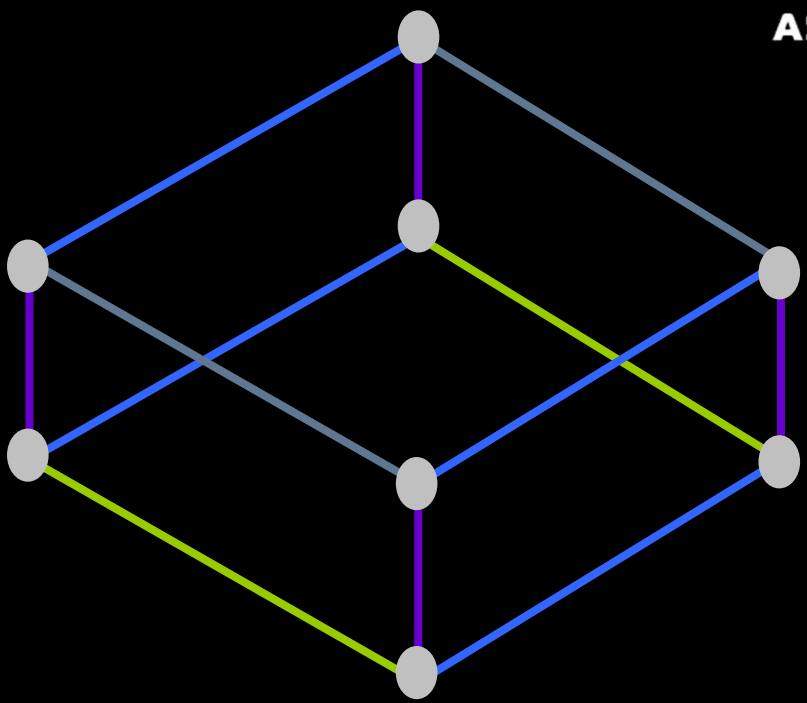
Actes de la Conférence sur les Avancées des Systèmes Décisionnels

ASD'2014



LES SYSTÈMES DÉCISIONNELS

Fondements et Applications

Éditeurs

Jamel FEKI

Faiez GARGOURI

Omar BOUSSAID

8ème édition

Conférence sur

Les **A**vancées des **S**ystèmes **D**écisionnels

ASD 2014

ASD 2014

Actes de la 8ème édition

Conférence sur

les **A**vancées des **S**ystèmes **D**écisionnels

Edités par

Jamel feki, Faiez Gargouri, Omar Boussaid

29-31 mai 2014

Hammamet, Tunisie

Préface

Les technologies des entrepôts de données et de l'analyse en ligne sont au cœur des systèmes décisionnels modernes. Consciente du rôle des recherches dans cette thématique, la communauté scientifique internationale ne cesse d'accorder une importance de plus en plus grandissante, voire privilégiée, à ce domaine ce qui s'est traduit par l'apparition de manifestations scientifiques venant enrichir le panorama des rencontres entre chercheurs et industriels.

Forte de son succès graduel et dans le prolongement des éditions précédentes (Agadir–Maroc 2006, Sousse-Tunisie 2007, Mohammedia-Maroc 2008, Jijel-Algérie 2009, Sfax-Tunisie 2010, Blida-Algérie 2012 et Marrakech-Maroc 2013), ASD fait peau neuve et s'est convertie depuis sa 7ème édition en 2013 en ***Conférence sur les Avancées des Systèmes Décisionnels***. Cette nouvelle édition ASD 2014 est accueillie cette année par la Tunisie.

ASD 2014 ambitionne de consolider les expériences conduites par les chercheurs, industriels et utilisateurs issus de communautés travaillant sur les systèmes décisionnels. L'objectif de cette huitième édition de la conférence, en particulier après le succès des précédentes éditions, est de contribuer à dynamiser davantage la recherche dans ce domaine et à créer une synergie entre les chercheurs, essentiellement mais non exclusivement maghrébins, travaillant dans leur pays ou dans des laboratoires de recherche à l'étranger. D'autre part, elle vise à renforcer les liens existants et à tisser de nouvelles relations afin de faire émerger une communauté thématisée *systèmes décisionnels.*

Ces actes regroupent les articles acceptés et présentés à cette nouvelle édition. ASD 2014 a reçu 35 soumissions d'articles en provenance de six pays (Algérie, Canada, France, Lybie, Maroc, Tunisie). Après évaluation par les membres du comité scientifique, composé par 70 chercheurs-experts internationaux du domaine, 14 articles longs et 9 articles courts ont été retenus. Ces papiers couvrent différents thèmes de recherche et d'application sur les systèmes décisionnels.

ASD 2014 est organisé par l'Institut Supérieur d'Administration des Affaires de Sfax (ISAAS) et le Laboratoire Miracl, de l'université de Sfax et a reçu leur soutien ainsi que le soutien de différentes institutions publiques d'enseignement et de recherche que nous tenons à remercier : l'Ecole doctorale EGI de la Faculté des Sciences Economiques et de Gestion de Sfax ; le Centre de Recherche en Informatique, Multimédia et Traitement Numérique des Données de Sfax ; le Laboratoire ERIC de l'université Lyon 2 ; l'université HASSAN II Mohammedia Casablanca ; la Faculté des Sciences et Techniques de Mohammedia ; l'association AER et toutes les autres institutions qui ont aidé de loin ou de près pour la réussite de cette manifestation.

Le succès de cette nouvelle édition d'ASD n'aurait pas été réalisé sans la coopération étroite des trois comités de pilotage, scientifique et d'organisation, que nous tenons également à remercier très chaleureusement.

Nous sommes très reconnaissants de leur soutien.

Nous voulons remercier l'ensemble des auteurs qui ont soumis à cette édition d'ASD. Nous félicitons ceux dont les articles ont été acceptés. Nous encourageons les autres auteurs des papiers non retenus à persévérer et à poursuivre leurs efforts.

Les éditeurs
O. BOUSSAID, J. FEKI et F. GARGOURI,

Présidents de la conférence

- FEKI Jamel (MIRACL, Université de Sfax, Tunisie)
- GARGOURI Faïez (MIRACL, Université de Sfax, Tunisie)

Présidents du comité d'organisation de la conférence

- Yasser HACHAICHI, (Université de Sfax, ISAA, Tunisie)

Comité de pilotage

- BEN ABDALLAH Hanêne, MIRACL, Université King Abdulaziz, Arabie saoudite
- BENTAYEB Fadila, ERIC, Université Lumière Lyon 2, France
- BOULMAKOUL Azedine, Université Hassan II, Maroc
- BOUSSAID Omar, ERIC, Université Lumière Lyon 2, France
- FEKI Jamel, MIRACL, Université de Sfax, Tunisie
- GARGOURI Faiez, MIRACL, Université de Sfax, Tunisie

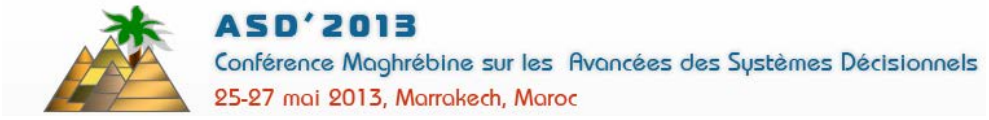
Comité scientifique

- ABDI Mustapha K., Université d'Oran, Algérie
- AHMED NACER Mohamed, USTHB Alger, Algérie
- AHMED OUAMER Rachid, Université Tizi Ouzou, Algérie
- AL-MALAISE AL-GHAMDI Abdullah, FCIT, King Abdulaziz University, KSA
- ALIMAZIGHI Zaia, USTHB Alger, Algérie
- ASFARI Ounas, Université Lyon2, France
- AYACHI Sonia, ISG, Sousse, Tunisie
- BADACHE Nadjib, CERIST Alger, Algérie
- BADARD Thierry, Université Laval, Canada
- BADIR Hassan, ENSAT, Maroc
- BADRI Abdelmajid, Université Hassan II, Maroc
- BAOTHMAN Fatmah, FCIT, King Abdulaziz University, KSA
- BELLATRECHE Ladjel, ENSMA Poitiers, France
- BEN ABDALLAH Hanene, FCIT, King Abdulaziz, KSA
- BEN AYED Mounir, FS, Sfax, Tunisie
- BEN BLIDIA Nadjia, Université de Blida Algérie
- BEN YAGNLANE Boutheinane, IHEC, Carthage, Tunisie
- BEN YAHIA Sadok, FS, Tunis, Tunisie
- BENHARKAT Nabila, INSA de Lyon, France
- BENSLIMANEl Djamel, Université Lyon1, France
- BENTAYEB Fadila, Université Lyon 2, France
- BIMONTE Sandro, Cemagref, Clermond-Ferrand, France

- BOUAZIZ Rafik, Université de Sfax, Tunisie
- BOUFAIDA Mahmoud, Université de Constantine 2, Algérie
- BOUFAIDA Zizette, Université de Constantine 2, Algérie
- BOUFARES Faouzi, LIPN Paris France
- BOUKHALFA Kamel, USTHB, Alger, Algérie
- BOUKRAA Doulkifli, Univesrité de Jijel, Algérie
- BOULMALKOUL azedine, Université Hassan II, Maroc
- BOURAMAOUL Abdelkrim, Université de Constantine 2, Algérie
- BOUSSAID Omar, Université Lyon 2, France
- CHAABOUNI Jami, FSEG, Sfax, Tunisie
- CHKIR Ali, FSEG, Sfax, Tunisie
- DARMONT Jérôme, Université Lyon2, France
- EL HEBIL Farid, INPT Rabat Maroc
- EL-MOUADiB Faraj, FIT, Benghazi, Lybie
- ELAMMARI Mohamed, FIT, Benghazi, Lybie
- ELFILALI Sanaa, EMSI FSBM, Maroc
- FAVRE Cécile, Université Lyon 2, France
- FEKKI Jamel, Université de Sfax, Tunisie
- GARGOURI Faiez, Université de Sfax, Tunisie
- GHOZZI Faiza, Université de Sfax, Tunisie
- HACHAICHI Yasser, Université de Sfax, Tunisie
- HAMMAMI Mohamed, Université de Sfax, Tunisie
- HARBI Nouria, Université Lyon 2, France
- HIDOUCI Walid, ESI Alger, Algérie
- JERBI Houssem , University College Dublin, Ireland
- KABACHI Nadia, Université Lyon1, France
- KAZAR Okba, Université de Biskra, Algérie
- KHROUF Kaïs, ENET'Com, Tunisie
- LALAM Mustapha, Université de Tizi-Ouzou, Algérie
- LEMIRE Daniel, Université du Québec à Montréal, Canada
- SLIMANI Yahya, FSTunis, Tunisie
- MELIT Ali, Université de Jijel, Algérie
- MEZIANE Abdelkrim, CERIST Alger, Algérie
- NABLI Ahlem, Université de Sfax, Tunisie
- OULAD HAJ THAMI Rachid, ENSIAS Rabat, Maroc
- RAVAT Frank, Université de Toulouse, France
- REGUIEG F Zohra, Université de Blida, Algérie
- SEKHRI Larbi, Université d'Oran
- SIDHOM Sahbi, Université de Nancy, France
- TAGHEZOUT Noria, Université d'Oran, Algérie
- TESTE Olivier, Université de Toulouse, France
- ZAROUR Nasreddine, Université de Constantine 2, Algérie
- ZEGOUR Djamel Eddine, ESI d'Alger, Algérie

Comité d'organisation :

- HACHAICHI Yasser (ISAA, Université de Sfax, Tunisie)
- MOALLA Mohamed Sahbi (ISET, Université de Sfax, Tunisie)
- CHAABANE Mohamed Amine (ISAAS, Université de Sfax, Tunisie)
- NABLI Ahlem (FS, Université de Sfax, Tunisie)
- BEN KRAIEM Maha (MIRACL, Université de Sfax, Tunisie)



Sommaire

Session 1 : Entrepôts de données et analyse en ligne

Session 2 : Sémantique et ontologies décisionnelles

Session 3 : Sécurité et gestion des ED

Session 4 : Modélisation multidimensionnelle

Session 5 : Systèmes Décisionnels et Applications

Near Real-Time Space-Time-Path Warehousing for Hazardous Materials Transportation Trajectories

Azedine Boulmakoul *, Lamia Karim **, Ahmed Lbath***

*,**Computer Science Department, Faculty of Sciences and Technology (FSTM),
University Hassan II, Mohammedia, Morocco
*azedine.boulmakoul@gmail.com
** lkarim.lkarim@gmail.com
*** Computer Science Department, Laboratoire LIG, University Joseph Fourier
Grenoble, France
ahmed.Lbath@ujf-grenoble.fr

ABSTRACT
In recent years, a significant portion of material transported is harmful to human and environment. Thus, the transportation of hazardous materials and its potential consequences raise public interest typically when there is a release due to an accident. In this paper, we introduce HazMat Space Time Path Data Warehouse that can be used for near real time decision making in different applications domain, using MongoDB as a NoSQL database for scalable, fault-tolerant and distributed space time paths big data storage and processing system. The system components are integrated into an interoperable software infrastructure respecting intelligent transport systems architecture. This infrastructure is distributed and based on a service-oriented architecture. It is also scalable by integration of MongoDB with Hadoop for large-scale distributed data processing. In this work, we also give an assessment of the performance, scalability and fault-tolerance of using MongoDB with Hadoop, towards the goal of identifying the right architecture and software environment for HazMat spatio-temporal data analytics.

1. Introduction

Dangerous goods or hazardous materials (HazMats) include explosives, toxics gases, flammable liquid and solids, oxidizing substances, and hazardous wastes. HazMat events and shipping data are needed in different location services to take near real time decisions like determining a route which minimizes the likelihood that the risk will be greater than a set threshold. Because hazmat accidents are generally being regarded as low probability and high consequence events. This kind of accidents/incidents attracts public attention as the damage to human being's health, deaths, economic and environment losses are high. Nowadays, it becomes more and more important to combine different applications fields with spatial and time related information. "80% of All Information is Geospatially Referenced" (Fitzke et al., 2010) and valuable real-time geo-tagged data are produced by indoor and outdoor location sensing. The analysis of such spatio-temporal big data raises opportunities for many innovative applications and has multiple challenges as short latency, scalability, performance, query processing, high-precision positioning, and privacy preservation.

Near Real-Time Space-Time-Path Warehousing for Hazardous Materials Transportation Trajectories

Location Based Services growth and need a near real time hazardous space time path data warehouse to analyze and make decision from spatio-temporal captured data. In the other hand, GeoStream data grow so large that is difficult to capture, store, manage, share, analyze and visualize using classical models and databases. Also, visualizing the implicit information of hazardous transportations trajectories data is very important for analyzing human activities and is of great value in the decision making process.

Our objective is to provide a hazmat space time path data warehouse that can be used for near real time decision making in different applications domain. It is based on the unified trajectories meta-model to be very adaptive to various locations based services.

The remainder of the paper is organized as follows: section 2 will present related works on trajectories data warehouses. Section 3 provides an overview of space time path presentations and presents the proposed HazMat space time path warehousing conceptual schema. Section 4 presents the proposed system for hazmat space time path data warehouse. Finally, section 5 will provide conclusions of our proposed data warehouse.

2. Related works

Data warehouses were developed for decision support. They include information from various transactional systems of the company. Data warehouses have emerged around 1990 in response to the need to gather all company's information in a unique database for analysts and managers (Doucet et al., 2001). All data, including their history, are used in many fields, such as: data analysis, decision support and in other applications (Benitez et al., 2001). Spatial data warehouses are based on the concepts of data warehouses and additionally provide support to store, index, aggregate, and analyze spatial data (MacEachren, 2001). The research in this field mostly focuses on conceptual models for spatial data warehouses and SOLAP as a client application on spatial data warehouses (Fubédard et al., 2001). Spatial Eye is an example of spatial data warehouse. Spatio-temporal data warehouses (Elzbieta et al., 2008) complete the spatial data warehouse by including both spatial and time components as there is a need to include the temporal aspects as well. The GeoPKDD trajectory data warehouses (Damiani et al., 2007) aim at extracting user-consumable forms of knowledge from large amounts of raw spatio-temporal geographic data. (Salvatore et al., 2007) discussed the problem of storing and aggregating in a trajectories data warehouse, and they contributed a novel way to compute an approximate presence aggregate function, which algebraically combines a bounded amount of measures stored in the base cells of the data cube.

The Unified Moving Object Trajectories' Meta-model (Boulmakoul et al., 2012) describes a general meta-model that could be used by different application domains; it can also use an object approach and integrates previous trajectories models described in (Güting et al., 2004, Meng et al., 2003, Wolfson et al., 1998, Yan et al., 2010). Using the space-time event ontology, the meta-model models space according to OGC Spatial Data Model (OGC, 2008), Observation domain of trajectory, according to OGC Sensor Meta Model and OGC Feature Type, physical and virtual activities between the beginning and the end of space time path (Shaw, 2011), sensors used for collecting moving object's traces, and movement patterns using composite region of interest. Simone et al. (Simone et al., 2011) provided a solution

named St-Toolkit for designing and implementing trajectory data warehouse based on semantic trajectory model, introduced by (Spaccapietra et al., 2008), in relational environment.

However, there are several limitations in current data warehouse tools to cope with spatio-temporal data as found in the literature (Vaisman et al., 2009). Data warehouses provide a decision support system for large stores of historical data, but are not yet adequate to support the hazmat space time path data warehouse from different facets: raw, structured, semantic, based on region of interest and also activities and none of them uses NoSQL database to store different kind of trajectories data warehouses.

3. HazMat Space Time Path Data Warehouse

A hazardous material is a substance which by its physico-chemical characteristics, toxicological, or by the nature of the reactions that may occur, may present risk to humans, property and / or the environment. The risk of transporting hazardous material is the result of a transport accident or incident on health and environment. Transportation accidents of dangerous goods take place in few minutes in an unpredictable place. Given that consequences of an incident are often considerable, location based services purpose is to help against the immediate consequences of the disaster and the random nature of first aid. Several steps must be taken in near real time to avoid the damage. Aside, restrictive regulations based the training of staff (drivers) application of strict driving and traffic rules, an obligation tanks approval for vehicles signaling and hazardous products transported.

We propose in the following the use of a data warehouse in near time to help in case of an eventual risk, by sending alerts to people near to the disaster geographical location, to prohibit some roads and find backup routes.

Several research efforts have been carried out on management of trajectories. Some include modeling and representing trajectories, raw trajectory is the recording of the positions of an object at specific space time domain (GeoStream data), for a given moving object (e.g. person, vehicle) and a given time interval, it is presented as a sequence of geometric position in 2D spatial system $[(x_1, y_1, t_1), (x_2, y_2, t_2), \ldots (x_m, y_m, t_m)]$ representing the movement as a sequence of positions at time $t_1, t_2, \ldots t_m$. Structured trajectory (Spaccapietra et al., 2008), defined as a raw trajectories structured into segments corresponding to meaningful steps in the trajectory trace (e.g. travel). And in (Spaccapietra et al., 2008) provides a semantic view of trajectory, which enables applications to associate whatever semantics they want with trajectories. However, this approach is only applicable to transactional schema. Indeed, no work has been published using trajectories as semantic objects with activities on multidimensional data modeling. Other recent approach describes trajectories in both spatial and temporal contexts based on Region of Interest (Giannotti et al., 2007) by defining spatial neighborhood and temporal tolerance. The "aquarium" (Hongbo et al., 2007) of the relevant time-space unit describes anything having spatial and temporal extent as paths (for instance, people, plants, animal).

