigme est appelé CloudIoT. Dans CloudIoT, la quantité de données à stoker et à traiter augmentent de jours en jours et de façon très rapide, ce qui nécessite un mécanisme qui permet d'équilibrer la charge de stockage et de traitement entre les différentes machines virtuelles du CloudIoT. L'équilibrage de charge est considéré comme l'un des problèmes clés, il est nécessaire pour répartir la charge de travail entre plusieurs nœuds afin de garantir qu'aucun nœud n'est surchargé (Sidhu et Kinger. 2013). Les algorithmes qui utilisent le principe du premier arrivé premier servi ont marqué leur limite dans l'équilibrage de charge, puisque ils ne permettent pas d'atteindre l'objectif de la répartition de charge entre les nœuds dynamiques ou entre les Machines Virtuelles du Cloud. Ainsi, certaines Machines Virtuelles seront surchargées tandis que d'autres seront sous-chargées (Chien et Loc. 2016).

Dans ce papier, nous proposons un modèle qui permet l'équilibrage de charge entre les différentes machines virtuelles d'un CloudIot. Le modèle proposé se compose de quatre essentiels composants: Le Classiffieur qui, est chargé de constituer cinq classes de Machines Virtuelles (MV); de la plus performante à la moins performante classe des machines virtuelles, en utilisant pour cela une des techniques de classification supervisée. Le routeur qui, en premier lieu, calcule les distances entre la tâche reçu et les centres de gravité des cinq classes, puis il envoi la tâche à la classe la plus proche par rapport aux distances précédemment calculées. L'équilibreur local dont le rôle est d'équilibrer la charge entre les différentes Machines Virtuelles, appartenant à la même classe, toute en prenant en considération le pire temps d'exécution (PTE)d'une tâche et sa taille. Enfin, le composant équilibreur général qui, est chargé d'équilibrer la charge entre les classes en migrant les tâches d'une classe surchargée vers une classe la moins chargée. Les résultats obtenus à travers l'étude de cas montrent que notre modèle permet effectivement l'équilibrage de charge entre les différentes Machines Virtuelles d'un CloudIoT.

Notre papier est structuré comme suit: après l'introduction, la section 2 sera consacrée aux travaux voisins. Dans la section 3 nous allons présenter d'abord l'architecture générale du modèle proposé, puis le fonctionnement et le détaille de chaque composant seront expliqués dans section 4. Quant à la section 5, elle sera consacrée à l'étude de cas afin de monter la validité et le fonctionnement des différents composants proposés, et nous terminerons ce papier par une conclusion dans la section 6.

2 Travaux voisins

L'équilibrage de charge est une technique importante pour améliorer les performances des systèmes. Les demandes des utilisateurs doivent être réparties de façon égale entre les machines ou nœuds afin d'obtenir une réponse rapide (Cardellini et al., 1999). L'objectif de l'équilibrage de charge est d'accroître l'utilisation des ressources, de réduire le temps de réponse et la surcharge de certains nœuds.

L'équilibrage de charge peut être réalisé par différentes approches avec différents degrés d'efficacité. Nous pouvons distinguer entre deux types d'équilibrage de charge: statique et dynamique. Ces approches dépendent de la manière dont un algorithme d'équilibrage de charge est basé pour prendre sa décision, c'est-à-dire qu'il est basé sur l'état actuel du système

ou non (Alakeel, 2010). L'équilibrage de charge statique peut être défini par une situation dans laquelle une charge de calcul est partitionnée entre des nœuds par un algorithme exécuté avant l'exécution du programme, c'est-à-dire l'assignation d'un ensemble de requêtes à un ensemble de ressources; alors que dans l'approche dynamique, l'algorithme d'équilibrage de charge est basé sur l'état actuel du système, de sorte que les demandes sont déplacées dynamiquement d'un nœud surchargé à un nœud sous-chargé, suivant la disponibilité des ressources système (Olejnik, et al., 2009). Il existe deux catégories d'algorithmes d'équilibrage de charge dynamique: les algorithmes centralisés et distribués. Dans l'algorithme d'équilibrage de charge dynamique distribué, tous les nœuds du système doivent exécuter l'algorithme et interagir entre eux pour réaliser l'équilibrage de charge. Cette interaction peut être coopérative ou non coopérative. L'interaction entre les nœuds dans les algorithmes d'équilibrage de charge dynamique distribués génère plus de messages par rapport aux algorithmes dynamique centralisés. Cependant, même avec la plus grande complexité d'exécution; les algorithmes dynamiques ont le potentiel d'offrir de meilleures performances que les algorithmes statiques (Olejnik, et al., 2009).

Plusieurs travaux relatifs à l'équilibrage de charge dans le Cloud Computing ont été proposés ces dernières années. Dans ce papier, nous citons les travaux les plus récents. Ainsi, dans (Chien et al. 2016), les auteurs ont proposé un algorithme d'équilibrage de charge qui, permet d'améliorer les performances de l'environnement Cloud en fonction de la méthode d'estimation du temps de fin de service. Ils ont réussi à améliorer le temps de service et le temps de réponse de l'utilisateur.

Le problème de consommation d'énergie a été abordé par Yakhchi,et al., (2015). Ainsi, ils expliquent que la consommation d'énergie est devenue un défi majeur dans les infrastructures de Cloud Computing. Pour cela, Ils ont proposé une nouvelle méthode d'équilibrage de la charge, appelée ICAMMT. Cette dernière, permet de gérer la consommation d'énergie dans les centres de données du Cloud Computing.

Pour satisfaire les besoins des utilisateurs, les auteurs Kapoor et al. (2015) ont proposé un algorithme qui vise à maximiser la satisfaction des utilisateurs, en minimisant le temps de réponse des tâches et en améliorant l'utilisation des ressources grâce à une allocation équitable des ressources Cloud.

Fahim et al., (2016) ont proposé une nouvelle conception d'équilibrage de charge. Leur modèle est basé sur deux phases principales: la première, est la pré-estimation du temps d'exécution des instructions d'une tâche, en se basant sur les méthodes qui calculent le pire temps d'exécution. Quant à la deuxième phase, elle consiste à ordonnancer les tâches selon des niveaux de classification afin d'allouer des tâches à des machines virtuelles avec la garantie de l'égalité au niveau de la répartition de la charge et de diminuer la dégrée de déséquilibrage de charge en prenant en considération les caractéristiques des tâches avant leurs allocations.

Al-Rayis et Kurdi. (2013) expliquent que, les équilibreurs de charge peuvent être déployés en fonction de plusieurs architectures différentes. L'architecture d'équilibrage de charge centralisée, qui se base sur seul un équilibreur de charge central, permet de maintenir l'ensemble du système à un état équilibré, en sélectionnant la ressource de Cloud qui, devrait prendre en charge le travail reçu. Dans l'architecture d'équilibrage de charge hiérarchique, un équilibreur de charge principal (parent) reçoit toutes les demandes de travail, puis les répartit vers d'autres équilibreurs de charge connectés (enfants) où chaque équilibreur de charge de l'arborescence peut utiliser un algorithme différent.

Les auteurs Hemam et al., (2017) ont proposé une nouvelle approche basée sur l'idée de cloner un service Cloud sur un ou plusieurs nœuds, lorsque le nombre de requêtes utilisateur sera important à un instant donné. Ils ont proposé un algorithme qui calcul la moyenne de la charge des nœuds afin de sélectionner les nœuds les moins chargés.

A la différence de ces travaux, notre travail consiste à équilibrer la charge au niveau de chaque classe de façon décentralisée, puisque les équilibreurs locaux sont indépendants entre eux, ce qui permet un gain en temps de service et de réponse d'un coté, et ils permettent de sélectionner la machine virtuelle qui correspond le mieux à la tâche reçu d'un autre coté. Quant à l'équilibreur central, il permet de mettre en contacte les équilibreurs locaux des classes les surchargées et les moins-chargées afin que ces dernières collaborent entre eux pour migrer les tâches d'une classe à une autre.

3 Architecture proposée

Cette section décrit l'architecture proposée (Cf Figure 1) . Cette dernière est constituée de deux types d'équilibreur de charge: les équilibrages locaux qui se trouvent au niveau de chaque classe de machines virtuelles, et un équilibreur central.

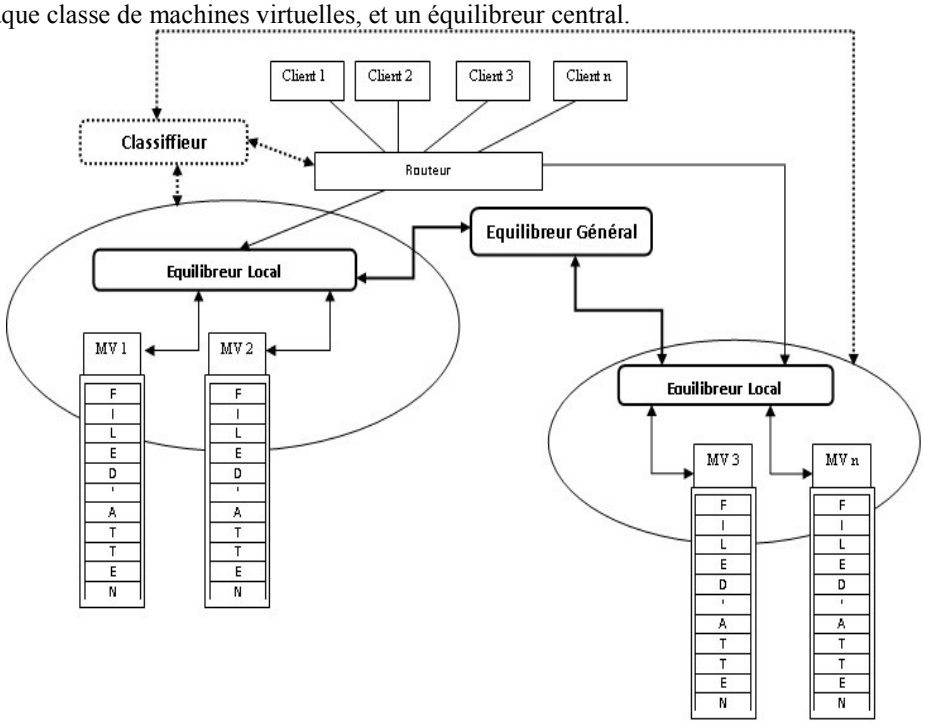


FIG. 1 – *Architecture proposée.*

En plus de ces deux composants, notre architecture contient le composant Classiffieur, dont son rôle est de classer les machines virtuelles en utilisant la méthode K-means, et le composant routeur qui, permet d'envoyer les tâches des clients vers les classes qui leur conviennent. Les classes d'appartenance des machines virtuelles sont: Très-Rapide (TR), Rapide

(R), Moyenne (M), Lente (L), Très-Lente (TL). Notre architecture est composée, aussi, des clients, des machines virtuelles et leurs files d'attente.

3.1 Client

Les clients (les objets) envoient leurs tâches au routeur. Les tâches se composent d'un ensemble d'instruction, et se caractérisent par leurs pire temps d'exécution qui est exprimé en Millions d'Instructions Par Seconde (MIPS) et leur taille.

3.2 Machine virtuelle

Une machine virtuelle se compose d'un ensemble de logiciels et matériels géographiquement distant, elle permet d'exécuter les tâches émissent par les clients.

3.3 File d'attente

La file d'attente est une structure qui permet d'ordonner les tâches, en attente d'exécution, selon leur ordre d'arrivé. Sa structure permet un accès direct aux tâches.

3.4 Classiffieur des MVs

Le composant Classiffieur permet de classer les MVs en cinq classes en fonction de la vitesse du CPU et la taille de la RAM en utilisant la méthode K-means (Fischer, 2014). Ce composant est déclenché à chaque fois que le nombre de machines virtuelles change.

3.5 Routeur

Ce composant est responsable de router les tâches des clients vers les classes. Pour cela, en fonction des caractéristiques (pire temps d'exécution et taille) d'une tâche, et en fonction des caractéristiques des classes. Le routeur sélectionne la classe adéquate pour l'exécution d'une tâche donnée.

3.6 Equilibreur local

Chaque classe possède son propre équilibreur de charge appelé équilibreur local. Son premier rôle est d'affecter les tâches, reçues de la part du routeur, aux MVs correspondantes. Il permet aussi d'équilibré la charge à l'intérieur de la classe à chaque fin d'exécution d'une tâche.

3.7 Equilibreur général

De la même façon que l'équilibreur local, ce composant assure l'équilibrage de charge entre les classes. Il est déclenché périodiquement, ou à la demande d'un équilibreur de charge local.

4 Fonctionnement du modèle proposé

Comme le montre la figure 2, les composants du modèle proposé interagissent entre eux afin de garantir une charge équitable entre les différentes machines virtuelles, ce qui permet de réduire le temps de réponse et de service.

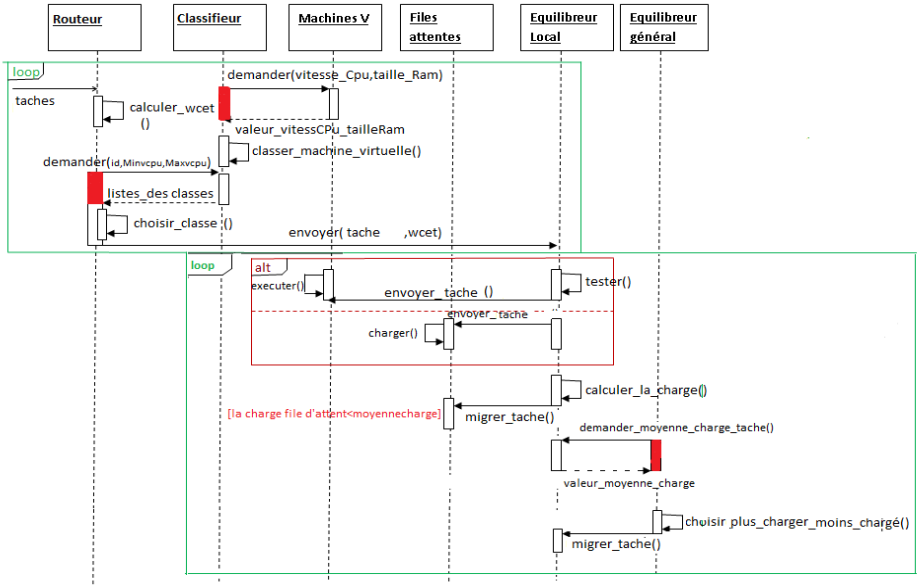


FIG. 2 – *Diagramme de Séquence du Modèle.*

Ainsi, le fonctionnement de notre modèle est devisé en trois niveaux. Le premier niveau permet de classer les machines virtuelles en 5 classes par rapport à la vitesse des processeurs (très rapide, rapide, moyen, lent et très lent) et à la taille de la mémoire de ces MVs. Ce niveau est réalisé en appliquant la méthode de classement K-means, cet algorithme est exécuté la première fois, et à chaque fois où le nombre de machines virtuelles change. Le routeur permet d'affecter une tâche à une classe des MVs qui lui corresponde.

Le deuxième niveau se situe à l'intérieur de la classe. Une fois que le pire temps d'exécution de la tâche est calculé par le routeur, le composant équilibreur de charge local affecte cette tâche à la MV qui lui correspond. A un moment donné, il va y avoir un déséquilibre de charge entre à l'intérieur de la classe car le choix de MV est basé sur le pire temps d'exécution d'une tâche par rapport à la vitesse du processeur de la MV, c'est-à-dire que probablement certaines tâches vont terminer avant leur temps estimé, ce qui provoque un déséquilibre entre les MVs de la même classe. C'est dans cette situation que l'équilibreur local se déclenche afin d'équilibrer la charge entre les MVs de la même classe.

Le troisième niveau est l'équilibrage de charge entre les classes. Ainsi, l'équilibreur général est responsable de l'équilibrage de charge entre les classes. Pour cela, il sélectionne les classes surchargées afin de migrer les tâches de ces dernières vers les classes les moins chargées.

4.1 Classiffieur

Le Classiffieur applique la méthode de K-means. Ainsi, il permet de classer toutes les MVs du système en cinq classes (TR, R, M, L, TL). Les MVs sont caractérisées par deux critères (Vitesse CPU et Taille RAM). La méthode commence par affecter les cinq premières MVs aux cinq classes respectivement puis, à chaque itération elle calcule la distance entre la MV et le centre de gravité de la classe.

$$D_{xy} = \sqrt{(Xcpu - Ycpu)^2 + (Xram - Yram)^2} \quad (1)$$

Où D_{xy} est la distance entre deux nœuds.

$$\left.\begin{aligned} Ci: CPUi &= \frac{\sum_{j=1}^{n} MVi(cpu)}{n} \\ Ci: RAMi &= \frac{\sum_{j=1}^{n} MVi(ram)}{n} \end{aligned}\right\} \quad (2)$$

Où Ci (CPU, RAM) est le centre.

Le Classiffieur affecte la MV à la classe où la distance est la minimale, jusqu'à ce que la structure des classes soit la même.

Les classes sont nommées TR, R, M, L, TL, et chacune d'elle est caractérisée par son centre de gravité Ci.

4.2 Routeur

Le Routeur commence par calculer le pire temps d'exécution d'une tâche en se basant sur les travaux de Fahim, et al. (2016), puis en fonction de ce temps calculé, de la taille de la tâche et en fonction du centre de gravité Ci de chaque classe, il sélectionne la classe la plus adéquate à la tâche.

4.3 Equilibreur local

Il reçoit la tâche, émise par le routeur, avec son pire temps d'exécution afin de l'affecter à la MV qui lui convient. A chaque arrivé ou départ d'une tâche, cet équilibreur assure l'équilibrage de charge local de sa classe. Pour cela, il calcul la charge de chaque MV (MV_c) selon la formule (3).

$$MV_c = Vp * \sum_{j=1}^{m} Tj \quad (3)$$

Où Vp est la vitesse du processeur, Tj est le pire temps d'exécution de la tâche et m est le nombre de tâches en attente d'exécution. Puis il calcul la moyenne de charge de la classe selon la formule (4), où n est le nombre de MVs.

$$Moy_c = \frac{\sum MV_c}{n} \quad (4)$$

Après il sélectionne les MVs surchargées, ç-à-d ceux dont leur charge est supérieure à la moyenne, afin de migrer certaines tâches vers les MVs les moins chargées. Le choix de la tâche à migrer se fait en se basant sur deux critères essentiels qui sont: la taille de la tâche et son pire temps d'exécution. Afin de minimiser le cout de migration, l'équilibreur local doit choisir la tâche dont la taille la plus petite possible mais le pire temps d'exécution est le plus

élevé. Pour sélectionner une tâche qui satisfait ces deux critères, on utilise la méthode TOPSIS (Yezza, 2017).

4.4 Equilibreur général

De la même façon que l'équilibreur local, l'équilibreur général calcul la moyenne de charge des classes afin de migrer les tâches des classes les plus chargées vers les classes les moins chargées. Dans ce contexte, nous pouvons distinguer deux cas. Le premier cas est lorsque la classe réceptrice est inferieure à la classe émettrice, dans ce cas la tâche candidate doit être choisie parmi celles ayant le pire temps d'exécution le plus petit. Dans le cas contraire, il choisira la tâche dont le pire temps d'exécution est le plus élevé.

5 Etude de cas

Cette section sera consacrée à la validation de notre modèle proposé à travers une étude de cas détaillée afin de clarifier le fonctionnement de notre approche.

5.1 Scenario 1: Classification des MVs

Dans ce scenario, nous considérons que le système est composé de 20 MVs où chaque MV à ses propres caractéristiques, comme le montre le Tableau 1 :

ID MV	Nom MV	Vitesse CPU (MIPS)	Taille RAM (G.O)	ID MV	Nom MV	Vitesse CPU (MIPS)	Taille RAM (G.O)
0001	MV1	5.0	4	0011	MV11	1.8	2
0002	MV2	6.0	5	0012	MV12	2.2	4
0003	MV3	3.2	2	0013	MV13	2.1	3
0004	MV4	3.5	3	0014	MV14	2.0	5
0005	MV5	4.2	3	0015	MV15	6.1	5
0006	MV6	2.5	4	0016	MV16	3.3	6
0007	MV7	6.2	3	0017	MV17	4.8	2
0008	MV8	4.1	4	0018	MV18	4.0	7
0009	MV9	4.0	4	0019	MV19	2.9	3
0010	MV10	4.0	3	0020	MV20	3.1	8

TAB. 1 – *Liste des machines virtuelles de notre environnement.*

L'algorithme K-means, se base sur le calcul des distances entre les MVs (Formule 1), et les centres de gravité des classes Ci (Formule 2).

Après l'exécution de cet algorithme on obtient cinq classes de MVs. Ces classes de MVs sont indiquées dans le Tableau 2. Comme cité plus haut, les classes sont nommées respectivement, Très Rapide, Rapide, Moyenne, Lente, Très Lente, et chaque classe est caractérisée par son centre de gravité (Ci).

Classe1 (TR)	Classe2 (R)	Classe3 (M)	Classe4 (L)	Classe5 (TL)
C1	C2	C3	C4	C5
MV1, MV2, MV7, MV15	MV16, MV18, MV20	MV3, MV11, MV13, MV19	MV6, MV12, MV14	MV4, MV5, MV8, MV9, MV10, MV17

TAB. 2 – *Classification des machines virtuelles en 5 classes (K-moyennes).*

5.2 Scenario 2 : Affectation des tâches aux classes

A l'arrivé de la tâche, le routeur va calculer le pire temps d'exécution (PTE) en exécutons l'algorithme WCET (Fahim et al. 2016). Soit le Tableau 3 des tâches ci dessous. Ce tableau présente l'ensemble des tâches caractérisées par leur taille et leur PTE.

Tâche	Taille (MI)	PTE	Tâche	Taille (MI)	PTE
T1	105	4.2	T11	52	2.2
T2	127	5.1	T12	410	8.1
T3	210	6.4	T13	119	4.8
T4	240	6.6	T14	221	6.4
T5	158	5.9	T15	265	7
T6	721	12.3	T16	220	6.4
T7	320	7.3	T17	147	5.7
T8	410	8.1	T18	152	5.8
T9	110	4.6	T19	149	5.7
T10	58	2.4	T20	36	1.1

TAB. 3 – *Liste des tâches avec leur PET.*

Apres avoir calculé le PTE, le routeur envoi les tâches aux équilibreurs locaux en fonction du centre de gravité des classes et des caractéristiques des tâches (PTE, taille). Le Tableau 4 ci-dessous montre le résultat de distribution des tâches sur les différentes classes:

Classe1 (TR) C_1	Classe2 (R) C_2	Classe3 (M) C_3	Classe4 (L) C_4	Classe5 (TL) C_5
T2, T3, T4, T5, T6, T7, T8, T12, T14, T15, T16, T17, T18, T19	T1, T9, T13		T10, T11	T20

TAB. 4 – *Résultat de distributions des tâches entre les cinq classes.*

5.3 Scenario 3 : Affectation des tâches aux MVs d'une classe

A l'arrivé de la tâche à la classe avec son PTE calculé précédemment, l'équilibreur local affecte la tâche à la MV correspondante en se basant sur la formule (4). Ainsi pour minimiser le temps de réponse et de service, l'équilibreur local doit prendre en considération la charge moyenne de la classe. Le Tableau 5 indique comment les tâches sont distribuées, par l'équilibreur local, entre les MVs de la classe TR.

Classe TR	MV	Exécution	File d'attente						
	MV 1	T2							
	MV 2	T3	T17	T19	T20				
	MV 7	T4	T6	T7	T8	T12	T14	T15	T16
	MV 15	T5							

TAB. 5 – *Etat des tâches de la Classe TR.*

5.4 Scenario 4 : Migration des tâches intra-classe

L'équilibreur local de la classe TR commence par calculer les charges de chaque MV de sa classe (MV1, MV2, MV7, MV15) selon la formule 3 (cf histogrammes en bleu de la figure 3). , puis il calcul la moyenne de charge globale de la classe TR selon la formule 4.

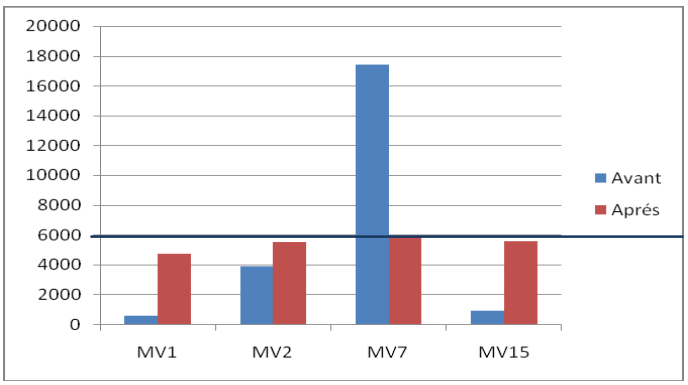


FIG. 3 – *Migration intra-classe*.

Pour équilibrer la charge entre les MVs, le composant équilibreur local doit migrer les tâche de MV7 (MV surchargée) vers les autres MVs et qui sont moin-chargées. Dans le cas où la tâche à sélectionner va être migrée de MV7 vers MV1 le composant utilise la méthode TOPSIS afin de choisir la tâche dont le pire temps d'exécution est le plus élevé d'un côté et ayant la plus petite taille possible, pour minimiser le temps de transfert.

Dans ce scénario, et en appliquant toutes les étapes de la méthode de TOPSIS, l'équilibreur va sélectionner la tâche T8 pour la migrer vers la MV1. La Figure 4 montre les étapes d'application de l'algorithme TOPSIS.

Etape 1

	Taille	W
T6	721	12.3
T7	320	7.3
T8	410	8.1
T14	221	6.4
T15	265	7
T16	220	6.4

Etape 2

Poids	0.1	0.2
	Taille	W
T6	0.7367095	0.6141485
T7	0.3269723	0.3644946
T8	0.4189333	0.4044392
T14	0.2258153	0.3195569
T15	0.2707739	0.3495154
T16	0.2247935	0.3195569

Etape 3

Poids	0.1	0.2
	Taille	W
T6	0.07367095	0.12282969
T7	0.03269723	0.07289892
T8	0.04189333	0.08088785
T14	0.02258153	0.06391138
T15	0.02707739	0.06990308
T16	0.02247935	0.06391138

Etape 4

A+	0.07367095	0.06391138
A-	0.02247935	0.12282969

Etape 5

Poids	0.1	0.2		
	Taille	W	E+	E-
T6	0.07367095	0.12282969	0.058918	0.0511916
T7	0.03269723	0.07289892	0.041948	0.0509655
T8	0.04189333	0.08088785	0.036028	0.0462171
T14	0.02258153	0.06391138	0.051089	0.0589184
T15	0.02707739	0.06990308	0.046977	0.053126
T16	0.02247935	0.06391138	0.051192	0.0589183

Etape 6

	S*
T6	0.46491367
T7	0.54852748
T8	0.56194349
T14	0.53558371
T15	0.53071203
T16	0.53508633

Etape 7

T8	0.56194349	le max S*

FIG. 4 – *Les étapes TOPSIS pour choisir la tâche T8*.

Après migration de quelques tâches de la MV7 vers les autres MVs, on obtient une classe des MVS équilibrée comme le montre les histogrammes de la figure 3.

5.5 Scenario 5 : Migration des tâches interclasses

Périodiquement, l'équilibreur général envoi une demande aux équilibreurs locaux leurs demandant la charge moyenne de leur classe. Ainsi, Les équilibreurs locaux communiquent la charge moyenne de leur classe à l'équilibreur général, afin qu'il assure un équilibrage inter-classe. Pour cela, il va migrer des tâches des classes surchargées vers les classes les moins chargées, toute en prenant en considération les performances des classes lors de la sélection des tâches

De la même manière que l'équilibreur local, l'équilibreur général calcule la moyenne de la charge des classes (comme l'indique les histogrammes en bleu de la figure 5), et après la migration des tâches de la classe C_1 vers les autres classes, on obtient un état équilibré entre les classes comme l'indique les histogrammes en vert de la figure 5.

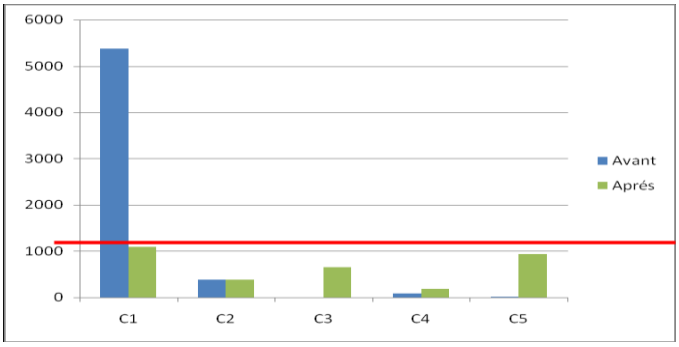


FIG. 5 – *Equilibrage de charge interclasses*.

6 Conclusion

Dans ce papier, nous avons abordé le problème de l'équilibrage de charge entre les MVs d'un CloudIoT, en proposant une architecture basée sur quatre composants essentiels. Le fonctionnement de notre modèle pour équilibrer la charge se déroule en sur trois phases principales: la première phase consiste à classifier les MVs en cinq classes, la deuxième phase est l'affectation des tâches aux classes et ceci en fonction des caractéristiques des tâches (PTE et Taille) et le centre de gravité des classes, quant à la troisième phase, elle permet d'équilibrer la charge intra-classe et interclasse.

Comme perspectives, nous envisageons de valider notre modèle en effectuant des simulations dans un environnement CloudIot réel afin de montrer la pertinence de notre modèle d'un coté et d'approfondir l'étude de l'impacte du nombres des clients (objets), le temps nécessaire affecter les tâches aux classes, et bien d'autres paramètres qui nous permettrons de ressortir les points positifs et négatifs de notre proposition.

Références

Alakeel, A.M. (2010). A Guide to Dynamic Load Balancing in Distributed Computer Systems. *International Journal of Computer Science and Network Security*, 10 (6): 153-160.

Al-Rayis, E. et H. Kurdi (2013). Performance Analysis of load balancing architectures in cloud computing. In: IEEE European Modeling Symposium, Manchester, UK.

Botta, A., W. Donato, V. Persico, et A. Pescapé (2015). Integration of Cloud Computing and Internet of Things: a Survey. *Future Generation Computer Systems*, 56: 684-700.

Cardellini, V., M. Colajanni, et P.S. Yu (1999). Dynamic Load Balancing On Web-server Systems. *IEEE Internet Computing*, 10: 28.39.

Chien, N.K., et H.D. Loc (2016). Load balancing algorithm based on estimating finish time of services in cloud computing. In: 18th International Conference on Advanced Communication Technology, Pyeongchang, South Korea.

Fahim, Y., E.H. Benlahmar, L. Elhoussine, et A. Eddaoui (2016). Une nouvelle conception d'équilibrage de charge dans le Cloud Computing. In: 4ème Journée sur les Technologies d'Information et de Modélisation, Casablanca, Morroco.

Fischer, A., (2014). Deux méthodes d'apprentissage non supervisé: synthèse sur la méthode des centres mobiles et présentation des courbes principale. *Journal de la Société Française de Statistique*, 155(2): 2-35.

Hemam, S.M., O., Hioual, et O., Hioual (2017). Load balancing between nodes in a volunteer Cloud Computing by taking into consideration the number of Cloud services replicas. In: 3rd International Conference of Cloud Computing Technologies and Applications (CloudTech), Rabat, Morocco.

Kapoor, S., et C. Dabas, (2015). Cluster based load balancing in cloud computing. In: Eighth International Conference on Contemporary Computing (IC3), Noida, India.

Mell, P., et T. Grance (2009). *The NIST definition of cloud computing*, National Institute of Standards and Technology, version 15.

Olejnik, R., I. Alshabani, B. Toursel, E. Laskowski, et M. Tudruj (2009). Load balancing metrics for the SOAJA framework. *Scientific International Journal for Parallel and Distributed Computing, Scalable Computing: Practice and Experience*, 10 (4): 1-10.

Sidhu, A.K., et S. Kinger (2013). Analysis of Load Balancing Techniques in Cloud Computing. *International Journal of Computers & Technology*, 4(2): 737-741.

Yakhchi, S., S. M. Ghafari, M. Yakhchi, M. Fazeli, et A. Patooghy (2015). ICA-MMT: A load balancing method in cloud computing environment. In: 2nd World Symposium on Web Applications and Networking (WSWAN), Sousse, Tunisia.

Yezza, A. (2017). La méthode TOPSIS expliqué pas à pas. Rapport proposé Mais 2015, mis à jour Avril 2017.

Summary

CloudIoT is a new paradigm that has emerged as a result of the integration of Cloud Computing and the Internet of Things. It provides a set of intelligent services and applications, which can strongly influence on our daily lives. In the CloudIoT, physical sensors are responsible for capturing and transmitting data to the cloud for processing and storing them. In this context, the amount of data to be processed through the CloudIoT is, usually, increasing, and in this case, a load balancing mechanism is needed to distribute the captured data between the different Virtual Machine (VM) of the CloudIoT.

In this paper, we propose an approach that allows load balancing between the different virtual machines of a CloudIot. The proposed approach is composed of four necessary components: The Classifier which is responsible for assigning Virtual Machines (VMs) to the corresponding classes, the router which attributes tasks to their corresponding class, the local balancer whose role is to balance the tasks between the different VMs of the same class and finally, the general balancer component which is responsible to balance the load between classes. The obtained results through the case study show that our approach effectively allows balancing the load between VMs.

Conception d'une architecture distribuée de stationnement intelligent basée sur les systèmes multi-agents et l'internet Des objets

Khaoula Hassoune *,**, Wafaa Dachry*, Fouad Moutaouakkil*, Hicham Medromi*

*(EAS) Equipe Architecture des systemes, Laboratoire de Recherche en Ingénierie
Université Hassan II, ENSEM
**Laboratoire Pluridisciplinaire de recherche et d'innovation (LPRI)
EMSI, Maroc
{Khaoula.hassoune,wafaa.Dachry}@gmail.com
fmoutaouakkil@hotmail.com
hmedromi@yahoo.fr

Résumé. Cet article présente une architecture dynamique de stationnement intelligent qui offre de nombreux services au conducteur basée sur les systèmes experts, les systèmes multi-agents et l'internet des objets. Nous avons intégré les différentes technologies afin de réaliser un système efficace, fiable, sécurisé et peu coûteux. Le paradigme du Cloud Computing a été sélectionné pour concevoir et développer l'architecture mentionnée, car il présente des caractéristiques intéressantes, telles que la réduction des coûts, l'agilité, l'évolutivité et l'élasticité, entre autres. Et pour améliorer l'intelligence de notre système, il est nécessaire d'intégrer l'aspect apprentissage dans nos agents pour assurer des services avec un degré maximal de précision, d'efficacité et de bonnes performances. Les agents sont des programmes informatiques sophistiqués qui agissent de manière autonome, dans des environnements ouverts et distribués, pour résoudre un nombre croissant de problèmes complexes.

1 Introduction

Les chercheurs se sont récemment tournés vers l'application de nouvelles technologies pour la gestion du stationnement par la conception et la mise en place de différents prototypes de systèmes de stationnement intelligent qui permet aux conducteurs de véhicules de trouver les places de stationnement en temps réel, de réserver et de payer. À l'avenir, la demande pour le service de stationnement intelligent augmentera en raison de la croissance rapide de l'industrie automobile. La gestion automatique des parkings par un suivi précis et une prestation de services aux clients et aux administrateurs est assurée par ces services émergents. Une solution efficace à ce service peut être fournie par de nombreuses nouvelles technologies.

Cet article décrit une architecture dynamique pour la gestion d'un système de stationnement intelligent basée sur les systèmes multi-agents et les systèmes experts. L'une des principales caractéristiques est l'utilisation des agents intelligents en tant que composants principaux qui se concentrent sur la distribution de la majorité des fonctionnalités du système.

Le papier est organisé comme suit: après une brève introduction, nous aborderons dans la deuxième section un état de l'art sur les systèmes de stationnement intelligent, puis, dans la troisième section, nous présenterons les concepts des systèmes multi-agents et des systèmes experts. Enfin, nous présenterons une architecture modulaire pour la gestion du stationnement intelligent. La dernière section conclut notre travail et donne un aperçu sur les perspectives.

2 Etat de l'art

Les Villes ont remarqué que leurs citoyens avaient des problèmes pour trouver une place de parking facilement surtout pendant les heures de pointe cela est dû à ne pas savoir où les places de stationnement sont disponibles à un moment donné. Même si elle est connue, de nombreux véhicules peuvent poursuivre un petit nombre de places de stationnement qui à son tour conduit à la congestion du trafic. Diverses approches et recherches sont faites pour surmonter les difficultés de parking. Par conséquent, de nombreux systèmes et technologies sont développés pour le stationnement.

Les auteurs (Yang et al. (2012), Yee et al. (2014), A.Poojaa et al. (2015)**)** présentent la conception et la mise en place d'un système de stationnement intelligent basé sur les réseaux de capteurs sans fil qui permettent aux conducteurs de véhicules de trouver les places de stationnement rapidement. En outre, (Orrie et al. (2015)) présente un système sans fil permettant de localiser les parkings à distance via un Smartphone. Ce système automatise le processus de localisation d'un emplacement de stationnement et de paiement.

Les auteurs de (Karbab et al. (2015)) ont proposé un cadre de stationnement de voiture évolutif et peu coûteux (CFP) basé sur l'intégration de réseaux de capteurs sans fil et de la technologie RFID. Ce framework inclue l'orientation du conducteur, le paiement automatique, la sécurité et la détection du vandalisme.

Dans d'autres études, les auteurs ont choisi de concevoir un système de stationnement intelligent automatique à l'aide d'Internet qui permet à l'utilisateur de trouver la zone de stationnement la plus proche et la place disponible dans cette zone (Basavaraju S R. (2015), Suryady et al . (2014), Gandhi et al. (2016)).

A partir de l'état de l'art, nous remarquons que les chercheurs ont favorisé certains services au détriment d'autres. Pour cette raison, nous proposons une nouvelle architecture basée sur des systèmes multi-agents et les systèmes experts. Nous devrions intégrer les deux technologies afin de réaliser un efficace, fiable, sûr et plus économique.

Le modèle proposé est une architecture modulaire multi-agent où tous les processus sont gérés et contrôlés par des agents capables de coopérer , de proposer des solutions et d'agir sur des environnements dynamiques pour résoudre des problèmes réels.

Différents types d'agents ont été utilisé dans l'architecture, chacun avec des rôles, des capacités et des caractéristiques spécifiques. Ce qui va faciliter la flexibilité de l'architecture en intégrant de nouveaux agents.

3 Architecture multi-agent globale du système de stationnement intelligent

À partir de l'état de l'art effectué précédemment on remarque que la plupart des auteurs ont proposé des systèmes limités en terme de :

- Centralisation :
  - le serveur est le seul maillon faible de ces systèmes, étant donné que tout le réseau est architecturé autour de lui.
  - Plusieurs utilisateurs se partagent des fichiers de données stockés sur un serveur commun ;
  - la gestion des conflits d'accès aux données doit être prise en charge par chaque programme de façon indépendante, ce qui n'est pas toujours évident.
  - Lors de l'exécution d'une requête, l'intégralité des données nécessaires doit transiter sur le réseau et on arrive à saturer ce dernier.
  - Il est difficile d'assurer la confidentialité des données

- Services : les systèmes proposés dans l'état de l'art offrent des services au détriment d'autres.

Pour cette raison on va proposer une architecture distribuée de stationnement Intelligent qui se base sur les systèmes multi agents et qui offre des performances en terme de :

- Distribution : La résolution des problèmes distribuées se préoccupe de la façon dont un problème donné peut être résolu par plusieurs modules (appelés noeuds) qui coopèrent en divisant et en partageant la connaissance à propos du problème et des solutions développées. Les systèmes multi-agents implémentent des stratégies de résolution basées sur le comportement d'un ensemble d'agents autonomes qui, éventuellement, peuvent déjà exister. Un système multi-agents peut être vu comme un ensemble faiblement interconnecté d'agents qui travaillent ensemble pour résoudre un problème en s'appuyant sur les capacités et les connaissances individuelles de chaque entité. Les agents sont autonomes et sont de nature hétérogène. Les systèmes multi-agents se caractérisent dès lors par :

  - la modularité, la vitesse (avec le parallélisme),
  - la fiabilité (due à la redondance).
  - le traitement symbolique (au niveau des connaissances),
  - la facilité de maintenance,
  - la réutilisation et la portabilité
  - la coopération (travailler ensemble à la résolution d'un but commun) ;
  - la coordination (organiser la résolution d'un problème de telle sorte que les interactions nuisibles soient évitées ou que les interactions bénéfiques soient exploitées) ;
  - la négociation (parvenir à un accord acceptable pour toutes les parties concernées).

- Services : Notre système doit offrir plusieurs services aux utilisateurs

- Gestion de places de stationnement (en temps réel, prédiction)
- Guidance (colonie des fourmilles)
- Réservation
- Gestion des profils utilisateurs
- Paiement
- Prédiction

3.1 Architecture proposée

Dans cette section, nous donnons un aperçu sur l'architecture multi-agent qui fournit un modèle de haut niveau pour la gestion du stationnement intelligent (Figure 1). Notre travail repose sur les approches de systèmes multi-agents et les systèmes experts en raison de leurs avantages.

La combinaison de ces deux approches englobera la coopération, la résolution de problèmes complexes, la modularité, l'efficacité, la fiabilité, la réutilisation et se situent sous l'utilisation conjointe du savoir en tant que modèles comportementaux des experts.

Notre solution proposée se concentre principalement sur l'analyse des requêtes des utilisateurs pour trouver un emplacement vacant en fonction de leurs préférences.

3.1.1 Les systèmes multi-agents

Un agent est une entité autonome, réelle ou abstraite, qui est capable d'agir sur elle-même et sur son environnement, qui, dans un univers multi-agents, peut communiquer avec d'autres agents, et dont le comportement est la conséquence de ses observations, de ses connaissances et des interactions avec les autres agents (Ferber, 1995).

L'agent a plusieurs caractéristiques :

Situé – l'agent est capable d'agir sur son environnement à partir des entrées sensorielles qu'il reçoit de ce même environnement;

Autonome – l'agent est capable d'agir sans l'intervention d'un tiers (humain ou agent) et contrôle ses propres actions ainsi que son état interne;

Proactif – l'agent doit exhiber un comportement proactif et opportuniste, tout en étant capable de prendre l'initiative au bon moment;

Capable de répondre à temps – l'agent doit être capable de percevoir son environnement et d'élaborer une réponse dans le temps requis;

Social – l'agent doit être capable d'interagir avec des autres agents (logiciels ou humains) afin d'accomplir des taches ou aider ces agents à accomplir les leurs.

Un système multi-agents est un système distribué composé d'un ensemble d'agents. Un SMA est caractérisé ainsi: " chaque agent a des informations ou des capacités de résolution de problèmes limités (ainsi, chaque agent a un point de vue partiel); " il n'y a aucun contrôle global du système multi-agents; " les données sont décentralisées; " le calcul est asynchrone.

3.1.2 Les systèmes experts

D'une manière générale, un système expert est un outil capable de reproduire les mécanismes cognitifs d'un expert, dans un domaine particulier. Il s'agit de l'une des voies tentant d'aboutir à l'intelligence artificielle.

Plus précisément, un système expert est un logiciel capable de répondre à des questions, en effectuant un raisonnement à partir de faits et de règles connues. Il peut servir notamment comme outil d'aide à la décision.

Tout Système Expert est composé de 3 principaux éléments : une base de connaissances, un moteur d'inférence et une interface graphique.

- Base de connaissances : est l'ensemble des données qui sont utilisées par le moteur d'inférence. Cette base est divisée en 3 parties :

  - Les standards d'engagement (connaissances de l'expert) : Cette partie représente les informations de base et de configuration du système : mesures (parfois en direct), lois, paramètres, données contractuelles.
  - Les règles d'inférence : Cette partie représente l'ensemble des règles logiques de déduction utilisées par le moteur d'inférence.
  - La base de faits (expérience) : Historisation et statistique des faits effectifs, des décisions et des buts.

- Moteur d'inférence : est un mécanisme qui permet d'inférer des connaissances nouvelles à partir de la base de connaissances du système. Il est basé sur des règles d'inférence qui régissent son fonctionnement. Il a pour fonction de répondre à une requête de la part d'un utilisateur ou d'un serveur afin de déclencher une réflexion définie par ses règles d'inférence qui utiliseront la base de connaissance. Il peut alors fonctionner en chaînage avant ou chaînage arrière. Il est a noter toutefois qu'il reste important que la base de connaissance reste indépendante du moteur d'inférence (sauf si elle contient les règles d'inférence elles-mêmes).

- Interface graphique : Même si son importance est de taille dans tout application cliente, elle l'est d'autant plus ici qu'un SE doit parfaitement s'intégrer à un milieu professionnel et aux habitudes de ses experts. Si celui-ci n'est pas capable de s'approprier naturellement le logiciel, c'est que l'interface graphique n'est pas correcte. La priorité est donc à l'intuitivité et à la représentation fidèle de l'environnement.

Dans cette section, nous donnons un aperçu de l'architecture multi-agents basée sur IOT Middleware, qui fournit un modèle de haut niveau pour la gestion intelligente des parkings (Fig. 1). Afin de comprendre en profondeur l'architecture commune, nous décrivons ci-dessous chaque couche de celle-ci.

- Système Expert: fournir des connaissances relatives à la réglementation des agents. Il s'agit d'un système expert et sa base de connaissances contient principalement des informations relatives à la performance environnementale et aux législations.

- Couche Communication: Cette couche assure la communication entre toutes les couches de l'architecture.

- Couche de traitement: Cette couche contient différents agents, qui peuvent être implémentés, répondant aux alertes de la couche de communication et aux demandes des utilisateurs.
-
- Couche de prétraitement des métadonnées: est responsable du prétraitement des données capturées dans l'environnement (capteurs, caméra, RFID). Il transmet les informations des sources de données au middleware IOT basé sur le cloud. Il complète la modélisation des données pour traiter et intégrer correctement diverses sources d'informations.

- L'écosystème FIWARE fournit des outils libres et gratuits qui peuvent être utilisés comme composants logiciels pour différentes applications ou architectures(Fiware Project). Il inclut des protocoles de sécurité lorsque l'information est échangée. L'IOT Middleware est établi dans le cloud et représente le composant principal de la plate-forme. il est divisé en deux composantes fonctionnelles, l'une en charge de la gestion des ressources (immatriculation des périphériques IOT réels, gestion des périphériques...) et l'autre exposant les ressources IOT virtualisées sous-jacentes par le biais de ses services IOT associés. l'accès aux données, qu'il s'agisse d'ensembles de données historiques ou de flux de données en temps réel.

- Couche Application: Il fournit les outils pour accéder aux services d'IOT Middleware. Il permet un nuage de services en facilitant le développement de nouvelles applications et de nouveaux services pour les conducteurs.

- Couche coordination: affichage des informations à l'utilisateur d'une manière adaptée en tenant compte des contraintes de son appareil. Et il est responsable de la transmission de la demande de l'utilisateur à un agent spécifique.

- Couche Interface: Il est responsable de la capture de la requête de l'utilisateur, ainsi que de l'affichage des résultats.

- Couche physique: Il s'agit des capteurs, caméras et tags RFID qui capturent les informations de l'environnement.

- Couche Action: Représente l'ensemble des comportements nécessaires au contrôle du stationnement.

3.2 Description de l’architecture proposée

La circulation sur les routes et les aires de stationnement est un sujet de préoccupation dans la majorité des villes. Pour éviter ces problèmes, notre principal objectif est de conce-

voir une architecture dynamique basée sur le cloud et des systèmes multi-agents qui permettent le déploiement de systèmes de stationnement,

La mise en œuvre de ce cadre évolutif et peu coûteux pour le stationnement automobile offrira de nombreux services au conducteur: guidance, paiement automatique, prédiction, connaissance pour prendre des décisions, sécurité et faible coût de mise en œuvre.

Notre travail est basé sur les systèmes multi-agent, les systèmes experts et l'internet des objets en raison de leurs avantages.

La combinaison de ces trois approches comprendra la coopération, la résolution de problèmes complexes, la modularité, l'efficacité, la fiabilité, la réutilisabilité et repose sur l'utilisation conjointe de diverses connaissances en tant que modèles comportementaux des experts.

Il existe différents types d'agents dans l'architecture, chacun ayant des rôles, des capacités et des caractéristiques spécifiques. Chaque agent a plusieurs fonctions qui fonctionnent en temps réel; les agents communiquent pour mieux effectuer les services de stationnement.

Agent Camera: Cet agent capture l'image à partir d'une séquence vidéo et l'envoie à l'agent détection

Agent capteur: Le rôle de cet agent est de détecter des événements ou des changements dans son environnement et d'envoyer les informations à la couche de prétraitement des métadonnées.

Agent RFID: utilise des champs électromagnétiques pour identifier et suivre automatiquement les étiquettes attachées aux objets. Les étiquettes contiennent des informations stockées électroniquement sur les conducteurs.

Agent de Transformation de Données: Responsable du prétraitement des données capturées dans l'environnement (capteurs, Caméra). Le prétraitement comprend la saisie efficace des mesures d'entrée pour déterminer l'état de chaque emplacement de stationnement et pour identifier l'état des conducteurs.

Agent filtrage de données: est un large éventail de stratégies ou de solutions pour affiner les ensembles de données en fonction des besoins d'un utilisateur.

Agent Manager: Il a pour rôle d'afficher les informations à l'utilisateur de manière appropriée en tenant compte des contraintes de son appareil. Et il est responsable de la transmission de la demande de l'utilisateur à un agent spécifique qui prendra soin du traitement de la demande.

Agent d'interface: Cet agent est chargé de capturer la requête de l'utilisateur et d'afficher les résultats.

Agent Barrière: Cet agent est responsable de l'ouverture de la barrière à l'entrée et à la sortie du parking sur décision de l'agent expert.

Agent de contrôle des communications: Son rôle est d'assurer la communication entre les agents du système.

Agent de réservation: Il a pour fonction de recevoir et de traiter la demande de réservation envoyée par l'utilisateur et de retourner la réponse au gestionnaire de l'agent après avoir communiqué avec l'agent d'information.

Agent identification de l'utilisateur: Permet d'identifier les différents profils qui veulent se connecter à l'application mobile et accéder au parking (abonnés, non abonnés, déjà réservés...). Il est toujours en contact direct avec la base de données par le biais de l'agent d'information.

Agent d'orientation: Il a pour rôle de recevoir et de traiter la demande d'orientation envoyée par l'utilisateur par le calcul du chemin le plus court et le plus adapté à l'utilisateur.

Agent de paiement: Cet agent vérifie le statut du paiement de l'utilisateur. Il est toujours en contact direct avec la base de données par le biais de l'agent d'information.

Agent de gestion des emplacements: Il donne une vue d'ensemble de l'état de chaque emplacement de stationnement en temps réel après avoir communiqué avec l'agent d'information.

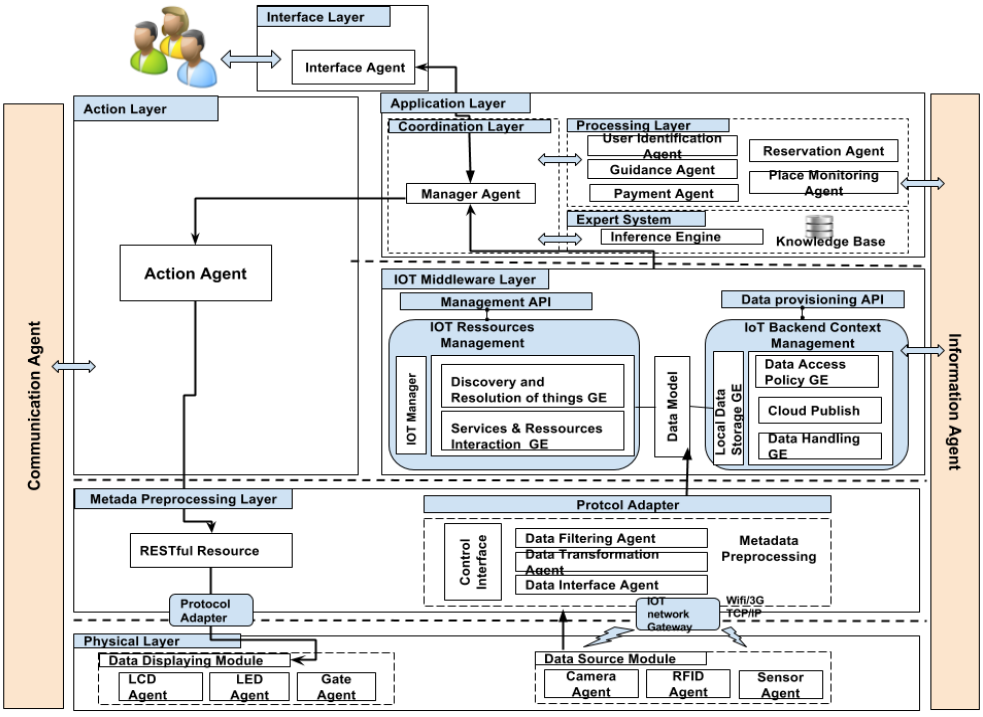


FIG. 1 – Architecture de stationnement intelligent basée sur les systèmes multi-agents et l'internet des objets

Système Expert: Cet agent fournit des connaissances relatives à la réglementation des agents. C'est un agent et sa base de connaissances contient principalement des informations relatives à la performance environnementale et aux législations.

Agent d'Action: représente tous les comportements nécessaires au contrôle des LEDs installées dans chaque parking, à la gestion du portail et à l'affichage des places disponibles.

4 Conclusion et perspectives

Comme un travail futur nous allons détailler chaque module de notre architecture, nous allons discuter les caractéristiques des agents et leurs comportements. Nous décrirons également l'implémentation du système proposé, y compris la communication entre les agents et l'interaction avec le système expert.

Dans cet article, nous donnons un aperçu des différents systèmes de stationnement mis en place par de nombreux chercheurs pour résoudre le problème croissant de la congestion de la circulation et aider à fournir un meilleur service public, réduire les émissions des voitures et la pollution. nous avons proposé une architecture multi-agents qui fournit un modèle de haut niveau pour la gestion intelligente du stationnement. C'est pour cette raison que nous avons utilisé différentes techniques modernes comme les systèmes experts, les systèmes multi-agents et l'internet des objets. Nous avons intégré les différentes technologies afin d'obtenir un système le plus efficace, fiable, sûr et peu coûteux.

Références

A.Poojaa (2015). Wsn based secure vehicle parking management and reservation system. National Conference on Research Advances in Communication, Computation, Electrical Science and Structures (NCRACCESS-2015)

Basavaraju S R. (2015) .Automatic Smart parking System using Internet of Things (IOT).International Journal of Scientific and Research Publications, Volume 5, Issue 12, December 2015

Gandhi, B. M. K., & Rao, M. K. (2016). A Prototype for IoT based Car Parking Management system for Smart cities. Indian Journal of Science and Technology, 9(17).

Jackson P. (1999), Introduction to Expert Systems, Harlow, England: Addison Wesley Longman, Third Edition.

Karbab, E., Djenouri, D., Boulkaboul, S., & Bagula, A. (2015, May). Car park management with networked wireless sensors and active RFID. In 2015 IEEE International Conference on Electro/Information Technology (EIT) (pp. 373-378). IEEE.

Orrie, O., Silva, B., & Hancke, G. P. (2015, November). A wireless smart parking system. In Industrial Electronics Society, IECON 2015-41st Annual Conference of the IEEE (pp. 004110-004114). IEEE.

Suryady, Z., Sinniah, G. R., Haseeb, S., Siddique, M. T., & Ezani, M. F. M. (2014, November). Rapid development of smart parking system with cloud-based platforms. In Information and Communication Technology for The Muslim World (ICT4M), 2014 The 5th International Conference on (pp. 1-6). IEEE.

Yang, J., Portilla, J., & Riesgo, T. (2012, October). Smart parking service based on wireless sensor networks. In IECON 2012-38th Annual Conference on IEEE Industrial Electronics Society (pp. 6029-6034). IEEE.

Yee, H. C., & Rahayu, Y. (2014). Monitoring Parking Space Availability via ZigBee Technology. International Journal of Future Computer and Communication, 3(6), 377.

Fiware Project, 2016. Core Platform for the Future Internet. Private Public Partnership Project (PPP), https://www.fiware.org. Consulted on May 2016

Summary

This paper presents a dynamic smart parking architecture that offers many services to the driver based on multi-agent, expert systems and IOT Middleware. We have integrated the different technologies together in order to achieve a system which is the most efficient, reliable, secure and inexpensive. The Cloud Computing paradigm has been selected for designing and developing the mentioned architecture, since it exhibits interesting features, such as cost reduction, agility, scalability and elasticity, among others. And to improve intelligence of our system it is necessary to integrate the learning aspect in our agents to ensure services with a maximum degree of accuracy, efficiency and good performance. Agents are sophisticated computer programs that act autonomously on behalf of their users, across open and distributed environments, to solve a growing number of complex problems.

A New Adaptive Routing Protocol for Internet of Things in Mobile Ad Hoc Networks

Nabil Nissar*, Najib Naja*
Abdellah Jamali**
* National Institute of Posts and Telecommunications, INPT Rabat, Morocco
{nissar, naja}@inpt.ma
** Hassan I University, Settat, Morocco
jamali.abdellah1@gmail.com

Abstract. Nowadays, Internet of things (IoT) is considered one of the newest emerging technologies. Objects are having identities and virtual personalities allowing them to send and receive data and to use intelligent interfaces to communicate within social environmental using the internet. This paper first analyses the mechanisms of some existing routing protocols such as AODV, DSR, DSDV, ZRP, and OLSR, which are widely used in Mobile Ad Hoc networks and their performance in IoT environments in order to find an appropriate routing mechanism for the IoT systems. Then we present a new signal strength-based approach for RREQs forwarding to better exploit the neighbor coverage knowledge and hence reduce routing overhead in MANETs in order fulfill the requirements of IoT environments in terms of average end to end delay, PDF, throughput and especially routing overhead. Simulation result shows that SAODV mechanism is more suitable for IoT because it performs better in terms of RREQ packets overhead, packets delivery ratio, throughput, and eventually, average end-to-end delay compared to the original AODV routing.

1 Introduction

Internet of Things is an emerging area and it refers the objects or devices interconnecting and cooperating in self-configuring wireless networks to connect all things at all times. In all applications of the Internet of Things, the data is sent from the source to the destination through a routing protocol which should be efficient enough in terms of Quality of Service (QoS), in order to guarantee the data transmission (Tan, 2010). On the other hand, Mobile Ad Hoc Networks represent a new paradigm of mobile wireless communication. They are self-organized networks which are deployed without the need for any fixed infrastructure. MANETs have attracted a lot of attention during the recent years. In the context of IoT, Mobile Ad Hoc Networks could represent scenarios such as people using mobile phones, a rescue team in an evacuation operation or soldiers in military applications, among other examples. MANETs are self-configuring, self-maintaining, self-healing, and self-repairing networks and such features are very suitable for mobile computing. The mobility of clients is an intrinsic characteristic of clients in MANETs which make even more challenging the deployment of these networks in real environments (Hoebeke et al. 2004, Asimakopoulou 2010, and Gutiérrez-Reina 2012). The design of routing protocols is one of the key components of MANETs. Due to the mobility of Nodes/clients and the increase in size of the network expansion and business, typical routing protocol can avoid traffic jamming, improve network performance and efficiency, and deal with such mobility conditions by acting

against possible changes and implementing mechanisms to re-establish broken communication routes. Another important issue related to the discovery process of routing protocols in MANETs, that can affect the QoS in the IoT systems, is the broadcast storm problem caused by the redundancy of request packets. Motivated by the above discussions, our research paper not only analyzes the particularity of Ad Hoc network technology in the Internet of Things, but also concentrates on limiting the number of signaling overhead incurred during discovering and maintaining routes to destinations in order to adapt AODV routing protocol to IoT systems. Therefore, we have suggested a novel signal strength based on-demand routing scheme (SAODV) for RREQ forwarding in order to improve AODV routing protocol performance particularly in terms of routing overhead. Finally, to demonstrate the performance of our novel scheme, we have compared on NS2 simulator five well known topology-based routing protocols: AODV, OLSR, DSR, DSDV and ZRP against the performance of SAODV using 'Random Waypoint Model'. This paper is structured as follows: Section II summarizes prior works related to reducing routing overhead in MANET. Section III introduces the detailed description of our proposed scheme SAODV. Section V presents our developed quantitative model to evaluate SAODV. NS2 experiments and simulation environment are presented in Section VI. Section VII provides the simulation results, analysis and discussions. Section VIII summarizes the main points of this study and discusses the prospects for continued work.

2 Literature review

Routing overhead occurs due to the control message dissemination process (RREQ, RREP, RACK, and RERR) in highly dynamic topologies of MANETs. Literature research has proven that usual reactive routing protocols generate a massive amount of routing traffic by simply flooding the whole network area at the time of route discovery process. Many protocols have been conducted to reduce signaling messages amount and routing overhead, each has their own advantages and disadvantages and have been compared under certain circumstances as in (Hanzo II 2007, Akkaya 2005 and Tian 2010). For instance, Pooja Gupta et al. estimated the execution of AODV, DSR and DSDV and their performance on Routing overhead, Packet delivery fraction ratio, Packet loss Percentage and Average end-to-end delay. AOMDV-IOT proposed by (Tian et al. 2010) is an extension of Ad Hoc On-demand Multipath Distance Vector routing protocol which implements an Internet Connecting Table (ICT) in order to adapt AOMDV with IoT requirements. Kavitha pandey et al. analyzed and compared the performance DSDV,AODV, DSR and ZRP according to Routing overhead, Average delay, Throughput and number of packets dropped. Y. Wang, et al. evaluated the efficiency of DIRCAST, OLSR and AODV according to Routing overhead, Packet delivery fraction ratio and Average end-to-end delay. Chilamkurti et al. displayed a relative investigation of DSR, Ex-DSR according to Routing overhead. Mixed routing protocols such as Landmark Routing Protocol (Wang Tao 2011) were proposed in order to narrow range control message routing dissemination delay characteristics and to improve the delay characteristics in IoT environments.

.

3 Saodv protocol design

3.1 Saodv Mechanism

The basic idea of SAODV is to improve the conventional AODV protocol efficiency to fulfill IoT requirements, especially in terms of routing overhead by limiting the number of route discovery messages to reduce the network overhead. In this approach, each client receiving the RREQ packet will decide whether the network is sparse or not, if it is the case, the received RREQ packet will be rebroadcasted according to the conventional AODV process in order to avoid clients isolation; otherwise, the received signal strength is compared with a predefined threshold value that is set according to clients transmission power. If the received signal strength is exceeding the threshold, the RREQ will be forwarded, and the network layer continues the route discovery process; if not SAODV will simply drop the RREQ packet. The major benefit of this method is to form routes with strong links where neighbor clients are within the consistent transmission range of each other. In high mobility cases, SAODV will have a lower probability of route failure due to link breakages and routing overhead will be reduced accordingly.

3.2 SAODV Algorithm

The following nomenclatures are used to describe SAODV:

- Avg (1): is a threshold value; it is defined as a measure of the average number of neighbor clients of the network:

$$\text{Avg}= \frac{\sum_{i=1}^{n} \text{NBi}}{n} \quad (1)$$

Ni: The i-th client neighbor client's number.
n: Total number of clients in the network.

- RSS (2): is the received signal strength calculated using two ray ground model

$$\text{RSS}= \frac{P_t * G_t * G_r * h_t^2 * h_r^2}{d^4 * L} \quad (2)$$

With:
Pr: Power received at distance d
Pt: Transmitted signal power
Gt: Transmitter gain (1.0 for all antennas)
Gr: Receiver gain (1.0 for all antennas)
d: Distance from the transmitter
L: Path loss (1.0 for all antennas)
ht: Transmitter antenna height (1.5 m for all antennas)
hr: Receiver antenna height (1.5 m for all antennas)

Step 1 *Waiting for a broadcast RREQ packet at Client J.*
Step 2 *Calculating number of neighbors of Client J (NBj).*
Step 3 *Calculating average number of*

neighbors Avg.
Step 4 *If packet RREQ received for the*
first time then If
NBj <=Avg (Sparse Network)
Use Standard AODV Protocol
Else If
NBj >= Avg (Dense Network)
Calculate RSS
If RSS >= Threshold
Rebroadcast the received RREQ.
Else
Drop the received RREQ.
End if
End_if
End_if
End of algorithm

According to the formal description of SAODV algorithm, if the network is dense, the recipient client can take its action by forwarding RREQ packet to clients with a signal strength value (RSS) that is greater than a predefined threshold. Contrariwise, if the simulated network is sparse, for better performance, our algorithm works as the standardized AODV and finds a route on the basis of minimum hop count.

4 Overhead evaluation model

Knowing that route discovery process and destination location discovery process are the main reasons why protocols generate overhead in MANETs, in this section we intend to quantify the number of signaling messages for AODV and SAODV protocols using a quantitative model, in terms of the number of clients, clients mobility, area size, etc. Initially, we introduce our overhead evaluation model, and then we describe the used parameters to evaluate and to determine generated overhead by AODV and our protocol SOADV for one communication.

4.1 Model principle

The model described in the following is depicted in Fig 1. We did a single communication study, in a MANET square network topology, between a source client src and a destination client dst. A message sent by the source can be relayed by one or several relay clients.

Let *CM (P)* be the set of control messages for a protocol P. Using AODV, *CM (AODV) = {RREQ, RREP, RERR}*. Regarding the specific reactive characteristics of AODV algorithm and its route discovery process, the number of generated RREP and RERR can be neglected compared to the generated RREQ messages. Therefore, the reduced form of total control messages forwarded per second N_{AODV} for one communication between source and destination clients using AODV protocol is shown in equation (3):

$$N_{AODV} = N(RREQ) \quad (3)$$

N (RREQ) value is the total number of RREQ packets forwarded by second during a single communication between src and dst using AODV routing protocol. *N (RREQ) (4)* de-

pends on two parameters: *f (RREQ)* and *nr(RREQ). f (RREQ)* is the sending frequency or the number of transmissions per second of RREQ packets, and *nr(RREQ)* is the number of relays for RREQ packets.

$$N(RREQ) = f(RREQ).nr(RREQ) \quad (4)$$

Let denote *l(RREQ)* the route length taken by a RREQ packet in meters, and *d(RREQ)* is the relays density for RREQ, which is the number of relay clients that transmit a message RREQ per meter. We have:

$$nr(RREQ) = d(RREQ).l(RREQ) \quad (5)$$

Next, in order to compute *N = (RREQ),* we establish the required formulas to estimate the values of three parameters: *l(RREQ), d(RREQ)* and *f(RREQ).*

4.1.1 Route length taken by RREQ : l(RREQ)

Using AODV protocol, *l(RREQ)* is being the average distance between the source src and the destination dst clients. Let X_src (respectively X_dst) be the client position of src (respectively dst). We consider a uniform distribution of clients on a square area. In this case, the average distance between src and dst is calculated using (6):

$$l(RREQ) = L = d \times E(|X\_src - X\_dst|) \quad (6)$$

With:

$$E(|X\_src - X\_dst|) = \iint_0^1 |x - y| dxdy \quad (7)$$

$d = \sqrt{a^2 + b^2}$, a and b are the lengths of of a rectangular area sides.
We can calculate L using (6):

$$L = (2 + \sqrt{2} + 5\log(\sqrt{2} + 1)) \, {}^{d}/_{15\sqrt{2}} = 0.36869 \ldots * d$$

4.1.2 Relays density: d(RREQ)

Let denote dt the density of clients in a square area per meter. In case of AODV, *d(RREQ)= dt* because messages are transmitted in a broadcast mode.

4.1.3 Sending frequency for the protocol AODV: f(RREQ)

Using AODV protocol, the RREQ packet transmission frequency relies directly on route failure frequency. In fact, AODV sends a new path discovery request in the network each time a route is broken, let *f(t, n)* denote route failure frequency which is the number of failure on the route between the source client src and the destination client dst per second. As a result, the sending frequency *f(RREQ) = f(t, n).* Consequently, to estimate the route break frequency, firstly the link failure probability between two direct communicating clients should be defined. When the two connected clients are no longer able to communicate directly, then the communication link is considered broken. Let P_{lb}(t) (10) denote the communication link break probability during the time interval t. Suppose the communication link breaks are independent events, which means that the probability of one communication link break on the route between src and dst occurs in no way affects the probability of another communication link breaks on the same route. With the previous hypothesis and knowing the communication link break probability P_{lb}(t), the route failure probability can be calculated during the time interval t denoted P_{rf}(t, n) (8), where n is the number of relay clients on the route:

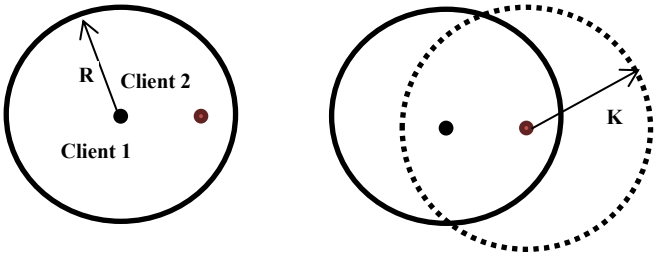
$$P_{rf}(t, n) = 1-(1 - P_{lb}(t))^n \quad (8)$$

In (9), the route failure frequency is derived from route failure probability as follows:

$$f(t, n) = f(RREQ) = \frac{P_{rf}(t,n)}{t} = \frac{1-(1-P_{lb}(t))^n}{t} \quad (9)$$

4.1.4 Link break probability $P_{lb}(t)$:

In order to calculate link breakage probability according to the distance between clients, figure.1 shows two neighbor clients (client 1 and client 2), the black cycle is the transmission range R of client 1, and client 2 may be located after an arbitrary time dt in any point within the blue cycle with a radius K.



(a) (b)

FIG. 1 – Position of relay node

Where $K = dt \times V$, and V is the maximum speed of client 2. Consequently, the link breakage probability can be calculated as follows:

$$P_{lb}(t) = \frac{\int \text{Area}_{\text{outside of circle k}}}{\int \text{Area}_{\text{circle k}}} \quad (10)$$

Where the numerator means the area outside the blue circle which does not overlap with the area of the black circle, and the denominator is the total surface of the blue circle.

4.2 Routing Overhead for SAODV

The RREQs sending frequency using our new algorithm SAODV is derived from equation (9). In addition, the probability that the received signal strength, at a client, will be greater than a predefined threshold is given by Equation (11), which requires Q-function to calculate this probability via the Gaussian process assuming that at any point in time clients move in a random way with a mean $Pr(d)$ and a fixed standard deviation σ used in formula (12) assumed to be known:

$$Q(z) = \frac{1}{\sqrt{2\pi}} \int_z^{\infty} exp(-\frac{x^2}{2})dx \quad (11)$$

$$P\,[Pr(d) > th] = Q[\frac{th-Pr(d)}{\sigma}] \quad (12)$$

As a result, routing overhead (13) of SAODV protocol according to the algorithm used for a communication from a source to a destination can be represented using (12) and (9):

$$N_{SAODV} = P(Pr>th).f\,(RREQ).d(RREQ).l \quad (13)$$

5 Experiments and results

In order to analyze the performance of our proposed algorithm SAODV in comparison with AODV, DSR, DSDV, OLSR and ZRP using NS-2, we present in the following the details of the simulation and the values of some specific parameters while varying network load and network mobility. In addition, we explain briefly the simulation methodology to generate them. Finally, the simulation procedure and results are described.

5.1 Network Load Analysis

5.1.1 Experiments Settings

Sittings of this experiment have been described in Table I.

Network Load	
Clients Number	10, 20, 30, 40,50
Network Size	600 m x 600 m
Mobility Model	Random WayPoint
Pause Time	30s
Speed	10 m/s
Simulation Duration	150s
Traffic Type	Constant Bit Rate (cbr)
Connection Rate	4 pkts/sec
Number of connections	5
Packet Size	512 bytes
Bandwidth	2Mbps
Protocols	SAODV, AODV, DSDV, OLSR, DSR, ZRP
Transmission Range	250m
MAC layer protocol	Distributed Coordination Function (DCF) of the IEEE 802.11

TAB. 1 – *Network load simulation parameters.*

5.1.2 Results

Next are the results of network load experiments:

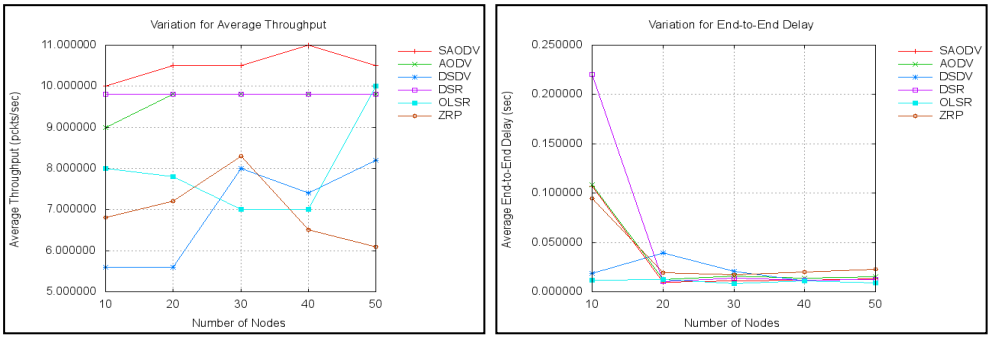


FIG. 2-3– Variation for throughput and delay using RWP model

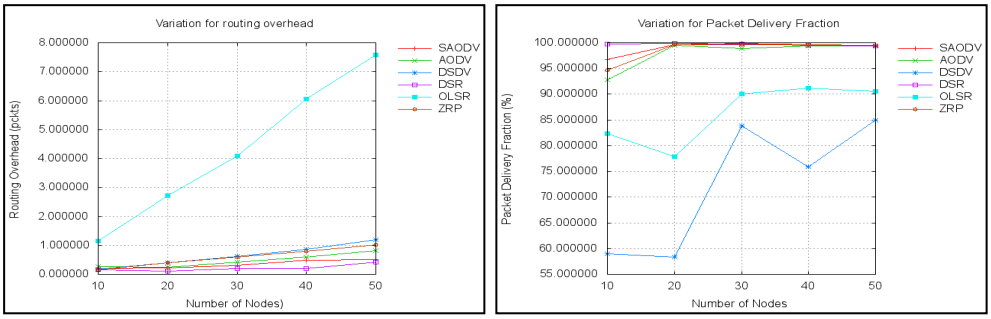


FIG. 4-5– Variation for overhead and PDR using RWP model

Fig. (2, 3, 4, and 5) show the performance results, in term of clients' density, obtained from the simulation based study on SAODV, AODV, DSDV, DSR, OLSR and ZRP using RWP model. Fig.2 depicts that SAODV, AODV, and DSR obviously outperform DSDV, OLSR and ZRP in term of average throughput, and it is due to their reactive characteristics with a great advantage of our new protocol SAODV. Moreover, based on the simulation result depicted in Fig.5, we notice that the effects of disconnected networks, due to the small density of clients, are obvious from the results of 10 clients per area for all protocols. Despite DSDV, PDR is gradually improving from above 85% in all cases where the density of clients is increased in comparison with of the network connectivity improvement. As a result, when the number of clients increases, packet delivery fraction increases, but it is still maximum in case of DSR and SAODV as compared to DSDV, AODV, OLSR and ZRP, and SAODV accordingly has a great impact in improving PDR in comparison with the other protocols while increasing clients number in the network and this is due to the fact that when a small number of clients in the network, most of the network clients are sparse, therefore our algorithm uses the standardized AODV for forwarding RREQ. On the other hand, while increasing the number of clients, the clients are in a denser network, and then our algorithm chooses only clients with high received signal strength for forwarding RREQ, which releases the bandwidth and improves the packet delivery rate.

According to Fig. (3, 4), SAODV has a relatively low delay and noticeably reduces routing overhead in comparison with the original AODV and the other protocols. This is because when the number of clients is relatively small, most of the network clients are sparse,

and our algorithm will choose to use the original AODV for forwarding RREQs in order to avoid disconnected links. However, while increasing the network density, our approach that is based on RSS greatly reduces the number of forwarding RREQs. Thus, network throughput increases, and the delay and routing overhead are reduced.

5.2 Mobility analysis

5.2.1 Experiments Settings

During mobility experiments, we tried to study the impact of mobility on the performance of our new algorithm SAODV in comparison with the other protocols; we used the following parameters as shown in (Table II).

Network Load	
Clients Number	30
Network Size	600 m x 600 m
Mobility Model	Random WayPoint
Pause Time	0s, 30s, 60s, 90s, 120s, 150s
Speed	10 m/s
Simulation Duration	150s
Traffic Type	Constant Bit Rate (cbr)
Connection Rate	4 pkts/sec
Number of connections	5
Packet Size	512 bytes
Bandwidth	2Mbps
Protocols	SAODV, AODV, DSDV, OLSR, DSR, ZRP
Transmission Range	250m
MAC layer protocol	Distributed Coordination Function (DCF) of the IEEE 802.11

TAB. 2 – *Mobility simulation parameters.*

5.2.2 Results

Here are the results of mobility experiments:

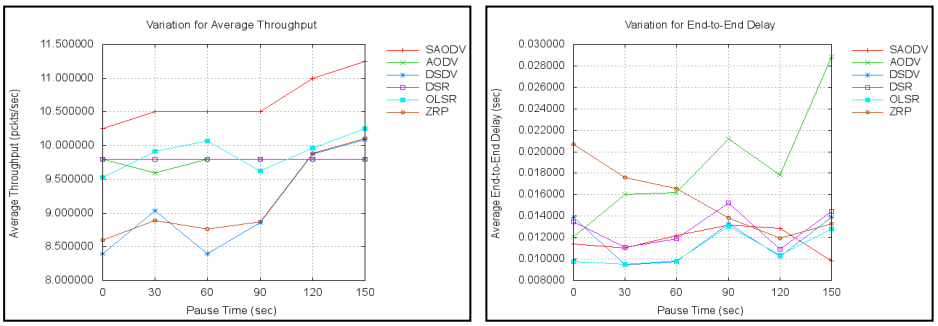


FIG.6-7– Variation for throughput and delay using RWP model

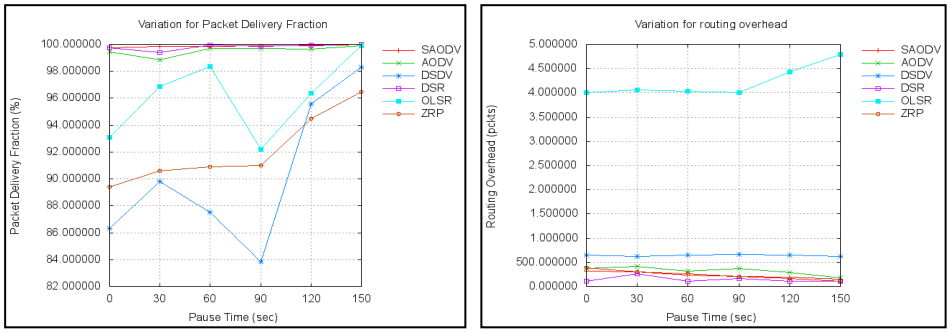


FIG.8-9– Variation for overhead and PDR using RWP model

From Fig. 6, it has been observed that the throughput value in low mobility rate exceeds the one of high mobility, but still, SAODV outperforms the original AODV and this is because the paths between source clients and required destinations frequently changed and reestablished in high mobility rates. Moreover, when clients mobility increased, more RREQ packets fail to reach their destinations. In such circumstances, more RREQ packets are generated and retransmitted, which lead to higher chance of collision due to the increase in control packets. SAODV, AODV, DSR, and OLSR outperform when mobility is very high. In general, the throughput increases when pausing time increases (low mobility) for all the simulated protocols, but it is maximum for DSR, SAODV, AODV, and OLSR. Fig. 9 depicts that the best performance is shown by SAODV as it delivers data packets at a higher rate of mobility (low pause time) in comparison to the other protocols. Moreover, at high rates of mobility, SAODV and DSR outperform all other protocols while DSDV is less efficient in comparison with the other simulated protocols since it drops downs to 80 % packet delivery ratio. Moreover, according to Fig. 7-8, while increasing clients movement (decreasing pause time); clients can quickly get the whole network client neighbors information, so this is when the average end-to-end delay is relatively smaller in case of SAODV. However, the average end-to-end delay increases for SAODV while clients are in high mobility to the relatively fast changes in the network topology, frequent and fast changes in the number of neighbors cause some errors while calculating the received signal strength value. Therefore, average delay increases for SAODV, but its performance is still better than AODV.

5.3 Confrontation with Overhead evaluation model

For a quantitative evaluation of the simulation accuracy, and in order to validate AODV and SAODV routing overhead results generated by simulation, we did its confrontation against results derived from the developed quantitative model. The number of clients is varying from 10 to 50 with an increment of 10 clients, while we set the pause time to 30 s, the network size to 600x600 square meters and the simulation duration to 150 seconds. Fig.10 depicts that the generated routing overhead, obtained by simulation, match with our quantitative model prediction for AODV routing protocol. In case of SAODV, Fig.11 shows that routing overhead results for both simulation and quantitative results exhibit the same

general trends with a small difference that is due to the fact of ignoring RERR et RREP messages in case of our developed quantitative model.

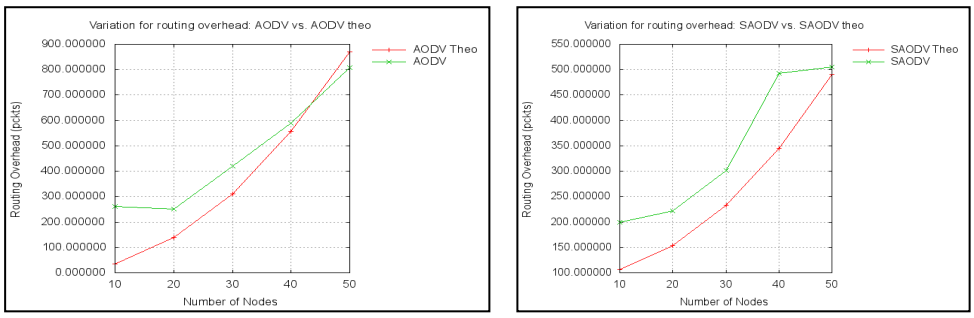


FIG.10-11– Variation for Overhead using AODV vs. AODV theo and SAODV vs. SAODV theo

References

Tan, L., Wang, N.: Future Internet: The Internet of Things. In: 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), August 20-22, pp. V5-376–V5-380 (2010)

Hoebeke, J., Moerman, I., Dhoedt, B., Demeester, P.: An overview of mobile ad hoc networks: Applications and chanllenges. Journal of Communications Networks 3, 60–66 (2004)

Asimakopoulou, E., Bessis, N., Varaganti, R., Norrington, P.: A Personalised Forest Fire Evacuation Data Grid Push Service – The FFED-GPS Approach. In: Asimakopoulou, E., Bessis, N. (eds.) Advanced ICTs for Disaster Management and Threat Detection: Collaborative and Distributed Frameworks, pp. 279–295. IGI (2010) ISBN: 978-1615209873

Gutiérrez-Reina, D., Toral, S.L., Johnson, P., Barrero, F.: An evolutionary computation approach for designing mobile ad hoc networks. Expert Systems with Applications 39, 6838–6845 (2012)

Hanzo II, L., Tafazolli, R.: A survey of QoS routing solutions for mobile ad hoc networks. IEEE Communications Tutorials & Surveys 9, 50–70 (2007)

Akkaya, K., Younis, M.: A survey on routing protocols for wireless sensor networks. Ad Hoc Networks 3, 325–349 (2005)

Tian, Y., Hou, R.: An improved AOMDV routing protocol for internet of things. In: International Conference on Computational Intelligence and Software Engineering (CiSE), pp. 1–4 (2010)

L. Atzori, A. Iera, and G. Morabito, “he internet of things: a survey,” ComputerNetworks,vol.54,no.15,pp.2787–2805,2010.

S.Y. Ni, Y.C. Tseng, Y.S. Chen, and J.P. Sheu, “The Broadcast Storm Problem in a Mobile Ad Hoc Network,” Proc. ACM/IEEE MobiCom, pp. 151-162, 1999.

Kevin Fall, Kannan Varadhan, "The NS Manual", http://www.isi.edu/nsnam/ns/ns-documentation.html, March 9, 2006.

Bai, F. & Helmy, A. (2004, June). A survey of mobility models in wireless ad hoc networks. Chapter 2, book on Wireless Ad Hoc and Sensor Networks..

Clausen, T., & Jacquet, P. (2003, October). Optimized link state routing protocol (OLSR)’’. Internet Request for Comments RFC 3626, Internet Engineering Task Force.

Johnson, D. B., Maltz, D. A., & Hu, Y.-C. (2003, April) ‘‘The dynamic source routing protocol for Mobile Ad hoc networks (DSR)’’ Internet Draft—draft-ietf-manet-dsr-09.txt, April 2003.

Perkins, C. C. E., & Bhagwat, P. (1994, October) Highly dynamic destination-sequenced distance-vector routing (DSDV) for mobile computers, ACM SIGCOMM’94, October 1994, pp. 234–244..

Mittal, S. & Kaur, P. (2009). Performance comparison of AODV, DSR and ZRP routing protocols in MANET’S international conference on advances in computing, control, and telecommunication technologies 978-0-7695-3915-7/09, 2009 IEEE.

Kavita pandey, abhishek swaroop,”A Comprehensive Performance Analysis of Proactive , Reactive and Hybrid MANETs Routing Protocols” IJCSI International Journal of Computer Science Issues, Vol. 8 Issue 6 No. 3 November 2011.

Y. Wang, C. Westphal, and J. J. Garcia-Luna-Aceves, “Using geographical coordinates to attain efficient route signaling in ad hoc networks,” in Proceedings of the Fourteenth International Symposium on a World of Wireless, Mobile and Multimedia Networks. IEEE Press, 2013.

Jiwon Park; Moh, S.; Ilyong Chung; , "A multipath AODV routing protocol in mobile ad hoc networks with SINR-based route selection," Wireless Communication Systems. 2008. ISWCS ‘08. IEEE International Symposium, vol., no., pp.682-686,2 1-24Oct. 2008doi:10.1109/ISWCS.2008.4726143.

J. Kim, Q. Zhang, and D.P. Agrawal, “Probabilistic Broadcasting Based on Coverage Area and Neighbor Confirmation in Mobile Ad Hoc Networks,” Proc. IEEE GlobeCom, 2004.

Z. Haas, J.Y. Halpern, and L. Li, “Gossip-Based Ad Hoc Routing,” Proc. IEEE INFOCOM, vol. 21, pp. 1707-1716, 2002.

Pooja Gupta, Dr Rajesh Kumar Tyagi,”A Significant Study And Comparison Of DSDV, AODV And DSR Protocols In MANET Using NS2” International Journal of Engineering Research & Technology vol. 2 Issue 3 March 2013.

Li Layuan, Li Chunlin, yaun Peiyan,”Performance evaluation and simulation of routing Protocols in ad hoc networks”, Computer Communications 30 1890-1898 2007.

Vijayalaskhmi M., Avinash Patel and Linganagouda Kulkarni, “QoS Parameter Analysis on AODV and DSDV Protocols in a Wireless Network”, International Journal of Communication Network & Security, Volume-1, Issue-1, 2011.

Y. Wang and J. J. Garcia-Luna-Aceves, “Dircast: Flooding-reduced routing in manets without destination coordinates,” in Proceedings of the 28th IEEE Conference on Military Communications, ser. MILCOM’09. Piscataway, NJ, USA: IEEE Press, 2009, pp. 2493–2498. [Online]. Available: http://dl.acm.org/citation.cfm?id=1856821.1857192

N. Chilamkurti, S. Zeadally, A. Vasilakos, V. Sharma, Cross-Layer Support for Energy Efficient Routing in Wireless Sensor Networks, Journal of Sensors, Vol.2009, Article ID 134165, 9 p.

J. Kim, Q. Zhang, and D.P. Agrawal, “Probabilistic Broadcasting Based on Coverage Area and Neighbor Confirmation in Mobile Ad Hoc Networks,” Proc. IEEE GlobeCom, 2004.

G. Varaprasad, “Power aware and signal strength based routing algorithm for mobile Ad hoc networks,” in proc. of Int. Conf. on Communication Systems and Network Technologies (CSNT), pp. 131-134, 3-5 Jun. 2011.

Ns . Giordano, R. Frank, A. Ghosh, G. Pau, and M. Gerla. Two Ray or not Two Ray this is the price to pay. In IEEE MASS 2009, pages 603{608, Macau SAR, China, October 2009.

“The Average Distance Between Points in Geometric Figures”, Steven R. Dunbar, The College Mathematics Journal, Vol 28, No. 3 (May 1977), pp 187 to 197.

C. Papamanthou, F.P. Preparata, and R. Tamassia. Algorithms for Location Estimation Based on RSSI Sampling. In Algorithmic Aspects of Wireless Sensor Networks, page 86. Springer, 2008.

Wang Tao. Introduction to wireless networking technology [M]. 1st edition.Beijing: Machinery Industry Press, 2011

D.B. Johnson and D.A. Maltz, Dynamic Source Routing in Ad Hoc Wireless Networks, in Mobile Computing, T. Imielinski and H. Korth, Eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996, pp. 153–181.

F. J. Ros, “MASIMUM - MANET Simulation and Implementation at the University of Murcia”, University of Murcia, http://masimum.dif.um.es/?Software:UM-OLSR, last accessed on Nov 2005

Résumé

Une caractéristique importante de l'Internet des objets c’est que chaque objet est capable de communiquer, et peut être un terminal. Les MANETs ont la capacité de prendre en charge un environnement de réseau mobile majeur tel que l'Internet des objets. Cet article se concentre sur la technologie de routage de réseau Ad Hoc mobile dans l'Internet des objets. Nous avons d'abord introduit un nouvel algorithme intelligent (SAODV) basé sur la puissance du signal reçu pour transmettre les paquets RREQs afin d'améliorer les performances du protocole de routage conventionnel AODV, notamment en terme de l’overhead, à fin d’adapter ce protocole à l'environnement Internet des Objets; ensuite, nous avons comparé la performance de notre nouveau protocole avec cinq protocoles bien connus de routage ad hoc. En évaluant les résultats, nous avons remarqué que l'algorithme proposé surpasse les autres protocoles dans le cas très dense où le trafic est très charge et génère moins de trafic rediffusé que

l'inondation pure utilisée par AODV traditionnel parce qu'on a moins des paquets RREQs redondants rediffusés. Pour de futures recherches, d'autres optimisations et d'autres améliorations pour réduire l'overhead de routage seront planifiés afin de cibler des types de trafic plus réalistes dans l'environnement Internet des Objets tels que VOIP, HTTP et FTP.