Multidimensional modeling is the foundation of data warehouses (Song et al., 2001). It is characterized by two primitives which are Facts and Dimensions. Those latter are used to construct the star schema (Freitas et al., 2002), the snowflake schema (Levene et al., 2003) or the constellation schema (Teste, 2001). In the proposed schema, figure 1, we present the proposed conceptual HazMat space time path warehousing using composite document nota-

tion. HazMat Space Time Path Data Warehouse sources and requirements are gathered from different spatio-temporal sensors and also location based service that store accidents and incidents.

In fact, our model is based on the unified trajectories Meta model (Boulmakoul et al., 2012) for designing and implementing hazmat space time path data warehouse. We use the star schema as a multidimensional model. Thanks to redundancy, we can provide horizontally-scalable systems as distribution of data across multiple machines is easy and does not cause problems. Space time path measures are given below:

Hazmat Space Time Path fact table represents the subject orientation and the focus of analysis. It typically contains measures that are attributes representing the specific elements of analysis. A dimension contains attributes that allow exploring measures from different perspectives. Hazmat Space Time Path data warehouse measures are:

- Average_duration: corresponds to the duration of hazardous space time path.
- Total hazmat accident: corresponds to total number of hazardous accidents.
- Total transported hazmat: corresponds to total quantity of hazardous materials transported.
- Frequent hazardous accident: corresponds to hazardous materials that are frequently transported in case of an accident.
- Frequent hazardous causes: corresponds to hazardous materials that are frequently transported in case of an accident.
- Frequent hazardous consequences: corresponds to frequent consequences of hazardous materials transported in case of an accident.
- Frequent accident region: corresponds to frequent regions where hazardous materials accidents occur.
- Frequent time period accident / incident: corresponds to frequent regions where hazardous materials accidents occur.
- Most safety region: corresponds to the most safety regions.
- Minimum_duration; corresponds to the minimum duration of the hazmat space time path.
- Maximum_duration: corresponds to the maximum duration of the hazmat space time path.
- Average_distance: corresponds to the average distance of the hazmat space time path.
- Average_speed: corresponds to the average speed of the hazmat space time path.
- Number_of_stops: contains number of stops in the hazmat space time path.
- Number_of_moves: contains number of moves in the hazmat space time path.
- Most_frequent_ROI: contains the most frequented region of interest in a hazmat space time path.
- Most_frequent_activity: contains the most frequented activity practiced in the hazmat space time path.
- Count_users: contains total number of moving objects taking the same hazmat space time path.
- Count trajectories: contains total number of trajectories of vehicles in a period.
- Shape: corresponds to the interpolated shape of the hazmat space time path.

In the proposed schema, we used hierarchy documents that contain several related levels. Principally, it is used for roll-up and drill-down operations. In the following, we describe dimensions of the proposed space time path conceptual schema:

Vehicle dimension: contains information about the tracked vehicle like reference number, type of vehicle, traveled kilometers, capacity, mobile sensor type and also different signalization. The vehicle can carry several hazardous materials. This latter is characterized by ONU number and quantity, and it belongs to a determined class according to ADR. The classification of dangerous goods (procedure for classifying solutions and mixtures, structure of the list of substances, classes of hazardous goods, nature of transported hazardous goods, physico-chemical and toxicological properties of dangerous goods).

In addition, we also warehouse the type of hazmat packaging with its entire characteristic (id, brand, volume, reusability, danger and handling labels) for a full Hazardous space time paths analysis. The vehicle dimension is related to the driver entity with his history of training. Recipient dimension: presents the company that supports the Hazardous materials on arrival. Carrier dimension: presents the company that transports Hazardous materials with or without transport contract. Filler dimension: presents the company that fills dangerous materials in the tracking vehicle. Charger dimension: presents the company that load packaged Hazardous materials in a vehicle. Accident / Incident dimension: To minimize damage and facilitate analysis of accidents / incidents during the transport of materials, the following information is warehoused: the estimated amount of lost product, the average retention, the material retention means, type of failure of retention means and description of the event. The near real time decision when an accident/ incident occur in a hazardous space time path requires a full knowledge about Specific weather conditions, Cause of the event, and Consequences of the event (Intervention of the authorities, estimated amount of material or environmental damage, Product loss, Bodily injury related to dangerous goods).

Event ontology has already been proven useful in a wide range of contexts, due to its simplicity and usability. The SHOE General Ontology defines an event as something that happens at a given place and time. In Dublin Core metadata standard, an event is defined as ‘a non-persistent, time based occurrence’. (Quine, 1985) described events as objects where objects are regions bounded in space and time.

Space Time Event: presents an event as an occurrence that happens in a small space and lasts a short time. From spatial point of view, it is a composition of Spatial Object. Each spatial object is characterized by a spatial reference and geometry to model a raw trajectory. A spatial object could be a point, line or a polygon to present raw trajectories. Semantic trajectory (Spaccapietra et al., 2008) expresses the application oriented meaning using four components (stop, move, begin and end). Stop, move, begin and end are no more spatio-temporal position, but semantic objects linked to general geographic knowledge and application geographic data. From semantic point of view, a semantic object is characterized by a Toponyme, and linked to the semantic begin move end and stop which respectively contains semantic information, time of begin and of the end for a given begin, move stop and end. To analyze the space time path data warehouse, the schema considers the presentation of spatio-temporal data and activities for each event. Activity dimension contains information about activity’s duration, time of begin and of end activity and also information about the kind of activity (Driving, Distributing, Stocking). As concerning the temporal aspects, it is detailed with others dimensions tables like Hour, Day, Month, Year and Time Period.

Thus, using the star schema combined with structuring dimension in several analysis perspectives, we provided a general HazMat space time path warehousing for different kind of Space time paths.

Additionally, hazmat space time path data warehouse schema presents hazmat trajectories in both spatial and temporal contexts based on Region of Interest and activities. The spatial

neighborhood is presented using Point of Interest dimension characterized by longitude, latitude and name. A hierarchy of spatial neighborhood is used for roll-up and drill-down operations. Area of interests has a shape and a surface. City region of interests to present a city as a region of interests, it contains about the city region of interest name, surface, first language and religion. In other applications domain, region of interests could be presented as a Voronoi network of interest or a modal network.

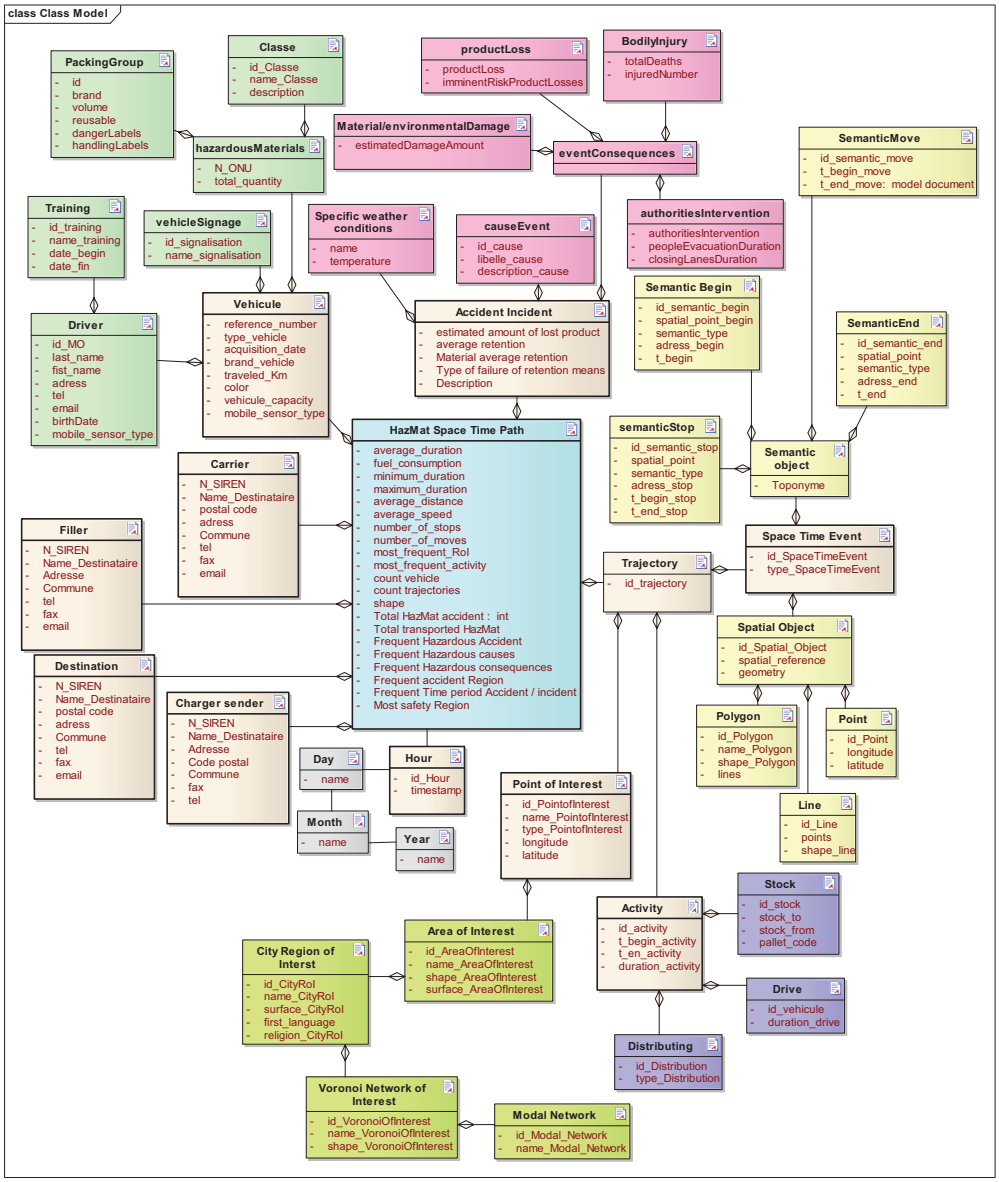


Fig. 1. *HazMat Space Time Path Data warehousing conceptual schema using composite documents.*

4. Proposed system for hazmat space time path datawarehouse

The system architecture of data warehouses is traditionally associated with the relational model as operational database and SQL as a query language. Whereas location based services requires more flexibility and agility in analyzing of hazmat space time path data warehouse information. In some key areas of business need and data processing, Document-oriented databases of the categories NoSQL offers an alternative powerful and extension to the traditional relational approach. The challenge of analyzing of the massive hazmat space time path data warehouse in near real time requires the use of a fast and scalable solution to bear the burden of voluminous data. MongoDB is chosen for its performances and scalability (Boulmakoul et al., 2013).

In this work, we provide a NoSQL database for the storage of hazmat space time path data warehouse. Other supports components are provided for collecting and visualizing of data and spatio-temporal events related to hazmat. All given components are integrated into an interoperable software infrastructure respecting intelligent transport systems architecture. This infrastructure is distributed and based on a service-oriented architecture. It is also scalable by integration of MongoDB with Hadoop for large-scale distributed data processing. In this work, we also give an assessment of the performance, scalability and fault-tolerance of using MongoDB with Hadoop, towards the goal of identifying the right architecture and software environment for HazMat spatio-temporal data analytics.

NoSQL databases

The NoSQL databases break the limitations of the relational model in terms of scalability and volume. Indeed, a recurrent problem of relational database is the loss of performance when you need to process a large volume of data. In addition, the proliferation of distributed architectures has brought the need for adapting natively solution mechanisms of data replication and load management. The acronym NoSQL signifies “Not Only SQL” (Mike, 2012). It is designed for storing data in a much simpler, flatter, and non-relational manner that allows data repositories to be scaled up. In a NoSQL database, there is no fixed schema so we can store, in the same entity, heterogeneous spatio-temporal data and activities generated by different kinds of locations sensors. NoSQL is a class of database management systems (DBMS) that do not follow all of the rules of a relational DBMS and cannot use traditional SQL to query data. The term is somewhat misleading when interpreted as "No SQL," and most translate it as "Not Only SQL," as this type of database is not generally a replacement but, rather, a complementary addition to RDBMSs and SQL.

Relational database scales up by getting faster hardware and adding memories whereas NoSQL, on the other hand, can take advantage of scaling out by spreading the load over many commodity systems. Consequently, NoSQL is an inexpensive database for scaling trajectories space time path data. Companies like (Google, Facebook, Twitter, Amazon, Twitter, Adobe, Viadeo) have left the relational world and all use NoSQL in one way or another because they have seen their needs in terms of load and data volume grow exponentially. Existing NoSQL solutions can be grouped into 4 main families: Key-values Stores, Column Family Stores, Document Databases, and Graph Databases.

MongoDB

MongoDB (from "humongous") is an open-source document database, and the leading NoSQL database. MongoDB (from "humongous") is an open-source document database and the leading NoSQL database developed by 10gen in 2009 (MongoDB, 2013). It is written in C++, document-oriented storage, full Index, rich document-based queries, and flexible aggregation and data processing. MongoDB may contain several databases. Using JavaScript for its query language, MongoDB supports both single and complex queries. Storing JSON documents, the basis documents format of many modern geospatial applications, makes it easy to build on top of MongoDB. MongoDB database benefits from ascending, descending, unique and geospatial indexes. To makes performance better, JSON is stored by MongoDB in an efficient binary format called BSON. BSON is a binary serialization of JSON documents and stands for Binary JSON. In general, document-oriented (e.g. MongoDB) are most directly relevant to business intelligent because of their more flexible and extensive search and retrieval functionality.To scale its performance on a cluster of servers, MongoDB uses a technique called sharding, which is the process of splitting the data evenly across the cluster to parallelize access.

This is implemented by breaking the MongoDB server into a set of front-end routing servers mongos), that route operations to a set of back-end data servers (mongod).

MongoDB queries examine one record at a time, which means that queries across multiple records must be implemented on the client or use MongoDB's built-in MapReduce (MR). Though MongoDB's MR can be executed in parallel at each shard, there are two major drawbacks: (1) the language for MR scripts is JavaScript, which is slow and has poor analytics libraries, and (2) the SpiderMonkey (Spider, 2013) Javascript implementation used by MongoDB, is not threadsafe, so only one MapReduce program can run at a time.

HADOOP

Hadoop (Jason, 2009) is the Apache Software Foundation top-level project that provides both distributed storage and computational capabilities. The Hadoop project provides and supports the development of open source software that supplies a framework for the development of highly scalable distributed computing applications. The Hadoop framework handles the processing details, leaving developers free to focus on application logic.

The Hadoop Core project provides the basic services for building a cloud computing environment with commodity hardware, and the APIs for developing software that will run on that cloud. The two fundamental pieces of Hadoop Core are the MapReduce framework, the cloud computing environment, and the Hadoop Distributed File System (HDFS).

Hadoop has been designed to run on multiple servers simultaneously. In practice, the data is spread across different servers, and Hadoop manages a replication systemso as to ensure a high availability of data, even when one or more servers are failing. The strength of Hadoop is to benefit from the computational power of multiple servers unmarked cluster. The parallelized processing is managed by MapReduce, whose mission is to distribute the treatments on different servers, and vice versa to aggregate the elementary results in an overall result.

The MapReduce (Alex, 2012) model simplifies parallel processing by abstracting away the complexities involved in working with distributed systems, such as computational parallelization, work distribution, and dealing with unreliable hardware and software. With this abstraction, MapReduce allows the programmer to focus on addressing business needs, ra-

ther than getting tangled up in distributed system complications. MapReduce decomposes work submitted by a client into small parallelized map and reduce workers. In traditional applications, the built-in aggregation functionality provided by MongoDB is sufficient for analyzing data [13]. However, storing and analyzing the collected spatio-temporal data of trajectories need more complex data aggregation. This is the reason to use Hadoop as a powerful framework for complex analytics queries in our system architecture.

Verma et al. [24] evaluate both MongoDB and Hadoop MapReduce performance for an evolutionary genetic algorithm.

There are other NoSQL databases that provide Hadoop support. Cassandra is a peer to peer key-value store that has the ability to replace Hadoop's HDFS storage layer with Cassandra (CassandraFS). HBase is an open source distributed column oriented database that provides Bigtable inspired features on top of HDFS. HBase includes Java classes that allow it to be used transparently as a source and/or sink for MapReduce jobs. Our choice of MongoDB is motivated by the need for a document-oriented store for HazMat space time paths visualization on the map.

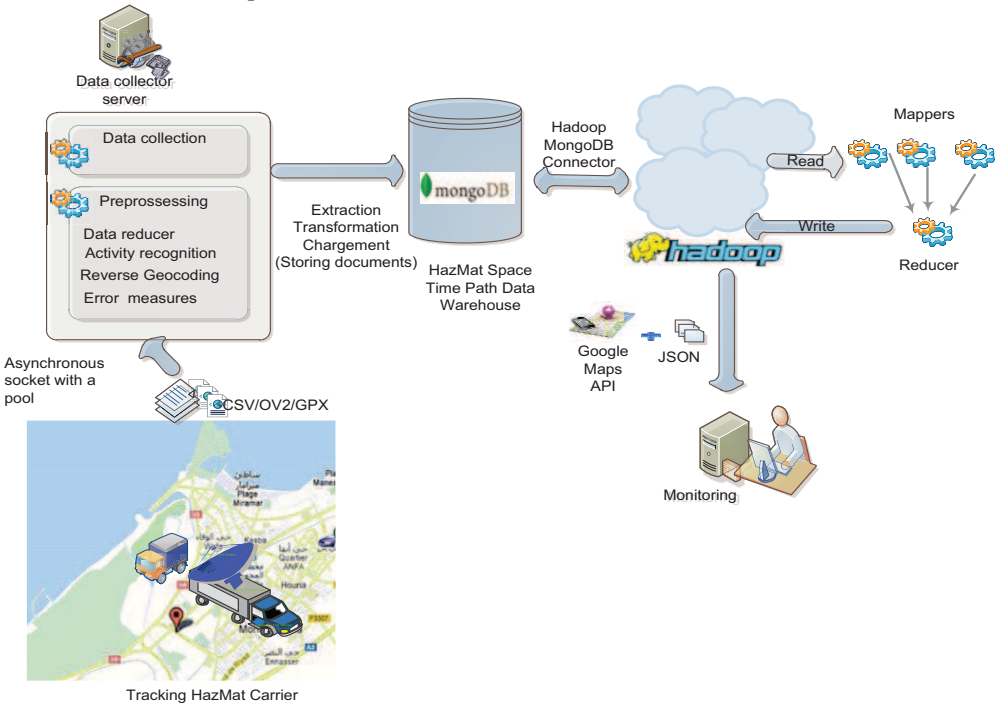


Fig. 2. *System architecture for HazMats Space Time Path Data Warehousing.*

In figure 2, we present the proposed scalable architecture for HazMats Space Time Path Data Warehousing respecting intelligent transport systems architecture. The first stage is to collect spatio-temporal data of trajectories, as GPX, OV2, or CSV files from different GPS enabling devices of drivers and vehicles. We use asynchronous .Net sockets for collecting data to data collector server. Then the collected data is processed using Data reducer, Error measures, Reverse Geocoding, and Activity recognition services. The Extract Transform and Load phase from heterogeneous data will be discussed in future article to respect the paper

pages limit. After that, data could be stored on MongoDB data base, processed within Hadoop via one or more MapReduce jobs. Output from these MapReduce jobs can then be written back to MongoDB for later querying and ad-hoc analysis. Since results format returned from MongoDB are in JSON (JavaScript Object Notation) with no needed conversion, and also JSON is much faster than other XML based technologies. We use JSON format, in the proposed framework, to monitor Dangerous Goods Transport Space Time Path Data in browsers.