A review & a new approach in MANET networks for the IoT environment

Mouad Benzakour*, Abdellah Jamali* Nagib Naja**

*Networks, Mobility and Modeling Laboratory, Faculty of Science and Technology-Settat, Hassan 1st University, SETTAT 577, Morocco
mouad.benzakour.uhp@gmail.com
abdellah.jamali@uhp.ac.ma
** National Institute of Posts and Telecommunications INPT-Rabat
naja@inpt.ac.ma

Abstract. The Mobile ad hoc networks (MANET) is a challenge for all the community of new technologies. All future developments in the field of communication in wireless networks are mainly based on the performances of Mobile Ad-Hoc Networks. Most of the published papers do not present the relation with the current state of the technological movements and the evolution to know the needs and the necessary for realized them. This paper presents a state of the art, applications and characteristics of Mobile Ad-hoc network (MANET), the techniques of routing and classification of existing variants routing protocols of Mobile Ad-hoc Networks such as the reactive, proactive and hybrid protocols. It gives also a comparison between existing variants of MANET protocols AODV, DSR, DSDV and OLSR using NS2 simulator based on different parameters: routing overhead, packet delivery ratio and packet loss rate. Furthermore, we examine the impact of mobility, density and distance between nodes on the behavior of these protocols. The main problem in AODV protocol is the intensive use of packets to control traffic and network stability. We initiated a new approach for AODV protocol to reduce routing overhead and improve quality of service (QoS). The idea is to reduce RREQ broadcast packets by using the PPO "Power of the Packets Overheard" which consists of rebroadcasting the packets according to the PTL process "Power Threshold Learning". The new IoT-AODV protocol for the IoT environment with the new approach will be detailed in a future document.

1 Introduction

MANET is an autonomous system “V.Uma Devi and A.Thilaka (2017)” of interconnection of mobile nodes, devices or terminals by wireless links to communicate with each other without having recourse to fixed infrastructure and implement beforehand, each node of the MANET network plays two roles, first of a host that allows it to communicate with other nodes to exchange information and the second of a router to build the network and dynamically discover its neighbors. The fundamental characteristics on MANET networks is can be created anytime and anywhere without infrastructure or human intervention and offer the most advantages of wireless ad-hoc network such as scalability, more adaptive to mobility context, high number of connections and homogeneous system for large types of devices. Routing is the important issue to deepen a search in Mobile Ad-hoc Networks, the routing protocols are designed to adaptively cater for dynamic changes in topology while maximizing throughput and packet delivery ratio and minimizing delay, routing load and energy consumption “Fahim Maan and Nauman Mazhar”. Therefore, the development of dynamic routing protocols that can efficiently find an optimum routes without neglect the false negative result that would impact the network more than before. For that, the actually researches

focus on routing protocols for efficiency manage and more profitability. A Mobile Ad-Hoc networks (MANET) pulls the tension of several military, government and civil organizations by its flexibility and resilience to natural disasters. Some more examples of possible uses of mobile ad-hoc networks include; sensors networks (sense and send back information), vehicle nodes. In order to find the most adaptive routing protocol for MANET topologies, the behavior of routing protocols needs to be analyzed at varying node speeds and mobility, number of traffic nodes, network size, as well as node density. In this paper, we aim to present a review of routing protocols classification and comparison performance based on routing overhead, Packet delivery ratio and packet loss rate metrics of an AODV, DSR, DSDV and OLSR routing protocols on Random Waypoint mobility model "Bhavyesh Divecha and al". To do that, we use NS2 simulator to analyzing the operation and performances of MANET protocols.

The rest of the paper is organized as follows. Section 2 gives a routing protocols classification. Section 3 presents the routing protocols operation in MANET. Section 4 define the performance evaluation metrics of the routing protocols. NS2 experiments, simulation environment, results, analysis and discussions are presented in Section 5. Section 6, introduces the operation of our proposed approach and discusses the simulation results and performance evaluation.

2 MANET routing protocols classification

There are different ways to classify the routing protocols "Nasrin Hakim Mithila (2013)". In this paper, we will use the most relevant method: classification by categories.

2.1 Proactive routing protocols

Proactive protocols continuously maintain routes to all nodes in one or more routing tables. It based on periodic updates to form the routing tables in the network. This information is kept on the routing tables and is updated periodically by the routing protocol. Each row in routing table has the next hop for reaching a node/subnet and the cost of this route. When the network topology becomes different, it becomes necessary that routing tables are update as per the changes that occurs "Navneet Kaur, Amandeep Verma (2017)". Examples of such protocols: DSDV, OLSR.

2.2 Reactive routing protocols

Reactive routing protocol is also called on demand routing protocol. Reactive protocols are based on demand for data transmission. These protocols set up routes when demanded. They do not begin route discovery by themselves, until they are requested. Routes are only discovered by broadcasting route query or request messages whenever they are actually needed to forward packets from source to destination. They can decrease routing overhead when the traffic is low and do not need to find and maintain routes when there is no traffic and no need to update route information regularly. The examples of reactive protocols are: AODV, DSR.

2.3 Hybrid routing protocols

This is the combination of the proactive and reactive protocols which work well for the networks with a small number of nodes. While the hybrid reactive/proactive protocols are used to achieve the high performance as the number of nodes increases. It is a key of idea to use a reactive routing at the global network level while the employing a proactive in a node's local neighborhood "Behra Rajesh Umashankar, Rakhi Kumari Purnima (2014)".

3 Routing protocols operation in MANET

To route the packet from one end to another end is a crucial task. The main goal of any routing protocol is to establish an optimal and efficient path between mobile nodes. In order to carry out their tasks, the routing protocols are based on topological information and position information to locate and achieve the destination.

3.1 Dynamic Source Routing (DSR)

Dynamic Source Routing (DSR) is a reactive routing protocol for multi-hop wireless mesh networks and is based on a method known as source routing with link state algorithm. It forms a route on-demand when a transmitting computer requests one and store it in route cache. Each intermediate node that broadcasts a route request (RREQ) packet adds its own address identifier to a list carried in the packet. The destination node generates a route reply (RREP) message that includes the list of addresses received in the route request and transmits it back along this path to the source "S. A. Ade & P.A.Tijare (2010)". Route maintenance in DSR is accomplished through the confirmations that nodes generate when they can verify that the next node successfully received a packet. When a finite number of retransmissions fail, the node generates a route error (RERR) message that specifies the problematic link transmitting it to the source node.

3.2 Ad-hoc On-demand Distance Vector

The Ad hoc On Demand Distance Vector (AODV) routing algorithm is a reactive routing protocol. AODV is capable of both unicast and multicast routing. It is an on demand algorithm, meaning that it builds routes between nodes only as desired by source nodes in order to minimize the number of broadcasts. AODV uses sequence numbers to ensure the freshness of routes. The AODV protocol uses route request (RREQ) messages broadcasted through the network in order to discover the paths required by a source node "Route Discovery Process". A node receiving the RREQ may send a route reply (RREP) if it is either the destination or if it has a route to the destination with corresponding sequence number greater than or equal to that contained in the RREQ. For route maintenance process, the route changes can be detected by different ways: failure of periodic HELLO packets, failure of disconnected indication from the link layer level, failure of transmission of a packet to the next hop.

3.3 Destination-Sequenced Distance Vector (DSDV)

Destination-Sequenced Distance-Vector routing protocol (DSDV) is a proactive table-driven routing protocol for ad hoc mobile networks based on the distributed Bellman-Ford algorithm. The improvement made to the Bellman-Ford algorithm includes freedom from loops in routing tables by using sequence numbers. The route labeled with the highest sequence number is always used and replace the existing one on the routing table. To minimize the traffic generated, there are two types of packets in the system. One is known as "full dump", which is a packet that carries all the information about a change. However, at the time of occasional movement, another type of packet called "incremental" will be used, which will carry just the changes, thereby, increasing the overall efficiency of the system **"**S. A. Ade & P.A.Tijare (2010)**"**.

3.4 Hybrid routing protocols

Optimized Link State Routing (OLSR) is a point-to-point proactive routing protocol where the routes are always immediately available when needed. OLSR may optimize the reactivity to topological changes by reducing the maximum time interval for periodic control message transmission. Where the most communication is concentrated between a large numbers of nodes. OLSR reduce the control overhead forcing the Multipoint Relay (MPR) to propagate the updates of the link state **"**S. A. Ade & P.A.Tijare (2010)**"**. OLSR employs three mechanisms for routing: (1) periodic HELLO messages for neighbor sensing, (2) control packet flooding using Multi-Point Relay (MPR), and (3) path selection using shortest path first algorithm.

4 Performance metrics

4.1 Routing overhead (RO)

Routing overhead is the number of routing packets used because of the frequent links breakages which lead to frequent path failures and route discoveries.

4.2 Packet delivery ratio (PDR)

Packet Delivery Fraction describes the ratio between the total numbers of packets or data bits originated by the application layer sources and the number of packets received by the sinks at the final destination "Nabil Nissar and al (2015)". The higher the delivery ratio, better is the performance of the routing protocol. PDR is determined as:

4.3 Packet loss rate (PLR)

Packet loss is the failure of one or more transmitted packets to reach their destination. It can occur anywhere along the path between source and the destination. The packet loss rate is calculated by dividing loss packet to total number of packets transmitted **"**S. A. Ade & P.A.Tijare (2010)**"**.

5 Simulation and result analysis

5.1 Simulation parameters

For simulation, we will present different scenarios to test the performance of the routing protocols using NS2 simulator, first we will check the evaluation parameters with a high mobility of the nodes in the ad-hoc network and with different speed values, then we will increase the number of nodes in the network and the distance to be able to analyze the performances in different probable situations of real case in a Mobile Ad-hoc Network. The comparison will focus on the performance of the four routing protocols defined previously. The following table display the parameters used to simulate the routing protocols.

Network parameters		Node parameters	
Network size	300, 500, 800, 1000 m²	Channel type	WirelessChannel
Nodes number	10, 20, 30, 40, 50, 70	Radio-propagation model	TwoRayGround
Pause time	5, 20, 40, 80, 100, 150 s	Network interface type	WirelessPhy
Speed	5, 10, 20, 40, 50 m/s	MAC type	802_11
Simulation duration	200 s	Interface queue type	DropTail/PriQueue, CMUPriQueue
Mobility models	Random WayPoint (RWP)	link layer type	LL
Traffic parameters			
Traffic type	Constant bit rate (CBR) over UDP	Packet size	512 bytes
Connection rate	5 pkts/s	Routing protocols	DSR, AODV, DSDV, OLSR
Max number of connections	8, 15, 20, 30, 40, 50	Evaluation parameters	Routing Overhead, PDR, PLR
Max packet in queue	50		

TAB. 1 – *Simulation parameters*

5.2 Result analysis

Routing overhead. The routing load resulted from the considered routing protocols have been presented in Figs. 1, 2, 3 and 4. The simulation results that all the protocols have a similar behavior in the low density levels. On the contrary, when node density levels increase in the simulation zone, the DSR and DSDV protocols take over and maintain an acceptable rate of routing packets, while the AODV and OLSR protocols increase the routing overhead Fig. 1. In Fig. 2, when the nodes increase of speed, the three protocols AODV, DSR, DSDV keep their routing overhead stable, however, there is an increase in the number of control packet rated OLSR.

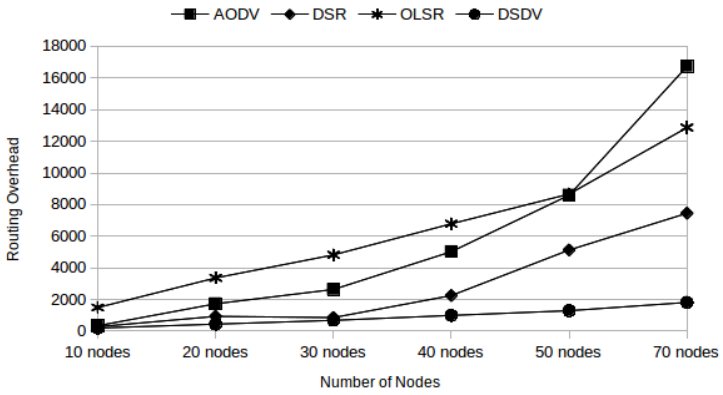


FIG. 1 – *Routing overhead by varying number of nodes*

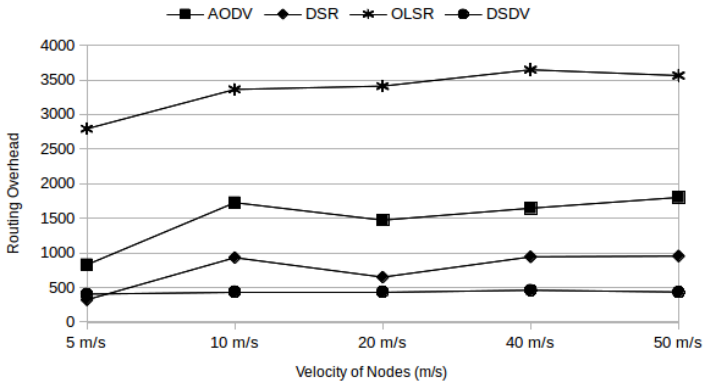


FIG. 2 – *Routing overhead by varying velocity of nodes*

In the case of changing the pause time and the size of the network, the protocols use a close number of control packets except for the OLSR protocol which keeps the same behavior and surpasses AODV, DSR and DSDV Figs. 3, 4.

This difference in performance is due to the routing mechanisms used by each protocol, the AODV and OLSR protocols used different types of packet to reach the destination and maintain the network stability, which explains the rate of control overhead in the different scenarios. While for DSR and DSDV protocols, the situation changes, DSR is known as a beaconless protocol in which no HELLO messages are exchanged between nodes for their notification of their neighbors in the network “Navneet Kaur, Amandeep Verma (2017)”. On the other hand, DSR almost always has a lower routing overhead than the other protocols, and this can be attributed to the caching strategy used by DSR. DSDV uses a full periodic updates and a partial updates when the network topology changes, and with the regular updates of sequence numbers to avoiding the formation of loops.

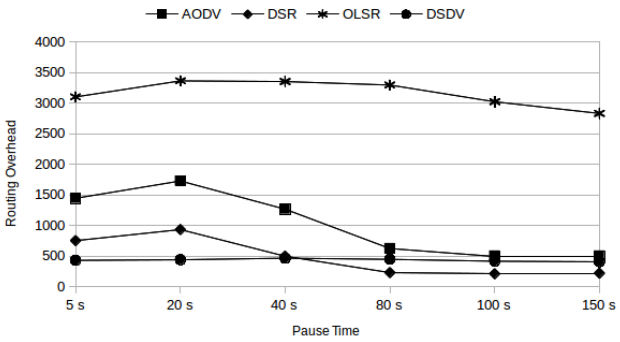


FIG. 3 – *Routing overhead by varying pause time*

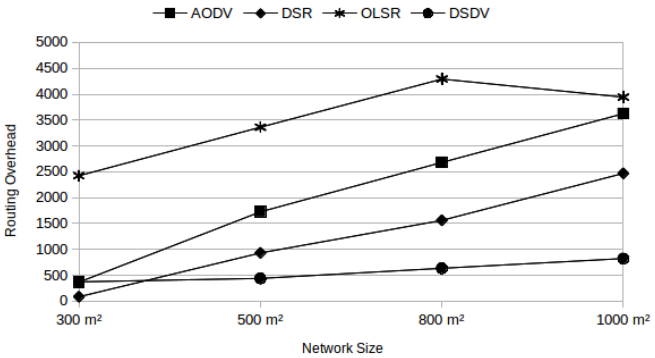


FIG. 4 – *Routing overhead by varying network size*

Packet delivery ratio. The Figs 5, 6, 7, 8 present different scenarios, the three protocols: DSR, AODV and OLSR, effectively deliver between 80 and 100 % of data packets Figs. 5, 6, 7. DSDV represents the lowest rate of packet delivery ratio. The behavior of protocols in different scenarios does not change with changing the number of nodes, velocity and pause time.

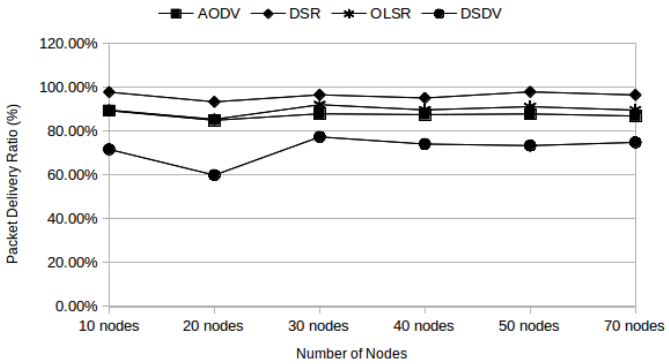


FIG. 5 – *Packet delivery ratio by varying number of nodes*

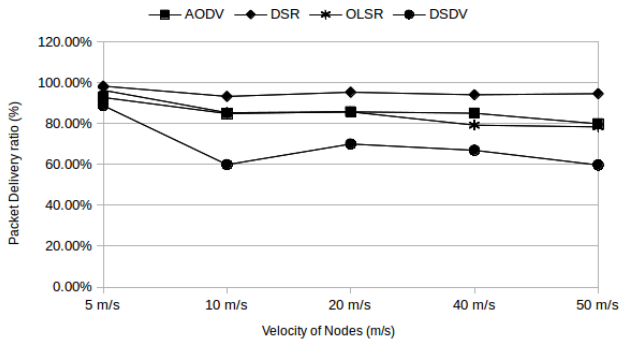


FIG. 6 – *Packet delivery ratio by varying velocity of nodes*

As an exception, in the scenario of varying the size of the network, it can be seen that the packet delivery rate low for AODV and DSR, for OLSR and DSDV the decrease is greater Fig. 8. As a result, when the number, velocity, pause time of nodes increases, packet delivery fraction increases, but it is still maximum in case of DSR and AODV as compared to DSDV and OLSR because DSR protocol always looks for the most fresh and reliable route when needed and does not look for it from the routing table. The performance of DSDV is degrading due to increase in the number and velocity of nodes, the load of routing tables exchange becomes high and this performance will decrease as the number of nodes increases.

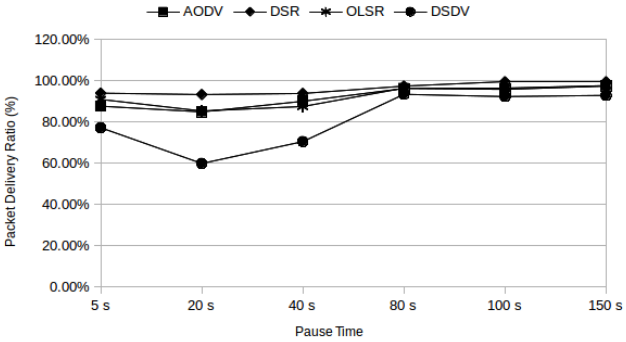


FIG. 7 – *Packet delivery ratio by varying pause time*

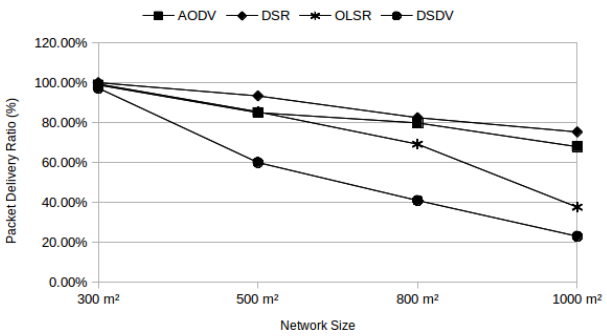


FIG. 8 – *Packet delivery ratio by varying network size*

Packet loss rate. In all simulation scenarios, the results indicate that AODV loses more packets than DSR, DSDV and OLSR. Mobility, node density and velocity are the dominants cause for AODV packet loss rate, they are responsible for nearly all the packets loss Figs. 9, 10, 11. In Fig. 12, in the case of variation of the network size, we note that the packet loss rate has increased in DSDV compared to other protocols. On the other hand, for AODV, the packet loss rate is close to the two DSR and OLSR protocols in large areas.

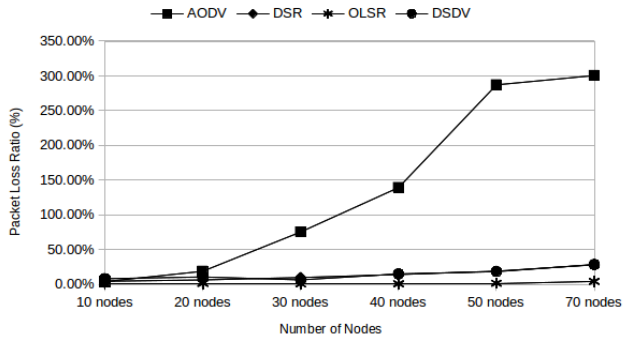


FIG. 9 – *Packet loss rate by varying number of nodes*

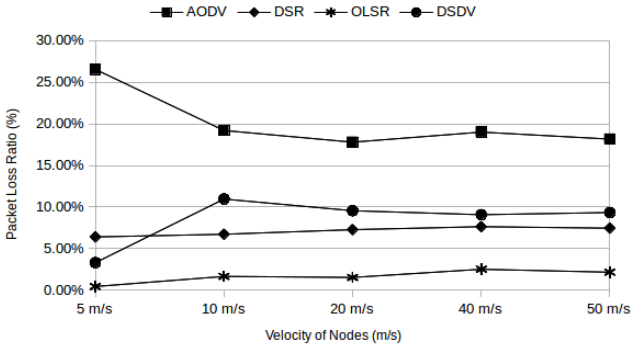


FIG. 10 – *Packet loss rate by varying velocity of nodes*

The high mobility impact significantly the performance of AODV protocol Fig. 11, with the low pause time, the AODV protocols generate more loss packets as and when the pause time rise, the packet loss rate decrease. For other protocols, the DSR and OLSR perform well in high and medium mobility followed by DSDV protocol.

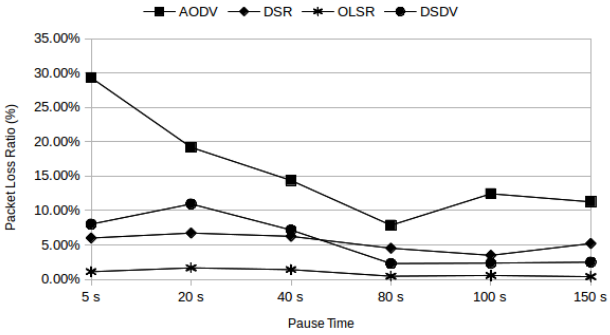


FIG. 11 – *Packet loss rate by varying pause time*

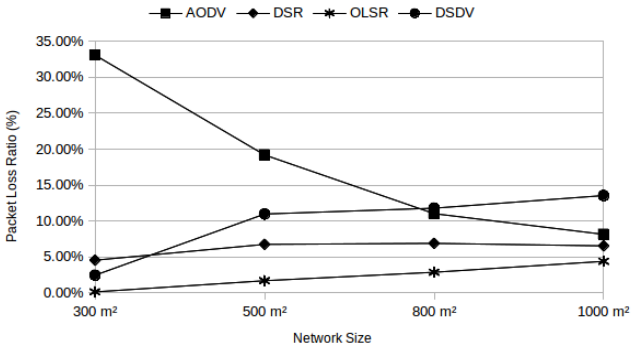


FIG. 12 – *Packet loss rate by varying network size*

The highly rate of packet loss in AODV is related to huge number of control packets exchange between nodes to make a route to the destination after breakage or in the first communication is establish. The main issue in AODV is using the RREQ broadcast methods to find the destination and HELLO messages to refresh neighbor's node list. In case of other protocols, the DSDV and OLSR use the routing tables to maintain the routes to available destination and updates them periodically.

6 Proposed approach

To mitigate the impact of undesirable traffic on the network performances, this paper propose a new approach to reduce the mobile ad hoc routing overhead in based on the received power of the packets from the destination. To do that, we will first set a threshold power according to predefined parameters, after, we will evaluate the power of the transit packets. If the PPO "Power of the Packets Overheard" is lower than the threshold power then the packet request (RREQ) is broadcasting to the nodes. On the other hand, the request packet is not sent to the destination with a higher power. This new RREQ forwarding mechanism will improve AODV protocol performance, especially in terms of routing overhead. The process is resume by Fig. 13:

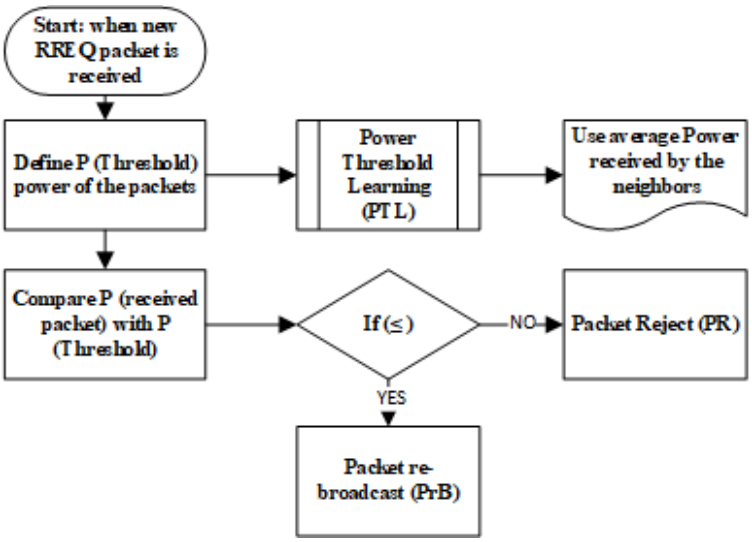


FIG. 13 – *Schema of the proposed approach*

To define PTL "Power Threshold Learning" process, the proposed formula is:

$$P_{threshold} = P_{avg} = \frac{\sum_{i=1}^{NB} P_i}{NB}$$

7 Conclusion

To have a high performance routing protocol in the MANETs network, it must be able to operate with the least resources and work in the strong conditions of mobility, density and traffic load. Note that no single MANET routing protocol is best for each situation, which means that network analysis and environmental requirements are essential for selecting an effective protocol. Our comparative study results that the AODV protocol uses more routing packets to manage and maintain the stability and homogeneity of the network than DSR, DSDV and OLSR. Our purpose is to benefit from strong points and to improve weak points of the protocol. The proposed approach focuses on reducing the spread of RREQ packets without impacting protocol performance in a high mobility context. In future work, we will implement the approach based on the algorithm of the AODV protocol. Then study the results obtained and make a comparative study to demonstrate the validity of our proposal.

References

Nabil Nissar, Najib Naja, Abdellah Jamali (2015). *A review and a new approach to reduce routing overhead in MANETs*, 1119–1139.

Navneet Kaur, Amandeep Verma (2017). *A State-of-The-Art of Routing Protocols for Mobile Ad-Hoc Networks*" Kaur et al., International Journals of Advanced Research in Computer Science and Software Engineering june 2017, 55-58.

V.Uma Devi and A.Thilaka (2017). *A State-of-the-Art Survey on MANET Protocols*", International Journal of Pure and Applied Mathematics, Volume 116, No. 10 2017, pages 471–480.

Behra Rajesh Umashankar, Rakhi Kumari Purnima (2014). *A Comparative Study of Topology and Position Based Routing Protocols in Mobile Ad Hoc Networks*. (IJARCST 2014) Vol. 2, Issue 2, Ver. 1.

Fahim Maan, Nauman Mazhar. *MANET Routing Protocols vs Mobility Models: A Performance Evaluation*".

Bhavyesh Divecha, Ajith Abraham, Crina Grosan and Sugata Sanyal. *Impact of Node Mobility on MANET Routing Protocols Models*. Mumbai University, India; School of Technology and Computer Science, Tata Institute of Fundamental Research, India.

S. A. Ade & P.A.Tijare (2010). *Performance Comparison of AODV, DSDV, OLSR and DSR Routing Protocols in Mobile Ad Hoc Networks*. 545-548.

Nasrin Hakim Mithila (2013). *Performance analysis of DSDV, AODV and DSR in Wireless Sensor Network*. (IJARCSEE) Volume 2, Issue 4, April 2013.

Kaoutar Ourouss, Najib Naja, Abdellah Jamali (2016). *Efficiency analysis of MANETs routing based on a new double metric with mobility and density models*. AICCSA2016: 1-8.

Résumé

Les réseaux mobiles ad hoc (MANET) constituent un défi pour toute la communauté des nouvelles technologies. Tous les développements futurs dans le domaine de la communication dans les réseaux sans fil sont principalement basés sur les performances des réseaux Ad-Hoc mobiles. La plupart des articles publiés ne présentent pas la relation avec l'état actuel des mouvements technologiques et l'évolution pour connaître les besoins et le nécessaire pour les réaliser. Cet article présente un état de l'art, les applications et les caractéristiques du réseau mobile ad hoc (MANET), les techniques de routage et de classification des variantes existantes des protocoles de routage des réseaux mobiles ad hoc tels que les protocoles réactifs, proactifs et hybrides. Il donne également une comparaison entre les variantes existantes des protocoles MANET: AODV, DSR, DSDV et OLSR à l'aide du simulateur NS2 et se basant sur les paramètres suivants: les frais généraux de routage, le taux de délivrance des paquets et le taux de perte de paquets. Ensuite, nous examinerons l'impact de la mobilité, la densité et la distance entre les nœuds sur le comportement de ces protocoles. Le principal problème du protocole AODV est l'utilisation intensive des paquets pour contrôler le trafic et la stabilité du réseau. Nous avons lancé une nouvelle approche pour le protocole AODV afin de réduire les frais généraux de routage et ainsi améliorer la qualité de services (QoS). L'idée est de réduire les paquets de diffusion RREQ en utilisant le PPO "Power of the Packets Overheard" qui consiste à rediffuser les paquets selon le processus PTL "Power Threshold Learning". Le nouveau protocole IoT-AODV destiné à l'environnement IoT avec la nouvelle approche sera détaillé dans un futur document.

A Taxonomy of challenges in Internet of Things (IoT)

Aboubaker Saddik LAIREDJ *, Khelifa Benahmed *
Fateh Bounaama **
*lairedjboubakerdz@gmail.com
Benahmed_khelifa@yahoo.fr
fbounaama2002@yahoo.fr

Abstract. The world is rapidly getting connected. Internet of Things is the next revolution in the field of information technology, it is a paradigm that has gained more popularity. At a conceptual level; IoT refers to the interconnecti-vity among our everyday devices such as personal computers, laptops, tablets, smartphones. All devices that exchange information, take decisions, invoke actions and provide service to the end users, thus there are intelligent communication between each other. In fact the essential purpose of Internet of Things is to make it easy for people to be connected to each other, anywhere and everywhere in fast paced , with whatever objects and in multiple paths and networks and any services. In this paper , we spot the light primarily on the challenges in the IoT.
Keywords—Internet of Things (IoT); Challenges; Security

1 Introduction

Internet of things (IoT) is an integrated part of the future internet and could be defined as a dynamic global network infrastructure with self-configuring capably. The IoT is simply the network of interconnected things-devices, things-things. The electronic embedded devices have sensors and software capable of connectivity which enable these objects to connect and exchange data. The IoT is a great connectivity to a heterogeneous set of objects using wired or wireless communications. IoT is the key to all areas of knowledge building, from business, government and management of educational technologies. The IoT will quickly become the main vector of growth and employment: agriculture, automotive, health, transportation, not to mention the smart cities or homes automation. However; the problem for now is that; we are swimming in full unknown. The main issues of the IoT were discussed in this paper, notably the need to help them know the technologies, standards, actors, and to equip themselves with reliable and accessible solutions (hardware, network, etc.). The other stake concerns obviously the data. With so many connected objects, these will still multiply with all the risks that this implies (theft, loss ...). Internet of Things is defined as a network of physical identifiable intelligent objects that can detect and communicate with other intelligent objects using the Internet, illustrated in Fig 1, linking billions of electronic devices globally and exchanging information with these devices. Since devices such as mobile phones, PCs, laptops, and tablets have sensors and actuators, they can make intelligent decisions and transmit useful information to the required entities. The ultimate goal of the IoT is to create an intelligent planet where physical things are converted into intelligent objects by installing appropriate hardware and applications within them connecting to the Internet (Internet Socie-

ty.2017).These can communicate with other intelligent objects in a transparent manner and provide services to end users (SparkFun Electronics.2003). The communication between the intelligent objects is done at the protocol of one machine to another machine (M2M). Since the Internet of things has objects that communicate with each other, it's architecture is totally different from the architecture of the networks. The challenges that Internet of Things face are the scalability, the interoperability, the reliability and the service quality. News about the privacy of Internet connected devices, surveillance concerns and fears of confidentiality has already attracted public attention. Technical difficulties and new political, legal and deve-lopment challeng-es are emerging.

The remainder of this paper is organized as follows, sections 2 presents an overview of the architecture of the internet of things with two modes: virtual one and physical one, in section 3, we present the design challenges in IoT, we organized it in three areas challenges, the reliability, building, and the security. In section 4 we concluded this paper.

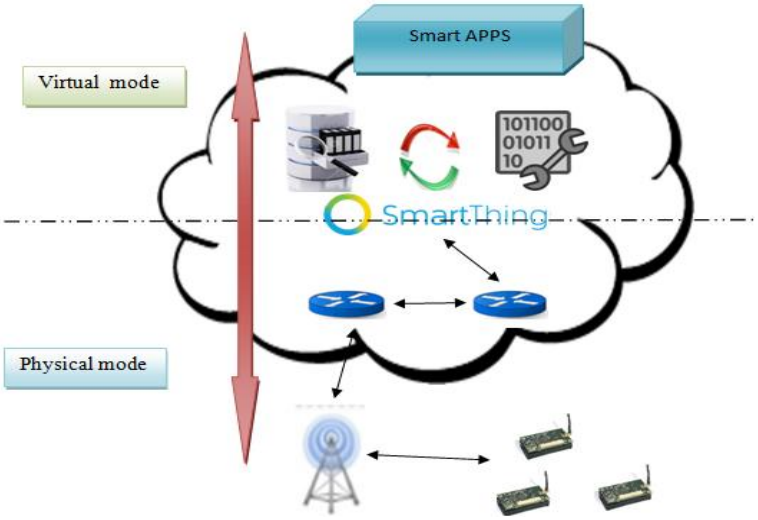


FIG. 1 – *Internet of Things*

2 IoT Architecture

The IoT architecture is composed of three layers; Figure 2 shows the application data, the network, and the PHY / MAC layer or detection. The first application layer uses computing technologies and intelligent applications to receive and retrieve data that will be analyzed and processed to provide services a space for users and IoT. The second is the network lay-ers, which are responsible for the operations of the routing and addressing network and also the connection between the application layer and the PHY/MAC layer (Mendez et all.2017). This last layer must have the secure connection of technologies. The last PHY/MAC layer is known as the physical layer, it contains the physical devices and materials. Actually it is the detection layer because of the sensors integration into the physical aspects. In this layer, the embedded sensors are responsible for collecting the information devices and sending them to the network layer.

During deployment and realize IoT globally, millions of users simultaneously communicate, it's a whole other technology, a massive technology. The considerations include heterogeneous networks. Table 1 summarizes the three layers above, in which we find that the most well known problems in the IoT network layer are presented in the second layer "communication, network layer" where there are interconnection and network sensors (Electronic-sign) the networking and interconnecting are heterogeneous devices and applications to a large scale with an exchange of data efficiently, and the security and the presence of detection are the key problem of the IoT, and they are detailed in the next section.

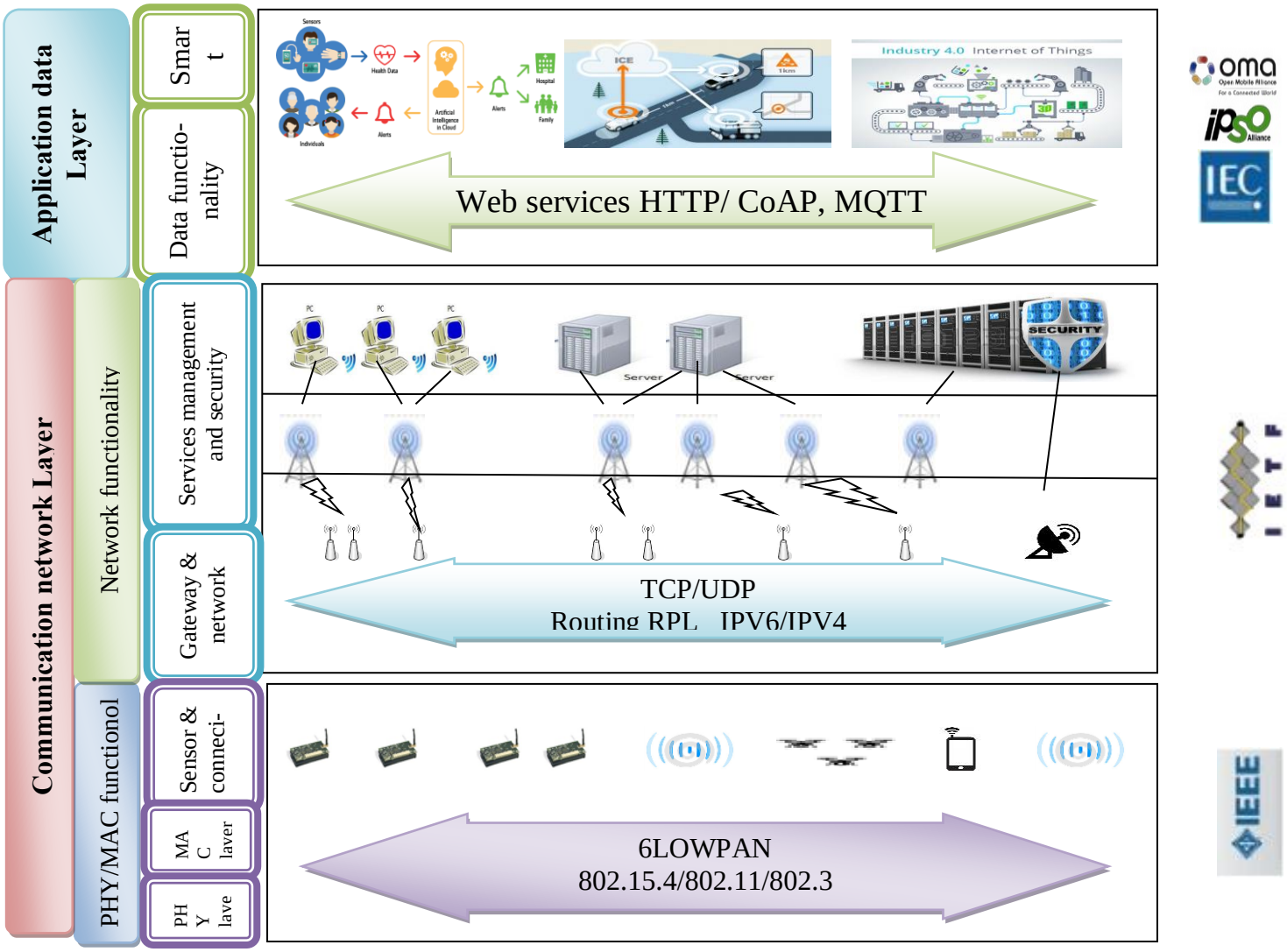


FIG 2– *IoT Architecture*

Layers	Services	Property	Parameters
Application data Layer	Smart apps	Application services /internet, the web	Service ID, cryptography, port
	Data functionality	Management data, Authorization	
Communication Network Layer	Network functionality	Security	Data routing, transmission control ,protection
		Gateway and Network	
Physical layer	PHY/MAC functional	Sensors and connectivity	Signal conversion, Laison type
		Physical Space	

TAB. 1 –*Characteristics of layer link in IoT*

3 Challenging issues

With the rapid growth of Internet technology (the future Internet) and the vast development of networking, it is reasonable to expect a new generation of Internet (Internet of IoT), it has both advantages and promises (A. Aijaz , A. Aghvami. 2015), as well as for IoT we posed a decisive questions. Since there are many challenges in construction, safety, and reliability. The IoT requires a great effort to intervene. The most important problems and challenges are listed below in detailed manner.

3.1 The Security Framework

Today with the birth of IoT, Security in touch, people communicate with each other, exchange data with services, nevertheless; the IOT observed a serious problem .which is the provision of security could be difficult because the automation of large devices has been increased, which has created a new big security problems. Indeed, security and privacy are considered as the IoT low link, and to ensure security, one must first ensure the reliability, resiliency, and stability of Internet applications and services which are essential to promote the trust to use the Internet. (Internet Society.2017) as Internet users, we must have a great confidence that the Internet security in IoT is basically tied to the ability of users to trust their environment and the question posed here is what is the security of IoT services?. Security can be classified by two concepts:

3.1.1 The Security structure

The Internet of Things provides different expensive security mechanisms for different layers (Hui Suo et al. 2012).There are also (Subho, Tripathy. 2015):

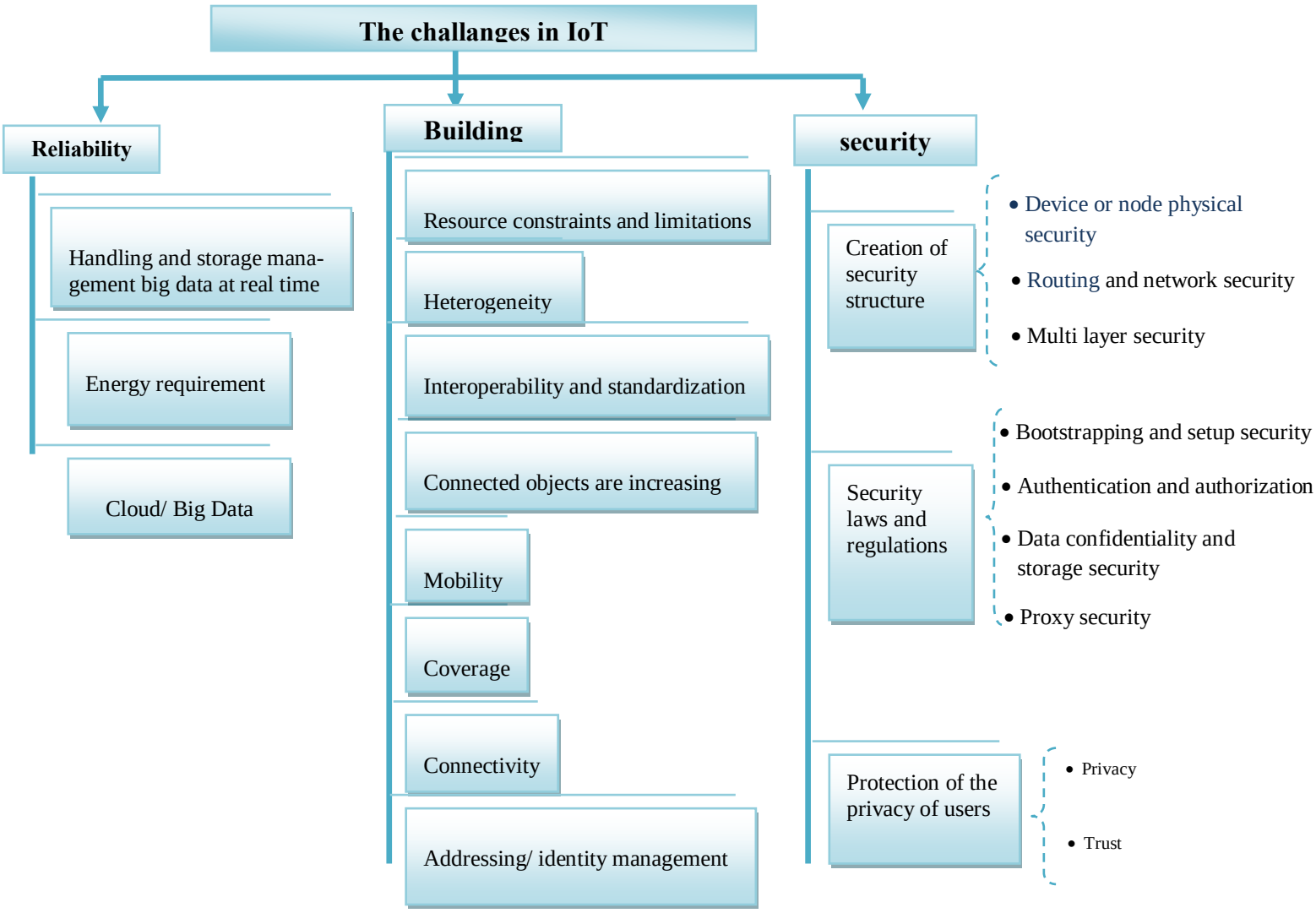


FIG 3– *Taxonomy of challenges in IoT.*

- **Physical security of the device or node**: Element environment the environment element in the IoT is open, the IoT nodes are very easily accessible, they have the possibility of being falsified or cloned;
- **Routing and network security**: Intrusion is a critical point in the IoT network routing, the intrusion of the detection systems are essential to detect when a network has been attacked and compromised by an attacker, the corresponding algorithms should be in a place to avoid further damage to the general network. In addition, affected nodes should be isolated and investigated for security breach;

- **Multi-layer security**: IoT is a dynamic infrastructure design in the world, combining several concepts, the structure of the IoT is divided into three layers: perception layer, network layer, and application layer (Kumar, Yogesh .2015). Low powered devices latching with the gateway might run on 6LoWPAN and then in their network they might need to adapt 802.15.4 link layer security, so the protocol stack should be flexible and accustomed with multiple security solutions while still maintaining the normal securi-ty requirements;

3.1.2 The Security laws and regulations

The different security mechanisms that have been proposed for different layers rely solely on assumptions, not by the experience of the people in the real world of the IoT (Daniel Elizalde.2016). There are also (Subho, Tripathy. 2015):

- **Bootstrapping and setup security**: While in the process of latching into a network in the bootstrapping phase, the information of nodes should be kept a secret. The typical cryptographic information including the preshared keys and other private keys, after the bootstrapping phase is done safely.though the device may participate in establishing shared keys;
- **Authentication/Authorization of Access control and Accounting**: Any node communicates with a server, it should be properly authenticate itself using certificates as well as the destination node to prevent open access to node data. And also justified authorization and access control rules should be implemented on these nodes to limit them from super user access, and detailed accounting information at the end of the transaction that must be registered;

  Data confidentiality and storage security: Data while being transmitted over the network should be made secure by using the light-weight crypto algorithms tailored to these resources constrained networks. In the IoT every node can't store a huge data, they have a less memory (Bagci et all. 2013). The greatest challenges in research to ensure the confidentiality of data in IoTare (Miorandi et all.2012) :
  - Build a reliable mechanism to manage and control access to data streams that are managed by the IoT devices,
  - Definition of communication methods by appropriate queries to retrieve information from a data stream,
  - Definition of suitable smart objects identity management system,
- **Proxy Security**: The proxies become a major target because data are transformed from one form to another. This bridging infrastructure is thus expected to provide decent transport level of security by implementing something like a secure;

3.1.3 Protection of the privacy of users

With a large number of sensors connected to the Internet and integrated into everyday objects that reveal our habits, our state of health , our geographic location and other types of information, we are deprived of the user qualifies user's right to be sensitive,in which there it will need to be robust mechanisms that can ensure the confidentiality of the data and increase confidence (Miorandi et all.2012) :

- **Privacy**: defines the rules under which data referring that refers to individual users may be accessed. The fundamental reasons in the IoT and is making privacy the technologies used;
  - Definition of a general model for privacy in IoT,
  - To carry out innovative techniques capable of supporting high heterogeneity under the IoT scenario,
  - Define solutions to make a balance between the need for localization of applications, and tracking other applications,
- **Trust**: Trust in the IoT it is used in a large scale of different performance. If it discusses about the IoT; in which are the objects interconnects between them each other without the interaction of humans ,we will realize that it confidence is the most critical notion on which there is no general information in the field of the IoT and in information science computerized the science of the computerized information. Although it possesses a very recognized importance. Different definitions are possible depending on the perspective adopted adopted prespective;
  - The introduction of an easy trusted language supports the IoT interoperability requirements,
  - The definition of a mechanism based on an access control of data flows,
  - Build a system capable of managing the identity of objects, within a framework of gaining total and general trust and more flexible,
  - The use of Internet objects in Distributed Denial of Service (DDoS) attacks is a major issue in the security of them, privacy and integrity (AlSaudi et all.2018),

3.2 Challenges of Building in IoT

In the large field of creation or realization of IoT which is based on many interconnected objects, possess a wide variety of uses in various environments space, and meet various requirements with an absence of a single platform or unified standard can avoid the challenges. There are many challenges in the structures that have prevented the IoT, all challenges are listed below in detail.