5. Hazmat space time path data warehouse analysis

The proposed HazMat Space Time Path Data warehouse in a NoSQL database MongoDB is designed for query and analysis the big volume of spatio-temporal data. MongoDB supports a rich, ad-hoc query language of its own. Therefore, in a scalable, fast and agile way, complicated HazMat Space Time Path Data warehouse analytical queries can be reduced to nearly line Mongo queries as there is no joins (documents are embedded).

In the following some examples for querying the HazMat Space Time Path Data warehouse in MongoDB:

- Find all trajectories related to the vehicle X

```
db.HSTP.find( { 'Vehicule.reference_number': X} , { Trajectory :1} )
```

- Find all trajectories related to the vehicle X between two dates

```
var start = new Date(2013, 3, 1);
var end = new Date(2010, 4, 1);
db.HSTP.find( { 'Vehicule.reference_number': 12 , 'Timeperiod.begin_Period' : {"$gte": start , "$lt": end}} , { Trajectory :1} )
```

- List the hazardous products transported by Vehicule X between two dates

```
db.HSTP.find( { 'Vehicule.reference_number': 12 , 'Timeperiod.begin_Period' : {"$gte": start , "$lt": end}}, { 'Vehicule.hazardousMaterials' :1} )
```

- List the hazardous products transported by Vehicule X between two dates in the point of interest named ‘Massira’

```
db.HSTP.find( { 'Vehicule.reference_number': 12 , 'Timeperiod.begin_Period' : {"$gte": start , "$lt": end}, 'POI.name_PointofInterest: "Massira" ' }, { 'Vehicule.hazardousMaterials' :1} )
```

6. Conclusion

In this paper we have proposed hazmat space time path data warehouse conceptual model for establishment of a decisional database that capture HazMat trajectories, shipments and occurring incidents or accidents.

The proposed system can be exploited in different applications domains and is able to handle in near real time GeoStream amount of spatio-temporal data of hazardous trajectories system from different moving objects and analyzing them in a scalable, fast and agile way.

Decision Makers can mine hazmat space time path data warehouse in MongoDB database using Hadoop framework and its MapReduce paradigm to benefit from the maximum of performance and scalability.

The perspective of this work is, in the short term, to implement more complex aggregate functions to perform the space time path data warehouse analytical operations, and to contin-

ue experimentations of the proposed hazmat space time path data warehouse by increasing the load and using the MapReduce paradigm in a cloud computing environment.

References

Alex Holmes. 2012. Hadoop in Practice. Manning Publications Co.

Benitez, E., Collet, C. and Adiba, M. 2001. Entrepôts de données : caractéristiques et probléma-tique. Revue TSI, 20(2).

Boulmakoul, A., and Karim, L. 2013. A framework for scalable NoSQL storing moving objects' trajectories. Conférence Maghrébine sur les Avancées des Systèmes Décisionnels, ASD'2013.

Boulmakoul, A., Karim, L., and Lbath, A. 2012. Moving Object Trajectories Meta-Model and Spatio-temporal Queries. International Journal of Database Management Systems; volume 4, Number 2, p. 35-54.

Damiani, M. L., Vangenot, C., Frentzos, E., Marketos, G., Theodoridis, Y., Veryklos, V., and Raffaeta, A. 2007. Geographic privacy aware Knowledge Discovery and Delivery.

Doucet A. and Gangarski, S. 2001. Entrepôts de données et Bases de Données Multidimension-nelles, Chapter 12 Book: Bases de Données et Internet, Modèles, langages et systèmes. Hermès editions.

Elzbieta, M., Esteban, Z. 2008. Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications. Data-Centric Systems and Applications. Springer; 1st ed. 2008. Corr. 2nd printing edition (April 6, 2011).

Fitzke, J., and Greve, K. 2010. Frei oder umsonst? - Nutzergenerierte Geoinformation zwischen Freiheit und Kostenlosigkeit. In: Angewandte Geoinformatik - 22. GIT-Symposium. 1. ed., Wichmann, Berlin; p. 732–741.

Freitas, G., Alberto, M., Laender, H.,and Luiza, M. 2002. Getting Users Involved in the Develop-ment of Data Warehouse Application, In Proc. of the 4th International Workshop (DMDW), To-ronto, Canada, p. 3-12.

Fubédard, Y., Merrett, T. and Han, J. 2001. Fundamentals of spatial data warehousing for geo-graphic knowledge discovery. Geographic Data Mining and Knowledge Discovery, London: Taylor and Francis, 53-73.

Giannotti, F., Nanni, M., Pedreschi, D., and Pinellin, F. 2007. Trajectory Pattern Mining. Interna-tional Conference on Knowledge Discovery and Data Mining; p. 330-339.

Güting, R.H., Behr, T., Almeida, V., Ding, Z., Hoffmann, F., and Spiekermann, M. Secondo. 2004. An extensible dbms architecture and prototype. Technical report.

Hongbo, Y., Shaw, S. 2007. Revisiting Hägerstrand's time-geographic framework for individual activities in the age of instant access", In H. Miller (ed.) Societies and Cities in the Age of Instant Access. Dordrecht, The Netherlands: Springer Science, p.103-118.

http://www.spatial-eye.com/Engels/Applications/Spatial-DWH/page.aspx/117.

Jason Venner. 2009. Pro Hadoop. Build scalable, distributed applications in the cloud.

Levene, M. and Loizou, G. 2003. Why is the Snowflake Schema a Good Data Warehouse Design? In Information Systems Vol. 3, N°28, p. 225-240.

MacEachren, A. M. and Kraak, M. 2001. Research challenges in geovisualization. Cartography and Geographic Information Science.

Meng, X., and Ding, Z. 2003. DSTTMOD: A Discrete Spatio-Temporal Trajectory Based Moving Object Databases System. DEXA, LNCS 2736, Springer; p. 444-453.

Mike, L. 2012. Planning for Big Data. O'Reilly Media. chapter 8 The NoSQL Movement. ISBN: 978-1-449-32967-9.

MongoDB 10gen. 2013. Available from: http://www.mongodb.org.

OGC 07-022r1 Version: 1.0. 2008. Available from: httpnt Systems and Machine Learning.

Quine W. V. O. 1985. Events and reification, In LePore E., McLaughlin B. P.(Eds.) Actions and events: Perspectives on the philosophy of Donald Davidson, Oxford, p.162–171.

Salvatore, O., Renzo, O., Alessandra, R., and Alessandro, R. 2007.Trajectory Data Warehouses: Design and Implementation Issues. Journal of Computing Science and Engineering, Vol. 1, No. 2, December 2007, p. 211–232.

Shaw, S. 2011. A Space-Time GIS for Analyzing Human Activities and Interactions in Physical and Virtual Spaces. Center for Intelligent Systems and Machine Learning.

Simone, C., Macedo, J., and Spinsanti, L. 2011. St-Toolkit: A Framework for Trajectory Data Warehousing. AGILE 2011, April 18-22.

Song, I., Medsker, W. 2001. An Analysis of Many-to- Many Relationships Between Fact and Dimension Tables in Dimension Modeling. In Proc. of the International Workshop on Design and Management of Data Warehouses, Vol 6, Interlaken, Switzerland, p. 1-13.

Spaccapietra, S., Parent, C., Damiani, M.D., Macedo, J.A., Porto, F., and Vangenot, C. 2008. A Conceptual view on trajectories. Data and Knowledge Engineering; p. 26–146.

Spider Monkey. 2013. https://developer.mozilla.org/en/ SpiderMonkey.

Teste, O. 2001. Towards Conceptual Multidimensional Design in Decision Support Systems. In Proc. of the 5th East-European Conference on Advances in Databases and Information Systems (ADBIS), Vilnius, Lithuania, p. 77-88.

Vaisman, A., and Zimányi, E. 2009. What is spatio-temporal data warehousing? In DAWAK.

Wolfson, O., Xu, B., Chamberlain, S., Jiang, L. 1998. Moving objects databases: Issues and solu-tions. Proceeding of the 10th International Conference on Scientific and Statistical Database Man-agement (SSDBM), USA, IEEE Computer Society; p. 111-122.

Yan, Z., Parent, C., Spaccapietra, S., and Chakraborty, D. 2010. Hybrid Model and Computing Platform for Spatio-Semantic Trajectories. 7th Extended Semantic Web Conference, Heraklion, Greece.

Réduction du nombre des prédicats pour les approches de répartition des entrepôts de données

Mourad Ghorbel*, Karima Tekaya**
Abdelaziz Abdellatif***

*Université de Tunis El Manar, Faculté des Sciences de Tunis,
Département informatique, URAPOP, 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. Dans le domaine des entrepôts de données, la plupart des approches de répartition se basent essentiellement sur les techniques de fragmentation et d'allocation des tables. Ces approches exploitent communément en entrée les prédicats extraits des requêtes OLAP les plus utilisées dans le processus de partitionnement. Etant donné que le nombre de prédicats est en augmentation continue, et vu l'impact négatif qu'engendre cette augmentation sur le nombre de partitions générées, il devient intéressant de le réduire avant de procéder au processus de fragmentation. Dans cet article, nous proposons une solution basée sur un algorithme de classification permettant de diminuer le nombre des prédicats pour les approches de répartition des entrepôts de données. La solution proposée englobe trois phases: une phase de codification des prédicats sous forme de matrices binaires, une phase de classification de ces prédicats par l'algorithme k-means et une phase finale pour la réduction du nombre de prédicats. Nous avons validé notre solution sur un entrepôt de données réel issu du benchmark APB1.

1 Introduction

L'accroissement du volume des données dans les systèmes d'information décisionnels (SID) est de nos jours une réalité à laquelle chaque entreprise doit faire face. Dans ce contexte, l'entreprise doit établir des solutions novatrices qui amélioreront la gestion des données et permettront de tirer parti des données massives afin d'assurer une croissance rentable. Pour ce faire, plusieurs techniques ont été adoptées notamment, l'exploitation des index, des vues matérialisées et des techniques de partitionnement. Aujourd'hui, l'augmentation du volume et la variété des données ainsi que la décentralisation des décideurs et des bases de données (BD) opérationnelles, conjuguées aux récentes avancées technologiques en matière de télécommunication appelle à l'élaboration du SID selon une toute nouvelle perspective : les solutions d'en-

trepôts de données (ED) répartis. Les principaux défis liés à la répartition des ED, comprennent la fragmentation des données des tables multidimensionnelles selon une liste de prédicats extraite à partir des requêtes OLAP (On Line Analytical Processing) distantes les plus utilisées et leur répartition sur les sites de l'entreprise selon les besoins des décideurs.

2 Problématique

Le processus de répartition commence tout d'abord par la fragmentation des tables de l'ED. La fragmentation peut être horizontale, verticale ou mixte. La Fragmentation Horizontale (FH) peut être primaire ou dérivée (Darmont (2006)). Dans notre travail, nous nous basons sur la FH des tables de l'ED. Son principe repose sur une liste de prédicats comme base de travail. Ces prédicats proviennent du jeu des requêtes posées par les utilisateurs (clause WHERE). En fonction de ces prédicats, des fragments de tables sont générés. Dans ce contexte, plusieurs solutions de fragmentation ont été proposées dans l'état de l'art. Le problème se pose lorsque le nombre de prédicats augmente. Son augmentation implique l'accroissement du nombre de partitions générées par la démarche de fragmentation utilisée. En conséquence, le processus de contrôle et de gestion des partitions devient de plus en plus complexe. Ceci, risque de diminuer l'efficacité des solutions proposées. 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 bénéficié de l'attention qu'il mérite en dépit de son importance.

3 Etat de l'art

Plusieurs solutions ont été proposées pour la fragmentation horizontale des ED relationnels. Dans ce qui suit, nous présentons un tour d'horizon sur quelques approches qui ont été développées dans ce contexte. Bellatreche et al. (2004) ont proposé une solution pour la fragmentation horizontale des tables multidimensionnelles de l'ED. Cette solution est une adaptation de l'approche basée sur les affinités du travail de Zhang et Orlowska (1995) développée initialement pour la fragmentation verticale des tables relationnelles. Bellatreche et al. (2011) ont ensuite, proposé un algorithme de sélection des meilleurs schémas de fragmentation en combinant un modèle de coût mathématique de Bellatreche (2008) avec l'algorithme du recuit simulé. Boukhalfa (2009) a proposé un ensemble d'approches permettant d'optimiser les ED. Ses approches d'optimisation reposent sur l'utilisation de trois techniques d'optimisation : la FH primaire, dérivée et les index de jointure binaires (IJB). Ziyati (2010) a formalisé le problème de sélection du meilleur schéma de fragmentation verticale comme un problème d'optimisation avec contrainte. Il a adapté l'algorithme génétique pour la fragmentation des ED. Sa solution commence par fragmenter le schéma relationnel d'un ED horizontalement, ensuite verticalement afin de réduire le coût d'exécution des requêtes. Barr (2010) a utilisé un algorithme basé sur les colonies de fourmis pour le partitionnement des tables de l'ED.
D'autre part, nous considérons la méthode de classification par l'algorithme k-means comme étant une méthode originale pour la FH. 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 présentons dans ce qui suit, deux travaux utilisant la classification comme technique de fragmentation. Mahboubi (2008) a proposé une

solution pour la FH des ED XML afin de les répartir sur une grille. Pour ce faire, il a exploité l'algorithme k-means. Sa solution englobe trois étapes : (1) codage des prédicats de sélection d'un ensemble de requêtes dans une matrice binaire qui représente le contexte de classification. Ensuite, (2) classification des prédicats par application de la technique des k-means qui permet de partitionner l'ensemble des prédicats en k classes disjointes et enfin, (3) construction des fragments. Le deuxième travail a été entrepris par Tekaya (2011) où elle a défini deux nouvelles notions : la corrélation sémantique et la corrélation géographique. La corrélation sémantique permet de fusionner par conjonction deux prédicats inclus dans la même clause WHERE. Une corrélation géographique permet de fusionner deux prédicats n'appartenant forcément pas à la même requête mais sont utilisés par la même localisation géographique. Une localisation géographique désigne le (ou les) site(s) sur lesquels nous envisageons allouer un magasin de données (MD). La solution proposée intègre l'aspect géographique dans le processus de fragmentation. elle englobe quatre phases : (1) détermination d'une liste de prédicats simples, (2) création de la matrice corrélation, (3) application de l'algorithme k-means pour la classification des prédicats et (4) génération des fragments horizontaux.
La diversité des solutions proposées pour la fragmentation horizontale des ED montre son importance. Cependant, l'efficacité de ces solutions risque de diminuer faisant face au problème d'augmentation des prédicats. A nos connaissances, ce problème n'a pas encore été considéré dans l'état de l'art.

4 Solution proposée

La solution que nous proposons pour réduire le nombre de prédicats admet comme entrée une liste de prédicats quelconques déduite à partir des requêtes proposées par les utilisateurs et produit comme résultat une liste réduite de prédicats.
Cette solution se compose de trois phases :

1. Phase de codification ;
2. Phase de classification ;
3. Phase de réduction du nombre de prédicats.

4.1 Phase de codification

Dans cette première phase, nous nous sommes inspirés du travail réalisé par Mahboubi (2008) pour la conversion des prédicats en codes binaires. Mahboubi (2008) a utilisé une matrice d'utilisation des prédicats (MUP) qui englobe les utilisations des prédicats par les requêtes. Les colonnes de la MUP englobent les prédicats les plus utilisés. Les lignes désignent les requêtes OLAP les plus fréquentes sur le système. Une cellule de la MUP englobe la valeur 1 si une requête donnée utilise un prédicat donné et la valeur 0, sinon.
Dans notre travail, nous adaptons cette matrice dans un contexte d'ED réparti. Les lignes englobent les prédicats les plus utilisés par les requêtes OLAP les plus fréquentes sur tous les sites de l'entreprise. Nous utilisons, tout d'abord, l'algorithme COMM_MIN de Ozsu et Valduriez (1999) pour réduire le nombre des prédicats. Les colonnes de la MUP englobent les sites géographiques de l'entreprise. Une cellule de la MUP englobe la valeur 1 si un prédicat donné est utilisé par un site donné et la valeur 0 sinon. La figure ci-dessous montre un exemple

de la MUP de notre solution.

$$\begin{pmatrix} & S1 & S2 & ... & Sm \\ \\ P1 & 1 & 1 & 0 & 0 \\ P2 & 1 & 0 & 0 & 1 \\ ..Pu & 0 & 1 & 1 & 0 \end{pmatrix}$$

4.2 Phase de classification

Comme entrée de cette phase (dans la partie pratique), nous utilisons la MUP. Elle englobe une description binaire de tous les prédicats utilisés dans leurs sites. Nous utilisons l'algorithme k-means de MacQueen (1967) pour la classification des prédicats des tables de dimensions contenus dans la MUP. C'est un algorithme de partitionnement de données relevant des statistiques et de l'apprentissage automatique. K-means est une méthode dont le but est de diviser des observations en K partitions dans lesquelles chaque observation appartient à la partition avec la moyenne la plus proche. Nous avons choisi d'utiliser cet algorithme car il est très populaire du fait qu'il est très facile à comprendre et à mettre en oeuvre de plus de sa simplicité conceptuelle et son application à tout type de données. La figure ci-dessous montre un exemple de classification. Chaque classe correspond à un site qui contient les besoins des utilisateurs.

$$\begin{pmatrix} C1 & C2 & ... & Cn \\ \\ P1 & P3 & P4 & P5 \\ P2 & S2 & S3 & P6 \\ S1 & & & S4 \end{pmatrix}$$

4.3 Phase de réduction du nombre de prédicats

Après la phase de classification, nous appliquons l'algorithme de réduction du nombre des prédicats. La variable P d'entrée (ligne 1) reçoit une liste de prédicats, la variable F (ligne 4) initialisée par l'ensemble des fragments des prédicats P et au fur et à mesure reçoit ou élimine les fragments des prédicats ajoutés ou supprimés et la variable P' de sortie (ligne 2) reçoit la liste finale des prédicats optimisée.
En premier lieu, nous éliminons les prédicats appartenant à tous les sites : nous utilisons une boucle for (ligne 6) qui nous permet de parcourir tous les prédicats donnés et l'instruction if (ligne 7) vérifie si le prédicat appartient à tous les sites ou non. S'il appartient à tous les sites, on l'élimine, sinon il reste dans la liste. Cela nous permet de minimiser le nombre de prédicats sans toucher les données et en conséquence, nous minimisons le nombre de fragments générés. L'algorithme 1 présenté ci-après détaille cette phase. Nous avons créé cet algorithme de complexité O(n) qui nous permet d'éliminer les prédicats appartenant à tous les sites donnés. L'accès à ces données peut se faire de n'importe quel site et ces prédicats n'entrent pas dans le processus de fragmentation de l'ED. Le fragment répond au besoin des utilisateurs du site donné et il reste spécifique à ces utilisateurs.

Algorithm 1 Algorithme d'optimisation du nombre de prédicats

Require: P : Liste des prédicats

Ensure: P' : Liste de prédicats plus minimale

$P' \leftarrow P$

$F \leftarrow$ chaque prédicat (P_i) correspond à 1 fragment (F_{p_i})

$nb = entier;\ i = entier;\ j = entier;$

for $(i = 0, i \leq nb, i++)$ **do**

  if $(P_i \in P)$ s'illustre dans tous les sites crées **then**

    $P \leftarrow (P - P_i)$

    $F \leftarrow (F - F_{p_i})$

    $P' \leftarrow (P - P_i)$

  end if

end for

end

FIG. 1 – *Algorithme de réduction du nombre des prédicats*

En second lieu, nous avons utilisé les notions de corrélation sémantique et géographique proposés par Tekaya (2011) pour la réduction du nombre de prédicats généré par l'algorithme d'optimisation du nombre des prédicats. Nous proposons de regrouper par conjonction les prédicats qui ont une interdépendance géographique dans un même prédicat. Ceci permet de regrouper les prédicats qui sont utilisés par les mêmes sites. De même, nous proposons de regrouper par conjonction les prédicats ayant une interdépendance sémantique, c'est à dire utilisés dans la même clause WHERE. Prenons par exemple les deux requêtes suivantes qui sont utilisées par une même localisation géographique sur une table de fait Vente ayant le schéma relationnel suivant : Vente(id_produit,id_client,id_temps,id_ville,Montant) :

(R1)
SELECT SUM(Montant)
FROM Vente
WHERE id_produit=100
AND id_temps=mai

(R2)
SELECT SUM(Montant)
FROM Vente
WHERE id_produit=200
AND id_temps=juin

Les prédicats "id_produit = 100" et "id_temps = mai" ont une corrélation sémantique parce qu'ils sont utilisés par la même requête R1. De même, les prédicats "id_produit = 200" et

"id_temps = juin" ont une corrélation sémantique parce qu'ils sont utilisés par la même requête R2.
D'autre part, les prédicats "id_produit = 100" et "id_produit = 200" ont une corrélation géographique parce qu'ils sont utilisés par deux requêtes appartenant à une même localisation géographique. De même, les prédicats "id_temps = mai" et "id_temps = juin" ont une corrélation géographique par ce qu'ils sont exécutés par deux requêtes appartenant à une même localisation géographique.
La solution proposée consiste à fusionner par conjonction les prédicats qui ont une corrélation sémantique et/ou géographique ce qui réduit le nombre de prédicats utilisés par le processus de fragmentation. Dans la section qui suit, nous présentons un exemple détaillé de notre solution, ainsi que son application sur un ED réel issu du banc d'essai APB1.

5 Validation expérimentale

Cette section englobe une validation expérimentale de notre solution. D'abord, nous présentons l'environnement expérimental du travail, ensuite, nous proposons un exemple d'application détaillé des différentes étapes de l'approche proposée et enfin, nous exposons les résultats expérimentaux obtenus.

5.1 Présentation de l'environnement expérimental

Pour valider notre travail, nous avons utilisé un ED réel issu du benchmark APB1. Sur cet ED, nous exécutons un ensemble de 19 requêtes réparties sur deux sites géographiquement distants. L'ED englobe 5 tables : une table de faits *Actvars* et quatre tables de dimensions, *Prodlevel*, *Custlevel*, *Timelevel* et *Chanlevel* (Spofford (1998)). L'APB1 benchmark, a été très populaire à la fin des années 1990. Il est toujours le plus utilisé par les travaux abordant le problème de la fragmentation des ED que nous considérons similaires à notre contexte de travail. Le choix du banc d'essai APB1 nous permettra de comparer nos résultats à ces travaux aussi bien au niveau technique que pratique de la solution. Le (tableau 1) résume les caractéristiques de chaque table.

Table	Nombre d'enregistrements	Taille d'un enregistrement
Actvars	24786000	74
Chanlevel	9	24
Custlevel	900	24
Prodlevel	9000	72
Timelevel	24	36

TAB. 1 – *Caractéristiques des tables de l'entrepôt de données*

Pour la répartition de l'ED, nous utilisons deux machines géographiquement distantes connectées par un réseau privé virtuel (VPN) (PHAM (2002)). Pour la génération des tables du banc d'essai APB1, nous avons installé Oracle 11g sur la machine 1 et Oracle 10g sur la machine 2.

5.2 Exemple d'application sur le benchmark APB1

A partir de l'ensemble des requêtes, nous avons collecté 18 prédicats de sélection sur les tables de dimension. Sur ces prédicats, nous allons appliquer les 3 phases de notre solution. Ci-dessous, la liste des prédicats choisie :

P_1 : id_mois = 'Novembre' P_2 : id_mois = 'Décembre'
P_3 : id_mois = 'Janvier' P_4 : id_ville = 'Tunis'
P_5 : id_ville = 'Sfax' P_6 : id_ville = 'Bizerte'
P_7 : id_client = '1' P_8 : id_client = '2'
P_9 : id_client = '3' P_{10} : id_produit = 'Chaussure'
P_{11} : id_mois = 'Février' P_{12} : id_mois = 'Mars'
P_{13} : id_mois = 'Avril' P_{14} : id_mois = 'Mai'
P_{15} : id_mois = 'Juin' P_{16} : id_mois = 'Juillet'
P_{17} : id_mois = 'Août' P_{18} : id_produit = 'Veste'

⋆ **Phase 1 : Codification des prédicats :** Dans cette phase, nous commençons par la conversion en code binaire des différentes utilisations des prédicats. Nous proposons ici comme exemple pour 3 sites, la MUP suivante :

$$\begin{pmatrix} & S1 & S2 & S3 \\ P1 & 1 & 0 & 0 \\ P2 & 1 & 0 & 0 \\ P3 & 1 & 0 & 0 \\ P4 & 1 & 0 & 0 \\ P5 & 1 & 0 & 0 \\ P6 & 1 & 0 & 0 \\ P7 & 1 & 0 & 0 \\ P8 & 1 & 0 & 0 \\ P9 & 1 & 0 & 0 \\ P10 & 1 & 1 & 1 \\ P11 & 0 & 1 & 0 \\ P12 & 0 & 1 & 0 \\ P13 & 0 & 1 & 0 \\ P14 & 0 & 0 & 1 \\ P15 & 0 & 0 & 1 \\ P16 & 0 & 0 & 1 \\ P17 & 1 & 1 & 1 \\ P18 & 1 & 1 & 1 \end{pmatrix}$$

⋆ **Phase 2 : Classification des prédicats :** Dans cette phase, nous exécutons l'algorithme des k-means via l'infrastructure expérimentale TANAGRA (RAKOTOMALALA (2004)) sur la MUP que nous avons proposé comme exemple dans la phase 1. Le résultat obtenu englobe la liste des classes suivante :

$$\begin{pmatrix} C1 & C2 & C3 \\ \\ P1 & P11 & P14 \\ P2 & P12 & P15 \\ P3 & P13 & P16 \\ P4 & P10 & P10 \\ P5 & P17 & P17 \\ P6 & P18 & P18 \\ P7 & S2 & S3 \\ P8 & & \\ P9 & & \\ P10 & & \\ P17 & & \\ P18 & & \\ S1 & & \end{pmatrix}$$

Les classes générées sont :
$C_1 : P_1 \bigwedge P_2 \bigwedge P_3 \bigwedge P_4 \bigwedge P_5 \bigwedge P_6 \bigwedge P_7 \bigwedge P_8 \bigwedge P_9 \bigwedge P_{10} \bigwedge P_{17} \bigwedge P_{18} \bigwedge S_1$
$C_2 : P_{11} \bigwedge P_{12} \bigwedge P_{13} \bigwedge P_{10} \bigwedge P_{17} \bigwedge P_{18} \bigwedge S_2$
$C_3 : P_{14} \bigwedge P_{15} \bigwedge P_{16} \bigwedge P_{10} \bigwedge P_{17} \bigwedge P_{18} \bigwedge S_3$
⋆ **Phase3 : Réduction du nombre de prédicats :** Dans cette étape, nous appliquons tout d'abord notre algorithme de réduction du nombre de prédicats, ensuite, nous procédons à la deuxième étape de la solution par la fusion des prédicats selon les interdépendances sémantiques et/ou géographiques. La liste des prédicats devient :
P'_1 : id_mois in ('Novembre','Décembre','Janvier')
P'_2 : id_ville in ('Tunis','Sfax','Bizerte')
P'_3 : id_client in ('1','2','3')
P'_4 : id_mois in ('Février','Mars','Avril')
P'_5 : id_mois in ('Mai','Juin','Juillet')
En effet, les prédicats P_{10}, P_{17} et P_{18} ont été éliminés par l'algorithme d'optimisation du nombre de prédicats car ils appartiennent à tous les sites donnés.
De plus, le prédicat P'_1 a remplacé les prédicats P_1, P_2 et P_3, le prédicat P'_2 a remplacé les prédicats P_4, P_5 et P_6, le prédicat P'_3 a remplacé les prédicats P_7, P_8 et P_9, le prédicat P'_4 a remplacé les prédicats P_{11}, P_{12} et P_{13} et le prédicat P'_5 a remplacé les prédicats P_{14}, P_{15} et P_{16} par l'utilisation des interdépendances géographiques. Les classes générées après réduction des prédicats sont :
$C_1 : P'_1 \bigwedge P'_2 \bigwedge P'_3 \bigwedge S_1$
$C_2 : P'_4 \bigwedge S_2$
$C_3 : P'_5 \bigwedge S_3$
Nous avons donc diminué à la fin le nombre de prédicats de 18 à 5 pour avoir 3 fragments. Pour la génération du schéma de partitionnement de l'ED, nous procédons par la conjonction des prédicats inclus par classe générée. Le résultat obtenu par classe constitue un des critères de partitionnement de la table de faits pour un site donné.

5.3 Présentation des résultats obtenus

Dans notre exemple, la dernière phase de classification a engendré 3 classes de prédicats. Nous avons utilisé la conjonction de ces prédicats par classe pour le partitionnement de la table de faits sur trois fragments selon la solution de Tekaya (2011). 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 ED centralisé, ensuite, partitionné sans réduction des prédicats, et enfin réparti en appliquant notre solution de réduction des prédicats. Les résultats obtenus sont récapitulés dans le tableau 2. Pour les requêtes numéro 1, 2, 5 jusqu'à 19, les temps d'exécution ont

Numéro de requête	Temps d'exécution centralisé	Temps d'exécution partitionné	Temps d'exécution réparti
1	1 :58 :34	0 :48 :12	0 :22 :10
2	1 :58 :01	0 :54 :54	0 :31 :43
3	5 :31 :88	2 :13 :18	5 :04 :16
4	5 :31 :08	4 :31 :24	8 :34 :45
5	7 :10 :25	3 :42 :85	1 :58 :70
6	9 :14 :33	7 :20 :11	4 :53 :62
7	10 :13 :14	9 :49 :62	7 :18 :78
8	14 :05 :13	13 :03 :91	10 :22 :83
9	15 :01 :03	10 :48 :65	8 :59 :17
10	16 :14 :78	13 :22 :14	10 :33 :45
11	7 :19 :11	5 :32 :92	4 :44 :89
12	7 :28 :17	4 :55 :69	3 :52 :51
13	8 :17 :21	6 :24 :28	5 :36 :77
14	9 :45 :02	8 :57 :95	7 :51 :09
15	6 :17 :20	5 :39 :78	4 :34 :30
16	7 :45 :34	5 :52 :74	4 :01 :09
17	9 :11 :97	8 :26 :85	7 :11 :51
18	8 :48 :68	7 :10 :35	5 :46 :03
19	10 :53 :93	8 :14 :47	7 :27 :42

TAB. 2 – *Temps d'exécution des requêtes utilisées en (mn :s :ms).*

diminué (figure 2), ce qui constitue pour nous un gain non négligeable. Le temps d'exécution global des requêtes dans un contexte réparti avec réduction des prédicats a diminué de 30% par rapport au contexte centralisé et de 10% par rapport au contexte partitionné sans réduction du nombre des prédicats(figure 3). Par contre, pour les requêtes 3 et 4, les temps d'exécution ont augmenté(figure 2). Ce sont les requêtes les moins bénéficiaires du partitionnement sont celles n'exploitant pas les partitions des tables et qui n'ont pas été considérées dans le processus de partitionnement. Ceci explique clairement l'augmentation remarquable des mesures obtenues. Dans ce contexte, quelques solutions sont envisageables pour la réécriture des requêtes OLAP distantes pour l'optimisation du temps d'exécution des requêtes OLAP distantes notamment

les travaux de Liang et Yu (2000) et/ou de Kalnis et Papadias (2001) où ils rajoutent une heuristique d'optimisation des requêtes. Vu les avancés dans les réseaux aussi, le temps de réponse est considérablement réduit, ceci n'a pas été mesuré dans notre solution.

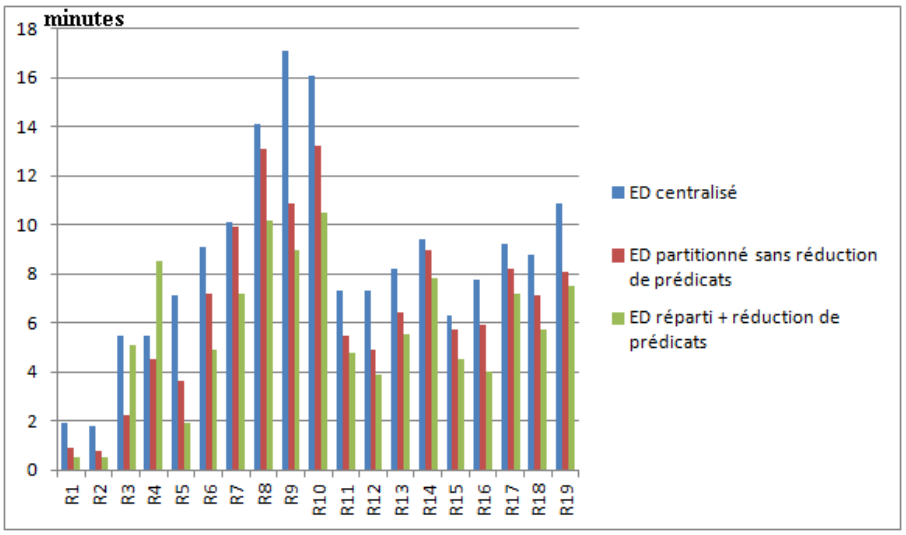


FIG. 2 – *Temps d'exécution des requêtes.*

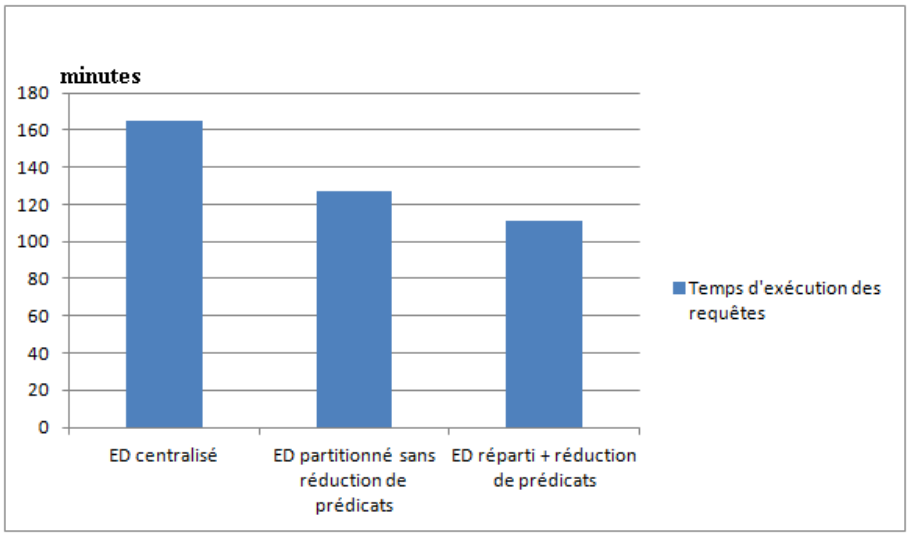


FIG. 3 – *Comparaison des temps d'exécution globaux.*

6 Conclusion

Dans cet article, nous nous sommes intéressés à l'optimisation des requêtes décisionnelles exécutées sur un ED modélisé en étoile. Nous avons proposé une solution pour la réduction du nombre de prédicats pour les approches de répartition des ED. Cette solution repose sur trois phases : phase de codification, phase de classification et phase de réduction des prédicats. Pour l'évaluation de notre solution, nous l'avons appliquée sur un ED réel issu du banc d'essai APB1 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 problème de l'allocation des données dans un contexte d'ED doit tenir compte en plus des contraintes classiques de répartition des BD notamment la contrainte d'accès à un MD à partir de tous les sites, le stockage, le délai de réponse et les fréquences d'utilisation d'un MD sur les différents sites de l'entreprise, la contrainte de chargement d'un MD dans tous les sites que nous n'avons pas considéré dans cette contribution. Ces contraintes vont être prises en considération dans nos prochains travaux de recherche.

• Dans la plupart des cas, la répartition d'un ED 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. Nous allons automatiser cette mise à jour dans nos prochains travaux.

Références

Barr, M. (2010). *Approche dirigée par les fourmis pour la fragmentation horizontale des entrepôts de données relationnels*. Mémoire de mastère, Ecole nationale Supérieure d'Informatique, Algérie).

Bellatreche, L. (2008). Bitmap join indexes and data partitioning. *Encyclopedia of Data Warehousing and Mining 2nd Edition*, 5,37,38.

Bellatreche, L., K. Boukhalfa, et P. Richard (2011). Primary and referential horizontal partitioning selection problems. *Concepts, Algorithms and Advisor Tool*.

Bellatreche, L., K. Karlapalem, et Q. Li (2004). Derived horizontal class partitioning in oodb: Design strategies, analytical model and evaluation. *Conceptual Modeling ER'98*, 465–479.

Boukhalfa, K. (2009). *De la conception physique aux outils d'administration et de tuning des entrepôts de données*. Thèse de doctorat, Université de Poitiers.

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.

Kalnis, P. et D. Papadias (2001). Optimization algorithms for simultaneous multidimensional queries in olap environments. *Data Warehousing and Knowledge Discovery*, 264–273.

Liang, W. Orlowska, M. et J. Yu (2000). Optimizing multiple dimensional queries simultaneously in multidimensional databases. *The VLDB Journal The International Journal on Very Large Data Bases 8*, 319–338.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *5th Berkeley Symposium on Mathematical Statistics and Probability*, 281–295.

Mahboubi, H. (2008). *Optimisation de la performance des entrepôts de données XML par fragmentation et répartition*. Thèse de doctorat, Université Lumière Lyon 2).

Ozsu, T. et P. Valduriez (1999). Principes of distributed database systems. *Prentice Hall*, 19–22.

PHAM, C. (2002). Vpn et solutions pour l'entreprise. *SaaS, Université de Pau et des Pays de l'Adour*.

RAKOTOMALALA, R. (2004). Tanagra : une plate-forme d'expérimentation pour la fouille de données. *Open Access Journal, Université Lumière Lyon 2*.

Spofford, G. (1998). Olap conseil apb-1 benchmark. *Guide d'installation*.

Tekaya, K. (2011). *Fragmentation et allocation dynamiques des entrepôts de données*. Thèse de doctorat, Faculté des sciences de Tunis.

Zhang, Y. et M. Orlowska (1995). Fragmentation approaches for distributed database design. *Information Sciences-Applications*, 117–132.

Ziyati, E. (2010). *Optimisation de requêtes OLAP en Entrepôts de Données : Approche basée sur la fragmentation génétique*. Thèse de doctorat, Faculté des Sciences Rabat.