3.2.1 Resource constraints and limitations

A vast majority of IoT devices have limited resources. The constraints could be in terms of available computational resources, on board memory (RAM and ROM), network bandwidth, energy availability, etc. Also, as these nodes have very little computation and storage capabilities, they are very limited in computing resources, memory storage, and energy (Subho, Tripathy. 2015). (Sahraoui.2016) .

3.2.2 Heterogeneity

Devices of various types with different capacities and belonging to networks of different natures, different communication and technologies, with all these forms of material and technological heterogeneities, it would be essential to put in place a well informed mechanisms (Sahraoui.2016) .

3.2.3 Interoperability and standardization

This is among the biggest challenges of achieving the Internet of objects. Actually Interoperability is the coexistence of the disjoint mechanisms of systems and the possibility of making them cooperate and interact flexibly. A recent trend is towards the standardization and unification of operational systems and protocols in the IoT and to present them in open source. This is to facilitate collaboration between connected objects, as well as coupling with external entities on the Internet (Sahraoui.2016) . (Wang et all .2017) .

3.2.4 Connected objects are increasing

It is expected that the number of intelligent objects that will populate the internet of the future will cross millions, and even billions. With this, the adoption of new mechanisms that effectively support continuous scalability in the number of connected objects is strongly recommended (Sahraoui.2016) . (Kumar, Yogesh .2015) .

3.2.5 Mobility

As the mobility of devices in the field area increases, will most often be mobile, there could be a frequent disruption in the physical connectivity between the devices and their local bridges. Peripheral devices that move to a new location have to maintain the service's continuity through secure handoff mechanisms. As a result, flexible mobility management solutions must put in a place to enable such objects to perform their missions efficiently regardless of the frequency and the speed of mobility (Sahraoui.2016) .

3.2.6 Coverage

We can say that the coverage on the IoT is a strong point to understand, in addition to the risk of hiding it in the IoT is a very hard and complicated because the number increases peripherals by communicating with each other and the data storage servers have appeared. Indeed, this interoperability is the critical function and promoter of the IoT products (Lon Berk.2014).

3.2.7 Connectivity

IoT applications require two forms of connectivity, and either of these has its own set of challenges. On both the physical level and the service connectivity. The IoT connectivity should be a forethought before deployment, not an afterthought. Having a scalable IoT network to connect devices and servers is crucial for a large-scale IoT applications. These are the types of Internet of Things challenges we've presented (SparkFun Electronics.2003).

3.2.8 Addressing and identity management

In IoT, billions of devices will be interconnected and the communication among them takes place through the network. As a result naming and management system is very difficult due to the need of unique identification of smart and dynamic objects. Coordinator nodes allocate local addresses to peer devices and these addresses do not follow a common standard. Moreover the addressing scheme of a field network it difficult to makes and isolate a rogue node (Subho, Tripathy. 2015).

3.3 Challenges of Reliability in IoT

Reliability is the most important quality of any computer system . Thus to create a reliable technology in IoT faces a lot of challenges , in reality there are different concepts such as data management, storage, quality of service. which are essential to the reliability of the IoT, but the standard technical tests which are defined a sufficient assurance of the latter, as the IoT tests and specific exams become increasingly widespread (*Anders P. Mynster.2017)*.

3.3.1 Handling and storage management big data in real time

One of the most important and difficult challenges of Internet of Things is handling big data in real time. Every day between the devices lots of data are being generated and there are a lots of information to be transferred from one place to another. Internet of Things needs a lots of storage capacity to handle them. So it is a must to check whether the exact data is being transferred or not. Data management has a very important role in the IoT (Hui Suo et al. 2012). (Kumar, Yogesh .2015).

3.3.2 Energy requirement

To achieve Internet of Things in reality there will is a need for the enormous amount of energy and this energy requirement can not be fulfilled by using the conventional power sources like a battery. Hence there is a requirement for going towards the non-conventional power sources to get the required power for working on devices the Internet of Things. This non-conventional power source is called the greening of Internet of Things. Moreover, these non-conventional power source should be adopted for the realization of Internet of Things in the near future (Bagci et all. 2013).

3.3.3 Quality of service

Depending on whether the application is critical or not, inter-object communications connected in and between IoT and ordinary Internet hosts may or may not require a minimum quality of service in terms of delays, flows, reliability (Sahraoui.2016).

4 Conclusion

IoT is a future technology for the next generation, it is a dynamic technology with wide usage in all possible fields of application. In this article, we have made the current status of classification of challenges and the main problems related to the IoT to frame it and high-

lighting the solutions of security, reliability, and construction the IoT technology, in addition to the basic objectives of the IoT design.

References

A. Aijaz , A. Aghvami,.(2015). Cognitive Machine-To-Machine Communications For Internet-Of-Things: A Protocol Stack Perspective" Ieee Internet Of Things Journal, Vol. 2, No. 2, Pp. 103-112

Anders P. Mynster,.(2017), [Online] Available At, Https://Testlab-Uk.Madebydelta.Com/News/Iot-And-Reliability/ [Accessed 12 Jun. 2017].

Daniel Elizalde, (2016) .Https://Iot-For-All.Com/People-Dont-Buy-Iot-Buy-Solution-Problem/ [Accessed 14 Jun. 2017].

Diego M. Mendez, Ioannis Papapanagiotou, Baijian Yang ,.(2017). Internet Of Things: Survey On Security And Privacy .

Daniele Miorandi, Sabrina Sicari, Francesco De Pellegrini, Imrich Chlamtac ,. (2012). Internet Of Things: Vision, Applications And Research Challenges 523–531.

Electronicdesign ,[Online] Available At :Http://Www.Electronicdesign.Com/ [Accessed 09 Jun. 2017].

John Brockman,.(2010).How Is The Internet Changing The Way You Think, [Online] Available At : Https://Www.Edge.Org/Responses/How-Is-The-Internet-Changing-The-Way-You-Think [Accessed 12 Jun. 2017].

I. E. Bagci, S. Raza, T. Chung, U. Roedig, And T. Voigt,.(2013). Combined Secure Storage And Communication For The Internet Of Things,In Sensor, Mesh And Ad Hoc Communications And Networks (Secon), 10th Annual Ieee Communications Society Conference On. Ieee,Pp

Internet Society (2017).[Online] Available At Http://Www.Internetsociety.Org/Who-We-Are/Mission/Values-And-Principles [Accessed 10 Jun. 2017]

Karen Rose, Scott Eldridge, Lyman Chapin,.(2015). The Internet Of Things: An Overview.

Hui Suo Et Al,. (2012). Security In The Internet Of Things: A Review, Ieee Computer Society, 648.651

Lon Berk,.(2014). [Online] Available At Http://Insurancethoughtleadership.Com/Coverage-Risks-From-The-Internet-Of-Things/, [Accessed 12 Jun. 2017]

Subho Shankar Basu, Somanath Tripathy, Atanu Roy Chowdhury, .(2015). Design Challenges And Security Issues In The Internet Of Things, Ieee Region 10 Symposium.

Somia Sahraoui,. (2016) . Mécanismes De Sécurité Pour L'intégration Des Rcsfs A Iot (Internet Of Things).

Santhosh Kumar Svn, Yogesh P,. (2015). Evolution Of Internet Of Things And Related Chal-Lenges , Number 9.

Spark-fu-nElectrnics,.(2003),[Online]AvailableAt :Https://Learn.Sparkfun.Com/Tutorials/Connectivity-Of-The-Internet-Of-Thing [Accessed 10 Jun. 2017]

Xiaoyan Wang, Benjamin Bu Sze, Marianne Vandecasteele, Yao-Hong Liu, Christian Bachmann, Kathleen Philips,.(2017). The Design Challenges Of Iot: From System Technologies To Ultra-Low Power Circuits.Ieice Trans. Electron.Vol.E100–C

Mohammed AlSaudi Ali, Dyaa Motawa, and Fahad Al-Harby,.(2018). Internet of Things and Distributed Denial of Service Mitigation

Chapter

8

E-Supply Chain Management: a competitive advantage

ASD'2018

Content

Process Mining for port container terminals: The state of the art and issues

Mouna AMROU MHAND*, Hassan BADIR*
Azedine BOULMAKOUL**

*SDET, National School of Applied Sciences, ENSA,
Abdelmalek Essaadi University, Tangier, 90 000, Morocco
Amroumouna@gmail.com
Badir.hassan @uae.ac.ma
**LIM, Faculté des sciences et techniques de Mohammedia
Université Hassan II Casablanca, Morocco
Azedine.boulmakoul@gmail.com

Summary. Container Terminals are random and dynamic complex systems, involving vital processes that cause many decision problems related to logistics planning and control issues. Process mining is an advantageous approach to acquire a better knowledge about those processes by analyzing the recorded event data. In this paper, a state of the art and issues of process mining for logistics environment, particularly container terminals, are presented. This work is motivated by the exploitation of process mining techniques in order to manage, power and rule marine logistics for assuring the best performance and to handle future endeavors in container terminals for to maximize service level as well as minimize the costs.

1 Introduction

Freight transportation plays a principal role in the modern economy as it allows goods exchange between distant countries. The most remarkable and firm technology for transporting freight, especially on long maritime routes, is containerization. The employment of containers for intercontinental maritime transport has impressively increased. Container terminals act as the standard unit-load notion for international freight. They basically serve as an interface between different modes of transportation such as domestic rail or truck transportation and deep sea maritime transport (Kim et Günther, 2007).

Container terminals involve important processes that cause many decision problems related to logistics planning and control issues. Hence, the necessity to supervise and analyze these processes in order to manage, power and rule marine logistics as well as assuring the best performance to handle future endeavors in container terminals.

Process analysis extremely fits to empower processes improvement.

Process mining is an advantageous approach to acquire a better knowledge about those processes from event data. It is an arising discipline that consists of offering comprehensive sets of means to produce fact-based insights.

This paper presents a survey of process mining for logistics environment, particularly container terminals. It is organized as follows : Section 2 covers a brief introduction to process mining, its main types, related work and overview of ProM framework. Section 3 brings in container terminals and entails port problems. It also describes logistic processes and gives a literature review of the studies involving container terminals problems in addition to a brief presentation of Tanger-Med port. Section 4 concludes and suggests future works.

2 Process Mining : Overview

Process mining is the bond between traditional model based process analysis and data centric analysis. It is an arising discipline that consists of offering comprehensive sets of means to produce fact-based insights as well as to assist process improvements and enhancements in a variety of application domains. The idea of process mining is to discover, monitor and improve real processes by retrieving valuable knowledge and process related information from event logs available in massive data volumes systems.

This discipline is more than a fusion of pre-existing approaches. For instance, existing data mining techniques are more data-centric to provide a coherent insight and understanding of the end-to-end processes in the organization. Business Intelligence tools center on dashboards and reporting rather than definitive business process insights. While process mining techniques count intensively on experts modeling. Process mining is not limited to process discovery. By tightly pairing event data and process models, it becomes possible to check conformance, detect deviations, predict delays and support decision making as well as suggest process redesigns.

2.1 Process Mining Types

Process mining types can be grouped into three categories (van der Aalst et al., 2007) :

— Discovery ; This first category consists of taking an event log and producing as a result a model without using any a-priori information. We cite the α-algorithm (van der Aalst et al, 2004) as an example, it takes an event log and produces a Petri net describing the behavior registered in the log.
— Conformance ; For this category, an existing process model is compared with an event log of the same process. It can be exploited to examine if reality is conformed to what is recorded in the log and vice versa.
— Enhancement ; it reposes on the concept of extending or improving an existing process model using actual information about the actual process recorded in

some event log. This type aims at changing or extending the a-priori model while the conformance checking type measures the alignment between model and reality. The first type of enhancement is repair ; it involves modifying the model to better match the reality. Another type of enhancement is extension ; it consists of adding a new perspective to the process model by cross-correlating it with the log. The figure below (FIG. 1) illustrates the interaction between the different process mining types.

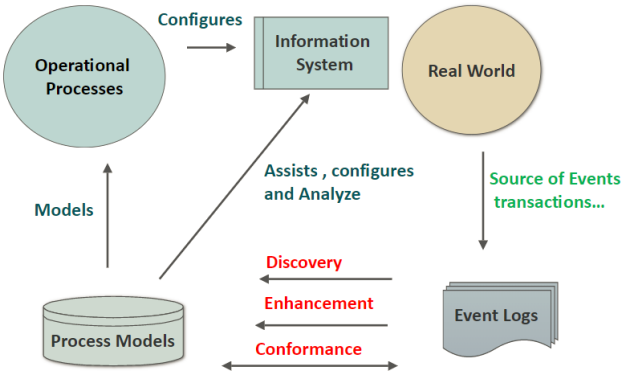


FIG. 1 – *Process Mining types interaction : discovery, conformance, and enhancement.*

2.2 Process Mining compares to data mining, BI and CEP

Data mining is data-driven and so is process mining. Nonetheless, mainstream data mining techniques are not process-centric (van der Aalst et al., 2007). Data mining is "the analysis of large data sets to identify unsuspected relationships and to summarize the data in new ways that are both understandable and useful to the data owner" (Van der Aalst, 2011). The input is generally a table and the output may be rules, clusters, tree structures, patterns, etc. There are a few data mining techniques that are similar to process mining. But, these methods do not take into account end-to-end processes. Process mining facilitates the application of data mining techniques to event data.

For Business Intelligence, Process mining is positioned under Business Intelligence. There is no evident definition for BI, it includes anything that has the intention to provide information that can be exploited to assist decision making. The focus is on querying and reporting associated with simple visualization illustrating dashboards.

As for complex event processing i.e. continuous processing of information. The principal aim is to identify significant events and respond in real time. CEP is particularly advantageous in case of the existence of plenty of low-level events. By reducing the flow of event data to feasible streams and logs. Thus, analysis becomes much easier.

2.3 Related work

The use of process mining for supporting businesses has been gaining in momentum. (Agrawal et al., 1998) was the first work that proposed applying process mining in the context of workflow management. The authors in (Van der Aalst et Weijters, 2004) addressed the issue of process mining in the context of workflow management using an inductive approach. In the context of applications of process mining techniques in different domain to support businesses, many works have been proposed. In (van der Aalst et al., 2007) the authors describe the application of process mining in logistic domain particularly in one of the provincial offices of the Dutch National Public Works Department, responsible for the construction and maintenance of the road and water infrastructure. They analyze the processing of invoices sent by the various subcontractors and suppliers from different perspectives. In (van Aalst et al., 2010) they propose a framework for auditing that employs process discovery and conformance checking. Another application in security, (Jans et al., 2010) based on a case company, think that the use of process mining techniques can have additional value to reduce the risk of internal fraud in companies. For Audits (Accorsi et Stocker, 2012), exploit process mining to demonstrate the feasibility of conformance checking as a tool for security auditing. In (Rebuge et Ferreira, 2012) they present an approach based on process mining techniques useful in healthcare environments. The methodology was applied for the emergency service in the Hospital of Sao Sebastiano and it consists of identifying regular behavior, process variants, and exceptional medical cases. (Zhong et al., 2013) presents a new redesigned solution for container terminal production processing system described by workflow and DFD the integrality and credibility of new system was proved using Petri Net analysis. (Lee et al., 2014) developed an intelligent system, using fuzzy association rule mining with a recursive process mining algorithm, to find the relationships between production process parameters and product quality. (Wang et al., 2014) introduces a comprehensive methodology for applying process mining in logistics covering the event log extraction and pre-processing as well as the execution of exploratory, performance and conformance analyses using as a case study Chinese port that specializes in bulk cargo. (Leemans et van der Aalst, 2015) presents a novel reverse engineering technique for obtaining real-life event logs from distributed systems to analyze the operational processes of software systems under real-life conditions, and use process mining techniques to obtain precise and formal models. The most recent study (de Alvarenga et al., 2018) proposes an approach to facilitate the investigation of huge amounts of intrusion alerts with the application of process mining techniques on alerts to extract information regarding the attackers' behavior and the multi-stage attack strategies they adopted.

2.4 ProM : A Process Mining Toolset

ProM (i.e. Process Mining) is a generic open source, process mining toolset available for downloading at [1]. ProM provides a platform of the process mining algorithms that is easy to use and easy to extend in form of plugins. The toolset has a pluggable architecture and supports a wide range of control-flow models including various types

1. http ://www.processmining.org

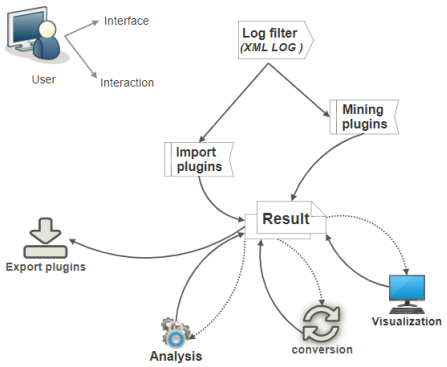


FIG. 2 – *Overview of the ProM framework.*

of Petri nets, event-driven process chains (EPCs), Business Process Modeling Notation (BPMN), and Business Process Execution Language (BPEL). ProM also supports models to represent rules (for example, LTL-based rules), social networks, and organizational structures.
The ProM framework reads files in the XML format through the Log filter. This sorts the events on timestamps when they exist for large datasets before the mining starts. Otherwise, the order in the XML file is preserved. Through the Import plugins a wide variety of models can be loaded. The Mining plugins do the mining and the result is stored in memory, and in a tab on the ProM interface. The framework allows plugins to function with each other's results. Commonly, the mining results contain some visualization or further analysis or conversion. The Analysis plugins take a mining result an analyze it and the Conversion plugins take a mining result and transform it into another format. (see FIG.2 for an overview of the ProM framework).

3 Container Terminals

A container terminal represents a complex system as it plays a vital role as a node in many supply chains since it is considered an area for container transshipment between different transport modalities (e.g. deep-sea, short-sea, inland waterway, road, and rail). A container terminal is made up of four major areas : quay area, transport area, yard area, and hinterland.
With the emanation of a new and recent family of deep-sea container vessels, the main ports are incited to reconsider and to expand their equipment and logistics. High-density, automated container handling equipment is a potential candidate to enhance container terminals performance and handle future endeavors in marine transportation. Yet, for these equipments to work efficiently, decision planning tool for integration and optimization becomes crucial.
We categorize container processes into three types : import/ export, transshipment and storage/ Handling :

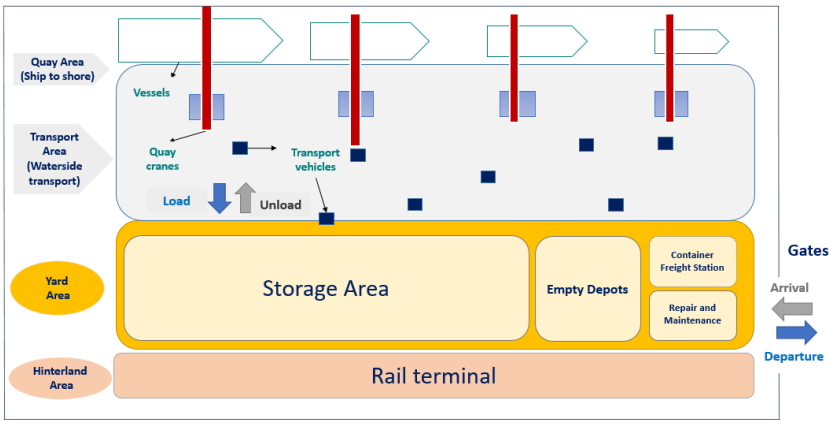


FIG. 3 – *Schematic container terminal layout.*

— Import/ Export ; this category consists of loading, unloading and anchoring of containers to their designated locations. This action is performed in the maritime operation area.
— Transshipment ; includes the receiving and shipping from/to other modes of transportation (trucks, trains) ; It comprises the storage operations and management of containers in the yard. The transshipment is carried out in the terminal storage area.
— Storage/ Handling ; involves storing the containers in the storage area and their transfer to the other modes of transport. The storage/ handling is executed in the land area.

In FIG. 3, we display a schematic layout of the container terminal. In the next section, we will highlight the important logistics processes in port container terminals.

3.1 Logistics Processes in Port Container Terminals

A maritime container terminal is a complex facility that is adjusted to a set of logistic processes. The logistic activities at a container terminal belong to further complex logistic processes and this is critical for a valuable management of the system as well as the choice of the modeling approach.
A rigid categorization of the decision dilemma in a container terminal abides by the next logistics processes (Vis et De Koster, 2003), (Steenken et al., 2004) : arrival of the ship, unloading and loading of the ship, transport of containers from ship to stack and vice versa, stacking of containers, and interterminal transport and other modes of transportation. These processes are detailed in the following : Each vessel has to make an advanced reservation according to its Expected Time of Arrival (ETA) when calling at the port. The entrance ship is based on the following aspects :

— Formality ; such as the cooperation between the shipping line and the port to operate within the port, thus giving a priority policy to the port entrance queue.

— Operationally ; which consists of pilot boat availability and berth spaces assignment.

Once the requirements are met, the ship can be maneuvered by one or two pilot boat captains into its berth. Otherwise, it has to wait outside the entrance to the roadstead. This process is called the arrival of the ship.

There are different types of vessels. These vessels can be ordered as follows : Mother vessels that are large container ships that cover transoceanic lines (up to 3,000 TEUs and no more than 14,000 TEUs). And feeders that are smaller ships that cover short and middle routes. They are used to connect the spokes to the transshipment hub (and vice versa). At the time a vessel is moored, unloading/loading operations can be started once the mechanical and human resources are assigned ; otherwise, the ship will be waiting in its dock position until the assignment of resources. The unloading/loading operations are performed by rail-mounted gantry cranes, moving containers between the ship and the quay area placed along the berth. The number of quay cranes that are assigned to the vessel is mainly restricted by the total number of cranes in the quay and the allowed number of cranes for each vessel, according to the physical requirements (i.e. the length of the vessel usually five for the longest vessel) and the logical constraints (i.e. interference between cranes operations). The performance of the discharge/loading process extremely depends on the availability of rail-mounted gantry cranes (RMGs) and their turnover speed. Accordingly, the magnificent deployment of these resources affects the overall completion time of each vessel. Lastly, the containers are moved forth and back between the berth area and the yard area by a fleet of vehicles, namely straddle carriers (SCs) or shuttle vehicles (AGV).

3.2 Maritime Port Container Terminals Problems

Modelling maritime container terminal is a complex activity and to study this large system, we must refer to specific means in order to solve assignment and scheduling problems. We distinguish between two main port container terminals problems according to the literature :

The first issue is named Berth Allocation Problem (BAP), it deals with assigning a berthing position and a berthing time to each vessel. It focuses on optimally allocating vessels to the berths or quay locations, the berthing time and the berthing position along the quay length of a vessel. BAP aims for maximizing the productivity of the vessel handling, maximizing service levels, minimizing the total handling of vessels as well as minimizing the costs.

Several studies have been conducted to discuss BAP. The subject of these studies includes : In (Golias et al., 2009), They formulated BAP as a multi-objective combinatorial optimization problem where vessel service is differentiated upon based on priority agreements. A genetic algorithms based heuristic is developed to solve the resulting problem. (Hendriks et al., 2010) present a robust optimization model for cyclic berth planning facing the problem of defining the arrival and departure times of each cyclically calling vessel on a terminal, take into consideration the awaited number of containers to be supported and the required quay and crane capacity to do so. (Buhrkal et al., 2011) consider the problem of allocating arriving ships to discrete berth locations at container terminals. The authors improve the performance of a model and

present a comparison from a computational perspective (Hendriks et al., 2012) address the problem of spreading a set of cyclically calling vessels over the various terminals and allocating a berthing and departure time to each of them with the objectives to balance the quay crane workload over the terminals and over time and to minimize the amount of interterminal container transport. and (Xu et al., 2012) examine a berth allocation problem in container terminals in which the assignment of vessels to berths is limited by water depth and tidal condition and analyze the computational complexity and develop efficient heuristics model for the static and the dynamic case.
The second issue is Quay Crane problem (QCP) that is classified at its turn into :
Quay Crane Assignment Problem (QCAP), it refers to assigning cranes to vessels in a way that all required transshipments of containers are fulfilled, according to the given berth plan and the available QC to serve the vessel without passing each other. QCAP have a strong impact strong impact on the vessels' handling times.
Quay scheduling problem(QCSP) (Kim et Park, 2004) implying to determine the sequence of discharging and loading operations that quay crane will perform so that the completion time of a ship operation is minimized. Hence, this work proposes a branch and bound (B & B) method to obtain the optimal solution of the quay crane scheduling problem.
Quay Crane Deployment Problem (QCDP), respecting the berth schedule assigned to the quay cranes to the incoming vessels. It aims for minimizing the number of used quay cranes and maximizing their utilization, under the restriction of finishing the loading/unloading operations, for each vessel, within the anticipated time of unberthing. Among the works that have covered the QCP, (Legato et al., 2008) suggest two phase approach for the resolving this problem ; an IP model is used to decide when and how many cranes must be assigned to each vessel then a heuristics approach to determine which specific crane should be assigned to a vessel. (Chang et al., 2010) using objective programming for dynamic berth allocation and quay crane assignments based on rolling horizon approach. (Lu et al., 2010) mixed integer programming model is proposed, and a simulation based Genetic Algorithm (GA) search procedure is applied to generate robust berth and QC schedule proactively. they took vessel arrival times and container handling uncertainty into consideration. (Giallombardo et al., 2010) combined the berth allocation and the quay crane assignment problems and developed a heuristic algorithm which combines tabu search methods and mathematical programming techniques.

3.3 Tanger Med Port Container Terminals

Tanger-Med port authority is considered a global logistics hub, located on the Strait of Gibraltar. In the following we introduce our study case, the Tanger-Med port and the maritime liaison as well as the complex distribution.
Through several ship owners, it provides regular services serving nearly 174 ports and 74 countries on the 5 continents. Thanks to its strategic position the port is 10 days from America and 20 days from China. It also serves nearly 35 ports and 21 countries in West Africa. Tanger Med port makes it possible to cross of the strait by ferry in less than 45 minutes. (see FIG. 4).

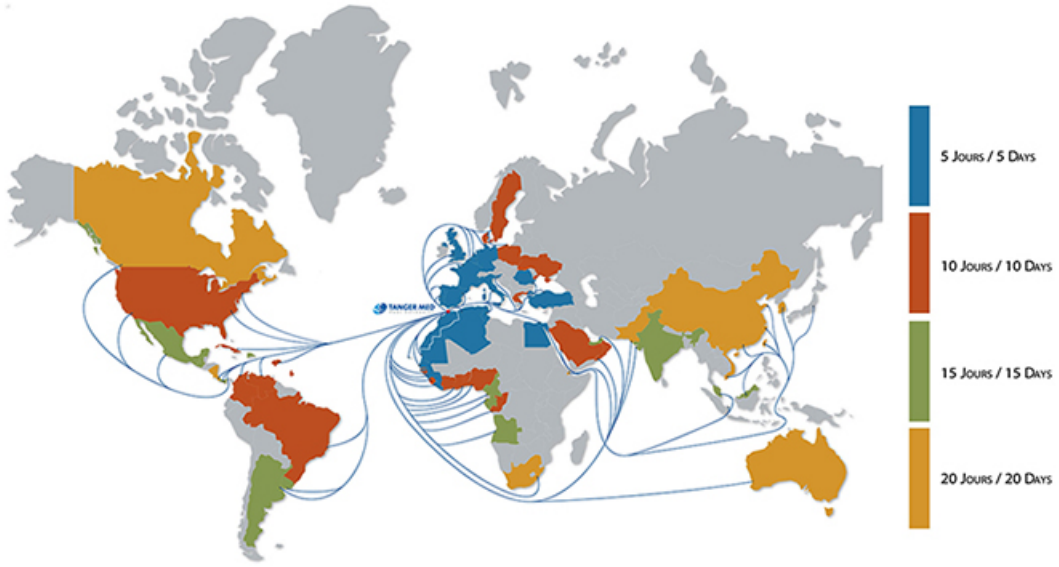


Fig. 4 – *Tanger-Med port connectivity.*

The Tanger Med port aspires to arise a powerful port platform combining transshipment activities, import export of extra logistics operation value. The Tanger Med port complex includes :

- The Tanger Med 1 ; consists of two container terminals, a railway terminal, hydrocarbons terminal, goods terminal, and vehicle terminal.
- The port Tanger Med 2 ; includes two container terminals.
- The Tanger Med Passengers Port ; involving the access zones and border inspections, the eight berths of boarding passengers and trucks, regulations zones, and the ferry terminal.
- Logistics Free Zone MEDHUB ; comprises actual 50 hectares of land surface as well as warehouses and offices for rent.
- The Tanger Med Port Center – TMPC ; a 30.000 m^2 of offices, banks...connected to the train, bus and maritime station.

Fig. 5 illustrates the complex's distribution.

Tanger Med port plays a vital role as a crucial platform of container transshipment on (Asia / Europe) and (Europe / Africa) routes, Tanger Med takes an essential part through connectivity and in the improvement and development of import and export traffic in Morocco. Moreover, it offers an important connecting point to the hinterland area due to its divers infrastructure like rail and highways.

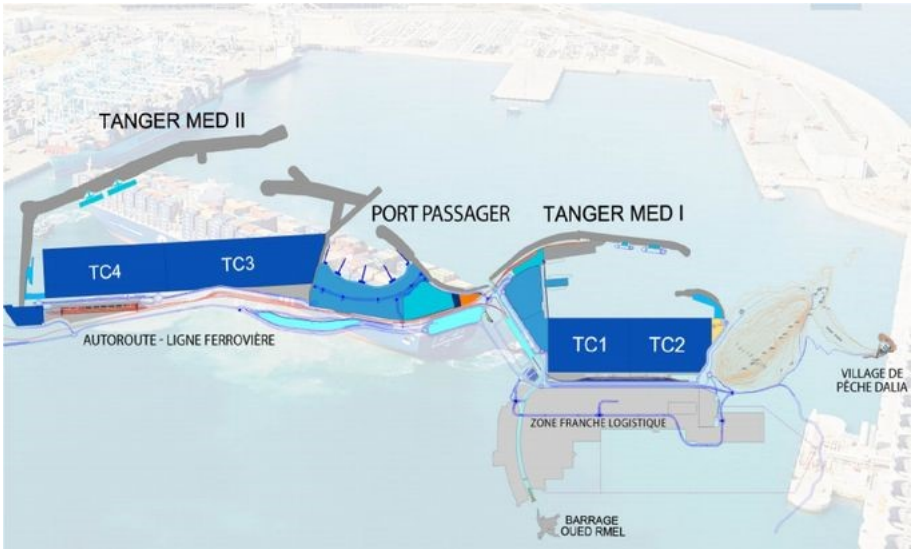


FIG. 5 – *Tanger-Med complex distribution.*

4 Conclusion

This paper addresses process mining for container terminals. It covered the state of the art and the issues. Most of the examined contributions refer to specific means (algorithms, optimization models...) to solve the encountered problems related assignment and scheduling.
The presented work the first theoretical frame of the tasks ahead. The main purpose of our research is to exploit event logs to discover the most suitable process model for each case. The applicability of our work will be demonstrated with a case study of Tanger-Med port.
This study promises to ensure the best logistics configuration for the port. Consequently, the extra costs are avoided, the service level is maximized and the handling time is minimized.

References

Accorsi, R. et T. Stocker (2012). On the exploitation of process mining for security audits: the conformance checking case. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pp. 1709–1716. ACM.

Agrawal, R., D. Gunopulos, et F. Leymann (1998). Mining process models from workflow logs. In *International Conference on Extending Database Technology*, pp. 467–483. Springer.

Buhrkal, K., S. Zuglian, S. Ropke, J. Larsen, et R. Lusby (2011). Models for the discrete berth allocation problem: A computational comparison. *Transportation Research Part E: Logistics and Transportation Review 47*(4), 461–473.

Chang, D., Z. Jiang, W. Yan, et J. He (2010). Integrating berth allocation and quay crane assignments. *Transportation Research Part E: Logistics and Transportation Review 46*(6), 975–990.

de Alvarenga, S. C., S. Barbon Jr, R. S. Miani, M. Cukier, et B. B. Zarpelão (2018). Process mining and hierarchical clustering to help intrusion alert visualization. *Computers & Security 73*, 474–491.

Giallombardo, G., L. Moccia, M. Salani, et I. Vacca (2010). Modeling and solving the tactical berth allocation problem. *Transportation Research Part B: Methodological 44*(2), 232–245.

Golias, M. M., M. Boile, et S. Theofanis (2009). Berth scheduling by customer service differentiation: A multi-objective approach. *Transportation Research Part E: Logistics and Transportation Review 45*(6), 878–892.

Hendriks, M., D. Armbruster, M. Laumanns, E. Lefeber, et J. Udding (2012). Strategic allocation of cyclically calling vessels for multi-terminal container operators. *Flexible Services and Manufacturing Journal 24*(3), 248–273.

Hendriks, M., M. Laumanns, E. Lefeber, et J. T. Udding (2010). Robust cyclic berth planning of container vessels. *OR spectrum 32*(3), 501–517.

Jans, M., N. Lybaert, et K. Vanhoof (2010). A framework for internal fraud risk reduction at it integrating business processes: the ifr^2 framework.

Kim, K. et H.-O. Günther (2007). Container terminals and terminal operations. *Container Terminals and Cargo Systems*, 3–12.

Kim, K. H. et Y.-M. Park (2004). A crane scheduling method for port container terminals. *European Journal of operational research 156*(3), 752–768.

Lee, C., G. Ho, K. Choy, et G. Pang (2014). A rfid-based recursive process mining system for quality assurance in the garment industry. *International Journal of Production Research 52*(14), 4216–4238.

Leemans, M. et W. M. van der Aalst (2015). Process mining in software systems: discovering real-life business transactions and process models from distributed systems. In *Model Driven Engineering Languages and Systems (MODELS), 2015 ACM/IEEE 18th International Conference on*, pp. 44–53. IEEE.

Legato, P., D. Gullì, et R. Trunfio (2008). The quay crane deployment problem at

a maritime container terminal. In *Submitted to the 22th European Conference on Modelling and Simulation*.

Lu, Z.-q., L.-f. Xi, et al. (2010). A proactive approach for simultaneous berth and quay crane scheduling problem with stochastic arrival and handling time. *European Journal of Operational Research 207*(3), 1327–1340.

Rebuge, Á. et D. R. Ferreira (2012). Business process analysis in healthcare environments: A methodology based on process mining. *Information systems 37*(2), 99–116.

Steenken, D., S. Voß, et R. Stahlbock (2004). Container terminal operation and operations research-a classification and literature review. *OR spectrum 26*(1), 3–49.

van Aalst, W. M., K. M. van Hee, J. M. van Werf, et M. Verdonk (2010). Auditing 2.0: Using process mining to support tomorrow's auditor. *Computer 43*(3).

Van der Aalst, W. M. (2011). Data mining. In *Process Mining*, pp. 59–91. Springer.

van der Aalst, W. M., H. A. Reijers, A. J. Weijters, B. F. van Dongen, A. A. De Medeiros, M. Song, et H. Verbeek (2007). Business process mining: An industrial application. *Information Systems 32*(5), 713–732.

Van der Aalst, W. M. et A. Weijters (2004). Process mining: a research agenda. *Computers in industry 53*(3), 231–244.

Vis, I. F. et R. De Koster (2003). Transshipment of containers at a container terminal: An overview. *European journal of operational research 147*(1), 1–16.

Wang, Y., F. Caron, J. Vanthienen, L. Huang, et Y. Guo (2014). Acquiring logistics process intelligence: Methodology and an application for a chinese bulk port. *Expert Systems with Applications 41*(1), 195–209.

Xu, D., C.-L. Li, et J. Y.-T. Leung (2012). Berth allocation with time-dependent physical limitations on vessels. *European Journal of Operational Research 216*(1), 47–56.

Zhong, W. Z., X. Q. Fu, et Y. P. Wang (2013). Petri net modeling: Container terminal production operation processing system analysis. In *Applied Mechanics and Materials*, Volume 409, pp. 1320–1324. Trans Tech Publ.

Résumé

Les terminaux à conteneurs sont des systèmes complexes aléatoires et dynamiques, impliquant des processus importants qui provoquent de nombreux problèmes de décision liés à la planification logistique et au contrôle. Le data mining des processus est une approche avantageuse pour acquérir une meilleure connaissance de ces processus en analysant les données d'événements enregistrées. Dans cet article, un état de l'art et problèmes pour l'environnement logistique, en particulier les terminaux à conteneurs, sont présentés. Ce travail est motivé par l'exploitation des techniques de data mining des processus afin de gérer, gouverner et de régir la logistique maritime dans le but d'assurer la meilleure performance pour faire face aux défis futurs dans le but de maximiser la qualité du service et minimiser les coûts.

Logistics Services Providers: The state of play of the Moroccan context

Latifa Fadile*, Mohamed El oumami**, Zitouni Beidouri***

*School of Technology Casablanca, Hassan II University of Casablanca, PO Box 8012, Oasis, Casablanca, Morocco
fadile_latifa@hotmail.fr

**School of Technology Casablanca, Hassan II University of Casablanca, PO Box 8012, Oasis, Casablanca, Morocco
mohoumami@gmail.com

***School of Technology Casablanca, Hassan II University of Casablanca, PO Box 8012, Oasis, Casablanca, Morocco
zbeidouri@gmail.com

Abstract. Logistics, nowadays, occupies an important part in every economy and every business entity and the current worldwide trend has pushed many companies to outsource their logistics functions to logistics services providers (LSPs), so as to focus on their core competencies and businesses.
This article is particularly interested in LSPs, identifies and categorizes the different existing types of these latter, highlights their situation in Morocco and finally discovers potential gaps for future research.

1 Introduction

The global context of the company expressed in particular in terms of increasing pressure to guarantee its survival, pushes decision-makers to seek to guarantee and maintain their competitive advantages by focusing more and more on their core business and to get rid of activities that impact their profits, including support functions. Logistics is among the functions that most companies agree to entrust professionals, especially when they know that they must, now, be agile and not only produce at low prices, and especially that this decision of outsourcing can be financially efficient comparing to investing in new human resources and infrastructures. This massive use of logistics services outsourcing has led to the emergence of a new actor, the LSP which now occupies a central place in the supply chain (SC) and has begun to diversify his offers, ranging from conducting operations to piloting the whole SC.

The rest of the manuscript is organized as follows. In the second section, the different classifications of the LSPs are presented. The third section sets out the state of play of LSPs in Morocco. Finally, the paper points out opportunities for future research.

2 Logistics services providers

For reasons of rationalization of practices and with a view to focus on core business, to introduce products and service innovation quickly (Lai, 2004), companies tend to outsource their logistics services (LSs). This massive use of logistics outsourcing has helped to the emergence of a new actor, the LSP which now has a major part in the SC and has begun to diversify his offers, ranging from conducting operations to piloting the whole SC.

The term LSP is applied as a synonym for similar terms such as outsourcer, carrier, forwarding company, transport company, logistics services company and third-party logistics provider (Forslund, 2012). And it has been defined in numerous ways in the literature. (Hertz & Alfredsson, 2003), for example, have stated that a LSP (the outsourcer) is an external provider who manages, controls, and carries out LSs on behalf of a company (the service user). For (Sink & Langley, 1997), a LSP is a service provider who is able to assume some or all of a company's LSs. And (Marchet, Melacini, Perotti, Sassi, & Tappia, 2017), in their definition, have added a very important point which is the added value provided by the LSP to a company's business.

In fact, the LSP should not be considered as an additional intermediary but he needs to be treated as a separate industry (Berglund, Laarhoven, Sharman, & Wandel, 1999). Actually, many authors in the literature have supported this vision. (Roveillo, Fulconis, & Paché, 2012), for instance, have looked at the LSP as a "logistics integrator", because his presence is of paramount importance throughout the company's SC (from the first supplier to the final customer), also because he is actively involved in managing the interfaces between its various components. For (Zacharia, Sanders, & Nix, 2011), a LSP has been seen as an "orchestrator". The term "orchestration", in this context, signifies the activity of managing and coordinating to facilitate the supply chain management (SCM) best practices.

Thus, the evolution of the definitions has followed the evolution of the services that the LSPs offer in order to succeed within a very competitive marketplace (Rushton, 2006). LSPs, nowadays; strive to become large in size with the ability to offer advanced logistics solutions. And the literature has conceptualized these developments by distinguishing different types of LSPs based on their ability to adapt their services to their customers and their ability to solve the logistics problems they face.

(Muller, 1993) peeps out to be the first to suggest two basic types of LSPs: operations-based third party logistics vendors and information-based third party logistics vendors. Later, the same author has modified this classification by suggesting the following four types of vendors (LSPs):

- Asset-based vendors: they refer to companies which provide physical LSs through the use of their own assets. It is generally about a trucks fleet or a group of warehouses or both.
- Management-based vendors: they refer to companies which are involved in providing - logistics management services through databases systems and services consulting. These companies do not own transportation or warehouse assets.
- Integrated vendors: they refer to companies that own assets, typically trucks, warehouses or a combination of both. However, they are not limited to the use of their own assets and will contract with other LSPs if required.
- Administration-based vendors: they refer to companies which mainly provide administrative management services such as freight payment.

This classification has been, in part, adopted by (Nemoto & Tezuka, 2002). These latter have allocated LSPs into two types: asset-based LSPs and non-asset-based LSPs. And (Fielser & Paché, 2008) have suggested a classification closer to that suggested by (Muller, 1993). In fact, they have distinguished three types of LSPs:

- Classical LSPs: they carry out physical operations related to the transport, handling, warehousing and storage of intermediate or finished products of their customers.
- Value-added LSPs: they include the management of industrial or commercial operations (for example delayed differentiation), administrative operations (for example invoicing) and informational operations (for example tracking and tracing of products).
- Dematerialized LSPs: they have no physical resources and they build their services by mobilizing resources from specialized subcontractors and ensuring their overall coherence through a total control of information flows.