Summary

In the domain of data warehousing, most approaches of distribution are essentially based on the techniques of fragmentation and allocation tables. These approaches exploit in input extracts predicates of OLAP queries most used in the partitioning process. Since continues increase of the number of predicates, and her negative impact, it becomes more and more interesting to reduce this increase before the fragmentation process. In this paper, we propose a solution based on a classification algorithm to reduce the number of predicates in the data warehouses allocation approaches. The proposed solution encompasses three phases: Coding predicates as binary matrices, a classification phase of these predicates by the k-means algorithm and a final phase to reduce the number of predicates. We validated our solution on a real data warehouse from the benchmark APB1.

Analyzing the behavior and text posted by users to extract knowledge

Soumaya Cherichi*, Rim Faiz*

* LARODEC, IHEC Carthage
University of Carthage
Carthage Presidency, Tunisia
soumayacherichi@gmail.com
LARODEC, IHEC Carthage
University of Carthage
Carthage Presidency, Tunisia
Rim.Faiz@ihec.rnu.tn

Abstract. With the explosion of Web 2.0 platforms such as blogs, discussion forums, and social networks, Internet users can express their feelings and share information among themselves. This behavior leads to an accumulation of an enormous amount of information. Among these platforms are so-called microblogs. Microblogging (e.g. Twitter1), as a new form of online communication in which users talk about their daily lives, publish opinions or share information by short posts, has become one of the most popular social networking services today, which makes it potentially a large information base attracting increasing attention of researchers in the field of knowledge discovery and data mining. Several works have proposed tools for tweets search, but, this area is still not well exploited. Our work consists of examining the role and impact of social networks, in particular microblogs, on public opinion. We aim to analyze the behavior and text posted by users to extract knowledge that reflect the interests and opinions of a population. This gave us the idea to offer new tool more developed that uses new features such as audience and RetweetRank for ranking relevant tweets. We investigate the impact of these criteria on the search's results for relevant information. Finally, we propose a new metric to improve the results of the searches in microblogs. More accurately, we propose a research model that combines content relevance, tweet relevance and author relevance. Each type of relevance is characterized by a set of criteria such as audience to assess the relevance of the author, OOV (Out Of Vobulary) to measure the relevance of content and others. To evaluate our model, we built a knowledge management system. We used a collection of subjective tweets talking about Tunisian actualities in 2012.

1 Introduction

In the current era, People are becoming more communicative through expansion of services and multi-platform applications, i.e., the so called Web 2.0 which establishes social and collaborative backgrounds. They commonly use various means including Blogs to share the diaries, RSS feeds to follow the latest information of their interest and Computer Mediated Chat (CMC) applications to hold bidirectional communications. Microblogging is one of the

most recent products of CMC, in which users talk about their daily lives, publish opinions or share information by short posts. It was first known as Tumblelogs on April 12, 2005, and then came into greater use by the year 2006 and 2007, when such services as Tumblr and Twitter arose. The problem that we face is how to find this information and transform data collections into new knowledge, understandable, useful and interesting in the context where it is located. Information retrieval systems solve one of the biggest problems of knowledge management (KM): quickly finding useful information within massive data stores and ranking the results by relevance.

Recent years have revealed the accession of interactive media, which gave birth to a huge volume of data in blogs and micro-blogs more precisely. These micro-blogs attract more and more users due to the ease and the speed of information sharing especially in real time.

Twitter has played a role in important events, but the service also allows people to communicate among a relatively small social circle, and a sizeable part of Twitter's success is because of this function.

Indeed a micro-blog is a stream of text that is written by an author. It is composed by regular and short updates that are presented to readers in reverse chronological order called time-line.

Today, the service called Twitter is the most popular micro-blogging platform. While micro-blogging services are becoming more famous, the methods for organizing and providing access to data are also improving. Micro-bloggers as well as sending tweets are looking for the last updates according to their interests. Finding the most relevant tweets to a topic depends on the criteria of micro-blogs.

Unlike other micro-blogging service, Twitter is positioned by the social relationship of subscription. And since the association is led, it allows users to express their interest in the items of another micro-bloggers. The social network of Twitter is not limited to bloggers and subscription relationships; it also includes all the contributors and data that interact in both contexts of use and publication of articles. We have analyzed the micro-blogging service Twitter and we have identified the main criteria of Twitter.

But the question arises what is the impact of each feature on the quality of results?

Our work consists in searching a new metric of features' impact on the search results' quality. Several criteria have been proposed in the literature Ben Jabeur et al. (2011) and Cha et al. (2010), but there are still other criteria that have not been exploited as audience which could be the size of the potential audience for a message: What is the maximum number of people who could have been exposed to a message?

We gathered the features on three groups: those related to content, those related to tweet and those related to the author. We used the coefficient of correlation with human judgment to define our score. For processing the content of tweets, we intend to use resources and linguistic methods

Our experimental result uses a corpus of thousand subjective tweets which are neither answers nor retweets, and we also collected a corpus of human judgments to find the correlation coefficient.

The remainder of this paper is organized as follows. In section 2, we describe the task Twitter Information Retrieval. In section 3 we present all the features that we have used to calculate our score. In Section 4, we discuss experiments and obtained results. Finally, section 5 concludes this paper and outlines future work.

2 Related works

Recent years have witnessed the advent of interactive media and especially Web 2.0. This led to a huge volume of data from blogs, discussion forums and commercial sites. Indeed, blogs, being the figurehead of Web 2.0, are characterized by their evaluative usage, in the sense that users are using them to express themselves freely and share their opinions on their interests.

A micro-blogging service is at once a communication mean and a collaboration system that allows sharing and disseminating text messages. In comparison with other social networks on the Web (for example Facebook, Myspace, LinkedIn, Foursquare), the microblogs articles are particularly short and submitted in real time to report a recent event. At the time of this writing, several micro-blogging services exist. In this paper, we will focus on the micro-blogging service Twitter which is the most popular and widely used. Twitter is characterized from similar sites by certain features and functionalities. An important characteristic is the presence of social relationships subscription. This directional relationship allows users to express their interest on the publications of a particular blogger. Twitter is distinguished from similar websites by some key features. The main one consists on the following social relationship. This directed association enables users to express their interest in other microbloggers' posts, called tweets, which doesn't exceed 140 characters. Moreover, Twitter is marked by the retweet feature which gives users the ability to forward an interesting tweet to their followers. A blogger, also called twitterer, can annotate his tweets using # hashtags or send it to a specific user through the user @ mentions. Finally, a tweet can also share a Web resource referenced by an URL.

Several works have focused on the analysis of data posted on microblogs, particularly in Twitter. Barbosa (2010), Go (2009) and Jiang (2011) propose approaches for sentiment classification of Twitter messages i.e. determine whether tweets express a positive, negative or neutral feeling. Positive and negative polarities correspond respectively to a favorable and unfavorable opinion. To solve this task the authors have used natural language processing and machine learning techniques.

Many studies have found that there is a high correlation between the information posted on the web and actual results. Doan et al (2011) have used tweets to analyze awareness and anxiety levels of Tokyo habitants the events of earthquakes tsunami and states of nuclear emergencies in Japan in 2011. Lampos (2010) have presented a method to measure the prevalence of H1N1 disease in the population of United Kingdom. They sought in the tweets the symptoms related to the disease. The obtained results were compared with real results from the Health Protection Agency. O connor (2010) also analyzed the tweets to predict public opinion and then compared the results with surveys.

Given the specificity of micro-blogs, looking for tweets is facing several challenges such as indexing the flow of items Sankaranarayanan et al. (2009), spam detections, diversification of results and evaluating the quality of tweets Nagmoti et al. (2010). We find that most approaches for information retrieval in micro-blogs don't take into account all the features to narrow the search. In fact, each feature has a unique impact on the other ones. Based on this observation and to improve the results of research, we will try to overcome these limitations by measuring the impact of these criteria. We will propose a measurement metric impact criteria for improving outcomes research. The search for tweets is an information retrieval task ad-hoc whose objective is to select the items relevant micro-blogs in response to a query. The definition of relevance in the search for tweets is not limited to textual similarity but

also takes account of social interactions in the network. In this context, the relevance of the items depends also on the tweets' technical specificities and the importance of the author.

Regarding the relevance of content, several studies have used Okapi BM25 algorithm Robertson (1998), other studies like work of Duan et al., (2010), have added new features such as tweets' quality ie the tweet that contains the least amount of Out of vocabulary (OOV) is considered as the most informative one. Also Duan et al, consider that the longer the tweet, the better amount of information it contains.

Our work consists of examining the role and impact of social networks, in particular microblogs, on public opinion. We aim to analyze the behavior and text posted by users to extract knowledge that reflect the interests and opinions of a population. We introduce in this paper our approach for tweet search that integrates different criteria namely the social authority of micro-bloggers, the content relevance, the tweeting features as well as the hashtag's presence. We present in the next section the main features of our criteria.

3. Analysis of short texts using NLP

Data analysis of social networks has become a major trend in the field of natural language processing. Thus, large communities NLP gave its fair share to the analysis of data microblogs. In recent years, major conferences have created workshops for data analysis in social networks. Several studies concerned with the analysis of short texts do not aim only to determine the polarity of the messages, but to use the messages to detect events or predict results.

Among the most important tasks for a ranking system tweet is the selection of features set. We offer three types of features to rank tweets:

- Content features refer to those features which describe the content relevance between queries and tweets.
- Tweet features refer to those features which represent the particular characteristics of tweets, as OOV and hashtags in tweet.
- Author features refer to those features which represent the authority of authors of the tweets in Twitter.

3.1. Content Relevance Features

The criterion "Content" refers to the thematic relevance traditionally calculated by IR systems standards. The thematic relevance is generally measured by one of several IR models. One of the models reference. Information Retrieval IR is the probabilistic model Jones et al. (2000) with the weighting scheme BM25 as matching request document function. For this reason, we have adopted this model for the calculation of the thematic relevance. Of course, it is made possible to calculate using any other IR model. BM25 is a search function based "bag of words", it allows us to organize all documents based on the occurrences of the query terms given in the documents. (cf section 2).

We used four content relevance features:

1. Relevance(T,Q): we used OKAPI BM25 score which measures the content relevance between the query Q and tweet T.

$$TF - IDF_(w,Ti) = TF_(w,Ti).IDF_(w,Ti)$$
$$= TF_{w,Ti}((\log_2 * \frac{N}{DF_w}) + 1)$$

Knowing that: w is a term in the query Q and Ti is the tweet i.

2. Popularity(Ti,Tj,Q): with i and j in n and i≠j : it used to calculate the popularity of a tweet from the corpus. It measures the similarity between the tweets in the context of the tweet's topic. We used cosine similarity, according to a study done by Sarwar et al. (2001) cosine similarity is the most efficient similarity measure ,in addition, it is not sensitive to the size of each tweet:

$$Cosine(Ti,Tj) = \frac{\sum_{w \in (Ti \bigcap Tj)} TFIDF_{w,Ti} * TFIDF_{w,Tj}}{\sqrt{\sum_{w \in Ti} (TFIDF_{w,Ti})^2 * \sum_{w \in Tj} (TFIDF_{w,Tj})^2}}$$

Knowing that w is a term in the query Q, Ti is tweet i, Tj is tweet j, i and j in n and i≠j.

3. **Length of tweet (Lg(Ti,Q)):** Length is measured by the number of characters that a tweet contains. It is said that more the tweet is long, more it contains information.

$$Lg(Ti,Q) = \frac{Lg(Ti) - MinLg(T)}{MaxLg(T)}$$

4. **Out of Vocabulary (OOV(Ti)):** This feature is used to roughly approximate the language quality of tweets. Words out of vocabulary in Twitter include spelling errors and named entities. This feature aims to measure the quality language of tweet as follows:

$$Quality(T) = 1 - \frac{NumberofOOV(Ti)}{Lg(Ti)}$$

With Number of OOV (Ti) is calculated as follows

```
String tweet[] = tweet.split(" ");
int count = 0;
for (int i = 1; i < tweet.length; i++)
if (checker.isNotCorrect(tweet[i]))
{
Number of oov ++;
}
```

The more number of out of vocabulary is small the more quality of tweet is better.

3.2. Tweet Relevance Features

We note that the thematic relevance depends solely on the item and query. Each tweet has many technical features, and each feature form selection criteria that we have exploited.

1. Retweet (Ti,Q): is defined as the number of times a tweet is retweeted. In a rational manner, the most retweeted tweets are most relevant. Retweets are forwarding of corresponding original tweets, sometimes with comments of retweeters. According to Duan et al. (2010), they are supposed to contain no more information than the original tweets.

$$Retweet(Ti,Q) = \frac{Retweet(Ti) - MinRetweet(T)}{MaxRetweet(T)}$$

2. Reply(Ti): An @reply is any update posted by clicking the "Reply" button on a Tweet, it will always begin with @username. This feature aims to calculate the num-

ber of reply to a tweet. Ultimately tweets that have received the most response are more relevant.

$$Reply(Ti, Q) = \frac{Reply(Ti) - MinReply(T)}{MaxReply(T)}$$

3. Favor(Ti): this feature aims to calculate the number of times a tweet is classified as a favorite. If a message is considered by many followers as a favorite, it means that it is relevant.

$$Favor(Ti, Q) = \frac{Favor(Ti) - MinFavor(T)}{MaxFavor(T)}$$

4. Hashtag Count(Ti):The # symbol, called a hashtag, is used to mark keywords or topics in a Tweet. It was created organically by Twitter users as a way to categorize messages. This feature aims to calculate the number of hashtags in tweet.

$$HashtagCount(Ti) = \sum \text{ of occurrences of hashtag}$$

5. Url count(Ti):Twitter allows users to include URL as a supplement in their tweets. This feature aims to estimates the number of times that the URL appears in the tweet corpus. According to [DAM 12], tweets containing URLs are more informative

$$URLCount(Ti) = \sum \text{ of occurrences of URL}$$

3.3. Author Relevance Features

Each blogger has specific characteristics such as number of follower and number of mention We said that users who have more followers and have been mentioned in more tweets, listed in more lists and retweeted by more important users are thought to be more authoritative.

1. Tweet Count(a):this feature represents the number of tweet posted by the author
2. Mention Count (author): A mention is any Twitter update that contains "@username" anywhere in the body of the Tweet , this means that @replies are also considered mentions. This feature aims to calculate the number of times an author is mentioned.
3. Follower(a):this feature represents the number of follower to the author
4. Following(a): this feature represents the number of subscriptions of the author (a) to other authors
5. Expertise(a): this feature was found by conducting a survey that asks people to rate the expertise of the blogger from 0 to 10.
6. RetweetRank (a): Retweet Rank looks up all recent retweets, number of followers, friends and lists of a user. It then compares these numbers with those of other users' and assigns a rank. Retweet Rank tracks both RTs posted using the Retweet button and other RTs (ReTweets) (e.g. RT @username).This feature is an indicator of how a blogger is influential on twitter.
7. TwitterPageRank(a): this feature represents the rank of author of the total twitter users using PageRank Algorithm

8. Audience (a): is the size of the potential audience for a message. What is the maximum number of people who could have been exposed to a message?

4. Metric Measure of the impact of criteria to improve search results

We introduce a research model that combines tweets relevant content, the specificities of tweets and the authority of bloggers. This model considers the specificities of tweets and the authority of bloggers as important factors which contribute to the relevance of the results.

The search for tweets is a task of information retrieval whose goal is to select the relevant sections in response to a user's request. To present an accurate list of articles, our model combines a score of content's relevance, a score of author's authority and a score of tweets' specificities. The objective of this combination is to provide a list of tweets that cover the subject of the request and are posted by major bloggers. After normalizing the feature scores, these three scores are combined linearly using the following formula:

$$\begin{aligned} Score(Ti,Q) &= scoreContent(Ti,Q) \\ &+\beta * scoreTweet(Ti,Q) \\ &+\gamma * scoreAuthor(Ti,Q) \end{aligned}$$

With score (Ti,Q) on [0, 2] and $\beta+\gamma=1$.

Where Ti and Q represent respectively, tweet and request. β and γ on [0,1] are a weighting parameter kermi and Faiz (2012). Scorecontent (Ti, Q) is the normalized score of the relevance of content. Scoretweet is the normalized score of the specificity of the tweet Ti and ScoreAuthor (a, Ti) is the normalized score of the importance of the author a corresponds to the blogger who published the tweet Ti.

We note that:

1. Scorecontent(Ti,Q)=Relevance(T,Q) + Lg(Ti) + Popularity(Ti,Tj,Q) + Quality(Ti);
2. ScoreTweet(Ti,Q)= Url count(Ti)+ Hashtag Count(Ti) + Retweet(Ti) + Reply(Ti) + Favor(Ti);
3. ScoreAuthor(a,Q)= TwitterPageRank(a) +Audience(a)+ Tweet Count(a) + Mention Count(a) + Expertise(a) + RetweetRank(a) + Follower(a) + Following(a).

5. Experimental Evaluation

We conducted a series of preliminary experiments on a collection of articles from Twitter, in order to evaluate the performance of our model.

5.1. Description of the collection

With the absence of a standard framework for evaluating information retrieval in microblogs, we collected a set of articles and queries. Our concern is that the database size is small. We describe in the following collection of articles and the approach for collecting relevance judgments.

5.1.1. Search Engine TWEETRIM

We built a search engine that we have called ``TWEETRIM'', which allows to calculate all scores and display the most relevant tweets according to these score. It has as input a query composed of three keywords and as output a set of relevant tweets relative to the query.

5.1.2. Tweet Set

We built a collection of articles, metadata about relationships subscription and reply. This corpus is collected manually ie a thousand blogs and thousands of tweets have been browsed. This collection contained a total of 3000 tweets published by 50 active Tunisian bloggers who are interested on the Tunisian news, we chose the period of March 4, 2012 until June 4, 2012.

5.1.3. Queries and relevance judgments

To perform queries and to collect the human judgement of relevance followed the following steps:

1. We collected 1000 queries on recent actualities in Tunisia from users,
2. then, we used the system that we have built which allows us to view the 10 results are especially relevant according to the score of the content,
3. and then, we asked 450 users to judge the 10 first results of each query.

We suppose that the content relevance already exists and we will improve our search result by varying our two other scores ScoreTweet and ScoreAuthor. We calculate the correlation coefficient between our scores and the corpus, which allowed us to find our weighting coefficients β and γ.

5.2. Results

5.2.1. Estimation of weights

We make a comparison within the values the values of correlation coefficients and from these results, we observe that the best correlation coefficient between βScoreTweet+γScoreAuthor with human judgment score = 0,161456763 when β = 0,4 and thus γ= 0,6.

5.2.2. Evaluation of our model

We compare, in Figure 2, the values of correlation's coefficients obtained by Tweet Features and Author Features with the parameters β, γ values respectively (1.0) and (0.1) obtained by experiments and the third configuration with β=0.4 and γ=0.6.

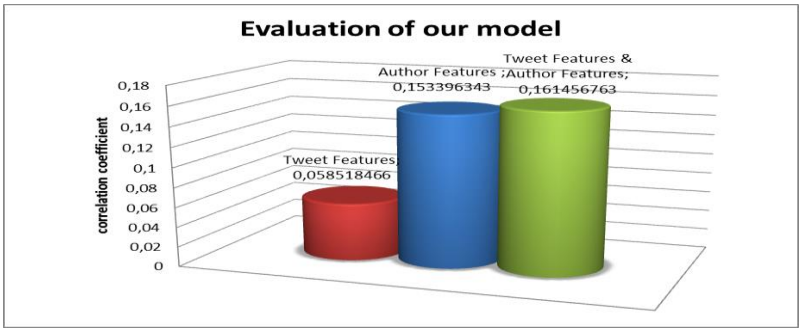


FIG.9 - *Comparing correlation coefficients*

We notice that the performance of the last 2 configurations are very close with a slight advantage for the combination ``Tweet Features \& Author Features" on the model based only on the specificities of the tweet and the importance of the author. We conclude that Author features have more impact on the search's results then Tweet features.

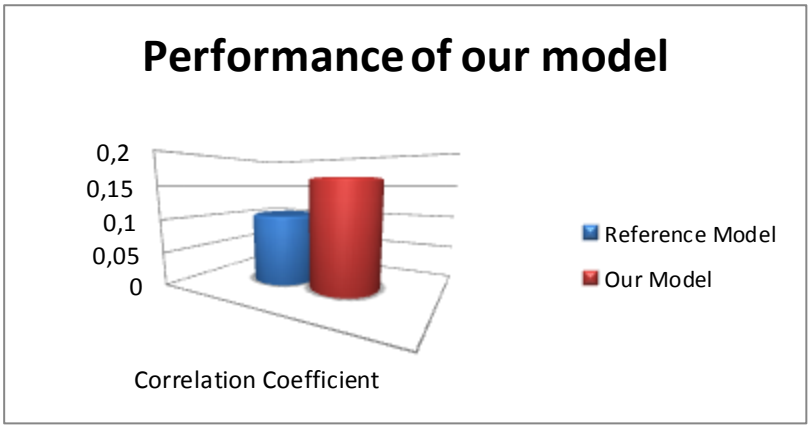


FIG.9 - *Comparing our model with reference model*

The reference model combines only the features linearly without weighting. This model gave us the correlation coefficient equal to 0.10 and our model gave us the correlation coefficient of 0,16. Can clearly be seen a 35% improvement in the satisfaction of our human judgment.

6. Conclusion

Research conducted under the auspices of knowledge management varies greatly in direction and scope. There are several approaches that have been proposed which are based on the features. Therefore the choice of characteristics is important to obtain a satisfactory result and close to the human judgment. We have proposed in this paper a new metric for Social Research on twitter. This has to integrate relevance of content, the specificities of tweets and the author's importance where we incorporate new features such as the audience. The primary experimental evaluation that we conducted on a collection of articles of Twitter shows the measurement that we propose allows a better assessing the impact of bloggers and tweets' technical specificities.

Looking ahead, we plan to conduct experiments under the Micro-blog Text REtrieval Conference (TREC) evaluation framework that will include a collection of many articles and queries for larger and whose relevance judgments are social. We also need to evaluate the influence of each feature independently. We plan to compare the performance of our model with other models for social searching of tweets.

References

Agarwal N., Liu H., Tang L., Yu P. S. (2008), « Identifying the influential bloggers in a community », Proceedings of the international conference on Web search and web data mining, WSDM '08, New York, NY, USA, 2008, ACM, p. 207-218.

Akermi I., And Faiz R., «Hybrid method for computing word-pair similarity based on web content.» In Proceedings of the International Conference on Web Intelligence, Mining and Semantics, WIMS'12 New York, NY, USA, ACM., 2012.

Balog K., De Rijke M., Weerkamp W., « Bloggers as experts : feed distillation using expert retrieval models », Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '08, New York, NY, USA, 2008, ACM, p. 753–754.

Barbosa. L and J. Feng. Robust sentiment detection on Twitter from biased and noisy data. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 36{44. Association for Computational Linguistics, 2010.

Barry S., "Web search: social & collaborative" in COnférence en Recherche d'Infomations et Applications CORIA 2011, 8th French Information Retrieval Conference CORIA : COnférence en Recherche d'Information et Applications Avignon, France, March 16-18, 2011. Proceedings CORIA, Editions Universitaires d'Avignon, 2011

Ben Jabeur L., Tamine L., Boughanem M., "Intégration des facteurs temps et autorité sociale dans un modèle bayésien de recherche de tweets",Conférence en Recherche d'Information et Applications (CORIA), Bordeaux, France, March 21-23, 2012. Proceedings, pp. 301–316, 2012.

Ben Jabeur L., Tamine L., Boughanem M., « Un modèle de recherche d'information sociale dans les microblogs : cas de Twitter » Conférence sur les Modèles et l'Analyse des Réseaux : Approches Mathématiques et Informatique (MARAMI 2011), Grenoble, 2011

Ben Jabeur L., Tamine L., « Vers un modèle de Recherche d'Information Sociale pour l'accès aux ressources bibliographiques » (poster). Dans : Conférence francophone en Recherche d'Information et Applications (CORIA 2010), Sousse, Tunisie, 18/03/2010-20/03/2010, Centre de Publications Universitaires, p. 325-336, mars 2010.

Cha M., Haddadi H., Benevenuto Krishna F., Gummadi P., "Measuring User Influence in Twitter: The Million Follower Fallacy Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media 2010, ICWSM 2010

Cherichi S., Faiz R., « Recherche d'information pertinente dans les microblogs: Mesure métrique de l'impact des critères pour améliorer les résultats de la recherche». Conférence Internationale sur l'Extraction et la Gestion des Connaissances – Maghreb, Hammamet, Tunisie, EGC-M 2012

Cherichi S. And Faiz R., New metric measure for the improvement of search results in microblogs. Proc. of the International Conference on Web Intelligence, Mining and Semantics (WIMS 2013), New York, NY, USA, 2013. ACM.

Cherichi S. And Faiz R., Relevant Information Discovery in Microblogs : New metric measure for the improvement of search results in microblogs. Proc. of INSTICC International Conference on Knowledge Discovery and Information Retrieval (KDIR 2013), Vilamoura, Portugal, 19-22 September 2013. ©SciTePress

Cherichi S., Faiz R., "Relevant information management in microblogs". In International Conference on Knowledge Management, Information and Knowledge Systems (KMIKS 2013), Hammamet, Tunisia, Avril 2013.

Damak F., Pinel-Sauvagnat K., Cabanac G., « Recherche de microblogs : quels critères pour raffiner les résultats des moteurs usuels de RI ? » COnférence en Recherche d'Information et Applications 2012 CORIA 2012 , Bordeaux, CORIA 2012 p. 371-328

Dong A., Zhang R., Kolari P., Bai J., Diaz F., Chang Y., Zheng Z., Zha H., «Time is of the essence : improving recency ranking using Twitter data », Proceedings of the 19th international conference on World wide web, WWW '10, New York, NY, USA, 2010, ACM, p. 331–340.

Duan Y., Jiang L., Qin T., Et Al., "An empirical study on learning to rank of tweets", COLING Proceedings of the 23rd International Conference on Computational Linguistics Proceedings of the Conference, 23-27 August 2010, Beijing, China, pp. 295–303, 2010. Tsinghua University Press, 2010.

Doan. S, B.K.H. Vo, and N. Collier. An analysis of Twitter messages in the 2011 Tohoku earthquake. Arxiv preprint arXiv:1109.1618, 2011.

Efron M., « Hashtag retrieval in a microblogging environment », Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval, SIGIR '10, New York, NY, USA, 2010, ACM, p. 787–788.

Go, A R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, pages 1{12, 2009.

Hiemstra D., « Using Language Models for Information Retrieval», PhD thesis, Enschede, January 2001.

Ingwersen P., “Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction”, Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 101–110, 1994 ACM/Springer 1994

Java A., Song X., Finin T., Tseng B., « Advances in Web Mining and Web Usage Analysis », Chapter Why We Twitter : An Analysis of a Microblogging Community, p. 118–138, Springer-Verlag, Berlin, Heidelberg, 2009.

Jiang. L, M. Yu, M. Zhou, X. Liu, and T. Zhao. Target-dependent Twitter sentiment classification. Proc. 49th ACL: HLT, 1:151{160, 2011.

Jones S., Walker K., Robertson S,. “A probabilistic model of information retrieval: Development and comparative experiments.” Information Processing & Management, 36(6) :779–808, 2000.

Lampos. V and N. Cristianini. Tracking the u pandemic by monitoring the social web. In Cognitive Information Processing (CIP), 2010 2nd International Workshop on, pages 411{416. IEEE, 2010.

Nagmoti R., Teredesai A., “Ranking Approaches for Microblog Search” Proceeding WI-IAT '10 Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01 Pages 153-157

OConnor. B, R. Balasubramanyan, B.R. Routledge, and N.A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In Proceedings of the International AAAI Conference on Weblogs and Social Media, pages 122{129, 2010.

Robertson S., Walker S., Hancock-Beaulieu M., « Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive », Text REtrieval Conference TREC, 1998, p. 199-210.

Sakaki T., Okazaki M., Matsuo Y., « Earthquake shakes Twitter users : realtime event detection by social sensors », Proceedings of the 19th international conference on World wide web, WWW ’10, New York, NY, USA, 2010, ACM, p. 851–860.

Sankaranarayanan J., Samet H., Teitler B. E., Liebermann M. D., Sperling J., « TwitterStand : news in tweets », Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS ’09, New York, NY, USA, 2009, ACM, p. 42–51.

Sarwar B., Karypis G., Konstan J., And Riedl J., “Item-based collaborative filtering recommendation algorithms”. In Proceedings of the 10th international conference on World Wide Web, WWW ’01, pages 285–295, New York, NY, USA, 2001. ACM.

The performance of the Apriori-DHP algorithm with some alternative measures

Faraj A. El-Mouadib[*]
Khirallah S. Al ferjani[**]
University of Benghazi
Faculty of Information Technology
[*]elmouadib@gmail.com
[**]kh214300@yahoo.com

Abstract. Nowadays, the explosive growth in data collection in many areas such as business, government, medical and etc… defeated human ability to understand it and digest it. The overwhelming data volumes presented new challenges to produce new tools and techniques to extract useful knowledge from such data. These challenges have resulted in the development of new tools and techniques of a fairly new field called Knowledge Discovery in Databases (KDD) and Data Mining (DM). One of the most widely studied and research task in the DM functionalities is Association Rules Mining (ARM) due to its use in business and commerce.
In this paper, we demonstrate the implementation of the well-known ARM algorithm APRIORI with one of its improvements namely; Direct Hashing and Pruning (DHP), Özel S. and Güvenir H. (2001) as a test bed. The two algorithms are implemented in a system called "ADAS" by the use of the MATLAB7.0 programming language. The objective is to evaluate the validity of using some of the suggested alternative interestingness measures namely; *Correlation, Conviction,* and *Odds ratio* in lieu of *Support-Confidence* frame work. The evaluation process is carried out by conducting 80 experiments on the implementation of the two algorithms. Finally, an extensive analysis and discussion of the results is given using the well-known mushroom database.

1 Introduction

Due to cheaper and larger storage capacities, there is a dramatic increase in the amount of collected data in many different formats. Nowadays, huge repository systems can have as many as 10^2 to 10^3 fields and 10^9 records Fayyad, U. M., et. al., (1996) that are very common in many businesses. So in fact, we are drowning in data, demanding information and starving for knowledge, because the numbers and sizes of databases far exceeds human capabilities to analyze and digest. Knowledge leads to power and success of decision making. Knowledge is the result of a new field known as Knowledge Discovery in Databases (KDD). KDD is defined as; the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. One of the most essential steps in the KDD process is Data Mining (DM) even though some people consider the two as synonymous. Generally, data mining tasks are grouped into descriptive and predictive, Han, J., et. al.

(2012). Extracted knowledge can come in many different forms such as; association rules, classification rules, clustering, discrimination rules and etc…

One of the most popular and widely used DM functionality is association analysis where many algorithms have been developed and used for such task Agrawal, R., et. al., (1993). Following the first algorithm AIS, Agrawal, R., and Srikant, R., (1994) for the discovery of association rules was the Apriori algorithm, which became the land mark for Association Rule Mining (ARM).

The Apriori algorithm and its variations (Apriori-based algorithms) suffer from two bottlenecks which are: the high cost to handling huge number of candidate sets and the need of multiple scans over the database. Also the two used measures namely; Support S and Confidence C, to filter out the real from the superficial association rule, have received some criticisms. For the two bottlenecks many improvements have been suggested i.e. Apriori_TID, Apriori_Hybrid and Direct Hashing and Pruning (DHP) Park, J. S., et. al., (1995). Dynamic Itemset Counting (DIC) Brin, S., et. al., (1997). The reduction of the number of records to be searched (i.e. Partitioning, and Sampling), Toivonen, H., (1996). For the criticisms of the used measure of interestingness, many researches in the field of ARM have proposed many alternative measures to Support and Confidence frame work.

In this paper, we concerned with the evaluation and the validity of some of the proposed alternative interestingness measures specifically: *Correlation*, *Conviction*, and *Odds ratio*. The evaluation is carried out in the form of experiments on Apriori-DHP with the suggested different interestingness measures. In the next section, we review the necessary background for studying the association rule mining and some of the related work. In Section 3, we present the APRIORI algorithm measures, criticisms to these measures and some of the proposed alternative interestingness measures. Section 4, we review of our test bed system to evaluate the validity of the alternative measures with and without the improvements of DHP to the APRIORI algorithm. In Section 5, we demonstrate the empirical results obtained from the ADAS test bed system to evaluate the validity of some of the alternative interestingness measures. In Section 6, we represent the results and in Section 7, we represent the conclusition and advise of some further research.

2 Association Rule Mining (ARM)

The ARM aims at the discovery association rules (finding interesting relationships among sets of items in a transactional database) Agrawal, R., and Srikant, R., (1994). One of the most expressive forms of knowledge representation is the "IF–THEN" rules due to its ease of human understandability and comprehension. Such form is used in association rules, discriminate rules, classification rules, etc… Due to the wide use of association rules in market basket analysis, the association rules have received considerable research and development attention [Agrawal, R., and Srikant, R., (1994), Agrawal, R., et. al. (1993). The early 90's had witnessed a lot of attention to association rules mining. As a result of the research new versions of the APRIORI algorithm were proposed and mainly on the fact that this algorithm uses prior knowledge of frequent itemset properties. The APRIORI algorithm has achieved better significance over previous ones due to its use of prior knowledge. Since the introduction of APRIORI many improvements have been suggested to make the algorithm more efficient in the sense of the reduction of the number of passes over the database. According to Fayyad, U. M., et. al., (1996), the problem of the performance has

sustained until the introduction of the (Frequent Pattern) FP-Tree algorithm Han J., et. al. (2000) that was best attempt to deal with this problem.

3 Association rules measures

Discovering association rules is considered to be one of the most important DM functionalities where many algorithms had been developed. Usually, not all of the discovered rules constitute a useful knowledge. So, the evaluation of all of the discovered rules is an important issue to separate good rules from superficial ones. The Apriori-based algorithms use two measures: *Support S* and *Confidence C* to evaluate the validity of the association rules. The efficiency of the algorithms that discover the association rules became a major issue because of the wide spread use of the association rules in market basket analysis.

3.1 Apriori criticisms

Since the introduction of the APRIORI algorithm in the early 90's, there have been some criticisms Liaquat M. et. al. (2004) to the *Support-Confidence* frame work that had been used in evaluating the interestingness of the discovered association rules. These criticisms are:

1. The measures of interestingness used in APRIORI, *Support* and *Confidence* are not suitable to capture such dependencies and are weak in expressing the notion of *Correlation*.
2. Sometimes, the *Confidence* measure gives untrue results especially when all transactions have the items in the consequent.

Here, we present two segments of transactional database examples in the form of a matrix to illustrate the above mentioned criticisms numerically. The first database segment is for the first criticism and the second is for the second criticism. These tables are:

		Transactions							
	Tid	*T1*	*T2*	*T3*	*T4*	*T5*	*T6*	*T7*	*T8*
Items	*X*	1	1	1	1	0	0	0	0
	Y	1	1	0	0	0	0	0	0
	Z	0	1	1	1	1	1	1	1

	Transactions						
	Tid	*T1*	*T2*	*T3*	*T4*	*T5*	*T6*
Items	*X*	0	1	0	0	0	1
	Y	1	1	1	1	1	1

Where *X*, *Y* and *Z* represent the items and *T1* … *T8* in the first able and *T1* … *T6* in the second table represent the transactions. The code of 1 means the existences of the given item in the transaction and 0 represents the lack of it. The above mentioned criticisms had encouraged researches in the field of association rule mining to propose alternative measures to *Support* and *Confidence* for rules interestingness.

3.2 Alternative measures

Since the introduction of the APRIORI algorithm in the early 90's, there have been quite a number of suggested alternative measures Liaquat M. et. al. (2004). Here, we give the definitions, notations and notions of some of the alternative suggested measures of interestingness in ARM.

3.2.1 Correlation measure

The *Correlation* (*Corr*) is a bivariate measure of association (strength) of the relationship between pairs of variables or pairs of itemsets. The range value of the *Correlation* is between -1 and 1 inclusive. The interpretation of the *Correlation* is; when the value of the *Corr* is -1 means that there is a negative correlation between the variables/ itemsets and when the value of the *Corr* is 0 means that there is a no *Correlation* between the itemsets. The *Correlation* value of 1, means that there is a positive correlation between the itemsets. The *Support*, *Confidence* and *Correlation* are calculated by:

$$Support = \frac{Number\_of\_transactions(X \cup Y)}{Total\_number\_of\_transactions} \quad (3.2.1)$$

$$Confidence = \frac{Number\_of\_transactions(X \cup Y)}{Number\_of\_transactions(X)} \quad (3.2.2)$$

$$Corr(X \rightarrow Y) = \frac{P(XandY) - P(X)P(Y)}{\sqrt{P(X)P(Y)(1-P(X))(1-P(Y))}} \quad (3.2.3)$$

The results of the calculations are depicted in table-1.

	$X \rightarrow Y$	$Y \rightarrow Z$	$X \rightarrow Z$
Support	25.0%	12.5%	37.5%
Confidence	50.0%	50.0%	75.0%
Correlation	0.577	-0.649	-0.383

Table-1: Calculation results of *Support*, *Confidence* and *Correlation* measures.

From table-1, we can see that the first criticism to the Support-Confidence frame work is true for this data set. The results for the second data set showed that the Support and Confidence values for the rule $X \rightarrow Y$ are: 0.33 and 100.00 respectively. The value of the Confidence gives the impression that all the transaction that contain the item *Y* also contain the item *X* which is not true for this data set. So, this data set supports the second criticism.

3.2.2 Conviction measure

The *Conviction* measure was introduced in Brin, S., et. al. (1997). This measure works like the *Correlation* where the antecedent and consequent are taken into consideration when measuring the association between two groups of itemsets.

For a rule on the form of $X \rightarrow Y$, the Confidence measure uses the conditional probability $P(Y|X)$, and does not take the probability of the consequence, $P(Y)$, into consideration. The *Conviction* measure was developed as an alternative to the Confidence and it uses the information of the absence of the consequent. The *Conviction* measure is calculated by:

$$Conviction(X \rightarrow Y) = \frac{P(X)P(\overline{Y})}{P(X and \overline{Y})} \qquad (3.3.1)$$

The range value of the *Conviction* measure is [0, ∞). The value of 0 represents a total independence between the items in the antecedent and consequent of the association rule. The upper bound value of ∞, means that the items in the antecedent and consequent are related on the magnitude of 100%. Table-2 depicts the results of calculating the *Conviction* measure, by the use of equation 3.3.1, along with the *Confidence* measure of the first example data.

	$X \rightarrow Y$	$Y \rightarrow Z$	$X \rightarrow Z$
Confidence	50.0%	50.0%	75.0%
Conviction	1.50	0.25	0.50

Table-2: Calculations results of *Confidence* and *Conviction*.