There is another classification given by (Hertz & Alfredsson, 2003), where they have distinguished between four types of LSPs:

- Standard LSPs: they perform the most basic operations of logistics such as picking, warehousing and distribution.
- Service developers: they provide advanced value-added services to their customers such as tracking and tracing, cross docking and specific packaging.
- Customer adapters: they offer services at the request of the customer. They improve LSs and do not develop new ones.
- Customer developers: they are the highest level of LSPs. They integrate themselves with customers and take over entire logistics functions.

In the same context, (Lai, 2004) through his study, has suggested that there are four types of LSPs, which are:

- Traditional freight forwarders: they have a low capability to carry out value-added LSs and technology-enabled LSs.
- Transformers: they achieve a medium level of capability to perform value-added LSs and possess a high level of capability in technology-enabled LSs and freight forwarding services.
- Full service providers: they possess a high level of capability in all of the three LSs: freight forwarding services, value-added LSs and technology-enabled LSs.
- Nichers: they are particularly weak in freight forwarding services and possess a medium level of capability in carrying out value-added LSs and technology-enabled LSs. They target the niche markets for value-added LSs and technology-enabled LSs in order to avoid head-on competition with either traditional freight forwarders or full service providers.

Conventionally, the lexicon of logistics terms proposes distinguishing five main types of LSPs (1PL, 2PL, 3PL, 4PL and 5PL) present on the market, according to the complexity of their system of offer:

- First Party Logistics (1PL):

This term is used for those manufacturers that carry out their logistics by themselves. They own all logistics assets and manage all their logistics operations in-house as shown in figure 1.

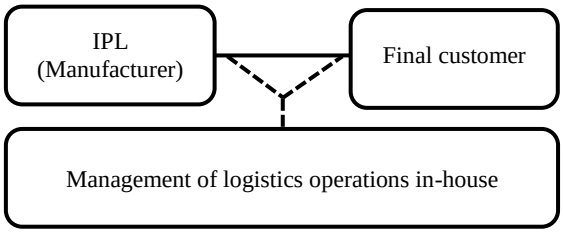


FIG. 1 – *Schematization of the role of First Party Logistics*

- Second Party Logistics (2PL):

When manufacturers began to extend their business geographically, it became tough for them to manage all the logistics operations by own. Then the concept of Second Party Logistics (2PL) came in the market. These 2PLs manage the simple execution of physical operations related to transport (Fielser & Paché, 2008), and consequently they offer a single function in the SC as depicted in figure 2.

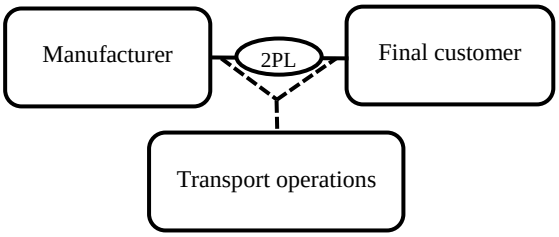


FIG. 2 – *Schematization of the role of Second Party Logistics*

- Third Party Logistics (3PL):

Thereafter, the 2PLs develop their capabilities in handling logistics functions and also integrate different services provided before separately, which lead to the emergence of a new type of LSPs which is the Third Party Logistics (3PL). These 3PLs can provide in addition to transport and warehousing, value-added operations such as cross-docking and delayed differentiation as indicated in figure 3.

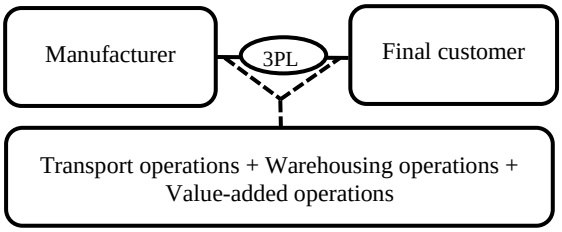


FIG. 3 – *Schematization of the role of Third Party Logistics*

- Fourth Party Logistics (4PL):

Fourth Party Logistics (4PL) is the next evolution of LSPs and it was developed on the basis of 3PL. 4PLs differ from the other providers (1PL, 2PL and 3PL) in the sense that they have no tangible assets. They have no physical means (trucks or warehouses) and their role is similar to that of logistics consultants who provide engineering services. The concept of 4PL should not be confused with that of LLP (Lead Logistics Provider). Actually, the two providers offer the same LSs but the existing difference lies in the means used. A 4PL, as we have seen, is a non-assets provider, whereas a LLP is a mixed-assets provider because he carries out his customers' LSs by using his own resources and those of other LSPs.

In the literature, a 4PL can be seen as a "transaction center" (Fulconis, Saglietto, & Paché, 2007), as a "supply chain integrator" (Håkansson & Shenota, 1995),(Rushton & Walker, 2007), as a "business process outsourcing (BPO) provider" (Mukhopadhyay & Setaputra, 2006) and as a "coordinator" (Rushton, 2006) by dint of his ability to manage the resources of its own organization with those of complementary LSPs as can be seen in figure 4.

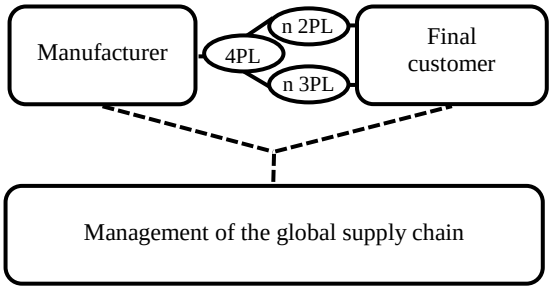


FIG. 4 – *Schematization of the role of Fourth Party Logistics*

This term of 4PL was filed by the consulting cabinet Accenture as a registered trademark in 1996. And the definition given at the time was as follows: *«the 4PL is an integrator that assembles its own resources, capabilities and technology and those of other service providers to design and manage complex supply chain ».*

And in 2000, the same cabinet proposed a segmentation of the action of 4PL in three operations and his position in relation to other providers. This is illustrated in figure 5.

From the figure below, it can be deduced that a 4PL can play three different roles depending on his relationship with his customers and other providers:

- Synergy Plus: in this role, the 4PL is placed alongside one or more LSPs (2PL and 3PL) towards several customers. It is about cooperation between the other LSPs and the 4PL, in a relationship that allows taking advantage of the resources and competencies of each one.
- Solution Integrator: in this role, the 4PL is considered as a solution integrator, because he can manage and build an integrated SC with many other LSPs (2PL and 3PL) towards a single customer.
- Industry Innovator: in this role, the 4PL synchronizes a group of customers in order to bring the SC to high efficiency, thanks to technologies and operational strategies.

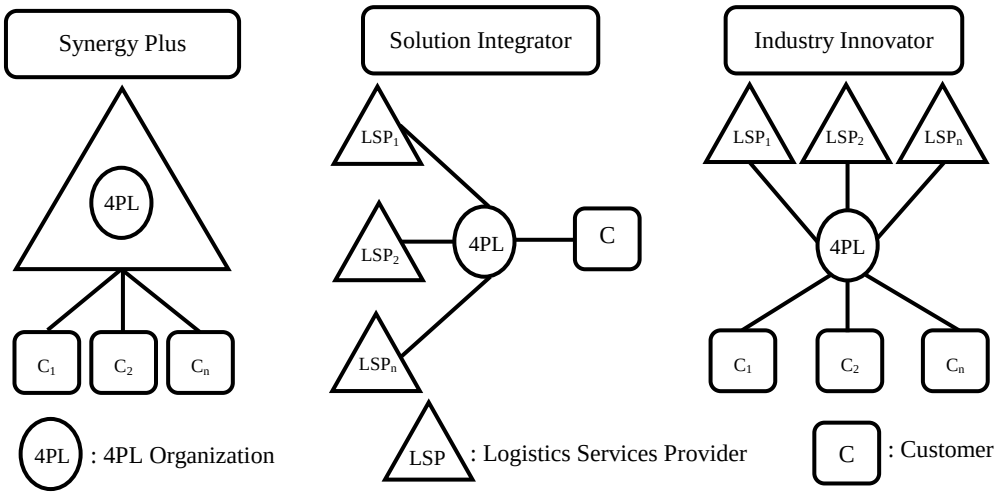


FIG.5 – *The three roles of a Fourth Party Logistics adopted from Accenture*

Referring to the definitions presented previously, we can deduce that a 4PL is a response to the multiplication of actors in the SC. He designs both the logistics architecture and the information system applying to the company's integrated processes. However, he does not execute in person the corresponding physical flows, which are entrusted to other LSPs. His principal mission is to coordinate the network of partners of the integrated company (suppliers, distributors, customers, 2PL, 3PL, etc.) as illustrated in figure 6.

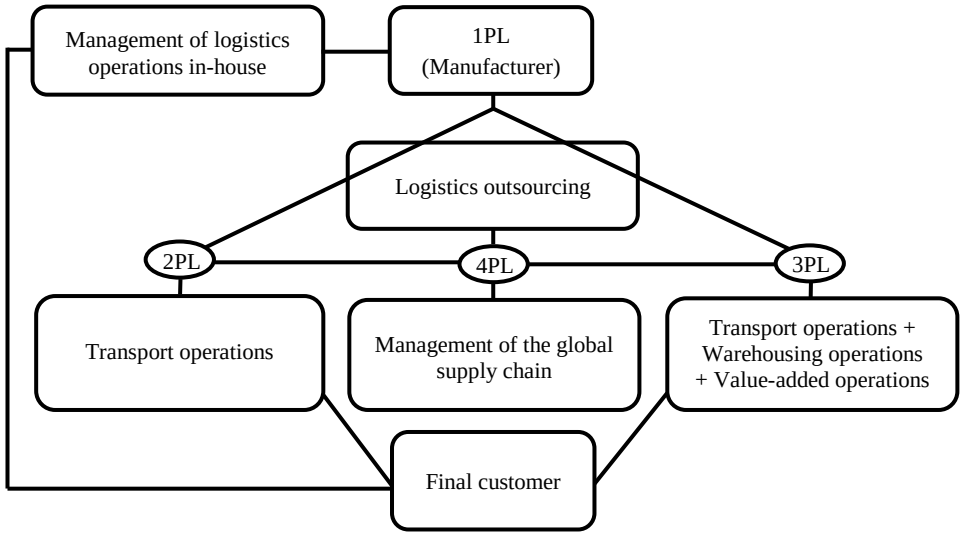


FIG.6 – *Schematization of the links between the different types of LSPs*

- Fifth Party Logistics (5PL):

The fifth and the final type of LSPs is the Fifth Party Logistics (5PL). 5PL is a new concept in logistics outsourcing and it is about the management of all parties of the SC in conjunction with e-business. Other terms in the literature are used to depict this type of LSPs, such as "virtual logistics services provider". The major focus of a 5PL is to offer automated and intelligent systems able to improve the performance of the SC and the key of success of this emerged type is the integration of information technologies and computer systems. Like the 4PL, a 5PL is almost wholly virtual. He possesses no typical assets, he has no physical presence but he forms a web-based system that provides information to the range of participants under his control (Hosie, Sundarakani, Tan, & Koźlak, 2012).

The LSPs' types present in the literature demonstrate the extent of the skills of the actors in this sector and the variety of situations encountered. While the LSP is primarily regarded as a "subcontractor" in a first phase, he became, in a later stage, a "co-designer" and even a "designer" and a "manager" of the SC, in an innovative and creative approach. Figure 7 gives a summary of these main LSPS' types.

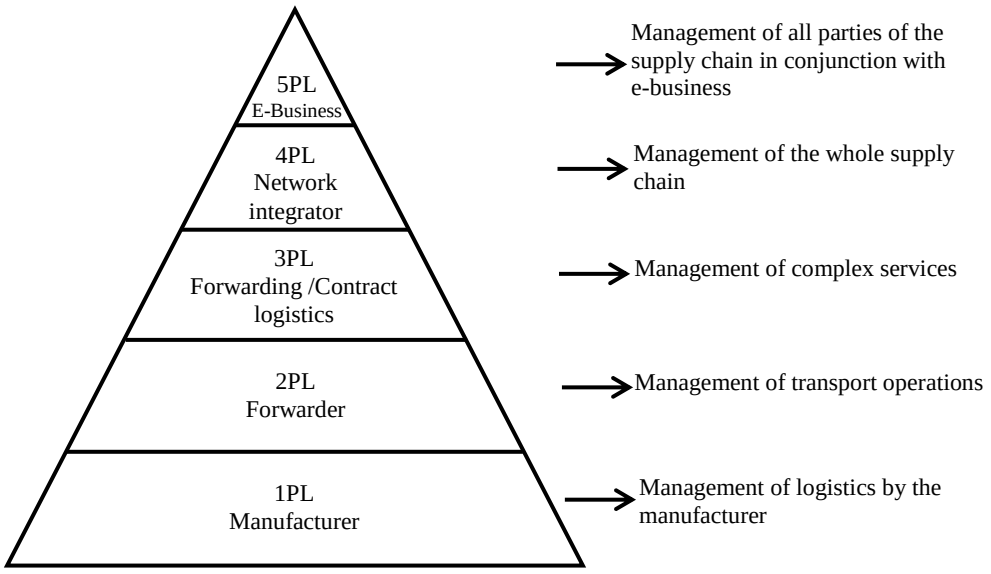


FIG.7 – *A summary of the main logistics services providers' types*

3 Panorama on logistics services providers in Morocco

Understand the evolution dynamics of LSPs, as defined in the Moroccan context, requires a brief overview of the state of logistics. Thanks to its geographical position, Morocco is projected to become the number one in logistics in Africa. Logistics is an essential part of Morocco's economic fabric. It contributes 5% of the Gross Domestic Product (GDP), employs about 300.000 people and makes a major contribution to the entire industrial and commercial fabric of the country, contributing to the growth of the country as a whole, as

well as to the balance of payments through exports and foreign direct investments (La confédération Générale des Entreprises du Maroc (CGEM), 2015).

The logistics sector is viewed as a key facilitator of Moroccan trade. In recent years, the Moroccan government has designated logistics as a strategic industry and has invested heavily in improving infrastructures (roads and railways, seaports, airports, platforms, etc). Morocco, which in 2010 had only a few dozen hectares of modern logistics platforms, now has nearly 550 ha developed in Casablanca, Tangier and other regions. In relation to this development, the contribution of public stakeholders was significant in terms of development as they proceeded to the ventilation of 87% of the developed area over the period 2010-2015. On the other hand, the contribution of private LSPs was more significant in the construction of logistics buildings with a share of 74% (La confédération Générale des Entreprises du Maroc (CGEM), 2016). As shown in figure 8, 9, 10 and 11.

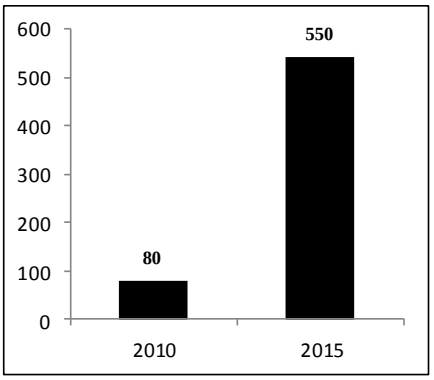


FIG.8 – *Evolution of the developed logistics area 2010-2015 (in ha)*

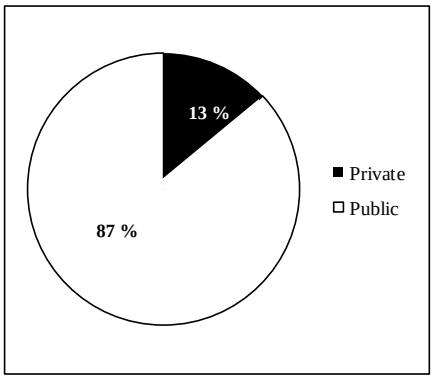


FIG.9 – *Ventilation of the developed area by type of developers since 2010*

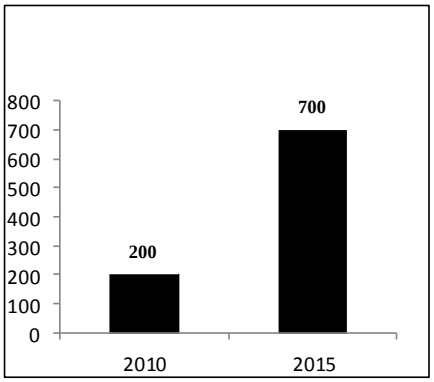


FIG.10 – *Evolution of the logistics area built in the Casablanca region in the period of 2010-2015 (in 1000 m²)*

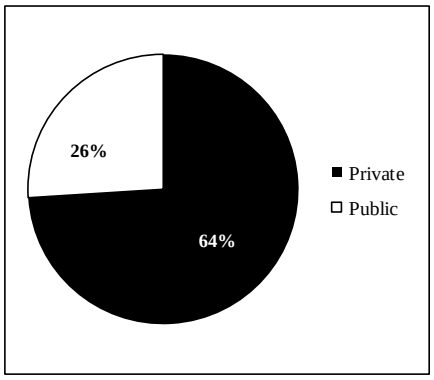


FIG.11 – *Ventilation of the built area by type of developers since 2010*

Source: Adopted from (La confédération Générale des Entreprises du Maroc (CGEM), 2016)

As regards the state of the LSP in Morocco, the concept, in its traditional sense given in the literature, is still new in Morocco and this type of company is still little used because the manufacturers remain mainly of small or average size and little inclined to externalize their logistics. It would seem, however, that the LS industry in Morocco is entering a phase of profound transformation in recent years under the pressure from major international or Moroccan companies.

In terms of developing the fabric of LSPs in the sector, since 2010 the national market has seen the advent of numerous international groups and a significant development of Moroccan LSPs. Actually, a large number of international and national specialized companies have developed their activities in Morocco. In fact, during the period 2010 to 2013, just over 5 000 logistics and transport companies were created, twice the number of business creation over the 2006-2009 period. (See figure 12).

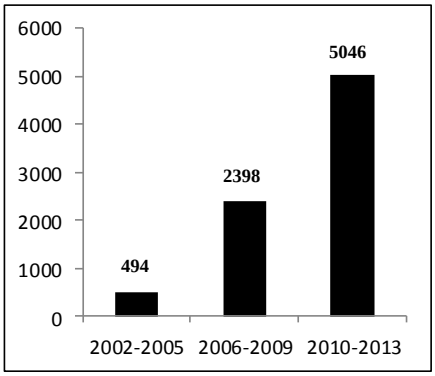


FIG.12 – *Creation of transport and logistics companies in Morocco (2002-2013) adopted from (La confédération Générale des Entreprises du Maroc (CGEM), 2016)*

It should also be noted that the Casablanca-Settat region ranks first in terms of the number of companies created (46%), followed by the Tangier-Tetouan-Al Hoceima region (15%) and the Rabat-Sale-Kenitra region (12%). (See figure 13 for more detail).

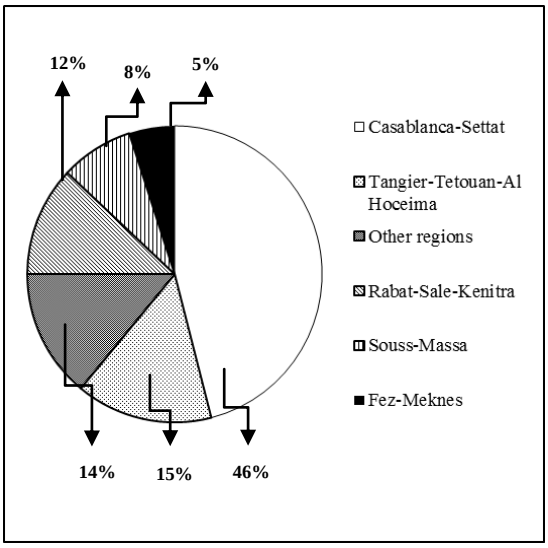


FIG.13 – *Distribution of transport and logistics companies in Morocco by region adopted from (La confédération Générale des Entreprises du Maroc (CGEM), 2016)*

Indeed, several LSPs present in the Moroccan market offer an integrated offer of LSs covering, in particular, transport, warehousing, order picking and other value-added services (labeling, copacking, etc.). The logistics market has undergone a marked evolution both by the multiplication of operators and the diversification of the offer, ranging from the simple provision of transport to the full support of the logistics functions and the customer's SC. Table 1 presents the main LSPs present in the Moroccan market and their proposed LSs.

Logistics Services Providers	Proposed Logistics Services									
	Transport			Warehousing				Shipping services		
	Execution	Organization	Freight Forwarding	Non-refrigerated storage	Refrige-rated storage	Order preparation	Value-added services	National Express	International Express	Fund return
ARAMEX										
BLUE EAGLE										
BOLLORE										
CTM										
DACHSER										
DHL										
GEFCO										
GEODIS										
GSTM										
ID LOGISTICS										
IPSEN LOGISTICS										
KUEHNE & NAGEL										
La Voie Express										
LOGICOLD										
LOGISMAR										
M&M										
MARBAR										
MORY										
MTR LM										
NUMILOG										
OMSAN										
RHENUS										
SCHENKER										
SDTM										
SJL										
SNTL										
TIMAR										
TRANSMEL										
UPS										
URBANOS										
ZIEGLER										

Service available before 2010 New services / New LSPs

TAB. 1 – The main LSPs present in the Moroccan market and their proposed LSs adopted from (La confédération Générale des Entreprises du Maroc (CGEM), 2016)

However, despite the development of LSPs and the diversification of their offers in the Moroccan market over the last years, there is still a lot to be done in this field, because it is growing in importance worldwide and especially because it still encounters obstacles that block its development and the National Federation of Road Transport (FNTR) has summarized these obstacles in nine essential points:

- The port cost is nearly 30% higher than regional competition.
- The prohibitive cost of land for setting up logistics platforms.
- The mistrust of shippers to communicate stocks, production rates and their customers.
- The small size of many shippers preventing them from bearing the costs of outsourcing their logistics.
- The weakness of the skilled workforce in this field.
- The lack of a comprehensive national strategy.
- The weakness of the purchasing of LS which, moreover, is not very diversified.
- Almost all companies offering a full range of LS are subsidiaries of European groups with a clientele of multinational companies.

4 Conclusion and future research

We have tried through this contribution, on the one hand to present in detail the different classifications of LSPs existing in the literature and we have noticed that the majority of articles read were empirical in nature which confirms that the literature on LSPs is weakly theorized. So, it could be beneficial to focus more on the production of theoretical articles so as to further enrich this area of research.

In the other hand, we have exposed the state of play of LSPs in Morocco and we have remarked that there is a total absence of studies in the literature concerning the Moroccan context, whether within companies (service users) or LSPs (outsourcers). Therefore, it would be interesting to do more studies in this respect in order to have an idea about the reality of LSPs in Morocco.

Finally, this paper may support researchers to understand the insufficiency in the logistics services providers' literature and to find the gaps for work to be accomplished in the future.

References

Berglund, M., Laarhoven, P. Van, Sharman, G., & Wandel, S. (1999). Third-Party Logistics: Is There a Future? *The International Journal of Logistics Management*.

Fielser, M., & Paché, G. (2008). La dynamique des canaux de distribution. *Revue Française de Gestion*, *34*(182), 109–133.

Forslund, H. (2012). Performance management in supply chains: logistics service providers' perspective. *International Journal of Physical Distribution & Logistics Management*, *42*(3), 296–311.

Fulconis, F., Saglietto, L., & Paché, G. (2007). Strategy dynamics in the logistics industry: a transactional center perspective. *Management Decision*, *45*(1), 104–117.

Håkansson, H., & Shenota, I. (1995). *Developing relationships in business networks*. *Routledge*.

Hertz, S., & Alfredsson, M. (2003). Strategic development of third party logistics providers. *Industrial Marketing Management*, *32*(2), 139–149.

Hosie, P., Sundarakani, B., Tan, A. W. K., & Koźlak, A. (2012). Determinants of fifth party logistics (5PL): service providers for supply chain management. *International Journal of Logistics Systems and Management*, *13*(3), 287.

La confédération Générale des Entreprises du Maroc (CGEM). (2015). *Zones Logistiques : Un Autre Choix D'Externalisation*.

La confédération Générale des Entreprises du Maroc (CGEM). (2016). *La stratégie logistique au Maroc: Bilan et perspectives de développement*.

Lai, K. hung. (2004). Service capability and performance of logistics service providers. *Transportation Research Part E: Logistics and Transportation Review*, *40*(5), 385–399.

Marchet, G., Melacini, M., Perotti, S., Sassi, C., & Tappia, E. (2017). Value creation models in the 3PL industry: what 3PL providers do to cope with shipper requirements. *International Journal of Physical Distribution & Logistics Management*, *47*(6), 472–494.

Mukhopadhyay, S. K., & Setaputra, R. (2006). The role of 4PL as the reverse logistics integrator. *International Journal of Physical Distribution & Logistics Management*, *36*(9), 716–729.

Muller, E. (1993). More top guns of third-party logistics. *Distribution*, *92*(5), 44–45.

Nemoto, T., & Tezuka, K. (2002). Advantage of Third Party Logistics in Supply Chain Management. *Technical Report*, 11–14.

Roveillo, G., Fulconis, F., & Paché, G. (2012). Vers une dilution des frontières de l'organisation: le prestataire de services logistiques (PSL) comme pilote aux interfaces. *Logistique & Management*, *20*(2), 7–20.

Rushton, A. (2006). *The Handbook of Logistics and Distribution Management*.

Rushton, A., & Walker, S. (2007). *International logistics and supply chain outsourcing: from local to global.*

Sink, H. L., & Langley, J. C. J. (1997). A managerial framework for the acquisition of third-party logistics services. *Journal of Business Logistics*, *18*(2), 163–189.

Zacharia, Z. G., Sanders, N. R., & Nix, N. W. (2011). The Emerging Role of the Third-Party Logistics Provider (3PL) as an Orchestrator • Supply Chain Risk Business Continuity Transport Vulnerability. *Journal of Business Logistics*, *32*(1), 40–54.

Logistics Services Providers: The state of play of the Moroccan context

Résumé

La logistique occupe aujourd'hui une place importante dans toutes les économies et toutes les entités commerciales et la tendance mondiale actuelle a poussé de nombreuses entreprises à externaliser leurs fonctions logistiques auprès des prestataires de services logistiques (PSLs) pour se concentrer sur leurs compétences et métiers.

Cet article s'intéresse particulièrement aux PSLs, identifie et catégorise les différents types existants de ces derniers, met en évidence leur situation au Maroc et découvre enfin les lacunes potentielles pour les recherches futures.

Closed Loop Supply Chain Network Design in the End Of Life pharmaceutical products

Mustapha AHLAQQACH*, Jamal BENHRA**
Salma MOUATASSIM***, Safia LAMRANI****

* PhD Student, LRI, OSIL Team ENSEM, CELOG-ESITH
Casablanca, Morocco
Ahlaqqach@gmail.com
** Research Director, LRI, OSIL Team ENSEM, Casablanca, Morocco
jbenhra@hotmail.com
*** PhD Student, LRI, OSIL Team ENSEM, Casablanca, Morocco
mouatassimsalma@gmail.com
**** PhD Student, LRI, OSIL Team ENSEM, Casablanca, Morocco
lamranisafia@yahoo.com

Abstract. Through this paper, we proposed a multi-objective model to design a sustainable closed-loop supply chain network taking the pharmaceutical industry as a case study. This model aims at generating economic gains, increasing the social responsibility of companies in terms of job creation and reducing the risk arising from the transport of End Of Life products (medical waste from the expiry of pharmaceutical products and their use in the healthcare centers). The multi-objective model expressed as mixed integer linear program was solved by an exact approach, this resolution allowed us to select the best compromise between the different objectives and to highlight the impact of social and societal responsibility on the design of closed loop supply chain networks.

1 Introduction

Designing a robust and efficient supply chain network (SCN) is now becoming a priority for the majority of companies that are organized as a global logistics network or part of such a network. This importance is expressed by Chopra & Mandel (2007) by the close connection between design and SCN. More than one study has considered SCN Design as the most important strategic decision in supply chain management (Klibi, Martel, & Guitouni, 2010). Nonetheless, theses studies are cost oriented and are not aligned with customer maturity seeking more and more for environmentally friendly services and products (Hong & Yeh, 2012). Alongside the cost and environmental factors, the social factor is becoming essential today and together they constitute the main pillars of sustainability defined by the World Summit of Sustainable Development (WSSD) as a balance between economic benefits, environmental protection and social developments. This sustainability has become more important nowadays due to the increase of social and environment impact of business process (Pishvaee, Razmi, & Torabi, 2014). Morocco is no exception to this international trend, particularly in pharmaceutical industries, where regulation is firm towards the availability of products, the creation of employment and control of the End Of Life (EOL) product. In fact, obsolete products in the SCN of these industries and medical waste (MW) coming from hospitals, which represent the EOL products of the studied Closed loop (CL) SCN, pose a high risk for the population (Ahlaqqach et al. 2017). Thus, these companies are required to track the EOL product while

generating economic profit, respecting the environment and creating employment opportunities. In this paper a multi-objective, multi-echelon, multi-product, sustainable supply chain network design including plants, warehouses, distribution centers, collection centers, external and internal incineration, recycling plants and landfills is considered. The main contributions of this paper that distinguish this study from related studies are as follows: (i) A sustainable multi-period, multi-product, closed-loop CLSCN that integrates inventory and location-allocation decisions; (ii) proposing a new environmental objective function to minimize the risk coming from the transport of hazardous materials (Hazmat) products; (iii) Three types of shipments are allowed in the model: direct shipment from plants to customers and indirect shipment from warehouse to customers and distribution centers to customers.

The structure of this paper is organized as follows: Section 2 provides an overview of the related literature. It is followed by the mathematical model in Section 3. Section 4 is devoted to the resolution of the problem and parameters tuning of each objective. Finally, conclusions and future research directions are provided in Section 5.

2 Literature review

As aforementioned, the pillars of sustainable development are economic, environmental and social. The first pillar was the most common factor considered in studies dealing with SCN Design. Govindan et al. (2015) presented a comprehensive literature review of more than 382 published papers in reverse logistic and CLSCN in scientific journals. The authors suggested utilizing new approaches in multi objective problems mainly applying more green, sustainable and environmental objectives. To align with this guidelines, several authors have proposed a number of Green Supply chain network (GSCN) models (Tognetti et al. , 2015), (Bing et al. , 2015), (Talaei et al, 2016), (MA et al. , 2016), (Rabbani et al., 2017). These latter have studied the trade-off between the economic objective and the environmental objective measured by emissions. However, GSCN does not cover the social dimension of sustainability. M. Zhalechian et al. (2016) presented a sustainable CL location-routing-inventory considering economic, environmental and social impact. The latter was measured by created job opportunities, however wasted energy, fuel consumption and CO_2 emission were the indicators used to assess the environmental impacts. Pedram et al.(2017) have developed a CLSCN model to cope between profit and job creation, the model take into account the EOL products. The proposed mixed integer linear programming (MILP) was solved with non-dominated sorting genetic algorithm II (NSGA II). None of the above-mentioned research works considers risk coming from the transport of Hazmat in the EOL products. As we mentioned the reverse logistic of pharmaceutical products, categorize by the World Health Organization (WHO) as MW, pose a high risk for the population. In fact, the United Nations Economic Commission for Europe (UNECE, 2014) classified MW as hazmat. Therefore, in this research, while the first objective function is defined to maximize the profit, and the second is developed to maximize job creation, another objective is to minimize risk coming from the transportation of Hazmat.

In the next section, the problem is mathematically formulated.

3 Model formulation:

3.1 Problem definition

The studied CLSCN, illustrated in Fig. 1, is a multi-echelon SCN including both forward and reverse networks. The structure contains manufacturer Centers (MFCs), warehouse centers (WHCs), distribution centers (DistCs), customers (Cs), collection centers (ColCs), safe landfill disposal centers (SLDCs), steel recycling centers (SRCs), in-house incinerator centers (HICs) and external incinerator centers (EICs) with multi-level capacities. In the forward flow, new products are manufactured by MFCs and shipped to WHCs, DisTCs and Cs. These products are shipped from WHCs to DistCs and Cs and from DistCs to Cs. Then in the reverse flow, the returned EOL products are first fully collected in ColCs and secondly shipped to SLDCs, RCs, HICs and EICs. The amount of returned products is determined as a predefined percent of obsolete products in MCFs, WHCs, DistCs and Cs inventories and MW generated by each customer category.

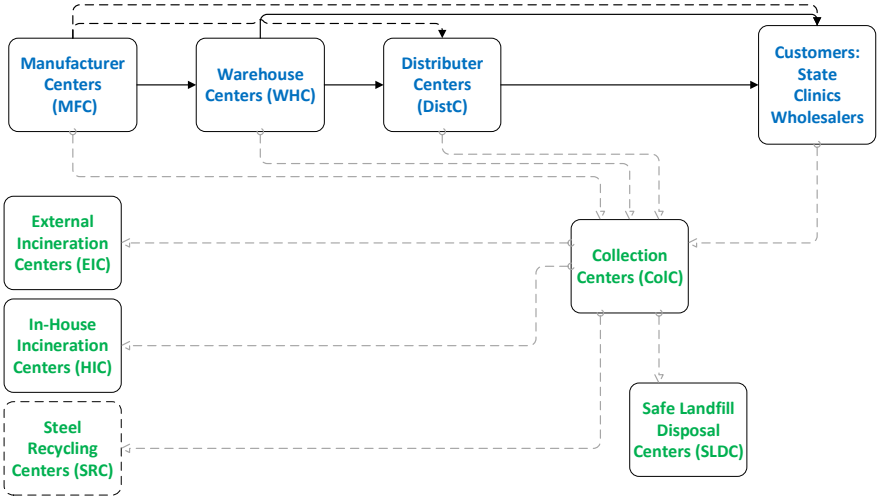


FIG. 1–*The structure of the CLSCN studied.*

We recall that we have three categories of customer: State market, wholesalers and clinics. Since the first customer is the one who accounts for the largest portion of sales and his need is defined on an annual basis, we have narrowed our case study to mono-period. But the formulation of the mathematical model is multi-period. Our approach in this model is to maximize profit and job creation and minimize risk coming from the transport of EOL products. The first objective will try to maximize sales and reduce costs, while the second will aim to increase employment opportunities by encouraging the opening of centers. On the other hand, the third one will opt for links giving off the minimum of risk during the transportation of MW. It should be pointed out that customer categories are assumed to be predetermined and constant in the network studied. Other assumptions are as follows:

1. All of the parameters are assumed to be deterministic.
2. Shortage is allowed but the customer service should be more than 70%.
3. Two main products are considered: Syringe and Saline bag.
4. Syringes are the only product involved in recycling.

3.2 Risk Calculation

Several studies have sought to determine the best risk models, in our study we will focus on risk calculation presented by Pradhananga, R. et al (2014), because it fits our case of Hazmat transportation and also of its simplicity. The main risk of transporting Hazmat comes from the possibility of accidents with significant consequences for human life and environment (Paredes-Belmar et al. , 2017). Thus the risk function in our case is related to the contamination following an accident on the road l, and will be expressed as follows:

$$R_{ij} = \sum_{(i,j)} \delta_{(i,j)} * \phi_{(i,j)} \quad (1)$$

δ_l is the probability of a Hazmat accident on the arc (i, j) and ϕ_l represent the population exposed to contamination during the accident on the arc (i, j).

3.3 Mathematical formulation:

The following notations, based mainly on Soleimani & Kannan (2015) model, are used in the formulation of the CLSCN model presented above:

Sets:

F = Existing and Potential MFC unit, indexed by ‘‘f’’.
W= Potential number of WHC, indexed by ‘‘w’’.
D = Potential number of DistC, indexed by ‘‘d’’.
C = Number of existing of customers categories, indexed by ‘‘c’’.
L= Potential number of ColC, indexed by ‘‘l’’.
S= Potential number of existing SLDC, indexed by ‘‘s’’.
R= Potential number of SRC, indexed by ‘‘r’’.
E= Potential number of EIC, indexed by ‘‘e’’.
H= Potential number of HIC, indexed by ‘‘h’’.
P= Number of product, indexed by ‘‘p’’.
A= The union (∪) of all sets indexed by ‘”a”; $A = F \cup D \cup W \cup L \cup S \cup R \cup E \cup H$.

Parameters:

D_{cpt}: Demand of product ‘‘p’’ by the customer ‘‘c’’ in period ‘‘t’’.
$PU1_{cpt}$: Unit price of product ‘‘p’’ at the customer ‘‘c’’ in period ‘‘t’’.
$PU2_{rpt}$: Unit price of product ‘‘p’’ at steel recycling center ‘’r’’ in period ‘‘t’’.
F_i: Fixed cost of the opening location ‘‘i’’.
FC_{fpt}: manufacturing capacity of MFC ‘‘f’’ of product ‘‘p’’ in period ‘‘t’’.
WC_{wpt}: WHC capacity in hours of WHC ‘‘w’’ of product ‘‘p’’ in period ‘‘t’’.
DC_{dpt}: capacity of DistC ‘‘d’’ of product ‘‘p’’ in period ‘‘t’’.
LC_{lpt}: capacity of ColC ‘‘l’’ of product ‘‘p’’ in period ‘‘t’’.
SC_{opt}: capacity of landfill center ‘‘s’’ of product ‘‘p’’ in period ‘‘t’’.
RC_{rpt}: capacity of SRC ‘‘r’’ of product ‘‘p’’ in period ‘‘t’’.
EC_{rpt}: capacity of EIC ‘‘e’’ of product ‘‘p’’ in period ‘‘t’’.
HC_{rpt}: capacity of HIC ‘‘h’’ of product ‘‘p’’ in period ‘‘t’’.
Ja_a : Number of jobs create with center a, aϵA.
Cja_a : cost of creation of jobs in center a, aϵA.
R_{ijpt}: Risk contamination of product ‘‘p’’ in road (i,j) in period ‘‘t’’.
Mc_{pt}: material cost of product ‘‘p’’ per unit supplied by suppliers in period ‘‘t’’.
Fc_{fpt}: manufacturing cost of product ‘‘p’’ per unit manufactured by ‘‘f’’ in period ‘‘t’’.

Lc_{lpt}: processing cost per unit of used product ‘‘p’’ at ColC ‘‘l’’ in period ‘‘t’’.
Sc_{spt}: processing cost per used product unit ‘‘p’’ at SLDC ‘‘s’’ in period ‘‘t’’.
Rc_{rpt}: processing cost per steel part of used product unit ‘‘p’’ at SRC ‘‘r’’ in period ‘‘t’’.
Nc_{fpt}: non-used manufacturing capacity cost of product ‘‘p’’ of ‘‘f’’ in period ‘‘t’’.
LNc_{lpt}: non-used processing cost of product ‘‘p’’ of ColC ‘‘l’’ in period ‘‘t’’.
Sc_{pt}: shortage cost of product ‘‘p’’ per unit in period ‘‘t’’.
Fh_{fp}: manufacturing time of product ‘‘p’’ per unit at MFC ‘‘f’’.
Lh'_{lp}: processing time of product ‘‘p’’ per unit at ColC ‘‘l’’.
WH_{wpt}: holding cost of product ‘‘p’’ per unit at the WHC ‘‘w’’ in period ‘‘t’’.
DH_{dpt}: holding cost of product ‘‘p’’ per unit at DistC store ‘‘d’’ store in period ‘‘t’’.
Tc_{ptij}: transportation cost of product ‘‘p’’ in period ‘‘t’’ between locations ‘‘i’’ and ‘‘j’’
RR_{pct}: return ratio of product ‘‘p’’ at each customer’s category « c » in period ‘‘t’’,
Rw_{pt}: predicted return ratio of product ‘‘p’’ at each WHC « w » in period ‘‘t’’,
Rd_{pt}: predicted return ratio of product ‘‘p’’ at each DistC « d» in period ‘‘t’’,
Rf_{pt}: predicted return ratio of product ‘‘p’’ at each MFC « f » in period ‘‘t’’,
BP_i: the minimum quantity to be served to a centre i,
M: is a large number.
Decision variables:
L_i: binary variable equals ‘‘1’’ if location ‘‘i’’ is open and ‘‘0’’ otherwise.
L_{ij}: binary variable equals ‘‘1’’ if links is established between centers ‘‘i’’ and ‘‘j’’.
Q_{ijpt}: flow of batches of product ‘‘p’’ from location ‘‘i’’ to location ‘‘j’’ in period ‘‘t’’.
R_{wpt}: the residual inventory of product ‘‘p’’ at WHC ‘‘w’’ in period ‘‘t’’,
R_{dpt}: the residual inventory of product ‘‘p’’ at DistC ‘‘d’’ in period ‘‘t’’.