From table-2, the *Support-Confidence* frame work shows that there is a very strong association between the itemsets *X* and *Y* for the rule $X \rightarrow Y$ while the *Conviction* measure shows a value of 1.5 which is very close to independence. For the rules $X \rightarrow Z$ and $Y \rightarrow Z$, the results had the same trend as for the rule $X \rightarrow Y$. For the second example data, the value of the *Conviction* measure for the rule $X \rightarrow Y$ is ∞, which is practically the same as for the *Support-Confidence* frame work.

3.2.3 Odds Ratio measure

The *Odds ratio* is a statistical measure that evaluates the ratio of the existence of an event in one group to the existence of the same event in another group, http://en.wikipedia.org/wiki/Odds-ratio and Westergren, A. et al., (2001). The *Odds ratio* for the rule $X \rightarrow Y$ is given by:

$$Odds\ ratio(X \rightarrow Y) = \frac{P(XandY)P(\overline{X} and \overline{Y})}{P(X and \overline{Y})P(\overline{X} and Y)} \qquad (3.3.2)$$

The range value of the *Odds ratio* measure is on the scale of [0, ∞). The interpretation of the range values is that; the value of 0 means that the itemset in the antecedent and the itemset in the consequent are independent. Otherwise they are related. The strongest association occurs when the value of the measure is equal to ∞. By considering the data given in the first example and applying the equation (3.3.2), the calculation of the Odds ratio measure and the *Support-Confidence* frame for all of the three association rules have resulted in the following:

	$X \rightarrow Y$	$Y \rightarrow Z$	$X \rightarrow Z$
Support	25.0%	12.5%	37.5%
Confidence	50.0%	50.0%	75.0%
Odd ratio	∞	0.00	0.00

Table-3: Calculation results of *Support*, *Confidence* and the *Odds ratio* measures.

The results of the *Support-Confidence* frame work had shown that there is a very strong association (25%, 50%) between the itemsets *X* and the itemsets *Y* for the association rule $X \rightarrow Y$ and the *Odds ratio* measure had resulted in a value of ∞ to indicate that there is a very strong association between the itemsets *X* and the itemsets *Y*. For the association rules $X \rightarrow Z$

and $Y \rightarrow Z$, the *Support-Confidence* frame work had shown that there is a very strong association (37.5%, 75%), (12.5%, 50%) respectively. But the *Odds ratio* measure of value of 0.0 for both of the association rules to indicate that the itemsets *X* and the itemsets *Z* are independent of each other and it is the same for the itemsets *Y* and the itemsets *Z* as well.

By considering the data given in the second example and applying the equation (3.3.2), the calculation of the *Odds ratio* measure for the rule $X \rightarrow Y$ is as follows:

$$Odds(X \rightarrow Y) = \frac{0.333 * 0}{0 * 0.67} = \frac{0}{0} = \infty$$

We have found out that the two measures, *Support-Confidence* frame work and *Odds ratio* had given practically the same result as far as the interpretation of the results concern.

4 Design and implementation

Here, we give a brief review of our test bed system to evaluate the validity of the alternative measures with and without the improvements of DHP to the APRIORI algorithm. This system APRIORI-DHP-AlternativeS (ADAS), see figure-1, consists of four sub-systems, each of which is slightly different than the others. MATLAB7.0 is used to implement all of the sub-systems. The MATrix LABoratory (MATLAB) is a programming language that is specialized in mathematical computations.

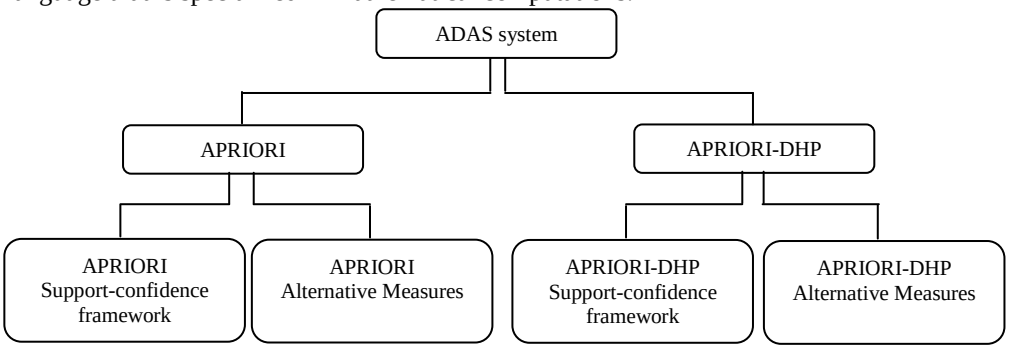


Figure-1: Main components of the ADAS test bed system.

5 Testing and experiments

Here, we demonstrate the empirical results obtained from the ADAS test bed system to evaluate the validity of some of the alternative interestingness measures. In the evaluation process two very well-known, in the field of association rules, data sets are used. The choice of using these data sets is based on their frequently use within the association rule research community. The first data set is the Mushroom data Bache, K. & Lichman, M. (2013), which was donated by Jeff Schlimmer and drawn from The Audubon Society Field Guide to North American Mushrooms (1981). G. H. Lincoff (Pres.), New York: Alfred A. Knopf. The second data set is the Chess data Bache, K. & Lichman, M. (2013), which was originally generated and described by Alen Shapiro and supplied by Peter Clark of the Turing Institute in Glasgow to the donor Rob Holte. Due to space limitation requirement, we will present only the Mushrooms experiment in this paper. The Mushroom set database consists of 8124 transactions, 108 different items and the average number of items per transaction is 23. The

size of this database is about 1.59 MB. This experiment has been conducted ten times, with a fixed threshold of 70 for the Support measure. The obtained results will be presented in a table format to exhibit the differences in results of applying the different alternative measures to the same data. A total of 80 experiments are conducted and the discussion of the results will be based on three criteria namely; number of produced rules, rule complexity (antecedent complexity and consequent complexity) and execution time. The results are organized according to the four different versions of the implemented algorithms with different levels of rule acceptance (*Confidence*) value of; 30, 37, 45, 52, 60, 67, 75, 82, 90 and 97.

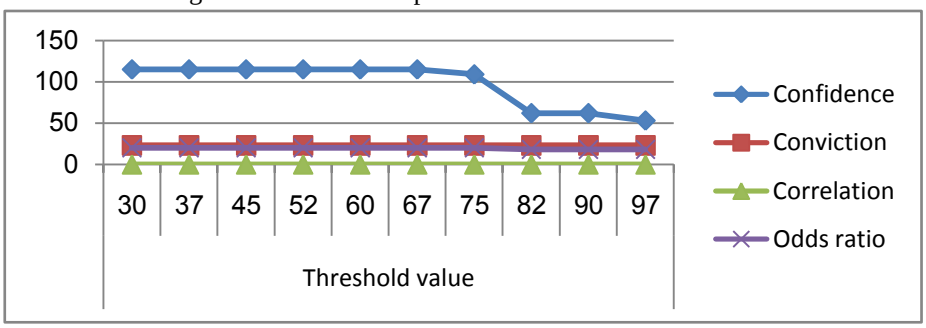
5.1 The experiments of APRIORI sub-system

This set experiment is to test the APRIORI sub-system of the ADAS test bed system. Table-5.1 depicts the numerical results for the number of rules for the APRIORI as well as the alternative measures with APRIORI. Figure-5.1 illustrates a plot of the results in table-5.1.

Table-5.1: Number of rules for the APRIORI sub-system and APRIORI with alternative measures.

Measure	Acceptance threshold value									
	30	37	45	52	60	67	75	82	90	97
Confidence	115	115	115	115	115	115	109	62	62	53
Conviction	23	23	23	23	23	23	23	23	23	23
Correlation	0	0	0	0	0	0	0	0	0	0
Odds ratio	20	20	20	20	20	20	20	18	18	18

Figure-5.1 illustrates a plot of the results in table-5.1.

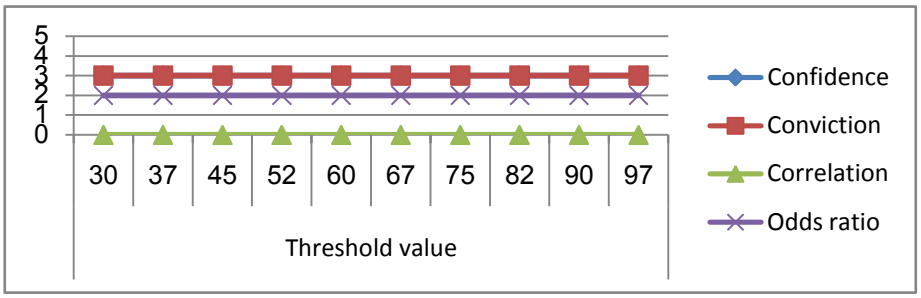


The numerical results of the APRIORI sub-system and the APRIORI with the alternative measures for the 80 versions of the experiment for the number of items in the antecedent of the rules are depicted in table-5.2. Figure-5.2 illustrates a plot of the results in table-5.2.

Table-5.2: Number of items in the antecedent for APRIORI and APRIORI with alternative measures.

Measure	Acceptance threshold value									
	30	37	45	52	60	67	75	82	90	97
Confidence	3	3	3	3	3	3	3	3	3	3
Conviction	3	3	3	3	3	3	3	3	3	3
Correlation	0	0	0	0	0	0	0	0	0	0
Odds ratio	2	2	2	2	2	2	2	2	2	2

Figure-5.2 illustrates a plot of the results in table-5.2.



The number of items in the consequent of the association rule for APRIORI sub-system and APRIORI with alternative measures sub-system is depicted in table-5.3. Figure-5.3 depicts a plot of the results in table-5.3.

Table-5.3: Number of items in the consequent of the association rule for APRIORI sub-system and APRIORI with alternative measures.

Measure	Acceptance threshold value									
	30	37	45	52	60	67	75	82	90	97
Confidence	3	3	3	3	3	3	3	3	3	3
Conviction	3	3	3	3	3	3	3	3	3	3
Correlation	0	0	0	0	0	0	0	0	0	0
Odds ratio	2	2	2	2	2	2	2	2	2	2

Figure-5.3 depicts a plot of the results in table-5.3.

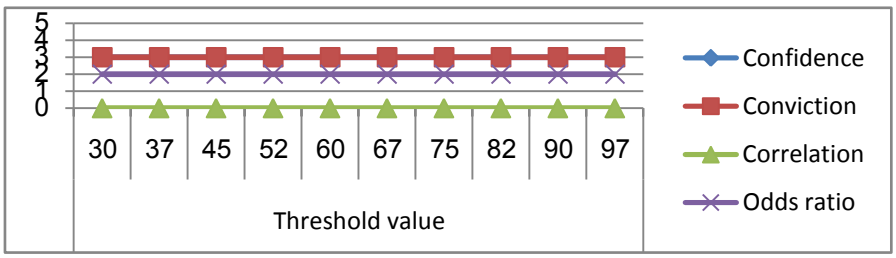
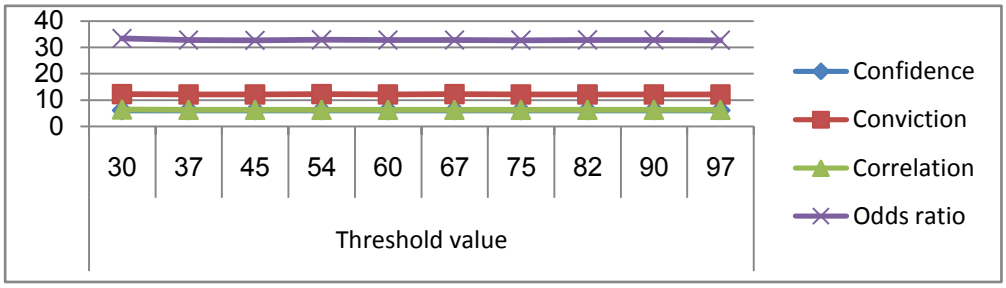


Table-5.4 depicts the numerical results of the 80 experiments for the execution time criterion for the APRIORI sub-system and APRIORI with alternative measures.

Table-5.4: Execution time in seconds for APRIORI and APRIORI with alternative measures.

Measure	Acceptance threshold value									
	30	37	45	54	60	67	75	82	90	97
Confidence	6.16	6.16	6.16	6.16	6.16	6.16	6.16	6.16	6.16	6.16
Conviction	12.32	12.18	12.23	12.30	12.23	12.30	12.19	12.22	12.19	12.19
Correlation	6.52	6.40	6.41	6.40	6.39	6.41	6.40	6.40	6.40	6.40
Odds ratio	33.42	32.87	32.73	32.91	32.85	32.87	32.77	32.89	32.80	32.77

Figure-5.4 depicts a plot of the results in table-5.4.



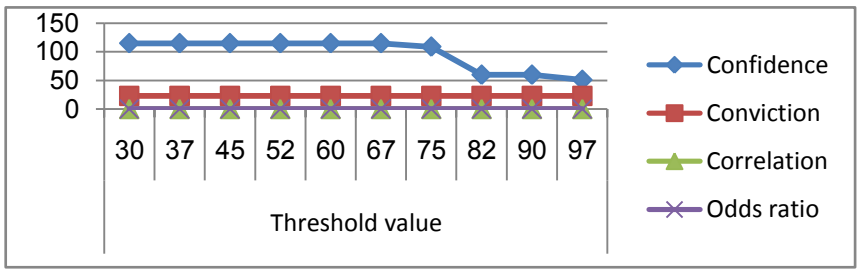
5.2 The experiments of APRIORI-DHP sub-system

This version of the experiment is to test the APRIORI-DHP sub-system of the ADAS test bed system with and without the alternative measures. Table-5.5 depicts the numerical results of the 80 experiments for the number of rules.

Table-5.5: Number of rules for the APRIORI-DHP sub-system and APRIORI-DHP with alternative measures.

Measure	Acceptance threshold value									
	30	37	45	52	60	67	75	82	90	97
Confidence	115	115	115	115	115	115	109	60	60	51
Conviction	23	23	23	23	23	23	23	23	23	23
Correlation	0	0	0	0	0	0	0	0	0	0
Odds ratio	0	0	0	0	0	0	0	0	0	0

Figure-5.5 illustrates a plot of the results in table-5.5.

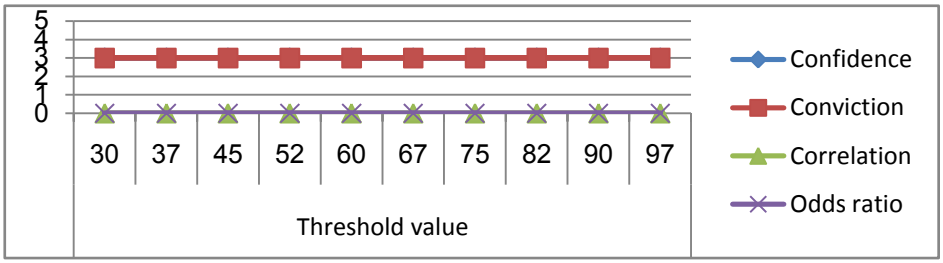


The numerical results of the APRIORI-DHP sub-system and APRIORI-DHP with alternative measures, in the 80 experiments for the number of items in the antecedent of the rules are depicted in table-5.6. Figure-5.6 illustrates a plot of the results in table-5.6.

Table-5.6: Number of items in the antecedent of the association rule for APRIORI-DHP sub-system and APRIORI-DHP with alternative measures.

Measure	Acceptance threshold value									
	30	37	45	52	60	67	75	82	90	97
Confidence	3	3	3	3	3	3	3	3	3	3
Conviction	3	3	3	3	3	3	3	3	3	3
Correlation	0	0	0	0	0	0	0	0	0	0
Odds ratio	0	0	0	0	0	0	0	0	0	0

Figure-5.6 illustrates a plot of the results in table-5.6.



The number of items in the consequent of the association rule for APRIORI-DHP sub-system and APRIORI-DHP with alternative measures is depicted in table-5.7. Figure-5.7 illustrates a plot of the results in table-5.7.

Table-5.7: Number of items in the consequent of the association rule for APRIORI-DHP sub-system and APRIORI-DHP with alternative measures.

Measure	Acceptance threshold value									
	30	37	45	52	60	67	75	82	90	97
Confidence	3	3	3	3	3	3	3	3	3	3
Conviction	3	3	3	3	3	3	3	3	3	3
Correlation	0	0	0	0	0	0	0	0	0	0
Odds ratio	0	0	0	0	0	0	0	0	0	0

Figure-5.7 illustrates a plot of the results in table-5.7.

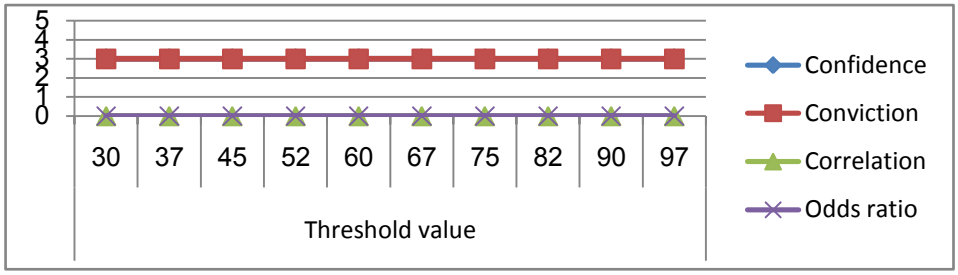
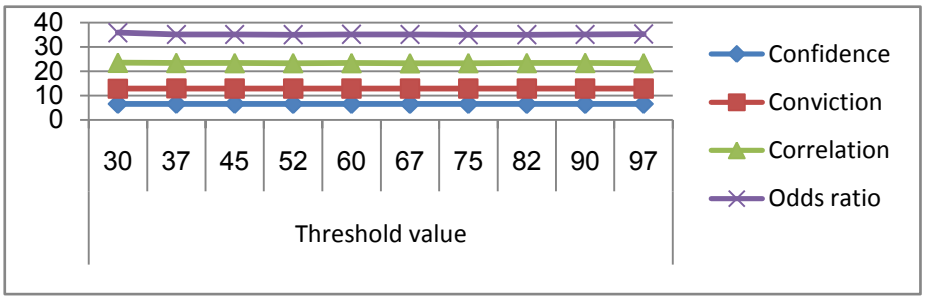


Table-5.8 depicts the numerical results of the 80 experiments for the execution time for the APRIORI-DHP sub-system and APRIORI-DHP with alternative measures. Figure-5.8 illustrates a plot of the results in table-5.8.

Table-5.8: Execution time in seconds for APRIORI-DHP sub-system and APRIORI-DHP with alternative measures.

Measure	Acceptance threshold value									
	30	37	45	52	60	67	75	82	90	97
Confidence	6.62	6.63	6.63	6.59	6.59	6.59	6.59	6.58	6.58	6.60
Conviction	12.88	12.95	12.92	12.90	12.91	12.94	12.93	12.90	12.92	12.89
Correlation	23.55	23.47	23.41	23.40	23.41	23.35	23.40	23.43	23.47	23.39
Odds ratio	35.88	35.15	35.12	35.06	35.16	35.14	35.04	35.03	35.10	35.30

Figure-5.8 illustrates a plot of the results in table-5.8.



6 Results

The goal of this study was set to evaluate the validity of some of the alternative interestingness measures namely; Conviction, Correlation and Odds ratio. The evaluation of the alternative measures was carried out in the implementation of the APRIORI algorithm and APRIORI-DHP algorithm. The two algorithms are implemented in a test bed system "ADAS" by the use of MATLAB7.0 programming language. We have tested our system via 80 experiments using Mushroom database. This database is of size 1.59MB.