The components of objective function were expressed by equations (2) to (11). We have considered three objective functions: Profit, risk and job creation. The profit is total sales minus total cost. So, Profit=W1= Sales to customer and recycle center (8) - Transportation Cost (7) – Shortage costs (6) – Non-utilized ColC capacity cost (5) – Non-used MFC capacity cost (4) – MFC, WHC and DistC operating Cost (3) – fixed opening center costs, Job creation cost, Purchased cost and Return product from customer cost (2). The risk is expressed as explained above by (9). The third objective aims to maximize the job opportunities, thus the total job opportunities expressed by the equation (10) depends on the decision of opening a center or not.

$$\sum_{a\in A}(F_a + Cja_a) * L_a + \sum_{f\in F}\sum_{t\in T}\sum_{p\in P} Q_{fpt}\, Mc_{pt} + \sum_{c\in C}\sum_{l\in L}\sum_{p\in P}\sum_{t\in T} Q_{clpt}\, PU1_{rpt} \tag{2}$$

$$\begin{aligned}&\sum_{f\in F}\sum_{w\in W}\sum_{p\in P}\sum_{t\in T} Q_{fwpt}\, Fc_{fpt+} \sum_{f\in F}\sum_{d\in D}\sum_{p\in P}\sum_{t\in T} Q_{fdpt}\, Fc_{fpt} + \sum_{f\in F}\sum_{c\in C}\sum_{p\in P}\sum_{t\in T} Q_{fcpt}\, Fc_{fpt} + \\ &\sum_{f\in F}\sum_{l\in L}\sum_{p\in P}\sum_{t\in T} Q_{flpt}\, Fc_{fpt} + \sum_{w\in W}\sum_{p\in P}\sum_{t\in T} R_{wpt}\, WH_{wpt} + \sum_{d\in D}\sum_{p\in P}\sum_{t\in T} R_{dpt}\, DH_{dpt} + \\ &\sum_{f\in F}\sum_{l\in L}\sum_{p\in P}\sum_{t\in T} Q_{flpt}\, Lc_{lpt} + \sum_{t\in T}\sum_{p\in P}\sum_{w\in W}\sum_{l\in L} Q_{wlpt}\, Lc_{lpt} + \sum_{t\in T}\sum_{p\in P}\sum_{d\in D}\sum_{l\in L} Q_{dlpt}\, Lc_{lpt} + \\ &\sum_{t\in T}\sum_{p\in P}\sum_{c\in C}\sum_{l\in L} Q_{clpt}\, Lc_{lpt} + \sum_{l\in L}\sum_{s\in S}\sum_{p\in P}\sum_{t\in T} Q_{lspt}\, Sc_{spt} + \sum_{l\in L}\sum_{r\in R}\sum_{p\in P}\sum_{t\in T} Q_{lrpt}\, Rc_{rpt} + \\ &\sum_{l\in L}\sum_{e\in E}\sum_{p\in P}\sum_{t\in T} Q_{lept}\, Ec_{ept} + \sum_{l\in L}\sum_{h\in H}\sum_{p\in P}\sum_{t\in T} Q_{lhpt}\, Hc_{hpt}\end{aligned} \tag{3}$$

$$\sum_{f\in F}\Big(\sum_{p\in P}\Big(\sum_{t\in T}\Big(\Big({}^{FC_{fpt}}\Big/_{Fc_{fp}}\Big)*L_f-\sum_{d\in D}\big(Q_{fdpt}\big)-\sum_{w\in W}\big(Q_{fwpt}\big)-\sum_{c\in C}\big(Q_{fcpt}\big)-\sum_{l\in L}\big(Q_{flpt}\big)\Big)Nc_{fpt}\Big)\Big) \tag{4}$$

$$\sum_{l\in L}\Big(\sum_{p\in P}\Big(\sum_{t\in T}\Big(\Big({}^{LC_{lpt}}\Big/_{Lh'_{lp}}\Big)*L_f-\sum_{s\in S}\big(Q_{lspt}\big)-\sum_{r\in R}\big(Q_{lrpt}\big)-\sum_{e\in E}\big(Q_{lept}\big)-\sum_{h\in H}\big(Q_{lhpt}\big)\Big)LNc_{lpt}\Big)\Big) \tag{5}$$

$$\sum_{c\in C}\left(\sum_{peP}\left(\sum_{t\in T}\big(D_{cpt}-\sum_{f\in F}Q_{fcpt}B_{fp}-\sum_{w\in W}Q_{wcpt}B_{wp}-\sum_{d\in D}Q_{dcpt}B_{dp}\big)Sc_{pt}\right)\right) \tag{6}$$

$$\begin{aligned}&\sum_{f\in F}\sum_{t\in T}\sum_{p\in P}\sum_{w\in W}Q_{fwpt}Tc_{ptfw}+\sum_{f\in F}\sum_{t\in T}\sum_{p\in P}\sum_{d\in D}Q_{fdpt}Tc_{ptfd}+\sum_{f\in F}\sum_{t\in T}\sum_{p\in P}\sum_{c\in C}Q_{fcpt}Tc_{ptfc}\\&+\sum_{f\in F}\sum_{t\in T}\sum_{p\in P}\sum_{l\in L}Q_{flpt}Tc_{ptfl}+\sum_{t\in T}\sum_{p\in P}\sum_{w\in W}\sum_{d\in D}Q_{wdpt}Tc_{ptwd}+\sum_{t\in T}\sum_{p\in P}\sum_{w\in W}\sum_{c\in C}Q_{wcpt}Tc_{ptwc}\\&+\sum_{t\in T}\sum_{p\in P}\sum_{w\in W}\sum_{l\in L}Q_{wlpt}Tc_{ptwl}+\sum_{t\in T}\sum_{p\in P}\sum_{d\in D}\sum_{c\in C}Q_{dcpt}Tc_{ptdc}+\sum_{t\in T}\sum_{p\in P}\sum_{d\in D}\sum_{l\in L}Q_{dlpt}Tc_{ptdl}\\&+\sum_{t\in T}\sum_{p\in P}\sum_{c\in C}\sum_{l\in L}Q_{clpt}Tc_{ptcl}+\sum_{t\in T}\sum_{p\in P}\sum_{l\in L}\sum_{s\in S}Q_{lspt}Tc_{ptls}+\sum_{t\in T}\sum_{p\in P}\sum_{l\in L}\sum_{r\in R}Q_{lrpt}Tc_{ptlr}+\\&\sum_{t\in T}\sum_{p\in P}\sum_{l\in L}\sum_{e\in E}Q_{lept}Tc_{ptle}+\sum_{t\in T}\sum_{p\in P}\sum_{l\in L}\sum_{h\in H}Q_{lhpt}Tc_{ptlh}\end{aligned} \tag{7}$$

$$\begin{aligned}&\sum_{f\in F}\sum_{c\in C}\sum_{p\in P}\sum_{t\in T}Q_{fcpt}PU1_{cpt}+\sum_{w\in W}\sum_{c\in C}\sum_{p\in P}\sum_{t\in T}Q_{wcpt}PU1_{cpt}+\sum_{d\in D}\sum_{c\in C}\sum_{p\in P}\sum_{t\in T}Q_{dcpt}PU1_{cpt}\\&+\sum_{l\in L}\sum_{r\in R}\sum_{p\in P}\sum_{t\in T}Q_{lrpt}PU2_{rpt}\end{aligned} \tag{8}$$

$$\begin{aligned}W_2=&\sum_{f\in F}\sum_{l\in L}\sum_{t\in T}\sum_{p\in P}R_{flpt}*L_{fl}+\sum_{w\in W}\sum_{l\in L}\sum_{t\in T}\sum_{p\in P}R_{wlpt}*L_{wl}+\sum_{d\in D}\sum_{l\in L}\sum_{t\in T}\sum_{p\in P}R_{dlpt}*L_{dl}\\&+\sum_{c\in C}\sum_{l\in L}\sum_{t\in T}\sum_{p\in P}R_{clpt}*L_{cl}+\sum_{l\in L}\sum_{s\in S}\sum_{t\in T}\sum_{p\in P}R_{lspt}*L_{ls}+\sum_{l\in L}\sum_{r\in R}\sum_{t\in T}\sum_{p\in P}R_{lrpt}*L_{lr}\\&+\sum_{l\in L}\sum_{e\in E}\sum_{t\in T}\sum_{p\in P}R_{lept}*L_{le}+\sum_{l\in L}\sum_{h\in H}\sum_{t\in T}\sum_{p\in P}R_{lhpt}*L_{lh}\end{aligned} \tag{9}$$

$$W_3=\sum_{a\in A}Ja_a*L_a \tag{10}$$

Subject to:

$$\sum_{l\in L}\sum_{t\in T}Q_{wlpt}/Rw_{pt}=\sum_{f\in F}\sum_{t\in T}Q_{fwpt}-\sum_{d\in D}\sum_{t\in T}Q_{wdpt}-\sum_{c\in C}\sum_{t\in T}Q_{wcpt}\quad\forall\, p\ in\ P, w\ in\ W \tag{11}$$

$$\sum_{l\in L}\sum_{t\in T}Q_{dlpt}/Rd_{pt}=\sum_{f\in F}\sum_{t\in T}Q_{fdpt}+\sum_{w\in W}\sum_{t\in T}Q_{wdpt}-\sum_{c\in C}\sum_{t\in T}Q_{dcpt}\quad\forall\, p\ in\ P, d\ in\ D \tag{12}$$

$$\sum_{l\in L}\sum_{t\in T}Q_{clpt}/RR_{pct}=\sum_{f\in F}\sum_{t\in T}Q_{fcpt}+\sum_{w\in W}\sum_{t\in T}Q_{wcpt}+\sum_{d\in D}\sum_{t\in T}Q_{dcpt}\quad\forall\, p\ in\ P, c\ in\ C \tag{13}$$

$$\sum_{l\in L}\sum_{t\in T}Q_{flpt}/Rf_{pt}=\sum_{d\in D}\sum_{t\in T}Q_{fdpt}+\sum_{w\in W}\sum_{t\in T}Q_{fwpt}+\sum_{c\in C}\sum_{t\in T}Q_{fcpt}\quad\forall\, p\ in\ P, f\ in\ F \tag{14}$$

$$\sum_{f\in F}\sum_{t\in T} Q_{flpt} + \sum_{w\in W}\sum_{t\in T} Q_{wlpt} + \sum_{d\in D}\sum_{t\in T} Q_{dlpt} + \sum_{c\in C}\sum_{t\in T} Q_{clpt} = \sum_{s\in S}\sum_{t\in T} Q_{slpt} + \sum_{r\in R}\sum_{t\in T} Q_{rlpt} + \sum_{e\in E}\sum_{t\in T} Q_{elpt} + \sum_{h\in H}\sum_{t\in T} Q_{hlpt} \quad \forall\, p\ in\ P, l\ in\ L \tag{15}$$

$$\sum_{f\in F}\sum_{t\in T} Q_{fwpt} + Rw_{pt-1} = Rw_{pt} + \left(\sum_{d\in D}\sum_{t\in T} Q_{wdpt} + \sum_{c\in C}\sum_{t\in T} Q_{wcpt} + \sum_{l\in L}\sum_{t\in T} Q_{wlpt}\right) \forall\, p\ in\ P, w\ in\ W \tag{16}$$

$$\sum_{f\in F}\sum_{t\in T} Q_{fdpt} + \sum_{d\in D}\sum_{t\in T} Q_{wdpt} + Rd_{pt-1} = Rd_{pt} + \left(\sum_{l\in L}\sum_{t\in T} Q_{dlpt} + \sum_{c\in C}\sum_{t\in T} Q_{dcpt}\right) \forall\, p\ in\ P, d\ in\ D \tag{17}$$

$$D_{cpt} \geq \sum_{c\in C}\sum_{t\in T} Q_{fcpt} + \sum_{c\in C}\sum_{t\in T} Q_{wcpt} + \sum_{c\in C}\sum_{t\in T} Q_{dcpt} \geq 0.7 * D_{cpt} \quad \forall\, p\ in\ P, c\ in\ C \tag{18}$$

$$\sum_{d\in D} Rd_{pt} + \sum_{d\in D} Rw_{pt} \geq 0.25 * \sum_{d\in D} D_{cpt} \quad \forall\, t\ in\ T, w\ in\ W, p\ in\ P, d\ in\ D \tag{19}$$

$$\left(\sum_{w\in W}\sum_{t\in T} Q_{fwpt} + \sum_{w\in W}\sum_{t\in T} Q_{fwpt} + \sum_{d\in D}\sum_{t\in T} Q_{fdpt} + \sum_{c\in C}\sum_{t\in T} Q_{fcpt}\right) * Fh_{fpt} \leq FC_{fpt} * Lf_f \quad \forall\, p\ in\ P, f\ in\ F \tag{20}$$

$$Rw_{pt} \leq WC_{wpt} * Lw_w \quad \forall\, p\ in\ P, w\ in\ W \tag{21}$$

$$Rd_{pt} \leq DC_{dpt} * Ld_d \quad \forall\, p\ in\ P, d\ in\ D \tag{22}$$

$$\left(\sum_{f\in F}\sum_{t\in T} Q_{flpt} + \sum_{w\in W}\sum_{t\in T} Q_{wlpt} + \sum_{d\in D}\sum_{t\in T} Q_{dlpt} + \sum_{c\in C}\sum_{t\in T} Q_{clpt}\right) * Lh'_{lp} \leq LC_{lpt} * Ll_l \quad \forall\, p\ in\ P, l\ in\ L \tag{23}$$

$$\sum_{l\in L}\sum_{t\in T} Q_{lspt} \leq SC_{lpt} * Ls_s\ , \sum_{l\in L}\sum_{t\in T} Q_{lrpt} \leq RC_{rpt} * Lr_r\ , \sum_{l\in L}\sum_{t\in T} Q_{lept} \leq EC_{ept} * Le_e\ , \sum_{l\in L}\sum_{t\in T} Q_{lhpt} \leq HC_{hpt} * Lh_h \quad \forall\, p\ in\ P, s\ in\ S, r\ in\ R, e\ in\ E\ h\ in\ H \tag{24}$$

$$L_{fj} * BP_j \leq \sum_{p\in P}\sum_{t\in T} Q_{fjpt} \leq M * L_{fj} \quad \forall f \in F, \forall j \in D \cup W \cup C \cup L \tag{25}$$

$$\sum_{f\in F} L_f \leq F\ , \sum_{w\in W} L_w \leq W\ , \sum_{d\in D} L_d \leq D\ , \sum_{l\in L} L_l \leq L\ , \sum_{s\in S} L_s \leq S\ , \sum_{r\in R} L_r \leq R, \sum_{e\in E} L_e \leq E\ , \sum_{h\in H} L_h \leq H \tag{26}$$

$$Jf_f * Lf_f \leq M * (\sum_{w\in W} L_{fw} + \sum_{d\in D} L_{fd} + \sum_{c\in C} L_{fc}) \quad \forall l \in L, \forall f \in F \tag{27}$$

$$Jw_w * Lw_w \leq M * (\sum_{f\in F} L_{fw}) \quad \forall l \in L, \forall w \in W \tag{28}$$

$$Jd_d * Ld_d \leq M * (\sum_{w\in W} L_{wd} + \sum_{f\in F} L_{fd}) \quad \forall l \in L, \forall d \in D \tag{29}$$

$$Jl_{lf} * Ll_l \leq M * (\sum_{l \in L} L_{fl} + \sum_{w \in W} L_{wl} + \sum_{d \in D} L_{dl} + \sum_{c \in C} L_{cl}) \qquad \forall l \in L, \forall l \in L \tag{30}$$

$$Ji_i * Li_i \leq M * \sum_{l \in L} L_{li} \qquad \forall i \in S \cup R \cup E \cup H \tag{31}$$

$$L_i, L_{ij} \in \{0,1\}\ R_{wpt}, R_{dpt}, Q_{ijpt} \in \mathbb{R}^+ \tag{32}$$

Waste products rate coming from each center to collect center is presented in constraints (11)-(14). Constraints (15)–(17) guarantee flow conservation in each center category. The customer's service level requirement of 70% is expressed in constraint (18). The days of supply in inventories in such industries have to insure 25% of customer demand as formulated in constraint (19). Constraints (20) and (24) are capacity constraints. Constraint (20) and (23) guarantee production capacity in manufacturing and collection centers respectively. Constraints (21) and (22) ensure the respect of holding capacity in WHC and distribution centers respectively. Constraint (24) expresses the capacity of landfill centers, SRC, EIC and in-house incineration centers. Constraint (25) manages the links between centers, when there are no flows between two centers, there should be no link between both centers. This constraint is valid for the other centers. Constraints indexed (26) manages the maximum number of allowable locations, the constraint avoid the use of more than potential center. Constraints (27) to (31) avoid the creation of job opportunities in a potential center when there is no flow between this center and other centers. Constraint (32) imposes a non-negative decision variable and characterizes the binary variable.

Our multi-objective function will be expressed as scalar function:

$$\alpha * W_1 - \beta * W_2 + \delta * W_3 \tag{33}$$

Where, α, β and δ are respectively the weight of profit, risk and job creation objectives. Instead of using the absolutes values of W_1, W_2 and W_3, respectively, we normalize them so they become comparable. We use normalization in Bronfman et al. (2016), where Y_i, U_i are the normalized objective function in case of minimization and maximization respectively. $W_{i\text{-}max}$, W_{imin} and W_i represent the maximum possible, minimum possible and actual value of each objective before normalization.

$$Y_i = \left[\frac{W_i - W_{imin}}{W_{imax} - W_{imin}}\right] \quad \& \quad U_i = \left[\frac{W_{imax} - W_i}{W_{imax} - W_{imin}}\right] \tag{34}$$

4 Solution methodology

4.1 Exact Approach: Experimentation and Results:

The experimentation of the model presented above is done on IBM ILOG CPLEX Optimization Studio 12.2. Our experiments are performed on an Intel® CORE Duo CPU with a 2.53 Ghz processor and 3 GB of RAM installed memory.

First, we present the results obtained on a real case study in order to validate the proposed model. Secondly, we illustrate the results obtained following the variation of the weights α, β and δ in order to seek an eventual compromise between the objectives studied.

4.2 Model experimentation:

Our experimentation is based on real case of a known pharmaceutical units based mainly in Casablanca city. The proposed case is a small-sized problem with 3 customers and 2 centers in each echelon. This choice is justified by the limitations of the exact approach, which experiences difficulty to deal with large instance, due to the complexity of the mathematical model. The study focuses on two main products manufactured by this company and which present more than 80% of sales. The products are: Syringes and serum. The customers of these products are divided on three categories: State hospitals, clinics and wholesalers. The annual customer demand and unit price at each customer is presented in the table 1. The computational parameters of the case are showed in table 2. The unit price and cost is in Moroccan Dirham (MAD). The fixed cost unit is Millions MAD (MMAD). The capacity unit in column 7 and 8 in table 2 is hour for MFCs and ColCs, holding capacity unit for WHCs and DistCs and production capacity unit for SLDCs, SRCs, EICs and HICs. Thanks to Google map we get addresses and coordinates of each center.

	Customer Demand per units		*Unit price at the customer*	
	Syringes(P1)	Serum(P2)	P1	P2
State hospital	1000000	5000000	60	12
Clinics	500000	1000000	62	15
Wholesalers	1500000	50000	65	17

Table 1 Annual customer demand and unit price of syringes and serum products.

	F_i **(MMAD)**	Ja_a	Cja_a **(KMAD)**	*Operating Cost* **(MAD)**		*Capacity* **(K units)**		Fh_{fpt}/Lh_{lp} **(10^{-3}hour)**		Nc_{fpt}/LNc_{lpt} **(MAD)**	
				P1	P2	P1	P2	P1	P2	P1	P2
MFC	20	500	2500	2	1	8	9	2	1	0.5	0.5
MFC	80	800	4000	4	3	8	8	2	1	0.5	0.5
WHC	2	100	500	0.5	0.75	2000	400				
WHC	4	150	750	1	1	6000	800				
DistC	1.5	40	200	1	1.25	800	100				
DistC	2	30	180	1.5	1.75	1000	150				
ColC	0.25	30	150	0.5	0.75	4	4	4	4	0.2	0.2
ColC	0.3	40	200	0.8	0.85	4	4	1	1	0.25	0.25
SLD	0	20	100	0.5	0.5	450	220				
SLD	0	10	50	4	5	300	220				
SRC	0	18	150	1	2	450	220				
SRC2	0	25	250	4	5	300	220				
EIC1	0	20	150	2	4	450	220				
EIC2	0	30	170	2.5	5	10	10				
HIC1	1	10	100	3	6	400	200				
HIC2	1.3	20	100	3.5	7.5	10	10				

Table 2 Computational study parameters.

The results from this experiment gave rise to the results shown in table 3. The experimentation validated the robustness of the model. Indeed, the resolution time of 0.22 seconds is

acceptable and the gap to relaxed solution is 10%. Also, the results obtained from this experimentation respect all the constraints linked to the model. The proposed CLSCND is relevant, we have not recorded any misbalancing between material flows coming from and going out each center. All capacities are respected and inventories are more than safety stock needed. Table 3 shows that the model chooses one MFC, one WHC, two ColCs, two SRCs and one HIC. The value of the profit objective is W1=84 Millions MAD, W2=200 people could contaminated following an accident during the risk transport of MW and W3=773 of job opportunities will be created. The solution suggested to manufacture both products at MFC1, one part of these products will be shipped to WHC1and the second part will be transported directly to customers. The obsolete products in MFC1 are shipped to Col2. The model proposed to handle safety stock in WHC1 which cover the quarter of annual demand. Beside the WHC1 is used to supply customer 1 and 2. The return products from WHC1 and customers are shipped to ColC1 and colC2. Finally, the collected products are shipped from ColC1 to SRC1 and SRC2, and from ColC2 to SRC1 and HIC1.

$\alpha=\beta=\delta$	*W1*	*W2*	*W3*	*MFC*	*WHC*	*DistC*	*LocC*	*SLDC*	*SRC*	*EIC*	*HIC*
1/3	84	200	773	1	1	0	2	0	2	0	1

Table 3 First experiment results

4.3 Experimentation of the weighting parameters:

4.3.1 Separate objectives

Firstly, we solved the presented model as three separate single-objective problems. The results are presented in table 4. As in the first experiment, W1 is expressed in Millions MAD, W2 in potential number of contaminated people, W3 in number of job created and centers in number of opening centers.

α	β	δ	*W1*	*W2*	*W3*	*MFC*	*WHC*	*DistC*	*LocC*	*SLDC*	*SRC*	*EIC*	*HIC*
1	0	0	168	295	698	1	1	0	1	2	1	1	0
0	1	0	-25.6	150	1743	2	2	2	2	0	2	0	1
0	0	1	5.54	520	1823	2	2	2	2	2	2	2	1

Table 4 Separate single-objective problems results

The results coming from this first experiment show the maximum value of W1 for $\alpha=1$, $\beta=\delta=0$, which are obvious because the solution is obtained regardless of other factors such as job creation and risk limitation. Consequently, it will focus on satisfying the customer demand to improve sales and decreasing costs in order to enhance the profit of the company. Therefore, the number of opening centers is at a minimum (43% of potential centers), thus the number of job created and the risk coming from the transport of the MW will not meet the goals of sustainable development. Whereas, in the second experiment we focus on the risk caused by the transport of the Hazmat, therefore, we recorded the minimum risk that can be generated by the network studied. However, the business experienced a deficit mainly due to the use of several capacities (70% of potential centers) and the choice of roads that are less risky but expensive in terms of the shipping cost. The third experiment, is focused on job creation, consequently we got the use of all capacities (94% of potential centers) except one HIC. The model behavior

in this case is predictable, as it will select the use of all the capacities in order to create the maximum jobs' opportunities. Nonetheless, it omits one HIC that will further strengthen the opportunities offered. This omission is explained by the constraint of the minimum quantity to be served in each center, noted BP_i, which imposes a minimum flow to guarantee the opening of a center. Since this condition is not verified, the model is forced to reject the choice of HIC despite the job opportunity offered.

4.3.2 Tuning parameters:

The second sequence of experiments aims at finding a compromise between the different objective functions, by tuning the weight values of each function and comparing two functions in an isolated way.

The first experiment ($\alpha+ \beta =1$, $\delta =0$) gave rise to the curves presented in FIG.2. As shown in FIG.2, increasing the value of α resulted in growing the profit generated by the CLSCN. The model limits the number of establishing center and chooses the minimum distance to serve each center without taking into account the risk arising from transportation of the Hazmat. Consequently, the risk grows when α value increase. We notice that for values of α more than 0.2 the profit function stabilize which mean that we don't generate big profit after this value. Whereas, for values of α greater than 0.6 the risk increases dramatically. Therefore we can conclude that the alpha values between 0.2 and 0.6 can be a very good compromise between profit and risk objectives.

The curve showed in FIG.3 comes from the second experiment ($\alpha+ \delta =1$, $\beta =0$), where the increasing the value of δ (decreasing the value of α) resulted in more job opportunities and decreasing the profit generated by CLSCN. In fact, the model establish a large number of centers in order to augment the number of created job when the function-objective give more importance to the social side (Value of $\delta>0.5$, $\alpha>0.5$). However, our model limits the establishing center to decrease cost when the profit side is more important (Value of $\delta<0.5$, $\alpha>0.5$). This explains the antagonistic behavior of the curves which present the standardized form of the profit objective and that of the job creation. The alpha values between 0.4 and 0.6 can be a very good compromise between profit and job creation objectives.

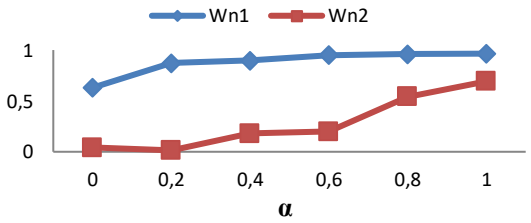


FIG. 2– *Comparison between Normalized profit Wn1 and Normalized risk Wn2*.

The third experiment ($\delta + \beta =1$, $\alpha =0$) gave rise to the curves presented in FIG.4. Both objectives have the same behavior with a dramatic fall in job opportunities for $\beta> 0.8$. We find it difficult to localize a zone of compromise between risk and the creation of opportunities. As a result we agreed, this particular case studied, to give the same weight for β and δ and launch a new experiment where we vary α with $\beta = \delta$.

The FIG.5 shows the behavior of the three curves according to α. This curve shows the possibility of a compromise between the three pillars of sustainable development for α value between 0.3 and 0.45. Thus, it can be concluded that the approach methodology followed in

this study may well seek a common areas between the various actors concerned by the EOL of Hazmat like MW.

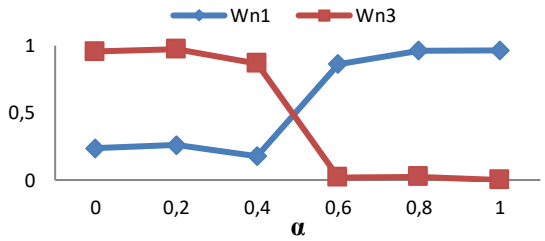


FIG. 3– *Comparison between Normalized profit Wn1 and Normalized job creation Wn3.*

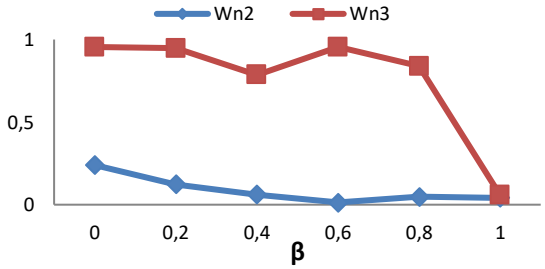


FIG. 4– *Comparison between Normalized risk Wn2 and Normalized job creation Wn3.*

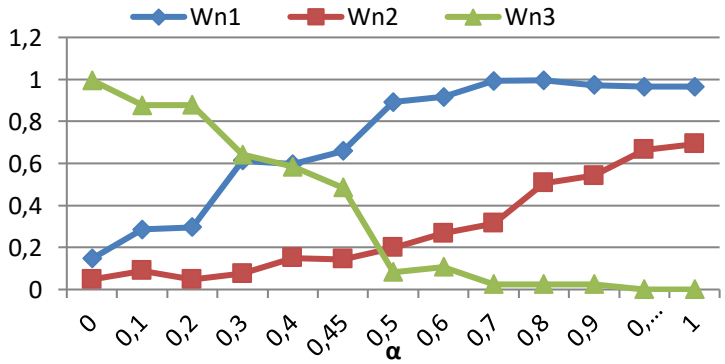


FIG. 5– *Comparison between Wn1, Wn2, Wn3. (δ = β = (1-α)/2).*

5 Conclusion

This research sheds new light on reverse logistics in the case of EOL pharmaceutical products, by proposing a multi-objective mathematical model with an exact resolution of a very complicated case of CLSCND. Our aim was to find a sustainable solution to a problem that humanity is experiencing. To this end, we have taken into consideration several realistic characteristics of the problem in order to make the model close to reality. This study developed a CLSCN model to cope with conflicting sustainable development objectives. The hazardous nature of the MW has led us to integrate the risk associated with the transport of these materials. The social responsibility of the pharmaceutical company is considered in CLSCN design

by the integration of job opportunity created in the objective function. In order to encourage the company to adopt our approach, we have integrated the profit aspect to show the possibility of coexistence between these conflicting objectives. The first experimentation with small instances gave a good result in a reasonable time. The second experiment allowed us to conclude that thanks to this model, we can seek a compromise between the different stakeholders in the project of EOL of pharmaceutical products. In our case the α value between 0.3 and 0.45 is the most user-friendly profit weight for such compromise

The reader should bear in mind that the study assumed that the parameters are deterministic. Except in reality, parameters are usually stochastic. Thus, for a better contribution in the framework of our scientific research, there is some guidance for future research. We will use metaheuristics to cope with large instances and non-deterministic approaches to deal with the nature of parameters. Also we will integrate the game theory approach to establish coalitions between distributors and customers group. In fact, This approach has been successful in creating such coalition (Mouatassim et al., 2016).

References

Ahlaqqach, M., Benhra, J., & Mouatassim, S. (2017). Optimization of routing for the collection and delivery of medical waste passing through a common warehouse. Logistique & Management, 25(1), 25-33.

Bing, X., Bloemhof-Ruwaard, J., Chaabane, A., & van der Vorst, J. (2015). Global reverse supply chain redesign for household plastic waste under the emission trading scheme. Journal of Cleaner Production, 103, 28–39.

Bronfman, A., Marianov, V., Paredes-Belmar, G., & Lüer-Villagra, A. (2016). The maxisum and maximin-maxisum HAZMAT routing problems. Transportation Research Part E: Logistics and Transportation Review, 93, 316–333.

Chopra, S., & Meindl, P. (2007). Supply chain management : strategy, planning, and operation. Pearson Prentice Hall.

Govindan, K., Soleimani, H., & Kannan, D. (2015). Reverse logistics and closed-loop supply chain: A comprehensive review to explore the future. European Journal of Operational Research, 240(3), 603–626.

Hong, I.-H., & Yeh, J.-S. (2012). Modeling closed-loop supply chains in the electronics industry: A retailer collection application. Transportation Research Part E: Logistics and Transportation Review, 48(4), 817–829.

Klibi, W., Martel, A., & Guitouni, A. (2010). The design of robust value-creating supply chain networks: A critical review. European Journal of Operational Research, 203(2), 283–293.

MA, R., YAO, L., JIN, M., REN, P., & LV, Z. (2016). Robust environmental closed-loop supply chain design under uncertainty. Chaos, Solitons & Fractals, 89, 195–202.

Mouatassim, S., Ahlaqqach, M., Benhra, J., & ELOUALIDI, M. (2016). Model based on hybridized game theory to optimize logistics case of blood supply chain. International Journal of Computer Applications, 145(15)(15), 37–48.

Paredes-Belmar, G., Bronfman, A., Marianov, V., & Latorre-Núñez, G. (2017). Hazardous materials collection with multiple-product loading. Journal of Cleaner Production, 141, 909–919.

Pedram, A., Pedram, P., Yusoff, N. Bin, & Sorooshian, S. (2017). Development of closed-loop supply chain network in terms of corporate social responsibility. PLoS ONE, 12(4).

Pishvaee, M. S., Razmi, J., & Torabi, S. A. (2014). An accelerated Benders decomposition algorithm for sustainable supply chain network design under uncertainty: A case study of medical needle and syringe supply chain. Transportation Research Part E: Logistics and Transportation Review, 67(September 2016), 14–38.

Pradhananga, R., Taniguchi, E., Yamada, T., & Qureshi, A. G. (2014). Bi-objective decision support system for routing and scheduling of hazardous materials. Socio-Economic Planning Sciences, 48(2), 135–148.

Rabbani, M., Saravi, N. A., & Farrokhi-asl, H. (2017). Design of a Forward/Reverse Logistics Network with Environmental Considerations, 4(2), 115–132.

Soleimani, H., & Kannan, G. (2015). A hybrid particle swarm optimization and genetic algorithm for closed-loop supply chain network design in large-scale networks. Applied Mathematical Modelling, 39(14), 3990–4012.

Talaei, M., Farhang Moghaddam, B., Pishvaee, M. S., Bozorgi-Amiri, A., & Gholamnejad, S. (2016). A robust fuzzy optimization model for carbon-efficient closed-loop supply chain network design problem: a numerical illustration in electronics industry. Journal of Cleaner Production, 113, 662–673.

Tognetti, A., Grosse-Ruyken, P. T., & Wagner, S. M. (2015). Green supply chain network optimization and the trade-off between environmental and economic objectives. International Journal of Production Economics, 170, 385–392.

Zhalechian, M., Tavakkoli-Moghaddam, R., Zahiri, B., & Mohammadi, M. (2016). Sustainable design of a closed-loop location-routing-inventory supply chain network under mixed uncertainty. Transportation Research Part E: Logistics and Transportation Review, 89, 182–214.

Résumé.

A travers ce document, nous avons proposé un modèle multi-objectif pour concevoir un réseau de chaîne d'approvisionnement en boucle fermée durable prenant l'industrie pharmaceutique comme cas d'étude. Ce modèle vise à générer des gains économiques, à accroître la responsabilité sociale des entreprises en termes de création d'emploi et à réduire le risque lié au transport des produits en fin de vie (déchets médicaux issus de l'expiration des produits pharmaceutiques et leur utilisation dans les hôpitaux). Le modèle multi-objectif exprimé sous la forme d'un programme linéaire mixte a été résolu par une approche exacte, cette résolution nous a permis de trouver le meilleur compromis entre les différents objectifs et de mettre en évidence l'impact de la responsabilité sociale et sociétale sur la conception des réseaux de chaînes d'approvisionnement en boucle fermée.

Optimisation par la simulation SED des moyens de manutention d'une ligne d'assemblage automobile à forte composante de main d'œuvre dans un contexte *Lean Manufacturing*: étude de cas réel

Safia LAMRANI *, Jamal BENHRA**
My Ali El OUALIDI ***, Mustapha AHLAQQACH ****

équipe OSIL, Laboratoire LRI, ENSEM, Hassan II University of Casablanca, BP : 8118, Oasis, Casablanca. Morocco.
* PHD student, safialamrani@gmail.com ;
** Research Director, jbenhra@hotmail.com ;
*** professeur habilité, eloualidi.ali@gmail.com
**** PHD student, ahlaqqach@gmail.com

Résumé : L'objectif de ce papier est d'illustrer et de documenter, via une étude de cas réelle, l'apport de la simulation à évènements discrets (SED) pour l'optimisation des moyens de manutention d'une ligne d'assemblage automobile à forte composante de main d'œuvre, dans un contexte de Lean Manufacturing. Le paramètre à minimiser est le nombre de balancelles nécessaires tout en maximisant le rendement de la ligne.

1 Introduction

Une ligne d'assemblage est un ensemble de postes de travail spécialisés disposés dans un ordre préétabli correspondant à la succession des opérations d'assemblage des composants d'un produit (Nourmohammadi & Eskandari, 2017). Il existe deux types de lignes de production manufacturière : les lignes d'assemblage et les lignes de fabrication. Le terme Ligne d'assemblage généralisée présente un mix entre les opérations de fabrication et les opérations d'assemblage. Une ligne d'assemblage peut aussi bien être manuelle, automatique ou hybride. (Saif, Guan, Wang, Mirza, & Huang, 2014)

Les lignes d'assemblage mobiles ont été introduites, pour la première fois, par Ford automobiles aux USA en 1913. Le temps de production d'un châssis d'automobile a été réduit de 12 hr 28 min à 1 hr 33 min. Avant la Ford modèle T, l'automobile était un produit réservé exclusivement aux riches. Grâce aux avantages du concept d'économie d'échelle, l'automobile est devenue un produit démocratisé abordable pour la classe moyenne. Jusqu'à nos jours, il est communément admis que si la demande d'un produit est suffisamment grande et stable pour une longue période de temps, il est généralement plus rentable d'adopter l'implantation linéaire (Clarke, 2005). Récemment, les lignes d'assemblage ont gagné en importance même dans la production de produits customisés (Mass Customisation). A cause du besoin en grand investissement requis lors de l'installation ou la conception d'une nouvelle ligne, la planification judicieuse de son implantation revêt une grande importance pour les industriels. (Boysen, Fliedner, & Scholl, 2007)

En littérature scientifique, le problème le plus abordé concernant les lignes d'assemblage est celui de leur équilibrage. (Bratcu, 2001) apporte une méthodologie de détermination systématique des graphes de précédence et d'équilibrage des lignes d'assemblage. (Corominas, Pastor, & Plans, 2008) étudient l'équilibrage d'une ligne d'assemblage de vélomoteurs avec prise en compte du niveau variable de compétences des travailleurs. Enfin, (Wickramasekara & Perera, 2016) apportent une approche améliorée pour l'équilibrage des lignes dans l'industrie du textile.

D'autres problèmes sont traités tels que la classification des lignes d'assemblage selon plusieurs paramètres (Saif et al., 2014). (Hager, Wafik, & Faouzi, 2017) proposent un processus combiné pour la conception des lignes d'assemblage qui inclue plusieurs aspects. (Rane & Sunnapwar, 2017) présentent une méthodologie pour réduire le temps de cycle et les pertes de temps dues aux principaux facteurs dont la manutention. L'objectif étant l'amélioration du rendement de la ligne d'assemblage sous des contraintes de coûts. (Fontanili, 1999) étudie l'intégration d'outils de simulation et d'optimisation pour le pilotage d'une ligne multi produits à transfert asynchrone.

Dans les lignes d'assemblage, L'investissement en moyens de manutentions peut atteindre 60 à 70% de l'investissement total. Le convoyage reste le moyen de manutention le plus adapté au transfert poste à poste. (Garcia-Diaz & Smith, 2008; Stephens & Meyers, 2013; Tompkins, White, Bozer, & Tanchoco, 2010)

(Halim et al., 2015) utilisent la simulation, sous DELMIA, pour la conception d'un système de manutention dans une ligne d'assemblage automobile. Le niveau de stocks WIP est réduit de 74% et la surface utile réduite de 18%. (Saffar, Jamaludin, & Jafar, 2017) présentent une étude d'amélioration du système de manutention dans une ligne d'assemblage automobile afin d'investiguer les changements ou les influences qui affectent la ligne.

Une multitude de productions scientifiques abordent les problématiques liées à la conception, à l'équilibrage, et à la simulation à évènements discrets (SED) des lignes d'assemblage Automobile. Néanmoins, à notre niveau de connaissance, nous constatons une relative rareté de données opérationnelles fournies par des études de cas réels. Le présent travail se base sur les données d'une étude de cas réelle d'une ligne d'assemblage automobile à forte composante de main d'œuvre. L'objectif étant de documenter et d'illustrer l'apport de la simulation SED pour la minimisation du nombre de balancelles du système de convoyage.

La suite du présent document est structurée comme suit : la seconde section présente la démarche adoptée lors de l'étude ; la troisième section donne la description physique de la ligne d'assemblage ; la quatrième section comporte la définition du problème et l'approche statique ; la section cinq explique la phase de simulation à proprement parler ; L'analyse des résultats est faite dans la section six ; et finalement, la conclusion et les perspectives de recherche sont donnés dans la section sept.

2 Démarche de l'étude

2.1 Approche étude de cas

Dans ce papier nous adoptons l'approche étude de cas(Clarke, 2005). (Yin, 2017) définit l'étude de cas, d'un point de vue recherche, comme étant : « une enquête empirique qui explore un phénomène contemporain dans son contexte en situation réelle, dans lequel les fron-

tières entre le phénomène et le contexte ne sont pas clairement évident, et dans lequel plusieurs sources de preuves sont utilisées ».

2.2 Démarche de simulation à évènements discrets

Pour mener cette étude de simulation, nous adoptons une approche en quatre étapes, représentées dans la FIG. 1: (1) Analyse du problème et compréhension du processus physique ; (2) modélisation et simulation ; (3) expérimentation sur le modèle et analyse des résultats ; (4) Rapport et conclusion. (Altiok & Melamed, 2010; Oualidi & Saadi, 2013)

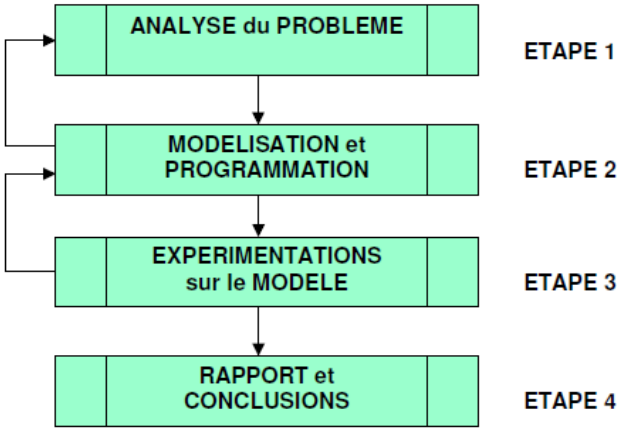


FIG. 1- Les 4 étapes d'une étude de modélisation et simulation SED

3 Description physique de La Ligne d'assemblage

L'analyse du problème et la compréhension du processus physique est la première phase pour la réalisation d'un projet de simulation des flux. C'est dans cette étape que l'on doit définir précisément ce que l'on veut mettre en évidence avec la simulation, et quelle précision on attend. Dans cette étape, il faut pouvoir fournir des données numériques et logiques au modèle. Celles-ci sont relatives à tous les éléments utilisés dans la simulation. Enfin, on doit disposer de documents graphiques afin de, d'une part, avoir une représentation géométrique du système étudié, et, d'autre part, avoir une représentation des flux. (Oualidi & Saadi, 2013)

Dans notre étude, nous utilisons, en guise de modèle géométrique, un plan de masse de la ligne d'assemblage étudiée. Réalisé sur AUTOCAD R V 2015 (Omura & Benton, 2017), ce plan synthétise le plus grand nombre d'informations pour établir notre modèle physique.

3.1 Le produit à assembler

La ligne de production étudiée est une ligne d'assemblage qui produit principalement des trains arrière de type multi-bras (FIG. 2 à droite). Les produits finis en sortie de la ligne d'assemblage sont de deux types : (1) Train arrière multi-bras avec la référence X74 ; (2)

Train arrière multi-bras avec la référence R8. Le train arrière d'un véhicule a pour fonction de supporter les deux roues à ses extrémités. Cet axe est disposé transversalement sous le véhicule. Il relie les roues au châssis. L'entrée principale de la ligne est une traverse (FIG. 2 à gauche). C'est une pièce d'appui, mise en travers pour assembler ou consolider l'ensemble train d'une voiture.

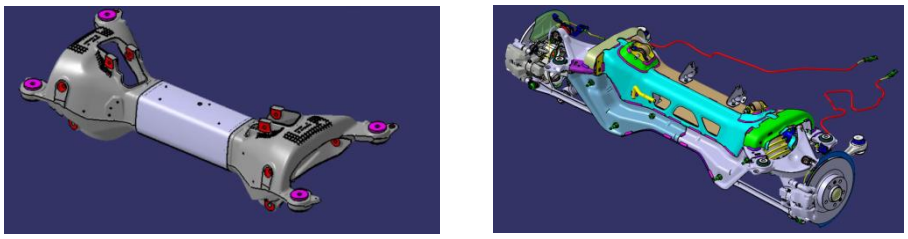


FIG. 2- Traverse (à gauche) ; Train arrière multi bras (à droite)

3.2 Les postes de travail de la ligne

Une ligne d'assemblage est un processus répétitifs : Le travail total y est divisé en plusieurs éléments (opérations indivisibles). Les opérations sont toutes suffisamment petites pour qu'une ou plusieurs opérations puissent être effectuées sur une station de travail. Une tâche est l'ensemble des opérations que doit réaliser un poste de travail. Un poste de travail (ou station) est une aire le long de la ligne qui requiert au moins un opérateur ou une machine (Dolgui & Proth J M, 2006).

Dans notre étude de cas, les postes de travail sont notés : OP10 ; OP20 ; OP30…etc. Dans le but d'assembler le train arrière, le produit doit passer par une multitude d'opérations successives dans la ligne de production. Ces opérations sont de natures variées, ils peuvent concerner : le chargement/déchargement du produit ; la mise en place des différents composants du train ; le vissage des différents points ; le contrôle de qualité du produit …etc. Les postes de travails peuvent être de plusieurs types : principal ; auxiliaire ; manuel ; automatique.

Les postes principaux sont ceux dans lesquels les opérateurs agissent directement sur le produit, ces opérations sont liées entre eux par un système de convoyage aérien (des balancelles suspendues portant le produit). Dans les postes auxiliaires, les opérateurs n'agissent pas directement sur le produit, mais travaillent sur les éléments constitutifs des trains arrière. La fonction principale de ces postes est de fournir quelques pièces élémentaires de montage pour alimenter les postes principaux.

Les postes manuels constituent 90% des postes de la ligne. Dans chaque poste de ce type, un ou plusieurs opérateurs effectuent les tâches attribuées aux postes, souvent à l'aide d'une variété d'outils.

Dans les postes automatiques, les tâches à effectuer sont entièrement réalisées par une machine, qui reçoit le produit, le traite de manière automatisée et le libère vers l'opération suivante. Ce type d'opérations constitue 10% des postes de la ligne. Les opérations entièrement automatiques dans la ligne sont des opérations de vissage qui nécessite une grande précision et rapidité d'exécution afin de garder un rythme élevé de production.

Un extrait des informations logiques et numériques relatives à la gamme opératoires est donné dans la FIG. 3. En plus du Takt time, on y trouve les informations suivantes : (1) le

nom de la station de travail selon l'ordre de traitement du produit ; (2) Le descriptif concis de la tâche à effectuer avec le nombre d'opérateurs du poste ; (3) la description de l'ensemble des opérations élémentaires constituants la tâche dans leur ordre d'exécution avec désignation de l'opérateur affecté à chaque tâche ; (4) les durées de chaque opération élémentaire en centième de minute (Cmin) ; (5) le traçage de ces durées sur le simogramme selon leur nature (manuel ; machine ; déplacement).

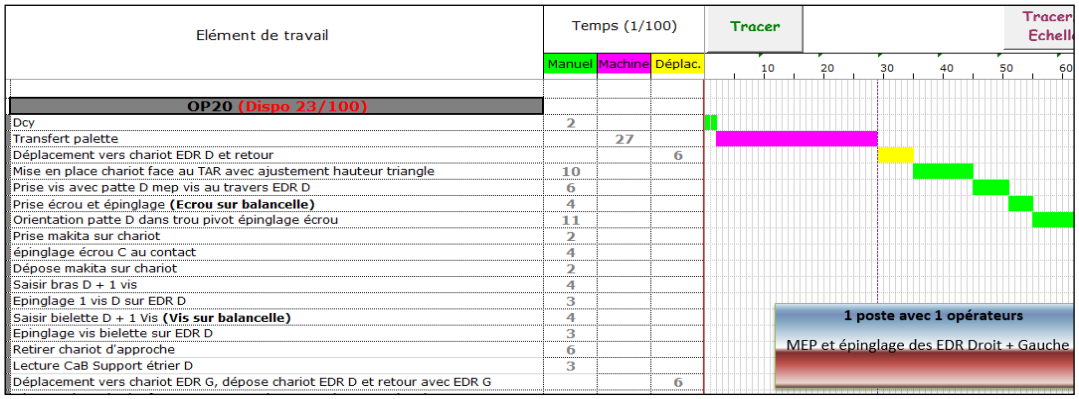


FIG. 3- Extrait du tableau contenant les données de la gamme

3.3 Moyens de manutention et logistique interne de la ligne

En entrée de la ligne, les traverses nues sont stockées dans une zone d'attente avant d'être chargées pour le traitement dans la ligne. Cette zone, qui alimente le chargement de la ligne, comprends deux grands conteneurs regroupant chacun six traverses. Chaque traverse est chargée par un opérateur à l'aide d'un préhenseur automatique sur une balancelle. Le système suspendu de manutention des traverses facilite leur manipulation et permet aux opérateurs de les déplacer à partir de leur stock initial vers le premier poste de travail de la ligne (OP10).

Avant d'entamer les opérations de traitement sur ce poste, les opérateurs doivent attendre l'arrivée de la balancelle à partir du poste (0P05), qui est un poste d'engagement automatique des balancelles. FIG. 4 donne une image d'une traverse posée sur une balancelle.

En parallèle avec le transfert de la balancelle, un opérateur s'occupe de la préparation des étiquettes à coller sur les traverses. Ces étiquettes contiennent les informations nécessaires concernant les différentes références du produit et permettent ainsi, aux opérateurs, de se focaliser sur les opérations de traitement. Au cours de ces opérations, les opérateurs peuvent se déplacer vers des chariots KANBAN afin d'apporter les éléments nécessaires dans les opérations concernées. L'équipe logistique s'occupe d'organiser et d'assurer la disponibilité de ces pièces élémentaires lors du besoin. Le produit à assembler visite l'ensemble des postes de travail de la ligne. À la fin du processus de montage, le produit final est alors prêt à être contrôlé et expédié.

FIG. 4- Traverse posée sur une balancelle (en jaune-gris)

Tout le long de la ligne d'assemblage, le transfert poste à poste est assuré, principalement, par un convoyeur aérien dans lequel circule les balancelles portant la traverse. Ce système de manutention est constitué d'une chaine attachée au plafond et des balancelles qui glissent avec cette chaine à l'aide des moteurs dédiés, ceci tout en portant le produit prêt à être traité par le poste de travail.

La ligne comprend trois types de transferts : (1) Transfert asynchrone (Stop and go) ; (2) transfert continu (Zone au défilé) et (3) Transfert avec bras robotisé et préhenseur. Le système de convoyage, utilisant les balancelles, concerne les deux premiers types de transfert.

4 Définition du problème

L'objectif de cette étude est d'explorer la possibilité de minimiser le nombre de balancelles requis tout en maximisant le rendement de la ligne. Ce rendement (RO) est calculé comme suit : $$\boldsymbol{R.O = (Q \times Tc)/To}$$

Tel que $\begin{cases} \mathrm{Q} = \text{la production hebdomadaire.} \\ \mathrm{Tc} = \text{Le temps de cycle de la ligne.} \\ \mathrm{To} = \text{Le temps d'ouverture qui est de 7440 min.} \end{cases}$

La FIG. 5 schématise les postes de travail de la ligne d'assemblage étudiée. Sur les dix-neuf postes de travail de la ligne, dix-sept postes requièrent les balancelles. Sont exclus donc les postes OP140 et OP150 qui utilisent le bras robotisé et le préhenseur pour la manutention du produit.

Selon l'analyse statique, en considérant un ratio de fabrication fixe de 50% pour chacune des deux références du produit, le nombre de balancelles doit être supérieur ou égal à dix-sept. Ce calcul trivial suppose que le système requiert au minimum une balancelle par poste de travail. En effet, par définition, les postes de travail d'une ligne d'assemblage travaillent simultanément lors du fonctionnement en cadence nominale de la ligne.

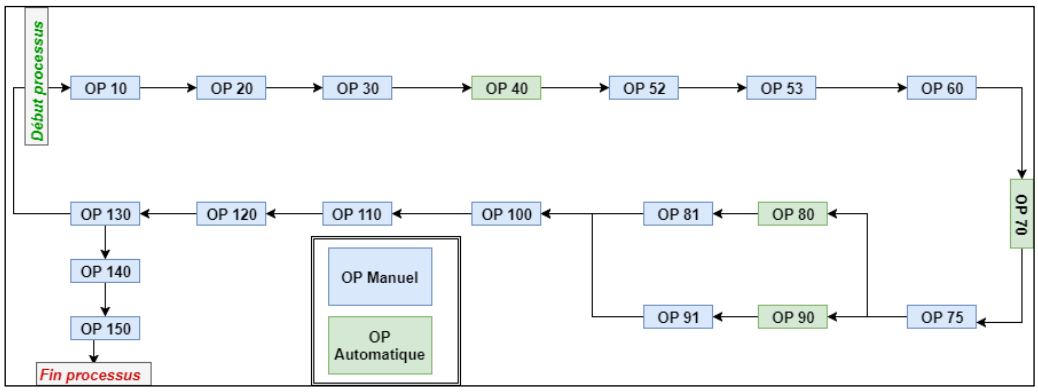


FIG. 5- schématisation des postes de travail de la ligne d'assemblage

5 Simulation de la ligne d'assemblage

La simulation SED est l'activation du modèle dans le temps, afin de connaître son comportement dynamique et prédire son comportement futur (Claver.J.F, Jacqueline G, 1996). Elle permet de tester différentes idées d'amélioration et de réorganisation en agissant sur les paramètres du système. Elle permet aussi de tester des idées de conception d'un système cible c'est-à-dire tester virtuellement des solutions alternatives sans être obligé de les mettre en œuvre préalablement(Altiok & Melamed, 2010). Cela représente un gain significatif en termes de temps et de coûts pour dimensionner les solutions proposées.

Pour la simulation de notre modèle, nous avons utilisé le logiciel Arena®. Il s'agit d'un outil graphique commercialisé par Rockwell automation® qui facilite la modélisation hiérarchique et l'animation des systèmes. Arena® connait un grand succès au regard de l'ampleur de son utilisation dans les domaines de l'industrie et de la recherche.

Après avoir récupéré les différentes données nécessaires pour notre étude, on doit synthétiser ces informations afin de construire un modèle de simulation représentatif de notre ligne d'assemblage. Le tableau (Tab. 1) résume, pour chacun des postes de travail de la ligne, les paramètres du modèle nécessaire à la simulation.

Opération	Auto/Man	Temps De cycle (minute)	Loi de traitement (min)	Distance Libre Avale	En-Cours Maxi
OP 10	MAN	1.77	TRIA(1,77; 1,87; 1,97)	3.25	1
OP 100	MAN	1.75	TRIA(1,75; 1,85; 1,95)	0	0
OP 110	MAN	1.8	TRIA(1,8 ; 1,9 ; 2)	0	0
OP 120	MAN	1.77	TRIA(1,77; 1,87; 1,97)	0	0
OP 130	MAN	0.5	TRIA(0,5 ; 0,6 ; 0,7)	13.41	4
OP 140	MAN	1.31	TRIA(1,31 ; 1,41 ; 1,51)	6.5	2
OP 150	MAN	1.8	TRIA(1,8 ; 1,9 ; 2)	Sortie	0
OP 20	MAN	1.76	TRIA(1,76; 1,86 ; 1,96)	0	0
OP 30	MAN	1.8	TRIA(1,8 ; 1,9 ; 2)	1.579	0
OP 40	AUTO	0.6	CONST(0,6)	1.611	0
OP 52	MAN	1.63	TRIA(1,63 ; 1,73 ; 1,83)	0	0
OP 53	MAN	1.49	TRIA(1,49 ; 1,59 ; 1,69)	0.901	0
OP 60	MAN	1.77	TRIA(1,77; 1,87; 1,97)	8.94	2
OP 70	AUTO	0.59	CONST(0,59)	5.199	1
OP 75	MAN	1.48	TRIA(1,49 ; 1,59 ; 1,69)	6,91 OP80 / 6,26 OP90	2 OP80 / 1 OP90
OP 80	AUTO	1.69	CONST(1,69)	0.338	0
OP 81	MAN	1.36	TRIA(1,36 ; 1,46 ; 1,56)	7.41	2
OP 90	AUTO	1.69	CONST(1,69)	2.432	0
OP 91	MAN	1.36	TRIA(1,36 ; 1,46 ; 1,56)	8.95	2
Total		27.92			

Tab. 1-paramètres du modèle de simulation

6 Analyse des résultats

Comme nous l'avons mentionné précédemment, sur la base de l'approche statique, le nombre de balancelles doit être supérieur ou égal à dix-sept stations. Cependant, cette approche triviale ne prend pas en compte le fait que les durées opératoires manuelles suivent une loi de probabilité Triangulaire. De même, les pertes d'équilibrage dans les différentes stations de la ligne ne sont pas considérées. Ces pertes induisent des variabilités entre les temps opératoires des postes.

L'aspect aléatoire qui caractérise les tâches manuelles est dû, dans la réalité du terrain, à la difficulté qu'à la main d'œuvre à tenir le même temps de cycle du fait de : la complexité de la tâche ; des erreurs éventuelles ; de la fatigue due à la répétitivité du travail…etc.

Dans notre modèle, ce comportement stochastique est exprimé par une loi triangulaire (TRIA) pour tous les postes manuels de la ligne. Cette loi permet de supposer que la durée du poste varie entre un minimum et un maximum. Notons que la durée idéale de la tâche correspond à la durée la plus courte (valeur minimale). Les postes automatisés, quant à eux, suivent une loi constante (CONST) qui ne présente aucune variabilité.

En second lieu, dans notre étude, la solution d'équilibrage nous a été imposée par la solution de conception de la ligne. Le calcul statique suppose que tous les postes de travail ont les mêmes durées opératoires qui correspondent à la durée de la station la plus lente. Celle-ci

doit être subordonnée au temps de cadencement de la ligne (Takt Time) imposé par la productivité hebdomadaire requise.

Suite aux résultats de la simulation, nous constatons que la ligne atteint son rendement maximal de 99.7% après l'introduction de 13 balancelles. Il est donc inutile, au regard de cet indicateur de performance, d'ajouter d'autres balancelles. Ce résultat est illustré dans la FIG. 6.

Ainsi, grâce à la simulation, nous avons démontré que treize balancelles, au lieu de dix-sept, sont amplement suffisantes pour assurer le rendement maximal de la ligne. Ce gain peut être expliqué par un phénomène de compensation entre, d'une part, la variabilité des tâches manuelles et les pertes d'équilibrage et, d'autre part, la disponibilité des balancelles.

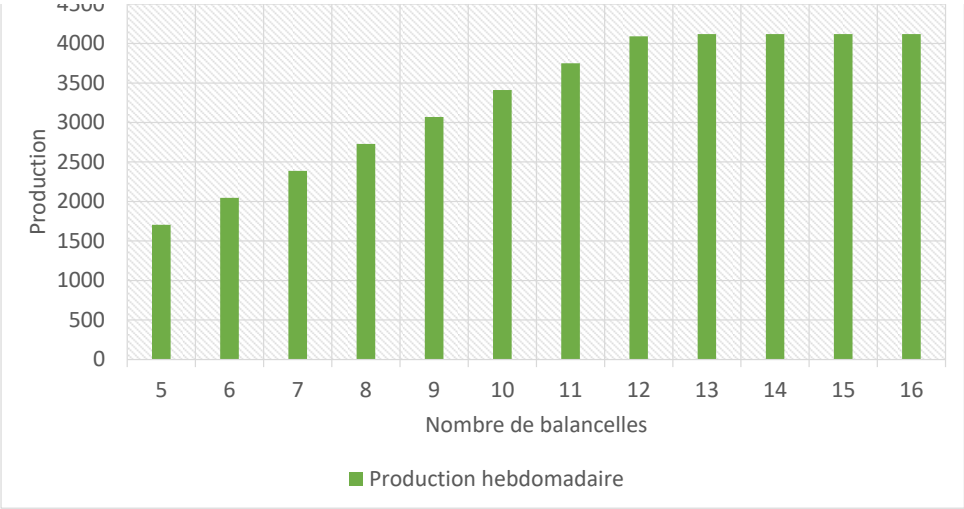


FIG. 6- production hebdomadaire en fonction du nombre de balancelles

7 Conclusion et perspectives

La présente étude s'est basée sur les données réelles d'une ligne d'assemblage automobile à forte composante de main d'œuvre dans un contexte de Lean Manufacturing. Elle a permis de démontrer l'apport de la simulation SED dans l'optimisation de l'investissement en équipements de manutention dans ce type de système de production manufacturière. Le système de transfert étant une solution de convoyage aérien, le paramètre à minimiser est le nombre de balancelles requis tout en maximisant le rendement de la ligne.

L'approche statique suggère que le nombre de balancelles soit supérieur ou égal aux nombre de postes de travail. Dans notre cas, dix-sept postes nécessitent des balancelles. Cependant, cette approche triviale ne prend en compte ni le caractère aléatoire lié aux opérations manuelles, ni la variabilité des temps de cycle des différentes stations dues aux pertes d'équilibrage. La simulation a donc permis d'explorer l'impact de ces deux facteurs sur le nombre requis des balancelles et sur le rendement de la ligne. Le gain de quatre balancelles est un résultat significatif qui justifie amplement l'intérêt d'avoir recours à la simulation pour optimiser l'investissement en moyens de manutentions lors de l'implantation ou la réimplantation des lignes d'assemblage.

Le lecteur devra noter que cette étude est basée sur des hypothèses. En effet, l'étude ne prend pas en compte un certain nombre de paramètres pouvant affecter le rendement de la ligne. Nous en citons : l'indisponibilité des équipements ; les arrêts dus aux changements d'outils ; les accidents ou incidents liés à la sécurité ; la rupture d'approvisionnement …etc. ces différents aspects présentent pour nous autant de perspectives de recherches futures.

References

Altiok, T., & Melamed, B. (2010). *Simulation modeling and analysis with Arena.* Academic press.

Boysen, N., Fliedner, M., & Scholl, A. (2007). A classification of assembly line balancing problems. *European Journal of Operational Research*, *183*(2), 674–693.

Bratcu, A. (2001). *Détermination systématiques des graphes de précédence et équilibrage des lignes d'assemblage. Thèse de Doctorat.* Université de Franche-Comté. France.

Clarke, C. (2005). *Automotive production systems and standardisation: from Ford to the case of Mercedes-Benz.* Springer Science & Business Media.

Claver.J.F, Jacqueline G, D. P. (1996). *Gestion de flux de production. modélisation et simulation.* Edition Hermès.

Corominas, A., Pastor, R., & Plans, J. (2008). Balancing assembly line with skilled and unskilled workers. *Omega*, *36*(6), 1126–1132. https://doi.org/10.1016/j.omega.2006.03.003

Dolgui, A., & Proth J M. (2006). *Les systèmes de production modernes*. Hermès Science publications.

Fontanili, F. (1999). *Intégration d'outils de simulation et d'optimisation pour le pilotage d'une ligne d'assemblage multiproduit à transfert asynchrone*. Paris 13. France. Retrieved from http://cat.inist.fr/?aModele=afficheN&cpsidt=199926

Garcia-Diaz, A., & Smith, J. M. (2008). *Facilities planning and design*. Prentice Hall.

Hager, T., Wafik, H., & Faouzi, M. (2017). Manufacturing system design based on axiomatic design: Case of assembly line. *Journal of Industrial Engineering and Management*, *10*(1), 111.

Halim, N. H. A., Yusuf, N., Jaafar, R., Jaffar, A., Kaseh, N. A. in, & Azira, N. N. (2015). Effective Material Handling System for JIT Automotive Production Line. *Procedia Manufacturing*, *2*(February), 251–257. https://doi.org/10.1016/j.promfg.2015.07.044

Nourmohammadi, A., & Eskandari, H. (2017). Assembly line design considering line balancing and part feeding. *Assembly Automation*, *37*(1), 135–143. https://doi.org/10.1108/AA-09-2016-122

Omura, G., & Benton, B. (2017). Installing and Setting Up AutoCAD. *Mastering AutoCAD® 2017 and AutoCAD LT® 2017*, 985–1014.

Oualidi, M. A. El, & Saadi, J. (2013). Améliorer la prise en charge des urgences : apport de la modélisation et de la simulation de flux. *Santé Publique*, *4*(25), 433–439.

Rane, A. B., & Sunnapwar, V. K. (2017). Assembly line performance and modeling. *Journal of Industrial Engineering International*, *13*(3), 347–355. https://doi.org/10.1007/s40092-017-0189-7

Saffar, S., Jamaludin, Z., & Jafar, F. A. (2017). Improving Material Handling System Performance in Automotive Assembly Line Using Delmia Quest Simulation. In *Asian Simulation Conference* (pp. 468–482). Springer.

Saif, U., Guan, Z., Wang, B., Mirza, J., & Huang, S. (2014). A survey on assembly lines and its types. *Frontiers of Mechanical Engineering*, *9*(2), 95–105. https://doi.org/10.1007/s11465-014-0302-1

Stephens, M. P., & Meyers, F. E. (2013). *Manufacturing facilities design and material handling*. Purdue University Press.

Tompkins, J. A., White, J. A., Bozer, Y. A., & Tanchoco, J. M. A. (2010). *Facilities planning*. John Wiley & Sons.

Wickramasekara, A. N., & Perera, H. S. C. (2016). An Improved Approach to Line Balancing for Garment Manufacturing. *Vjm*, *2*(1), 23–40.

Yin, R. K. (2017). *Case study research and applications: Design and methods*. Sage publications.

Summary

This paper aims to illustrate and document the benefit of simulation at discrete events in the process of optimizing of material handling equipment for an automotive assembly line into lean manufacturing context. Real data are collected via the CAD plan of a real case study of production line and stochastic manual operations time law. An air conveyer system is used to transfer the product from station to station. We aim to minimize the number of product circulating at the same time on the line in order to minimize material handling investment without loss in the production line ratio of output.

Game theory model applied to distribution network optimization: A confrontation between biform and cooperative game

Salma MOUATASSIM*, Mustapha AHLAQQACH**
Jamal BENHRA***

*OSIL Team, LRI Laboratory, ENSEM, Hassan II University of Casablanca, BP : 8118, Oasis, Casablanca. Morocco.
s.mouatassim@ensem.ac.ma
** OSIL Team, LRI Laboratory, ENSEM, Hassan II University of Casablanca, BP : 8118, Oasis, Casablanca. Morocco.
CELOG-ESITH
ahlaqqach@gmail.com
http://www.une-autre-page.html
*** OSIL Team, LRI Laboratory, ENSEM, Hassan II University of Casablanca, BP : 8118, Oasis, Casablanca. Morocco.
jbenhra@hotmail.com

Abstract. This work analyzes a two-step decision problem for regional distribution centers sharing the same product families. It gives a comparison of biform and cooperative strategies. For the biform game, at the first, in an uncertain environment, each distribution center must define the quantities to be ordered from the production units in order to maximize its own gain. When the demand is deterministic, the centers collaborate to meet their local demand. On the other hand, for the cooperative game, distribution centers collaborate in the two steps. They communicate demands information not only in the deterministic environment but also in the uncertain step. We study a case of bottling company in Morocco; we compute the costs generated by each strategy and analyze results.

1 Introduction

Companies find it difficult to manage their supply chain due to variations of market demand. Several factors affect the market and can sometimes be unpredictable and uncontrollable. Forecasting future demand in a relevant way can help companies to face this variation. However, in the case of unforeseeable events companies can resort to collaboration. The total distribution cost of a logistic coalition is generally between 9% and 30% lower than the sum of costs of each partner distributing separately (Vanovermeire *et al.*, 2014).

Distribution centers, often encounter the phenomenon of bullwhip effect (Tsiakis, Shah and Pantelides, 2001), this makes the real demand unpredictable. In addition to a relevant forecast, collaboration between distribution centers is a good solution in this case. It can take different forms; companies can collaborate on several levels and share all or some of their demand information.

The following work falls within this framework. It presents a comparison of two strategies of collaboration through biform and cooperative game. It considers a two-strep decision problem. At first, the demand is uncertain; distribution centers have to forecast quantities to be ordered from the production units. Secondly, the demand becomes deterministic.

A biform game can be interpreted as a non-cooperative game, but having cooperative games as results. Within the strategic framework, each center decides the quantity ordered at the production unities according to its own economic interest and the information it possesses on the forecast of local demand. When the demand is deterministic, the differences between the forecast and the real demand can then be compensated by collaboration between the different centers by exchanging products(Triqui Sari and Hennet, 2016). The coupling between the strategic game of the first stage and the cooperative game of the second stage generates a biform game (Brandenburger and Stuart, 2007).

A cooperative game means that companies communicate their information and aim to optimize the global profit. Demand is forecasted for all centers in the uncertain environment. The distribution of products is made when demand becomes deterministic. The cooperative game allows optimizing the gains and minimizing the overall cost (Charles and Hansen, 2008).

We find several works dealing with game theory. However, the biform game is rarely treated by researchers. In addition, most existing research works focus on cost allocation only. Our work broadens the field of study and takes into account demand forecasting in an uncertain environment and make a comparison between cooperative and biform game.

This paper is organized as follow: Section 2 presents a detailed analysis of the problem addressed in supply planning. Section 3 shows the application of the proposed approach to a real case and comparison between biform and cooperative game strategies. Section 4 concludes the work.

2 Supply planning

Distribution centers should face demand uncertainty. A two-stage decision approach is developed in order to solve this problem as a biform and cooperative game. The demand is supposed uncertain when distribution centers order products from suppliers. When the clients confirm their orders, the demand becomes deterministic. Distribution centers should then counterbalance the offset between forecast and real demand.

A good forecast is primordial for collaboration between centers, forecasting is the process of predicting future demand variation. Demand in most cases is unknown and always biased with uncertainty. A proper forecasting or any approximation algorithm will provide us with proper working data for the collaborative game.

Several forecasting methods and techniques do exist, we can name the single moving average SMA, the exponential smoothing forecasting, time series with ARIMA (Zhang, 2003), and many more (Dalrymple, 1975; Holt, 2004). For this we'll benchmark under professional software called GMDH shell for experimenting (Dag and Yozgatligil, 2012) with various forecasting techniques.

Neural forecasting or forecasting using artificial neural networks is a widely used technique that displayed its proficiency in dealing with large datasets and complex demand forecasting (Zhang, Eddy Patuwo and Y. Hu, 1998), the neural networks emulate the human brain through a complex set of graphical models and learning algorithms in order to predict

future data. Neural forecasting is used in demand forecasting and financial forecasting (Yao and Tan, 2001)(Benkachcha, Benhra and El Hassani, 2013), for the present work we'll be using neural networks through GMDH for data classification and forecasting based on the forecast performance and general RMSE (Chai and Draxler, 2014).

2.1 Biform game

2.1.1 Strategic game

Demand is supposed uncertain; each distribution center foresees its own demand, aiming to optimize its own gain function.

Once the demand is forecasted, the supply problem is resolved to know which quantity order and from which supplier. The objective is to optimize the total cost including transportation and purchasing cost.

The model used to optimize supplying cost in case of multi-suppliers and multi-customers is:

Data:

S: set of suppliers

C: set of distribution centers

D_j: forecasted demand

p_i: purchasing cost

t_{ij}: transportation cost

cap_i: capacity of supplier

Decision variables:

Q_{ij}: quantity ordered from supplier i by center j

Objective function:

$$\text{minimize} \sum_{i \in S} \sum_{j \in C} q_{ij} * (p_i + t_{ij}) \quad (1)$$

Constraints:

$$\forall j \in C \sum_{i \in S} q_{ij} \geq D_j \quad (2)$$

$$\forall i \in S \sum_{j \in C} q_{ij} \leq cap_i \quad (3)$$

$$\forall i \in S \, \forall j \in C \; q_{ij} \geq 0 \quad (4)$$

The objective function (1) aims to optimize the total cost, considering the transportation and purchasing cost. Constraint (2) ensures that no center receive more than its demand, while constraint (3) states that the total quantity transiting from supplier i to all distribution centers don't exceed its production capacity cap_i . Constraint (4) corresponds to the positivity of quantities q_{ij} ordered by center j from supplier i.

2.1.2 Cooperation

After the confirmation of clients order, demand is known with certainty. Distribution centers must compensate the offset between physical inventory and real sales. They collaborate and exchange products so as to meet their local demand.

2.2 Cooperative game

Companies communicate their demand information and aim to optimize the global profit. Demand is forecasted for all centers in the uncertain environment. Two possibilities are available in this case:

In the first configuration, we sum demand of all centers and then forecast the global demand. We reserve capacity from supplier while minimizing transportation and purchasing costs using MILP (1). Once the demand is deterministic, products are distributed according to the effective demand and the same MILP is used to optimize supply cost. For the second possibility, we use the history of each center to forecast its demand. We sum forecasted demand and reserve capacity following the same process.

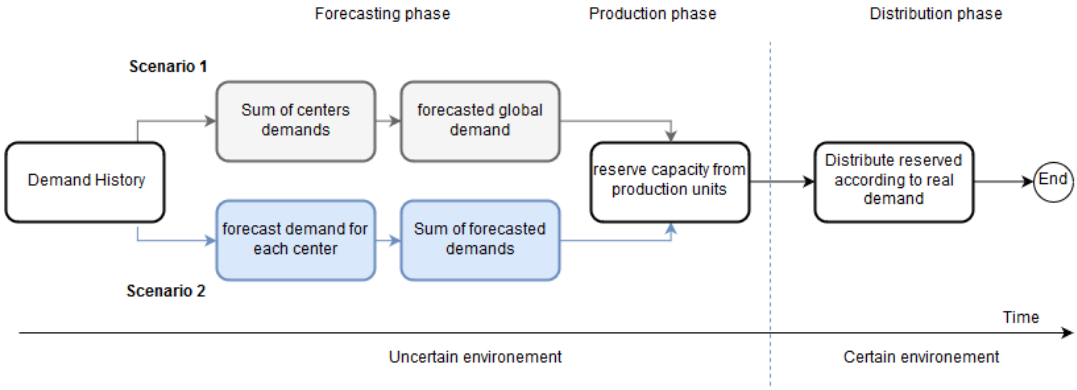


FIG. 1 – *Possible scenrios of cooperation.*

The forecast error is not the same for the two configurations. Demand variation is different for each center, the summation of demand history before forecasting increase considerably the forecast error (Prak, Teunter and Syntetos, 2017). We use the second possibility to minimize the gap between forecasted and effective demand.

In the following, we use biform game, at first, to manage the horizontal collaboration. We forecast demand for each center using neural networks through GMDH Shell. When demand becomes deterministic, distribution centers cooperate and exchange products to meet their local demand. We compute supply and cooperation cost. Secondly, we use cooperative game approach. We estimate each center demand to calculate the total forecast demand. In this case, future trade is anticipated in the first step. We reserve capacity from production units, while minimizing the future transportation cost. Once demand is known with certainty, we proceed to the distribution of products. Cost engendered by this strategy is calculated and compared to the biform strategy cost.

3 Case study

In this section, we study an application case of bottling company with four distribution centers and four production unities. We focus on the four distribution centers.

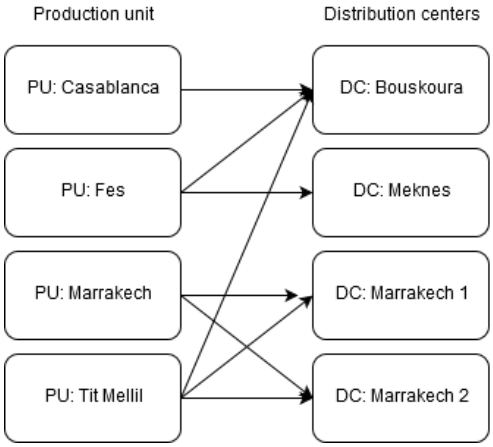


FIG. 2 – *Diagram of two echelons of distribution network.*

3.1 Cooperative game

Each distribution center forecasts quantity to order from suppliers. Once the demand is known, distribution centers collaborate in order to meet costumers demand.

3.1.1 Forecast

A history of three years of data is used to forecast a demand of a few months; we use neural network through GMDH Shell to forecast demand in this study.

The software provides forecast demands for each distribution center, and a confidence band of 5%. Other accuracy indicators are calculated and optimized.

	Bouskoura	Meknes	Marrakech 1	Marrakech 2
Forecast demand	737 820	183 220	303 296	229 855

TAB. 1 – *Forecast demand.*

3.1.2 Supply problem

After forecasting demands, supply problem is resolved in order to optimize transportation and purchasing costs. The mixed integer linear Programming model (MILP) of supply problem (1) is implemented and resolved using the following data for one month:

	Casablanca	Tit Mellil	Marrakech	Fez
Purchasing cost	18	14.4	14.46	17.7

TAB. 2 – *Purchasing cost.*

The result bellow is justified, distribution centers save more by purchasing products from Tit Mellil and Marrakech, as the purchasing cost proposed by these production unities is lower than the others. The difference of purchasing costs is very important comparing to the difference between transportation costs.

	Bouskoura	Meknes	Marrakech1	Marrakech2
Casablanca	0	0	0	0
Tit Mllil	737820	183220	0	0
Marrakech	0	0	303296	229855
Fès	0	0	0	0
Supply cost			**2.130.801**	

TAB. 3 – *Solution of supply problem.*

The real demand is a value of the confidence band. We consider a random number of confidence bands of each distribution center as its real demand.

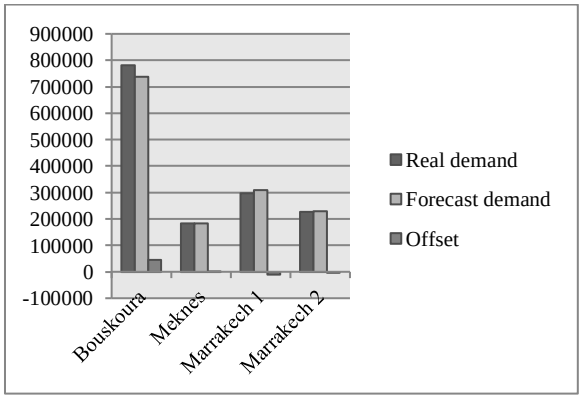


FIG. 3 – *Offset between real and forecast demand.*

3.1.3 Cost of the grand coalition

In order to take into account the costs of non-satisfaction of requests and the costs of losses in the case of non-use of product, a mathematical model of regulation of stocks is developed. The main purpose of this study is to satisfy the applications at less cost (Mouatassim *et al.*, 2016).

The mathematical model used to calculate the optimal cost of each coalition is presented below.

Sets:

N: Set of players

K: Set of coalitions

N_{1k} : Set of surplus players belonging to the coalition k

N_{2k} : Set of deficit players belonging to the coalition k

Data:

E_i : gap between the forecast demand and the real demand of a surplus player

D_j : gap between the forecast demand and the real demand of a deficit center

d_{ij} : distance between the player I and the player J

C : cost of transport per km

C_d : cost per lost product

C_{ns} : unit cost of non-satisfaction

Decision Variables:

Q_{ij} : entire quantity which passes from the center i to the center j

$$Xij = \begin{matrix} 1 & \text{if } Q_{ijk} \neq 0 \\ 0 & \text{else} \end{matrix}$$

Objective function:

$$\min C_k = \sum_{i=1}^{N1k} \sum_{j=1}^{N2k} c * d_{ij} * x_{ij} + \sum_{i=1}^{N1} PP_i * C_d + \sum_{j=1}^{N2} DNS_j * C_{ns} \quad (5)$$

Constraints:

$$\alpha = \begin{cases} 1, & si \sum_{i=1}^{N1k} E_i \geq \sum_{j=1}^{N2k} D_j \\ 0, & sinon \end{cases}$$

$$\beta = \begin{cases} 1, & \text{si } \sum_{i=1}^{N1k} E_i > \sum_{j=1}^{N2k} D_j \\ 0, & \text{sinon} \end{cases}$$

$$Ei - M * \beta \leq \sum_{j=1}^{N2k} Q_{ij} \leq E_i \qquad \forall i \in \{1 \dots N1\} \tag{6}$$

$$D_j - M * (1 - \alpha) \leq \sum_{i=1}^{N1k} Q_{ij} \leq D_j \qquad \forall j \in \{1 \dots N2\} \tag{7}$$

$$PP_i = E_i - \sum_{j=1}^{N2} Q_{ij} \qquad \forall i \in \{1 \dots N1\} \tag{8}$$

$$DNS_j = D_j - \sum_{i\_=1}^{N1} Q_{ij} \qquad \forall j \in \{1 \dots N2\} \tag{9}$$

$$\frac{Q_{ij}}{M} \leq x_{ij} \leq Q_{ij} \qquad \forall i \in \{1..N1\}\ \forall j \in \{1..N2\} \tag{10}$$

$$Q_{ij}, PP_i, DNS_j \geq 0; \quad x_{ij} \in \{0,1\}\ \forall i \in \{1..N1\}\ \forall j \in \{1..N2\} \tag{11}$$

The objective function allows minimizing the cost of each coalition, by taking into account the cost of transportation, the cost generated by the products non-used as well as the cost of non-satisfaction of requests. The constraints (6) and (7) shall ensure that the quantity transferred from a surplus player i does not exceed its surplus and that each player in deficit j does not receive more than its deficit, taking into consideration the different possible cases ($\sum_{i=1}^{N1k} Ei \geq \sum_{j=1}^{N2k} Dj$; $\sum_{i=1}^{N1k} Ei < \sum_{j=1}^{N2k} Dj$). The constraint (8) calculates the lost products or non-used and the constraint (9) calculates the non-satisfied requests, while the constraint (10) binds the two variables of decision x_{ij} and Q_{ij}. The last constraint (11) defines the binary nature of x_{ij} and the positivity of Q_{ij}, PP_i, DNS_j.

The implementation of this MILP provides cost generated by the grand coalition which is **769.124**

The total cost of management strategy through biform game is the sum of supply and co-operation cost: **22.073.925**

3.2 Cooperative game

We use the history of each center to forecast its demand. We can retain the previous results of forecasting. We reserve capacity from production units while minimizing the future supply cost.

	Casablanca	Tit Mllil	Marrakech	Fez
Reserved capacity	0	921040	533151	0

TAB. 4 – *Reserved capacity from each center.*

While demand is deterministic, products according to the reserved capacity are distributed. We optimize supply to minimize cost.

	Bouskoura	Meknes	Marrakech 1	Marrakech 2
Casablanca	0	0	0	0
Tit Mllil	693667	182656	0	19906
Marrakech	0	0	318944	214207
Fez	0	0	0	0
Total cost		**20.972.424**		

TAB. 5 – *Solution of supply problem.*

The total cost of management strategy through cooperative game is the sum of supply cost and reservation cost of non-used capacity: **21.870.548**

From the results obtained, we notice that the same total quantity was forecasted. But it is distributed differently in both models. We also find that the cooperative model, where future trade is anticipated in the first step, results in a higher profit. The absolute gain expected is of the order of **203.377**, which corresponds to a profit increase of about 1%. This augmentation is relatively important according to the literature (Triqui-sari, 2014). It is accompanied by an improvement of the overall service. Product transportation and its availability date are optimized.

4 conclusion

This work compares two strategies of a two-step decision-making problem. The first strategy uses a biform game, when the second is based on cooperative game. The problem is composed of two phase, an uncertain environment, which requires a good forecast; we use neural networks through GMDH Shell software. For the second step demand is deterministic, centers have to meet the local market demand.

Through a biform game, each center forecasts its own demand and optimizes its gain function. Once the customers confirm their demand, centers collaborate and exchange products so as to satisfy the effective demand. On the other hand, for the cooperative game, cen-

ters anticipate future trade. They communicate their demand history and aim to optimize the global profit. Demand is forecasted for all centers in the uncertain environment.

We applicate the two strategies to a real case of bottling company. The difference between costs generated by cooperative and biform game is significant giving the literature. Cooperation in the two steps improves the overall service and minimizes product transposition. Availability date of the products is optimized. Cooperation provides greater visibility of products movement and simplifies supply management.

References

Benkachcha, S., Benhra, J. and El Hassani, H. (2013) 'Causal method and time series forecasting model based on artificial neural network', *International Journal of computer applications*. Foundation of Computer Science, 75(7).

Brandenburger, A. and Stuart, H. (2007) 'Biform Games', *Management Science*, 53(4), pp. 537–549. doi: 10.1287/mnsc.1060.0591.

Chai, T. and Draxler, R. R. (2014) 'Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature', *Geoscientific Model Development*, 7(3), pp. 1247–1250. doi: 10.5194/gmd-7-1247-2014.

Charles, S. L. and Hansen, D. R. (2008) 'An evaluation of activity-based costing and functional-based costing: A game-theoretic approach', *International Journal of Production Economics*, 113(1), pp. 282–296. doi: 10.1016/j.ijpe.2007.08.008.

Dag, O. and Yozgatligil, C. (2012) 'GMDH: An R Package for Short Term Forecasting via GMDH - Type Neural Network Algorithms', *The R Journal*, XX, pp. 1–8.

Dalrymple, D. J. (1975) 'Sales forecasting methods and accuracy', *Business Horizons*, 18(6), pp. 69–73. doi: 10.1016/0007-6813(75)90043-9.

Holt, C. C. (2004) 'Forecasting seasonals and trends by exponentially weighted moving averages', *International Journal of Forecasting*, 20(1), pp. 5–10. doi: 10.1016/j.ijforecast.2003.09.015.

Mouatassim, S. *et al.* (2016) 'Model based on hybridized game theory to optimize logistics case of blood supply chain', *International Journal of Computer Applications*, 145(15)(15), pp. 37–48. Available at: http://www.ijcaonline.org/archives/volume145/number15/25358-2016910910.

Prak, D., Teunter, R. and Syntetos, A. (2017) 'On the calculation of safety stocks when demand is forecasted', *European Journal of Operational Research*. Elsevier B.V., 256(2), pp. 454–461. doi: 10.1016/j.ejor.2016.06.035.

Triqui-sari, L. (no date) 'Comparaison entre un Modèle de Jeu Biforme et un Modèle de Jeu Coopératif pour un Réseau de Distribution de Produits'.

Triqui Sari, L. and Hennet, J.-C. J. C. (2016) 'Cooperative inventory planning in a distribution network', *International Journal of Production Research*, 54(19), pp. 5916–5931. doi: 10.1080/00207543.2016.1189103.

Tsiakis, P., Shah, N. and Pantelides, C. C. (2001) 'Design of multi-echelon supply chain networks under demand uncertainty', *Industrial & Engineering Chemistry Research*. ACS Publications, 40(16), pp. 3585–3604.

Vanovermeire, C. *et al.* (2014) 'Horizontal logistics collaboration: decreasing costs through

flexibility and an adequate cost allocation strategy', *International Journal of Logistics Research and Applications*, 17(4), pp. 339–355. doi: http://dx.doi.org/10.1080/13675567.2013.865719.
Yao, J. and Tan, C. (2001) 'Guidelines for Financial Forecasting with Neural Networks', *Proceedings of International Conference on Neural Information Processing*, pp. 772–777. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.18.6874.
Zhang, G., Eddy Patuwo, B. and Y. Hu, M. (1998) *Forecasting with artificial neural networks:*, *International Journal of Forecasting*. doi: 10.1016/S0169-2070(97)00044-7.
Zhang, G. P. (2003) 'Time series forecasting using a hybrid ARIMA and neural network model', *Neurocomputing*, 50, pp. 159–175. doi: 10.1016/S0925-2312(01)00702-0.

Résumé

Ce travail analyse un problème de décision en deux étapes au niveau des centres de distribution régionaux partageant les mêmes familles de produits. Il compare deux stratégies, l'une se basant sur un jeu biforme et l'autre sur un jeu coopératif. Pour le jeu biforme, dans un premier temps, l'environnement est incertain, chaque centre de distribution doit définir les quantités à commander auprès des unités de production afin de maximiser son propre gain. Lorsque la demande est déterministe, les centres collaborent pour répondre à la demande locale. D'un autre côté, pour le jeu coopératif, les centres de distribution collaborent dans les deux étapes. Ils communiquent les informations concernant la demande, non seulement dans l'environnement déterministe mais aussi dans l'étape incertaine. Nous étudions un cas de société d'embouteillage au Maroc; Nous calculons les coûts générés par chaque stratégie et analysons les résultats.

Intégration stratégie et eSCM : une approche constructiviste

Said Bensbih*, Ferdaous Ajouami**, Mohamed Saad***, Otmane Bouksour*,
Abderrahmane Sbihi****, Naoufal Sefiani*****
*Université Hassan II, ENSEM, LMPGI, Casablanca
saidbensbih@gmail.com, bouksour2@gmail.com
**IDS - ENSA de Tanger
f.ajouami@gmail.com
***AUSIM, Casablanca
saad@casablanca-bourse.com
**** IDS-ENSA Tanger
sbihi@ensat.ac.ma
*****FST Tanger
tsefiani@gmail.com

Résumé. À partir de la littérature existante, ce travail reprend les concepts de stratégie et du e-Supply Chain Management (eSCM) afin de souligner l'importance du Business Model (BM) dans l'intégration de ces concepts. Une approche constructiviste est proposée avec une étude de cas dans le secteur hospitalier pour analyser un modèle conceptuel proposé afin de mettre en évidence les interactions entre les deux concepts mentionnés ci-dessus. Cette recherche est une contribution au travail visant à élucider l'importance d'une approche intégrée en alignant eSCM et stratégie avec le Business Model en tenant compte du contexte de la transformation numérique.

1 Introduction

Dans un contexte de transformation numérique, comprendre l'interaction entre les différents concepts entrant en jeu pour ce qui est du management d'une organisation et plus particulièrement la stratégie, le Business Model et le eSCM, s'avère une perspective de recherche à même de contribuer aux efforts d'accompagnement de cette importante vague du 21ième siècle que constitue l'évolution des technologies de l'information et de la communication, TIC..

Nous revenons, premièrement, sur la définition de ces concepts fondamentaux pour essayer d'en saisir le sens et proposer ensuite un modèle conceptuel qui puisse donner une perspective d'ensemble montrant la relation qui se dessine entre stratégie, BM et eSCM.

Nous commençons par la stratégie avec les approches prescriptives et descriptives pour la définir et constater, déjà à ce niveau que selon les auteurs, plusieurs perspectives sont déclinées.

Le même constat est fait pour le BM et le eSCM nous conduisant à proposer un modèle conceptuel reprenant l'essentiel des composantes de ces deux concepts en vue de les aligner avec la stratégie.

Nous montrons que dans une telle situation le constructivisme est recommandé en vue d'analyser ce phénomène vu l'importance d'une analyse approfondie de l'intervention de la multitude des parties prenantes contribuant à l'élaboration et la mise en œuvre d'une stratégie.

Une étude de cas dans le secteur de la santé avec plus précisément une initiative TIC relevant du eSCM est décrite pour un essai de validation du modèle conceptuel proposé.

Nous formulons enfin une conclusion et un aperçu des perspectives de recherche dans cette direction.

2 Analyse des concepts

2.1 Stratégie

De très nombreuses œuvres ont été consacrées à la stratégie et parmi elles des ouvrages considérés comme l'expression d'écoles ayant servi de référence aux académiciens, consultants et managers dont on citera Ansoff (1965) qui a été l'un des précurseurs, Andrews, 1971; Porter (1980, 1985) ; Mintzberg et al., 1998).

De même, une abondante production en études et articles de recherche reprenant parfois les principes de ces écoles ou abordant différentes dimensions de la stratégie ont enrichi ce domaine d'apports considérables (Ansoff , 1995 ; Christensen, 1982 ; Steiner, 1969 ; Porter, 1981 ; Mintzberg 1989 ; Huff, 1990 ; Lindblom 1959 ; Pettigrew, 1977 ; Edwards, 1977 ; Pugh et al., 1963 ; Khandwalla, 1970 ; Farjoun, 2002 ; Gavetti ; 2004).

Par ailleurs, nous notons, à partir des années 2000, l'émergence de nouvelles approches en stratégie tentant de proposer d'autres perspectives. Nous citons à ce propos Kim et Mauborgne (2005) « Blue Ocean Strategy », Raynor (2007) « Strategy Paradox », Kaplan et Norton (2004) « Strategy Maps », Lehmann-Ortaga et al. (2013) « Strategor », Johnson et al. (2014) « Stratégique ».

Nous considérons cependant que l'ouvrage de Mintzberg (1998) « Safari en pays stratégie », permet de prendre connaissance des principales écoles de pensée stratégique, au nombre de dix : trois écoles normatives (Conception, planification positionnement), six écoles descriptives (entrepreneuriale, cognitive, apprentissage, pouvoir, culturelle, environnementale) et une dernière école pouvant être une combinaison de toutes les autres (configuration).

Nous retenons pour notre travail les trois premières écoles prescriptives qui sont les plus abondamment citées par les chercheurs et pratiquées au niveau des organisations.

2.1.1 Ecole de la conception

Cette école fait appel à un process de conception. C'est la première des écoles dites normatives. Elle a été initiée principalement par Selznick (1957) qui introduit la notion de compétence distinctive. Chandler (1962) a travaillé sur la corrélation entre la structure de l'entreprise et la stratégie en prenant en considération son secteur d'activité.

Mais ce sont Learned et al. (1965) qui vont propulser cette école au rang de référence aussi bien pour le management que pour l'enseignement de cette matière. Pour eux, la stratégie est définie sous forme d'un modèle cherchant l'adéquation entre les forces et faiblesses de l'entreprise d'une part et les opportunités et les menaces que présente son environnement d'autre part. C'est ainsi qu'est né le très connu concept SWOT : (Strengths and Weaknesses / The opportunities and Threats).

Ce modèle de base LCAG, en référence à ses auteurs (Learned, Christensen, Andrews, Guth), est supposé permettre à la Direction Générale de formuler une stratégie unique et adéquate parmi plusieurs scénarii . La stratégie est ainsi une perspective élaborée selon un process délibéré et exprimée d'une manière plutôt simple et claire pour être diffusée, communiquée et mise en œuvre.

C'est Andrews (1971) qui va reprendre le travail de LCAG et plusieurs versions successives succèderont à son œuvre originale sans pour autant remettre en cause le principe de séparation entre la réflexion et l'action. La stratégie reste ainsi hypothéquée par la stabilité des critères pris en compte pour sa formulation.. Il est reproché cependant à l'analyse SWOT de ne pas être assez fine pour l'élaboration de facteurs de succès suffisamment pertinents (Trego et Zimmerman, 1980).

La complexité grandissante dans le management stratégique d'une organisation a cependant favorisé le recours accentué à une planification encore plus fine telle que préconisée par l'école de la planification.

2.1.2 Ecole de la planification

La stratégie selon cette école est élaborée par des planificateurs en suivant un process formel, explicite et complexe. Ils y trouvent un rôle prépondérant aux côtés de la Direction générale. Ansoff (1965) a consacré un de ses livres les plus influents à l'école de la planification.

Son modèle fondamental de planification stratégique tel que décrit par Steiner (1969, 1979, 1983) est subdivisé en trois grandes parties : données, planification et mise en œuvre. En fait l'outil d aide à la décision stratégique que constitue la matrice SWOT est repris pour être décliné avec détails et sous détails en étapes, listes techniques, objectifs, budgets et plans opérationnels. Même les stratégies sont réparties en sous-stratégies.

Les étapes de ce modèle fondamental comprennent une définition des objectifs qui peuvent être d'ailleurs séparées des stratégies (Shendel et Hofer, 1979), un audit externe permettant de prévoir et préparer les actions (Ackoff, 1983), un audit interne reflétant le principe de la stratégie par énumération (Jelinek et Amar, 1983), une évaluation, une programmation du processus et un plan directeur regroupant ceux qui sont opérationnels.

Cette planification minutieuse garantit d'après cette école une hiérarchie claire et précise entre une projection des activités comprenant stratégies et programmes et un contrôle des résultats au niveau des objectifs et des budgets.

Les principes de cette école sont profondément ancrés dans une conception mécaniste consistant en une suite complexe d'étapes successives selon un process maitrisé et conscient de planification formelle.

Tout en contribuant à l'élaboration de stratégies, cette école est très critiquée (Mintberg, 1994 ; Hayes, 1985) pour vouloir prédéterminer un environnement incertain.

L'école du positionnement propose alors des stratégies qui vont justement donner une place plus importante à l'analyse compétitive.

2.1.3 Ecole du positionnement

Avec cette école, les consultants ont retrouvé tous leur place. Ils sont les architectes d'un processus analytique qui contrairement aux affirmations générales des deux précédentes écoles vont accorder la plus grande importance aussi bien aux stratégies elles mêmes qu'à leur contenu. Le management stratégique prend les devants et cette école devient et reste la référence en élaboration et études stratégiques.

Bien que Shendel et Hofer (1979) aient déjà eu à traiter du contenu du management stratégique, c'est principalement Porter (1980, 1881,1985) qui s'est intéressé tout d'abord à la technique d'analyse concurrentielle en liaison avec les secteurs d'activité. L'analyse structurelle des secteurs permet en effet de dégager une vue sur les entrants, les fournisseurs, les clients et les produits.

Trois aspects définissent l'approche de Porter : La chaine de valeurs, le modèle d'analyse concurrentielle et les stratégies génériques. Pour lui, le choix des entreprises se limite à opter soit pour les meilleurs coûts soit pour une différentiation tout en jouant sur l'effet de la concentration.

Mintzberg (1998) considère que cette école dont l'une des stratégies génériques a pour objectif principal la maitrise des coûts, ne donne pas toute la place qu'il faut aux aspects sociaux et politiques au sein de l'entreprise. Hamel (1997) lui reproche par ailleurs de laisser peu de marge de manœuvre à la créativité et ne valorise pas les nouvelles expériences. Langley (1995) parle quant à elle de paralysie par l'analyse.

2.1.4 Hypothèses retenues

Nous pensons que la planification et le positionnement stratégiques sont toujours d'actualité dans le management des organisations avec un certain dosage en fonction du contexte visant parfois dans l'adoption d'une configuration adaptée prenant en compte des exigences entrepreneuriales et culturelles. De même, il est tout aussi pertinent que le pouvoir et l'apprentissage prennent une part importante dans l'émergence de stratégies en vue d'une pérennité nécessaire des organisations.

Ces hypothèses sont cependant à valider par une analyse approfondie des organisations concernées et scruter le BM adopté est un des moyens pour cette validation.

2.2 Business Model

Le concept du BM est discuté dans différents domaines tels que e-business, systèmes d'information, stratégie et management, Pateli et al (2003).

Il a fait l'objet de plusieurs définitions de la part des équipes de recherches dont une revue a été établie par Shafer & al (2005) pour le décrire comme étant « une représentation de la logique de base et des stratégies sous-jacentes des entreprises dans leur quête de création et de récupération de la valeur ».

Le BM est décrit ainsi soit tout simplement en donnant une idée de ce qu'il représente (Timmers, 1998; Magretta,2002) et la manière dont une entreprise fait son travail (Galper 2001; Gebauer et al, 2003), ou alors en le positionnant comme une approche de modélisation (Chesbrough et al, 2000 ;Hamel, 2000 ;Gordijn, 2002 ;Osterwalder, 2004; Osterwalder et al, 2005; George et al, 2009).

A signaler cependant que le concept reste malgré tout insuffisamment appréhendé, Linder et al (2000). Et alors que certaines revues de littérature semblent aller dans le sens d'une convergence de définitions se référant au BM comme une conception ou une architecture pour la création de valeur, la prestation et les mécanismes de sa récupération dans une entreprise (Teece, 2010), il est admis qu'il n'y a pas une convergence à ce niveau car le concept est souvent analysé suivant différentes perspectives (Zott et al, 2011).

Le BM continue aussi d'intéresser aussi bien les chercheurs que les organisations et une définition plus proche des praticiens le définit comme étant la manière selon laquelle une organisation créée, livre et se procure de la valeur (Osterwalder and al, 2010). Il est ainsi décliné en neuf composantes : l'offre, la cible, la relation client, les canaux, la structure des revenus, les activités clefs, les ressources clefs, les partenaires et la structure des couts.

C'est cette dernière déclinaison du BM qui nous semble rapprocher le plus les travaux de recherche aux pratiques des organisations. C'est en particulier les travaux portant sur eSCM qui pour nous croisent les éléments de la définition BM retenue.

2.3 eSCM

2.3.1 SCM

Nous commençons d'abord par analyser le concept SCM « Supply Chain Management. Nous constatons d'ores et déjà, comme pour les concepts stratégie et BM, que malgré la popularité du terme SCM, une confusion considérable entache son sens et sa définition (Mentzer et al, 2001; Skjøtt-Larsen, 1999).

Ainsi, Supply Chain « SC », est un réseau d'organisations de bout en bout dans le but de la création de valeur pour le client final, Christopher (1998). Le SCM est aussi une méthode d'intégration de l'ensemble des interactions dans une SC « Supply Chain » (Gunasekaran et al, 2004). SCM est d'après d'autres chercheurs une progression logique des développements dans la logistique, (Metz, 1998 ; Coyle et al, 2003), en est une extension, Cooper et al (1997), ou est tout simplement plus que de la logistique Giunipero et al (1996).

Cependant, Simchi-Levi et al (2000; 2003) admettent qu'ils ne voient pas de différence entre le SCM et logistique. Sur le plan théorique, Handfield et al (2004) notent une multitude d'approches et de perspectives.

Une étude exhaustive faite par Larson et al (2007) fait ressortir quatre groupes de réponses concernant la relation SCM/Logistique dont une retenue par Médan et al (2008) concluant que le SCM inclut la logistique et à laquelle nous adhérons.

2.3.2 eSCM

La convergence du SCM et d'Internet et le fait d'intégrer les TIC en vue de répondre à la nécessité de coordination de l'ensemble des opérations de la SC sont des ingrédients du concept Internet-enabled Supply Chain Management (eSCM), Gimenes et al (2004). Il ne s'agit donc pas d'une simple mise en place d'un front office web mais plutôt de pouvoir le gérer dans le cadre d'un BM qui inclut le SCM (Van Hoek 2001). De même, McGuffog et al (1999) ont considéré l'effet du e-commerce sur la planification collaborative SC.

VSC « Virtual Supply Chain » est un autre concept qui émerge cependant en parallèle à eSCM et qui se définit comme étant l'application d'un système d'information en temps réel visant à améliorer la communication tout au long de la SC (Gunasekaran et al, 2004) ou un réseau temporaire d'entreprises se mettant ensemble pour exploiter des opportunités rapidement changeantes (Strader et al,1998).

Nous retenons comme définition eSCM l'intégration de toute nouvelle technologie basée sur Internet en vue de faciliter une coopération et une collaboration entre l'ensemble des parties prenantes SCM et ce en adéquation avec la stratégie de l'organisation avec sa déclinaison BM.

Cette perspective est synthétisée au niveau du modèle conceptuel proposé ci-dessous (figure.1) :

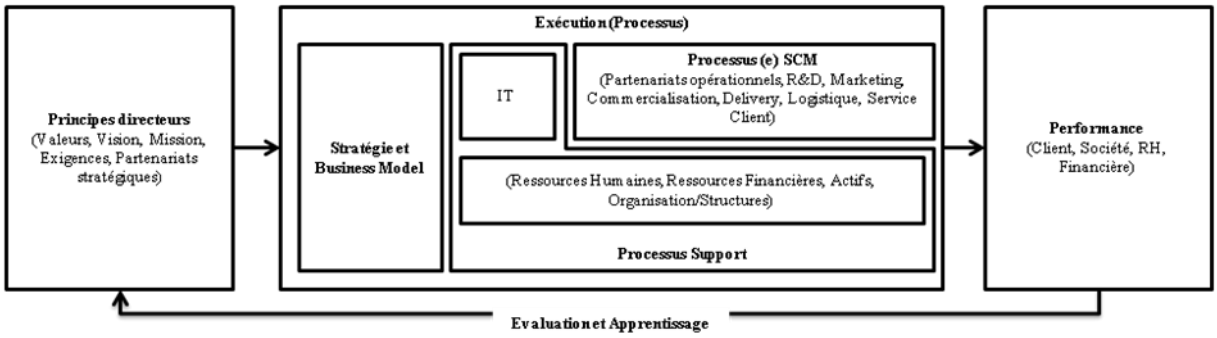


FIG. 1 – *Intégration stratégie/ e-SCM: modèle conceptuel*

Ce modèle constitue l'essence de notre travail et nous cherchons à le vérifier par une approche constructiviste.

3 Méthodologie

Dans le contexte d'une stratégie de transformation digitale avec une étroite corrélation du BM et du eSCM et plus généralement avec les autres concepts figurant dans le modèle proposé, il est très difficile d'analyser ce phénomène sans nous intéresser aux multiples interactions entre les différentes parties prenantes de l'ensemble de la chaine de valeur. En effet, ces multistakeholders contribuent, chacun à son niveau et de bout en bout, à la prestation finale, y compris les bénéficiaires des services voulus. Le chercheur y est inclus aussi.

Les interlocuteurs appartenant aux diverses organisations impliquées utilisent un langage pour construire du sens permettant d'explorer les voies selon lesquels ils se comprennent et influencent l'un l'autre (Barry et al ,1997 ; Ford et al, 1995), Giordano (1998), Samra-

Fredericks (2003), Hendry (2000), ainsi que les représentations qu'ils se font de l'entreprise et de son environnement, Liedtka et al (1996).

Nous proposons alors que l'approche constructiviste est la plus appropriéepour approfondir le sujet compliqué que nous abordons en procédant à des entretiens semi-structurés à différents niveaux de prise de décision et en veillant à respecter les recommandations (Yin , 1994 ;, Zerbe et al, 1987) ; Patton, 1990) est la plus appropriée.

Concernant le nombre de cas à étudier, des chercheurs vont dans le sens d'un cas unique permettant de recueillir le plus de détails au regard du contexte dans lequel le phénomène est analysé (Dyer et al, 1991; Voss et al, 2002 ; Narasimhan et al, 1998). Il est ainsi suggéré que le moins de cas à travailler offre l'opportunité d'observer en profondeur le phénomène.

D'autres chercheurs recommandent plutôt une fourchette entre quatre et dix cas pour saisir la complexité du monde réel sans aller au-delà du seuil proposé pour ne pas rendre difficile le traitement des informations recueillies, Eisenhardt (1989), Yin (1994).

4 Etude de cas

Dans le cadre des stratégies digitales élaborées par le gouvernement marocain et en particulier les volets concernant les administrations publiques, la transformation numérique de ce secteur apparait comme pré requis fondamental en faveur d'un service de meilleure qualité pour le citoyen. A ce premier objectif, est associé un souci de raccourcir les délais ainsi que la nécessité de simplifier les procédures. Par ailleurs, il s'agit aussi de servir les différents prestations tout en tenant compte des contraintes socioéconomiques, des lois en vigueur, de la réglementation, des normes et standards, du niveau d'éducation, de la perspective environnementale, du contexte culturel et des contributions et négociations avec les différentes parties prenantes.

Le secteur de la santé publique, vu son impact sur l'ensemble de la population mais aussi de par son importance comme critère essentiel d'incitation dans des domaines tels que les investissements, le tourisme, l'organisation d'événements et d'activités de toutes natures, des flux MRE pour n'en citer que quelques aspects, a été parmi les premiers à mettre en place des schémas directeurs informatiques. A signaler à ce stade que l'alignement des systèmes d'information avec les stratégies du Ministère en charge de la santé publique était déjà une pratique d'usage. Ces stratégies pouvaient être par exemple axées sur les prestations hospitalières mais aussi focalisées sur les soins de santé de base.

Les stratégies étant énoncées, nous retrouvons les composantes du BM avec la description du portefeuille de soins adaptés aux diverses populations cibles et des aménagements de couverture des frais associant une multitude d'organisations (mutuelles, assurances, organismes locaux et internationaux, dispositions d'accompagnement telles que AMO et RAMED). Des process sont aussi mis en œuvre pour assurer ces prestations ainsi que la mobilisation des ressources clés telles que le personnel médical, paramédical et administratif.

Les éléments de la SCM se trouvent réunis à partir de cette première description du BM et se manifestent dans l'organisation aussi bien centrale que locale avec l'objectif d'un fonctionnement harmonieux de l'ensemble du dispositif en vue d'une mise en œuvre efficiente de la stratégie adoptée.

Comme signalé précédemment, le recours aux technologies de l'information et de la communication s'inscrit dans les axes stratégiques comme étant le moyen de disposer de l'information nécessaire au fonctionnement de l'ensemble SCM et ce aussi bien au niveau sanitaire qu'administratif et autres entités de support.

Nous pouvons alors dire que le cas du secteur de la santé publique reflète d'une manière probante le concept eSCM qui se trouve bien ancré au niveau stratégique via les composantes BM comme indiqué ci-dessous. A noter que dans son ensemble, il est tout à fait observable que les stratégies de ce secteur s'adaptent au retour d'information collectées essentiellement via le eSCM et aboutissent le cas échéant à un repositionnement du portefeuille de soins et leur mise en œuvre et ce tel que relaté dans le modèle conceptuel proposé.

Pour affiner notre analyse, nous avons pris une des actions entreprises par le Ministère de la santé publique que nous plaçons comme composante eSCM : la prise des rendez-vous dans les hôpitaux par internet.

En fait, le choix de cette action a découlé principalement de notre participation (immersion) au jury d'une compétition nationale, eMtiaz, dont l'un des axes était justement la mise en œuvre de la stratégie eGOV par la réalisation d'applications mobiles facilitant l'accès du citoyen aux prestations publiques. Cette contribution nous a permis de relever l'importance d'une intégration Stratégie/eSCM via BM pour garantir les objectifs voulus.

En adoptant une démarche qualitative grâce à des entretiens semi structurés et approfondis mais aussi en analysant les documents et rapports disponibles mis à notre disposition, plusieurs constatations, étayées par notre immersion citée précédemment :

- L'application elle-même est effectivement utilisée par les citoyens mais nécessite encore des efforts pour que tout un chacun puisse y accéder. Ceci passe en particulier par l'association de partenaires tels que les opérateurs télécoms ;
- La mobilisation des ressources clefs en vue de faire un succès de cette application est fondamentale pour garantir l'adhésion des intervenants et leur disponibilité ;
- Les délais de l'obtention d'une consultation allant être raccourci incite à prévoir l'infrastructure et les ressources nécessaires pour un parcours médical garantissant les prestations qui en découlent ;
- Des aspects humains sont nécessairement à prendre en compte comme le niveau d'éducation et l'influence du secteur informel que nous retrouvons dans d'autres applications du même type ;
- L'implication du top management au niveau ministère mais aussi celle des autres autorités et collectivités concernées est très importante pour généraliser l'utilisation de cette application à l'instar d'autres initiatives mises en œuvre et relativement réussies ;
- Une réorientation stratégique dans le sens ressources hospitalières parait fort probable ;

5 Conclusions et perspectives

Alors que la transformation digitale gagne la plupart des domaines y compris des secteurs publics clefs tels que celui de la santé, il est important que le monde académique puisse contribuer davantage au niveau de l'accompagnement des différentes parties prenantes dans la chaine de valeur en vue de mieux intégrer cette transformation aux stratégies adoptées. Un

tel alignement stratégique avec le business modèle et le eSCM nécessite des travaux de recherche à même de mieux cerner des étapes clefs partant de la formulation à l'implémentation et proposant des méthodes et des outils pour une évaluation continue des initiatives digitales tout en définissant les rôles et les procédures les plus appropriées. Le cas de la mise en œuvre d'une application de prise de rendez vous en ligne pour les hôpitaux illustre la nécessité de travaux ultérieurs en vue d'analyser les différents aspects à considérer dans le but de garantir l'objectif final qu'est un soin de santé au service du citoyen qui soit de qualité, à temps et à un cout acceptable.

Le modèle conceptuel proposé est à affiner en approfondissant davantage l'étude de cas abordé et en étendant éventuellement l'analyse à d'autres cas.

References

BARRATT, Mark, CHOI, Thomas Y., et LI, Mei. Qualitative case studies in operations management: Trends, research outcomes, and future research implications. Journal of Operations Management, 2011, vol. 29, no 4, p. 329-342.

BERMAN, Saul J. Digital transformation: opportunities to create new business models. Strategy & Leadership, 2012, vol. 40, no 2, p. 16-24.

Bernardes, Maria Elisa Brandao, (2008) La construction sociale de la stratégie en contexte PME : une analyse en profondeur de quatre cas de diversification, Thése PhD HEC – Montréal, - Canada

CLARKE, Mike P. Virtual logistics: an introduction and overview of the concepts. International Journal of Physical Distribution & Logistics Management, 1998, vol. 28, no 7, p. 486-507.

COOPER, Martha C., LAMBERT, Douglas M., et PAGH, Janus D. Supply chain management: more than a new name for logistics. The international journal of logistics management, 1997, vol. 8, no 1, p. 1-14.

DANIELS, Norman et SABIN, James. Limits to health care: fair procedures, democratic deliberation, and the legitimacy problem for insurers. Philosophy & public affairs, 1997, vol. 26, no 4, p. 303-350.

EISENHARDT, Kathleen M. et GRAEBNER, Melissa E. Theory building from cases: Opportunities and challenges. Academy of management journal, 2007, vol. 50, no 1, p. 25-32.

GIMÉNEZ, Cristina et LOURENÇO, Helena R. e-SCM: internet's impact on supply chain processes. The International Journal of Logistics Management, 2008, vol. 19, no 3, p. 309-343.

GUNASEKARAN, Angappa et NGAI, Eric WT. Virtual supply-chain management. Production Planning & Control, 2004, vol. 15, no 6, p. 584-595.

GUNASEKARAN, Angappa et NGAI, Eric WT. Information systems in supply chain integration and management. European Journal of Operational Research, 2004, vol. 159, no 2, p. 269-295.

KANE, Gerald C., PALMER, Doug, PHILLIPS, Anh Nguyen, et al. Strategy, not technology, drives digital transformation. MIT Sloan Management Review and Deloitte University Press, 2015, vol. 14.

LARSON, Paul D., POIST, Richard F., et HALLDÓRSSON, Árni. Perspectives on logistics vs. SCM: a survey of SCM professionals. Journal of Business Logistics, 2007, vol. 28, no 1, p. 1-24.

LUSCH, Robert F. Reframing supply chain management: a service-dominant logic perspective. Journal of supply chain management, 2011, vol. 47, no 1, p. 14-18.

MCKONE-SWEET, Kathleen E., HAMILTON, Paul, et WILLIS, Susan B. The ailing healthcare supply chain: a prescription for change. Journal of Supply Chain Management, 2005, vol. 41, no 1, p. 4-17.

MINTZBERG, Henry, AHLSTRAND, Bruce, et LAMPEL, Joseph. Safari en pays stratégie. Paris, Village Mondial, 1999.

OSTERWALDER, Alexander et PIGNEUR, Yves. Business model generation: a handbook for visionaries, game changers, and challengers. John Wiley & Sons, 2010.

PORTER, Michael E. What is value in health care?. New England Journal of Medicine, 2010, vol. 363, no 26, p. 2477-2481.

PORTER, Michael E. et VAN DER LINDE, Claas. Toward a new conception of the environment-competitiveness relationship. Journal of economic perspectives, 1995, vol. 9, no 4, p. 97-118.

SHAFER, Scott M., SMITH, H. Jeff, et LINDER, Jane C. The power of business models. Business horizons, 2005, vol. 48, no 3, p. 199-207.

SCHMIDT, Rainer, ZIMMERMANN, Alfred, MÖHRING, Michael, et al. Digitization–perspectives for conceptualization. In : European Conference on Service-Oriented and Cloud Computing. Springer, Cham, 2015. p. 263-275.

WILLIAMS, Lisa R., ESPER, Terry L., et OZMENT, John. The electronic supply chain: Its impact on the current and future structure of strategic alliances, partnerships and logistics leadership. International Journal of Physical Distribution & Logistics Management, 2002, vol. 32, no 8, p. 703-719.

YIN, Robert K. Case study research and applications: Design and methods. Sage publications, 2017.

ZOTT, Christoph, AMIT, Raphael, et MASSA, Lorenzo. The business model: recent developments and future research. Journal of management, 2011, vol. 37, no 4, p. 1019-1042.

Integrating Strategy and e-Supply Chain Management : A Constructionist Perspective

Summary

Starting from the existing literature, this work takes up the concepts of strategy and e-Supply Chain Management to highlight importance of the business model for the integration of these concepts. A constructivist approach is proposed with a case study in the hospital sector to analyze a conceptual model to emphasize the interactions between the two concepts mentioned above. This research is a contribution towards work aimed at elucidating the importance of an integrated approach by aligning supply chain with strategy in line with business model considering digital transformation context.

Keywords : *Strategy,Constructionism,Digital Transformation,eSourcing Capability Model, Healthcare*

E-supply chain & sustainable development: When sustainability challenges the e-supply chain

Ferdaous Ajouami*, Said Bensbih**, Naoufal Sefiani***, Otmane Bouksour**, Abderrahmane Sbihi****
* LabTIC ENSA de Tanger
f.ajouami@gmail.com
**Université Hassan II, ENSEM, LMPGI, Casablanca
saidbensbih@gmail.com, bouksour2@gmail.com
***FST Tanger
tsefiani@gmail.com
**** ENSA Tanger
sbihi@ensat.ac.ma

Abstract: Combining the e-supply chain and sustainable development may seem impossible because of antagonism. Besides, sustainable development has become a social problem that leads to the quasi-systematic invocation of mutualisation at all costs, without taking into account the concrete conditions and difficulties that are multiple and probably reinforce each other. Thus, Management sciences are often aimed at solving paradoxical issues and emphasize on the need for this to adopt rather global approaches integrating all the actors of the sector concerned. So, our research underlines modes of cooperation between the e-supply chain and the sustainable development providing a relatively well-adapted theoretical and practical framework.

1 Introduction

Supply chain management (SCM) is a concept that received great attention from industrialist as strategic planning in design, maintenance and operation of supply chain process satisfaction of end user needs. Although the improvements have been achieved through the successfully SCM practice, some of organizations are neglected to take care the environmental issues such as global energy, global warming, reverse logistic, etc. Environmental, ecological concerns in global competition attracted researchers in variety of disciplines.

The growing body of literature on the subject demonstrates a widespread appeal especially with regard to the application of ISO 14001 or Environmental Management System (EMS) standards. Simultaneously, the public's environmental awareness has increased through formal and informal environmental education channels. As a result, a systematic approach, Green Supply Chain Management (GSCM), has been increasingly accepted and practices by forward-thinking organizations

Organizations are increasingly aware and concerned with the environmental and social impact of their business activities (Carter and Easton, 2011; Yu and Tang, 2011; Winter and Knemeyer, 2013). The focus on supply chains is a forward step into a broader adoption and

development of sustainability. Supply chain managers must address a complex assortment of factors that include the product and the process on both the upstream and downstream of the supply chain (Vachon and Klassen, 2006).

Environmental impact of business activities has become an important issue in the last years due to the growing public awareness of environmental, and the introduction of environmental legislations and regulations mainly in developed countries (Lau, 2011). Srivastava (2007) argues that “much research is needed to support the evolution in business practice towards greening along the entire supply chain”. However, in recent years, more and more companies are introducing and integrating environmental issues into SCM processes by auditing and assessing suppliers on environmental performance metrics (Handfield et al., 2005). In this way there is a gap between intention and behavior, to which behavioral theory can offer valuable insights. The importance of integrating sustainability issues in SCM research has been established for more than a decade and many researchers have produced guidelines, frameworks and research agendas identifying crucial topics. they seek to ensure that they have effective tools not only for measuring environmental performance of their suppliers but also to help select them for new projects/products or for carrying out action plans to improve their performance (Olugu et al, 2011; Naini et al, 2011).

2 SCM and sustainability: literature review

To move from the supply chain concept to the supply chain management (SCM), there must be a "management"; the existence of simple commercial links is not enough.

The issue of global management of the multi-actor supply chain is being addressed more numerous in the last fifteen years (Colin, 2005, Pache and Spalanzani, 2007); The application of social responsibility for the SCM is even more recent (Ciliberti et al., 2008b). For Pagell and Wu (2009) interest in green SC and now sustainable has grown over the last decade.

First of all from the environmental point of view, the integration of environmental concerns in the SCM may involve environmental collaboration or an environmental management system that can be supported by a system of social responsibility management (CSR management systems). For Vachon and Klassen (2008), environmental collaboration is composed of joint activities and cooperation aimed at finding solutions to environmental problems; This collaboration concerns suppliers and customers and may impact operational and environmental performance: “environmental Cooperation with primary suppliers and major customers, defined as encompassing joint environmental planning activities and cooperation in finding solutions to environmental Challenges, can have a significant positive impact on both manufacturing and environmental Performance" (Vachon and Klassen, 2008, p. 309). The dialogue, support for collaboration Supplier / customer, facilitates understanding of the environmental impact of the logistics chain (Lamming and Hampson, 1996 cited by Simpson et al 2007); the joined forces improve the environment (Florida 1996, Hall 2000 and 2001 cited by Simpson et al, 2007) This collaboration takes place around a focal company.

The CSR management system "CSR management system" allows to transfer socially responsible behavior from one partner to another of a supply chain and outlines the environmental (and social) principles to Respect: " such system(s) can be used to transfer socially responsible Supply chain, in particular to influence the practices of their business

partners and to provide a (Ciliberti et al., 2008b, p. 1580). These systems can be based on standards. For Castka and Balzarora (2008), the SC are increasingly difficult to control because networks are increasingly decentralized and independent; the development of standards is a way of reducing information asymmetries and reduce the complexity of SC management; the Multinationals play a role in the international propagation of these standards.

The effective management of these returns involves the existence of a reverse supply chain: “The Process of planning, implementing and controlling the efficient, cost-effective flow of raw Materials, in-process inventory, finished goods and related information from the point of consumption to the point of origin for the purpose of recapturing or creating value, or for proper disposal " (Rogers and Tibben Lembke, 1999 cited by Srivastava, 2008, p 538). If the Returns are re-integrated into the production process, the circuit becomes closed (" Loop "); Otherwise it remains open ("open loop").

Inverse logistics is not a symmetrical image of "go" logistics; it is more Reactive (Srivastava, 2008). For Srivastava (2008), the review of the literature and 84 interviews of stakeholders suggest that reverse logistics is more complex than Logistics: it must be more reactive, it must be driven by supply and dependent on the rate of return; it must therefore be the subject of a very specific analysis. For French and La Forge (2006), the flows characterizing the logistics of returns are very diverse. The collection of returns can be based on the structures of the "go" chain through the distribution centers of the "go" channel or on specific structures of the chain "Return" (Gou et al, 2008)

Reverse logistics is itself part of the "green management of the logistics chain" or "Green supply chain management which “encompasses environmental initiatives in : inbound logistics; Production or the internal supply chain; Outbound logistics; And in certain cases reverse logistics, including and involving materials Suppliers, service contractors, vendors, distributors and end users working together to reduce or eliminate adverse environmental impacts of their activities "(Rao and Holt, 2005, p 899).

The GSCM is to consider environmental objectives in Procurement decisions, product design and manufacture, distribution and integrating environmental problems related to the end-of-life of the product: "Integrating environmental Product sourcing, product sourcing Selection, manufacturing processes, delivery of the final product End-of-life management of the product after its useful life "(Srivastava 2007, quoted by Srivastava 2008, p. 536)

It is based on collaboration or control: « GSCP comprise a series of inter-organizational activities arising from two very different options for improving environmental management: mutual problem-solving versus inspection and risk minimization which are termed environmental collaboration and environmental monitoring, respectively ». (Vachon et Klassen, 2006, p 796). It has efficiency objectives: “green-supply is a potentially effective mechanism for supply chain managers to improve the organizations record on corporate social responsibility, minimize reputational risks, reduce wastes and increase flexibility in response to new environmental regulations” (Simpson et al, 2007, p 29)

Other authors prefer to talk about environmental management of SC ("environmental SCM "), or "environment-friendly SCs " (like Diniz and Fabbe-Costes, 2007):" the set of supply chain management policies held, action taken , and relationships formed in response to concerns related to the natural environment with regard to the design, acquisition, production, distribution, use, reuse, and disposal of the firm.s goods and services (Zsidisin et Siferd, 2001, Cited by Hagelaar and van der Vorst, 2004, p 30). The two concepts are very close, we find the performance objective also for this environmental management:

"Environmental Supply Chain Management attempts to restructure supply chains to improve their environmental performance" (Côté et al, 2008, p 1561).

If we add to the environmental dimension, the social dimension the SCM becomes sustainable (Sustainable Supply Chain Management). The call for papers on the theme of Sustainability and SCM by the International Journal of Production Economics resulted in 37 proposals for 10 publications in 2008 (volume 111); that by the Journal of cleaner 42 proposals were produced, of which were published in 2008 (in volume 116). Seuring, Sarkis, Müller and Rao (2008) define sustainable supply chain management in their editorial as management of information flow and cooperation integrating economic, environmental and social objectives and stakeholder expectations: « we define sustainable SCM as the management of material and information flows as well as cooperation among companies along the supply chain while taking goals from all three dimensions of sustainable developement, i.e. economic, environmental and social, and stakeholder requirements into account. (p 1545). These three dimensions of sustainable management are also to be found in Ciliberti et al (2008b): « sustainable SCM is defined as the management of supply chains where all the three dimensions of sustainability, namely the economic, environmental, and social ones, are taken into account (Ciliberti et al, 2008b, p 1580). For Pagell and Wu (2009), an SSC must be effective on the three dimensions of performance: « a sustainable supply chain is then one that performs well on both traditional measures of profit and loss as well as on an expanded conceptualization of performance that includes social and natural dimensions" (Pagell et Wu, 2009, p 38).

Seuring and Müller (2008) add physical and financial flows to the definition « the management of material, information and capital flows as well as cooperation among companies along the supply chain while taking goals from all three dimensions of sustainable development, i.e., economic, environmental and social, into account which are derived from customer and stakeholder requirements. In sustainable supply chain, environmental and social criteria need to be fulfilled by the members to remain within the supply chain, while it is expected that competitiveness would be maintained through meeting customer needs and related economic criteria » (Seuring et Müller, 2008, p 1700).

The concepts of environmental collaboration and sustainable supply chain management, are they put into practice by companies? Yes, but partially. The 191 articles studied by Seuring and Müller (2008) show that empirical studies do not reveal any real global coordination. Man and Burns (2006) observe the impact and the limited role of cooperation for sustainable development in SC in the Case of the paper supply chain; Efforts are often modest; they depend on public perception and the weight of partners. Nevertheless, Strand (2008) highlights the importance of trust in the supplier relationship for a number of Scandinavian companies known for their social commitment. The supplier / customer dialogue facilitates understanding the environmental impact of logistics' chain (Lamming and Hampson, 1996 cited by Simpson et al 2007); Joint efforts improve the environment (Florida 1996, Hall 2000 and 2001 cited by Simpson et al, 2007). The Harwood and Humby case study (2008) reveals that most organizations focus on a component of responsibility (social, environmental, or ethical). Ten case studies of environmental and social organizations in the United States of America by Pagell and Wu (2009) highlight practices of a sustainable SC from the best practices in a traditional SC.

3 Green supply chain “GSCM”: a competitive tool for companies

The SCM seeks to optimize the management of flows and stocks in the company and its environment, while minimizing costs and deadlines. We can distinguish two levels in the management of the chain logistics, the supply chain which is specific to the company and which concerns only its own activity, and the extended one which includes all the actors of the chain, suppliers and their subcontractors to customers. Recently, supply chain management takes into account the environmental parameters."Environmental pressures have caused green supply chain management (GSCM) to emerge as an important corporate environmental strategy for manufacturing enterprises" (Zhu et al 2012). Today, for companies, the main objective is to find a balance between the economic and ecological requirements. Previously, motivation was mainly related to levels of cost and service. From now on, the environmental impact represents a third dimension. In terms of sustainable development, companies must react to climate change and pollution. Thus, to reconcile economic and ecological requirements, companies seek to reduce environmental pollution generated by activities throughout the supply chain while optimizing the logistics chain operations.

Investments in GSCM can save resources, eliminate waste and improve productivity, can potentially reduce costs and increase efficiency and flexibility (Wilkerson 2005). Furthermore, this can ideally lead to the identification and creation of new opportunities for products and services in cooperation with up-stream and down-stream partners [...] (Kumar et al 2012).

Companies can find cost savings by reducing the environmental impact of their business processes. By re-evaluating the company's supply chain, from purchasing, planning, and managing the use of materials to shipping and distributing final products, savings are often identified as a benefit of implementing green policies.

The effective management of supply chains is one of the key areas for firms to gain a competitive advantage" (Dos Santos et Smith, 2008). Although, modern supply chains often have numerous problems often attributed to a lack of accurate and integrated data" (Seymour et al., 2008, p. 42).

In modern business environments, an effective supply chain management (SCM) is crucial to business continuity. Competition between supply chains has replaced the traditional competition between companies. Lean, Agile, Resilient and Green (LARG) paradigms are advocated as the foundation of a competitive SCM. To make a supply chain more competitive, capable of responding to the demands of customers with agility and capable of responding effectively to unexpected disturbance, in conjugation with environmental responsibilities and the necessity to eliminate processes that add no value, companies must implement a set of LARG SCM practices (...)" (Cabral et al, 2012).

4 From a traditional SCM to the E-SCM

Incorporating e-business approach in supply chain management has been proved as a competitive method for increasing values to be added and improving process visibility, agility, speed, efficiency, and customer satisfaction. Thus, E-Supply chain refers to the business activities that incorporate e-business approaches into supply chain processes. Also,

E-Supply chain management involves applying e-business technologies to assist and optimize value-adding activities in supply chains. A more detailed definition of e-supply chain management can be found in the description of Norris et al. [2]: "Electronic supply chain management (e-SCM) is the collaborative use of technology to enhance business-to-business processes and improve speed, agility, real-time control, and customer satisfaction. Not about technology change alone, e-SCM is about culture change and changes in management policy, performance metrics, business processes, and organizational structures across the supply chain."

A key feature of e-business equipped supply chain management is network centric. This focuses on connectivity, co-operation, co-ordination and information transparency. Networked supply chain partners share information, knowledge and other resources in real time. The networked relationships change the traditional supply chain information flows from linear transmission to end-toend connections, i.e. information can be transferred directly from any partner of the supply chain to another partner without distortion and delay.

5 E-supply chain and sustainability: an antagonist duo

We would expect that the integration of sustainability and e-supply chain management research would already have taken place. Yet, in a several refereed international academic journals, the authors showed that the ties between environment and e-SCM research is still not as strong as desired. The main exceptions are on either greening the supply chain or reverse logistics and closed-loop supply chain management. The conclusion was clear: the research on the impact of e-SCM on the environment is tiny when compared to other topics, let alone sustainability as a whole.

Addressing sustainability means that analysis is multi-objective and multi-disciplinary. Both are unappealing for different reasons. Multi-objective studies are more complex and often no clear-cut conclusions can be made, since conclusions depend on the decision-maker's preferences regarding the weights to be assigned to the three dimensions of sustainability. Therefore, no simple formulas or rules of thumb can be given, which may be a barrier for dissemination of the results. Summarizing to provoke a structural change it is essential to take into account the conflicting nature of the task and the context of e-supply chain management research. Otherwise, the analysis of sustainability issues in e-supply chain management research will long remain an add-on for special-interest groups instead of an integral part of mainstream research as it should be.

Logistics and supply chain management (SCM) are far reaching activities that have a major impact on a society's standard of living. In Western developed societies we have come to expect excellent logistics services and only tend to notice logistical and supply chain issues when there is a problem. To understand some of the implications to consumers of logistics activities, consider:

- The difficulty in shopping for food, clothing, and other items if logistical and supply chain systems do not conveniently bring all of those items together in one place, such as a single store or a shopping center.
- The challenge in locating the proper size or style of an item if logistical and supply chain systems do not provide for a wide mix of products, colors, sizes and styles through the assortment process.

- The frustration of going to a store to purchase an advertised item, only to find the shipment is late in arriving.

These are only a few of the issues that highlight how we often take for granted how logistics touches many facets of our daily lives. However, the various activities associated with logistics and e-SCM also have an impact on environmental sustainability.

6 Recommendations

The greening of logistics activities and e-supply chains means ensuring that these activities are environmentally friendly and not wasteful, and particularly focus on reducing carbon emissions across the entire supply chain.

The World Economic Forum WEF (2009) argued that a collaborative responsibility for greening the supply chain resides with three groups: logistics and transport service providers, shippers and buyers as recipients of such services, and both government and non-government policy makers.

WEF presented specific recommendations for these three groups, as follows:

- Transportation, vehicles and infrastructure networks Logistics and transport service providers should increase adoption of new technologies, fuels and associated processes where there is a positive business Logistics and Supply Chain Management;
- deploy network reviews of large closed networks to ensure efficient hierarchies and nodal structures,
- look to integrate optimization efforts across multiple networks,
- enable further collaboration between multiple shippers and/or between carriers and look to switch to more environmentally friendly modes within their own networks.

We suggest on our side that actions should come from several perspectives as the sustainable supply chain need commitment and contribution from several multistakeholders and fields as shown in the framework below (figure 1.)

Shippers and buyers should build environmental performance indicators into the contracting process with logistics service providers, work with consumers to better support their understanding of carbon footprints and labelling where appropriate and make recycling easier and more resource efficient. They should also support efforts to make mode switches across supply chains and begin to ‘de-speed’ the supply chain.

The same, policy makers should promote further expansion of integrated flow management schemes for congested roads and make specific investments in infrastructure around congested road junctions, ports and rail junctions, mode switches to rail, short sea and inland waterways, and consider re-opening unused rail lines, waterways and port facilities with government support.

Logistics and transport service providers should encourage wider industry commitment to improve existing facilities through retrofitting green technologies and work towards industry-wide commitments to boost investment in new building technologies, and develop new offerings in recycling and waste management, working collaboratively with customers. Policy makers should encourage industry to commit to improvements that consider the boundaries of possibilities with current and future technologies, through individual and sector-wide actions.

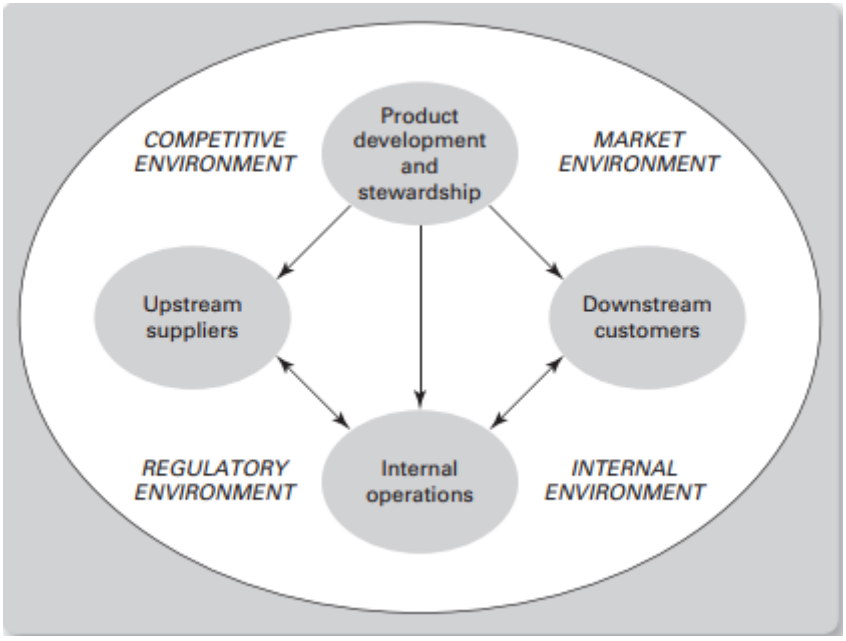


FIG. 1 – *A framework for a sustainable supply chain*

At the sourcing, product and packaging design levels: Shippers and buyers should determine how much carbon is designed into a product through raw material selection, the carbon intensity of the production process, the length and speed of the supply chain, and the carbon characteristics of the use phase. Shippers and buyers can take decisions that actively drive positive change up and down the supply chain. Shippers and buyers should agree additional standards and targets on packaging weight and elimination, and seek cross-industry agreements on modularization of transit packaging materials. They should also develop sustainable sourcing policies that consider the carbon impact of primary production, manufacturing and rework activities, and integrate carbon emissions impact into the business case for near-shoring projects.

7 Conclusion

Logistics and SCM have a major impact on the global economy as well as everyday life. The concepts of transportation or 'Go', and storage or 'Stop' activities enables the right products to be in the right place in an efficient and effective manner. However, while the trends of increased globalization, outsourcing and deeper relationships, more use of technology, lean and agile supply chain processes, and a one-way flow in the supply chain

have assisted logistics and SCM activities, they have also been detrimental from a sustainability perspective.

Emissions of greenhouse gases, use of fuel and other natural resources, other forms of pollution, and increased levels of waste from packaging are just some of these detriments. However, it is still under-developed and under-researched, particularly regarding trade-offs between a sustainable supply chain and current logistical and supply chain practices that involve long, global one way supply chains dependent on technology, outsourcing and time compression to meet ever-increasing customer demand for more and better products in a timely manner.

References

Carter, C.R.; Jennings, M.M. Logistics social responsibility: An integrative framework. J. Bus. Logist. 2002, 23, 145–180.

Carter, C.R.; Rogers, D.S. A framework of sustainable supply chain management: Moving toward new theory. Int. J. Phys. Distrib. Logist. Manag. 2008, 38, 360–387.

Gonzaléz, S.G.; Perera, A.G.; Correa, F.A. A new approach to the valuation of production investments with environmental effects. Int. J. Oper. Prod. Manag. 2003, 23, 62–87.

Gherra S, 2005, Développement durable, supply chain management et stratégie : le cas de L'éco-conception, Logistique et management, vol 13, n°1, pp 37-48

Harwood I, Humby S, 2008, Embedding corporate responsibility into supply : a snapshot of progress, European Management Journal, vol 26, pp 166-174

Hutchins MJ, Sutherland JW, 2008, An exploration of measures of social sustainability and their application to supply chain decisions, Journal of cleaner production, vol 16, pp 1688- 1698

Ijomah WL, Mcmahon CA, Hammond GP, Newman ST, 2007, Developmant of robust design-for-remanufacturing guidelines to further the aims of sustainable development, International journal of production research, vol 45, n° 18 et 19, pp 4513-4536

Jayaraman V, R Klassen, JD Linton, 2007, Supply chain management in a sustainable environment (editorial), Journal of Operations Management, vol 25, pp 1071-1074

Johansson O, Hellström D, 2007, The effect of asset visibility on managing returnable transport items, International Journal of Physical Distribution & Logistics Management, vol 37, n° 10, pp 799-815

Krajnc, D. and P. Glavic (2005). A model for integrated assessment of sustainable development. Resources, Conservation and Recycling. 43(2), p. 189-208.

Kocabasoglu C, C Prahinski, RD Klassen, 2007, Linking forward and reverse supply chain investments : the role of business uncertainty, Journal of Operations Management, vol 25, pp 1141-1160 18

Linton JD, R Klassen, V Jayaraman, 2007, Sustainable supply chains : an introduction, Journal of Operations Management, vol 25, pp 1075-1082

Maloni MJ, Brown ME, 2006, Corporate Social Responsibility in the Supply Chain : An Application in the Food Industry, Journal of Business Ethics, Springer, pp 35-52

Man R de, Burns TR, 2006, Sustainability : Supply chains, partner linkages, and new forms of self-regulation, Human Systems Management, vol 25, pp 1-12

Murphy PR, RF Poist, 2002, Socially responsible logistics : an exploratoty study, Transportation journal, vol 41, n°4, pp 23-35

Paché G, A Spalanzani, (Ed) 2007, La gestion des chaînes logistiques multi-acteurs, PUG

Pagell M, Wu Z, 2009, Building a more complete theory of sustainable supply chain management using case studies of 10 exemplars, Journal of supply chain management, vol 45, n° 2, pp 37-56

Rao P, Holt D, 2005, Do Green Supply Chains lead to Competitiveness and Economic Performance ?, International Journal of Operations & Production Management, vol 25, n° 9/10, pp 898-916

Schmidt M, Schwegler R, 2008, A recursive ecological indicator system for the supply chain of a company, Journal of cleaner production, vol 16, pp 1658-1664

Seuring S, Sarkis J, Müller M, Rao P, 2008, Sustainability and supply chain management . an introduction to the special issue, Journal of cleaner production, vol 16, pp 1545-1551

Résumé

Conjuguer le e-supply chain et le développement durable peut paraître impossible car antagonique. Le développement durable, de sa part, est devenu un problème de société qui conduit à l'invocation quasi systématique de la mutualisation à tout va, sans prendre en considération les conditions et les difficultés concrètes qui sont multiples et se renforcent probablement les unes les autres. Alors, Les sciences de gestion visent souvent la résolution de type de paradoxes et soulignent la nécessité d'adopter des démarches plutôt globales intégrant l'ensemble des acteurs de la filière concernée. C'est dans ce cadre que notre travail portera sur les modes de coopération de e-supply chain et les maillons du développement durable fournissant un cadre théorique et pratique relativement bien adapté.

INDEX

ASD'2018

ADVANCES OF DECISIONAL SYSTEMS

Marrakech