From the obtained results for the criterion number of rules, we would like to make the following comments:

1. For the Confidence and Correlation measures, the number of rules decreased when the threshold measure was increased. Such result was naturally expected.
2. For the Odds ratio measure, the number of rules decreased when the threshold measure was increased.
3. The measures Conviction had produced no rules, so the evaluation of such criterion is not possible.

7 Conclusion and future work

In conclusion, from our experience with the data and the measures that we have used, the Confidence and Correlation measures are better than the other measures.

From the obtained results for the criterion execution time, we would like to make the following comments:

- For the APRIORI sub-system:

  The best average and worst execution time is with the use of the Lift measure. The worst average execution time was with the Conviction measure.
- For the APRIORI-DHP sub-system:

The best average execution time is with the use of the Odds ratio measure. The worst and average execution time was with the Conviction measure. The Odds ratio measure had outperformed the other measures as far as the criterion of execution time is concerned.

From the results we had obtained, we would like to make the following points for future work:

- Conduct more experiments with different sets of data.
- Study the possibility to combine some of the alternative measures for better results.

- Study the possibilities of modifying the Support-Confidence frame work to overcome the criticisms.

References

Agrawal, R., and Srikant, R., (1994). Fast algorithms for mining association rules in large databases. *In Proceedings of 20th International Conference on Very Large Databases*, Santiago, Chile. Pages 478 - 499.

Agrawal, R., Imielinski, T., and Swami, A., (1993). Mining association rules between sets of items in large databases. *In Proceedings of International ACM SIGMOD Conference on Management of Data*, Washington, D.C. Pages 207 -216.

Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Brin, S., Motwani, R., Ullman, J. D., and Tsur, S., (1997). Dynamic itemset counting and implication rules for market basket analysis. *In Proc. ACM-SIGMOD Int. Conf. Management of Data*, Tucson, Arizona. Pages 255 - 264.

Fayyad, U. M., et. al., (1996). From Data Mining to Knowledge Discovery: An Overviews, *Advances in Knowledge Discovery and Data Mining*, AAAI Press/ MIT Press. Pages 1-34.

Han J., Pei J. and Yin Y. (2000), Mining Frequent Patterns without Candidate Generation. In *Proceeding Conference on the Management of Data*, ACM Press. New York, USA. Pages 1– 12.

Han, J., Kamber, M. and Pei J., (2012). *Data mining: concepts and techniques (3rd edition)*. Morgan Kaufmann Publishers is an imprint of Elsevier. 225Wyman Street, Waltham, MA 02451, USA.

http://en.wikipedia.org/wiki/Odds-ratio, Last visit December, 2013.

Liaquat Majeed Sheikh, Basit Tanveer, Syed Mustafa Ali Hamdani., (2004). *Interesting Measures for Mining Association Rules*. FAST-NUCES, Lahore.

Özel S. and Güvenir H. (2001). An Algorithm for Mining Association Rules Using Perfect Hashing and Database Pruning, in: *Proceedings of the Tenth Turkish Symposium on Artificial Intelligence and Neural Networks(TAINN'2001)*, A. Acan, I. Aybay, and M. Salamah (Eds.), Gazimagusa, T.R.N.C. (June 2001). Pages 257-264.

Park, J. S., Chen, M.S., and Yu, P.S., (1995). An effective hash-based algorithm for mining association rules. In: *Proc. ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'95)*, San Jose, CA. Pages 175–186.

Piatetsky-Shapiro, G., (1991). Discovery, analysis, and presentation of strong rules. *Knowledge Discovery in Databases*, Pages 229-248.

Toivonen, H., (1996). Sampling large databases for association rules. *Conf. Very Large Data Bases. Bombay*, India. pages 134-145.

Westergren, A. et al., (2001). INFORMATION POINT: Odd ratio. *Journal of Clinical Nursing*, 10. Blackwell Science Ltd, Pages 257- 269.

SEIR-SW : Un modèle de suivi de propagation d'épidémies

Fatima-ZohraYounsi*,***, Ahmed Bounekkar**,
Djamila Hamdadou***

*Laboratoire ERIC, Univ.Lumière Lyon 2,5 avenue Pierre
Mendès-France, 9676 Bron Cedex, France
Fatima-zohra.younsi@univ-lyon2.fr
**Polytech Lyon, Bâtiment ISTIL 15 Boulevard André Latarjet, 69622, France
ahmed.bounekkar@univ-lyon1.fr
http://polytech.univ-lyon1.fr/m-bounekkar-ahmed-706492.kjsp
***Laboratoire LIO, Univ. Oran,BP 1524, El-M'Naouer, 31000, Oran, Algérie
dzhamdadoud@yahoo.fr

Résumé. Les maladies infectieuses sont, chaque année, la principale cause d'une morbidité et d'une mortalité importante. Malgré les progrès en termes de traitement et de prévention, il n'y a finalement que très peu de maladies infectieuses éradiquées. Cela mène à une nécessité d'un suivi et d'un contrôle de la propagation des maladies infectieuses afin de mieux les comprendre. Dans ce contexte, nous nous intéressons, dans ce travail, à la modélisation du phénomène de propagation de l'épidémie de la grippe au sein de la population oranaise. Notre objectif est double : il consiste, d'une part, à comprendre comment l'épidémie se propage par l'utilisation du réseau social petit monde (Small World) et le modèle mathématique d'épidémie SEIR (Susceptible-Exposed-Infected-Removed), et d'autre part, prédire quelle sera l'évolution future de la maladie.

1 Contexte et problématique

Les maladies infectieuses représentent aujourd'hui un problème majeur de santé publique (Morens et al. 2004). Devant l'augmentation des résistances bactériennes, l'émergence de nouveaux pathogènes et la propagation rapide de l'épidémie, la prévention de la transmission de la maladie devient particulièrement importante et indispensable.

Face à une telle menace, la société doit se préparer à l'avance pour réagir rapidement et efficacement si une telle épidémie est déclarée. Ce contexte épidémiologique souligne la nécessité de coupler les champs disciplinaires de l'épidémiologie et de la modélisation du phénomène afin d'identifier et de mieux comprendre comment ces maladies sont transmises dans leur environnement et d'étudier quelles sont les stratégies de contrôle les plus efficaces face à la progression de cette épidémie.

La modélisation en épidémiologie a été utilisée dans l'évaluation des préventions, des programmes de contrôle et de lutte. Comprendre le phénomène de transmissiondes virus dans lespopulations humaines est une questionfondamentaleen épidémiologie. Les modèles tradi-

tionnels épidémiologiques supposent que la transmission d'une infection dans une population hôte homogène augmente bien avec le nombre d'individus et avec un risque d'infection au hasard **(**Anderson et May,79)(May, 97). Toutefois, la transmission d'une épidémie se fait par un contact direct entre un susceptible et un infecté. Autrement dit, dans notre étude le virus de la grippe est transmis suite à un contact direct. Récemment, la recherche a reconnu l'importance de l'étude du comportement de l'hôte (individu) et des moyens de contact entre les hôtes dans la transmission de parasites. Les deux concepts : comportement de l'individu et degré de contact entre les individus dans une population seront influencés par la structuration sociale de la population.

En épidémiologie et dans le cas des maladies infectieuses, nous cherchons à identifier les agents infectieux et à comprendre leur mode de propagation. Un certain nombre de questions se pose :

- Comment l'épidémie se propage d'un individu vers un autre ?
- Quels sont les facteurs qui favorisent cette transmission ?
- Peut-on prédire l'apparition future de l'épidémie ?

Répondre à ces questions nous permettra de proposer une approche hybride entre le modèle mathématique d'épidémie SEIR et le modèle réseau social petit monde (Small World) afin de comprendre le phénomène de transmission des virus grippaux dans les populations humaines et prédire l'évolution future de la maladie.Ce travail s'inscrit dans le cadre de l'élaboration d'un système d'information décisionnel dédié à la surveillance des épidémies où le simulateur SEIR-SW joue le rôle de moteur d'analyse en ligne (OLAP) pour simuler le phénomène de propagation d'une maladie..

L'article est structuré comme suit : La section 2 est consacrée à la présentation d'un état de l'art sur les modèles et les approches utilisées pour modéliser la propagation de l'épidémie de grippe. Le modèle proposé sera détaillé en section 3 suivi par des expérimentations présentées en section 4 et enfin nous terminons par une conclusion et quelques perceptives.

2 Etat de l'art

Dans la littérature, il existe une variété d'études en épidémiologie qui utilisent les différentes structures de réseaux afin de comprendre le phénomène de la propagation de la maladie à travers cesréseaux : les graphesaléatoires « random graphs »(Warren et al (2002);Volz(2008), Miller (2011)), le petit monde « Small-World »(Watts et Strogatz(1998), Kleinberg(2000), Vazquez (2006)) et le réseau Sans échelle « scale-free »(Pastor-Satorras et Vespignani(2001), Barthélemy et al.(2005)).Dans ce qui suit nous présentons quelques travaux de la modélisation mathématique d'épidémie de grippe:

Dans (Carrat et al., 2006), les auteurs proposent un modèle qui comporte deux niveau : *le niveau individuel*, dans lequel le risque d'infection par le virus de la grippe et la dynamique de l'excrétion virale sont simulés selon l'âge, le traitement et le statut vaccinal, le modèle mathématique étudié est le modèle SEIR, *et le niveau de communauté*, dans lequel les auteurs réalisent un réseau petit monde et modélisent les rencontres entre les individus par les graphes aléatoires Barabási-Albert. Afin de prédire la propagation d'une souche pandémique du virus de la grippe en Italie et l'impact des mesures de contrôle, les auteurs dans (Rizzo et al., 2008), ont développé un modèle déterministe SEIR avec une composante de simulation

stochastique. Les auteurs modélisent l'impact des mesures de contrôle telles que la vaccination, prophylaxie antivirale et mesures de distanciation sociale, etc. Les résultats de leurs simulations ont montré que les mesures telles que l'augmentation de la distance sociale et la fermeture des écoles, pourraient être utile pour retarder le pic épidémique et offrant ainsi plus de temps pour les vaccins qui seront produits. Dans (Basileu et al.,2010), les auteurs proposent un modèle de diffusion spatiale à base d'agents hybrides simulant la diffusion d'une pandémie, fondé sur les caractéristiques médicales de la pandémie ainsi que sur la structure socio-économique de la zone géographique concernée.Hsu et Shih (2010) ont focalisé leur étude sur la transmission de la grippe interhumaine, ils ont étudié les effets des activités de transport aérien sur une pandémie de grippe dans un réseau petit monde. Le modèle mathématique mis en place pour cette étude est le modèle SI (Susceptible–Infected). Dans une étude similaire, Yoneyamaet Krishnamoorthy(2012)ont élaboréun modèle qui tient compte à la fois le modèle SEIR basée sur les zones locales et le modèle de réseau pour la connexion globale entre les pays se référant aux données sur les voyageurs internationaux. Leurs intérêt est de reproduire la situation en utilisant les données du stade précoce de la pandémie et de prédire le futur passage en prolongeant le cycle de simulation, ils ont constaté que les résultats fournis de la simulation ont des tendances presque identiques en comparant avec la situation réelle. De leur part, Dorjee et al. (2013) ont fait une présentation des différentes méthodes et approches appliquées à la modélisation de la propagation de la grippe zoonotique chez la population humaine et animale, et un résumé sur les paramètres importants est nécessaires à la modélisation de la propagation du virus.Cette liste n'est certainement pas exhaustive et les applications concrètes des modèles mathématiques restent toujours possibles.

3 Notre approche : SEIR-SW

Dans cette partie nous présentons, en détails le modèle que nous proposons SEIR-SW (Susceptible, Exposed, Infected, Removed-Small World) pour modéliser le phénomène de propagation de la grippe dans larégion d'Oran, ce modèle est conçu par assemblage de deux principaux composantes : le modèle réseau social "petit-monde" et le modèle mathématique d'épidémie SEIR :

3.1 Le modèle réseau social "petit-monde"

Les réseaux complexes sont présents dans de nombreux domaines aussi divers les uns que les autres : biologie, sociologie, psychologie, informatique, etc. Dans cette partie, nous utilisons les réseaux complexes pour tenir compte des interactions permettant la transmission des maladies infectieuses entre les individus. De tels réseaux permettent de prédire l'issue d'une épidémie et donc d'aider à tester et à améliorer les politiques de santé publique. Un réseau social peut être défini comme : «*un ensemble de personnes ou de groupes de personnes possédant des schémas de contacts ou d'interactions entre eux* », Degenne (1994). C'est à partir de ce type de réseaux que la modélisation du monde réel a été introduite de façon empirique grâce à l'expérience de Milgram(Milgram 1967).

Dans un graphe, l'effet petit monde de Watts et Strogatz (1998), signifie que la plupart des nœuds sont connectés par un plus court chemin à travers le réseau, et qu'il y a un effet de regroupement (clustering) signifiant qu'il y a une grande probabilité pour que deux nœuds soient connectés directement à un autre s'ils ont un nœud voisin en commun. Le réseau petit monde occupe une place intermédiaire entre réseaux réguliers et aléatoires c.-à-d. il est ni

totalement aléatoire ni parfaitement régulier, la probabilité (p) de recablage (rewiring) joue un rôle important dans le passage d'un type de réseau vers un autre.

Dans ce travail, nous avons choisi le modèle petit monde (SW), car ce réseau est le plus proche de la réalité des contacts entre les individus ou les groupes d'individus. Il combine entre les deux caractéristiques d'un réseau réel : le fort coefficient de clustering et le petit diamètre. Dans le modèle de Watts-Strogatz (1998), le degré de nœud est fixe pour tous les nœuds du réseau. Or, dans la réalité chaque individu a un nombre de contacts différents d'un autre, cela nous amène à proposer une distribution de degré des nœuds suivant une loi de puissance (Lebhar 2005).Dans ce qui suit, nous formalisons l'approche de Watts-Strogatz (1998) sous forme d'algorithme décrivant les étapes de génération d'un SW.

L'algorithme de génération du réseau Small World (SW)

Entrée :*Nœud* le nombre initial de nœuds qui représente le nombre d'individus (N)
Entrée :*Degré* le nombre de degré de nœud (k)
Entrée :*Recablage* le paramètre spécial de recablage (p)
Sortie :*Réseau* Small World (G)
Debut
Copier les Nœuds dans List.Nœud
/**Création d'un réseau en anneau :Réseau.Anneau**/
Si (k<(N-1)/2) alors
Pour i ← 0 à Taille.List.Nœud faire
Pour j ← 1 à j ≤ k faire
d← (i+j)
Connecter List.Nœud [i] avec List.Nœud [d]
Fin
Fin
Fin Si
/**Recâbler les arrêtes au hasard avec une probabilité* (*p*)*/
Si Réseau.Anneau= existe alors
Pour i ← 1 à N faire
Pour j ← (i + 1) à (i+ k/2) faire
Si j > N alorsj ← j − N
Fin Si
Si p > C alors /* *C = une variable aléatoire uniforme entre 0 et 1*/*
Choisir Nœud[l] uniformément de l'ensemble des nœuds
Déconnecter: Nœud [i] et Nœud [j]
Créer un arrête Nœud [i] et Nœud [l]
Fin Si
Fin Fin Fin Si
FIN

3.2 Le modèle mathématique d'épidémie SEIR

Le modèle mathématique pour l'épidémiologie SEIR est une extension du modèle compartimental SIR (Susceptible- Infected -Removed) simple. Ce modèle est fréquemment utilisé pour étudier l'évolution d'une épidémie. Il est aussi recommandé dans le cas d'étude d'une maladie qui n'est pas caractérisée par une période de latence. En effet, dans le cas de la

grippe, le modèle SEIR est le plus approprié. Il permet de modéliser les différents statuts de la maladie et divise la population en quatre compartiments : susceptibles, exposés, infectés et retirés au cours de l'infection. Le passage d'un état vers un autre se fait selon des probabilités α, β, γ. Le processus SEIR est illustré par la figure (FIG.1).

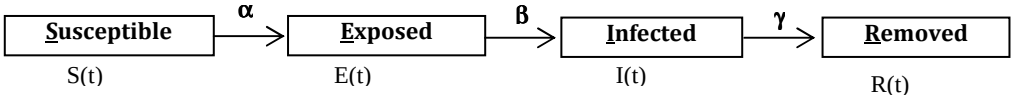


FIG. 1 – Représentation schématique de la transmission de la maladie dans le modèle SEIR

Avec : α: Taux de transmission de la maladie
β: Taux de latente ($1/\beta$: la période de latence)
γ: Taux de guérison ($1/\gamma$: la durée d'infection)

De par leur simplicité conceptuelle, les modèles compartimentaux peuvent être aisément adaptés à plusieurs situations épidémiologiques en faisant varier le nombre de catégories dans lesquelles la population est divisée (Allard, 2008).Un individu susceptible infecté par le virus de la grippe, passe par une période de latence, qui est nécessaire pour passer de l'état de contamination, à celui de contagieux. Cette période précède l'apparition des symptômes. L'individu reste infecté pendant une durée qui s'appelle la période d'infection puis il passe à la phase de retiré.

Dans la présente étude, nous optons pour le modèle SEIR car durant la saison 2009-2010, un seul type de virus grippal a circulé : les virus de type A. Aucune souche de type B n'a été détectée durant la période de surveillance de la grippe saisonnière selon le bilan annuel (INSP,2010). En effet, nous supposons que les immunisés ne redeviennent pas susceptibles car ils sont immunisés de cette souche après leurs guérison. Afin de créer notre modèle SEIR-SW, nous avons dû poser quelques hypothèses :

H1 : La taille de la population égale à N, supposée fixe ;

H2 : La variable de temps t est de type discret, tel que $t \in T$ ou T est la durée totale;

H3 : La période de temps Δt peut représenter des jours ;

H4 : A chaque instant t, la population est partitionnée en quatre classes aléatoires : Ps : ensemble d'individus susceptibles, Pe : ensemble d'individus exposés, Pi : ensemble d'individus infectés et Pr : ensemble d'individus retirés ;

H5 : Nous admettons que chaque individu susceptible dans une période Δt soit infecté puis guéri ;

H6 : Un individu infecté ne peut plus redevenir susceptible.

3.2.1 Algorithme SEIR-SW

Dans le modèle SEIR-SW, on prend en considération la notion de voisinage ainsi que la distribution de degré des nœuds qui suit une loi de puissance. Nous proposons un algorithme SEIR-SW qui se déroule comme suit:

L'algorithme SEIR-SW proposé

Entrée :*Réseau* Small World (G)

Entrée : αtaux de transmission, β taux de latente, γtaux de guérison

Sortie :*List.infecté*Liste des infectés

Début

List.infecté←ϕ

Infecter quelques Nœuds
Pour I(t)≠0 faire
Nœud.S devient Nœud.E avec une probablité 1- $(1- \alpha)^{k}$
Nœud.E devient Nœud.I après une période t_1~E $(1/\beta)$
  Ajouter Nœud.I à List.infecté
Nœud.I devient Nœud.R après une période t_2~E $(1/\gamma)$
  G évolue à G'
Fin
 Retourner List.infecté
FIN

3.2.2 Dynamique et seuil épidémique (le taux de reproduction de base)

Le taux de reproduction de base (R_0) est défini comme : « *un concept clé en épidémiologie. On le définit « heuristiquement » comme le nombre moyen de nouveaux cas d'infection, engendrés par un individu infecté moyen (au court de sa période d'infectiosité), dans une population entièrement constituée de susceptibles*» Sallet (2010). Il joue un rôle très important pour la prédiction, car il est relié par les trois paramètres qui peuvent diminuer l'évolution d'épidémie : la transmission, le nombre de contacts d'un individu et la période d'infectiosité (contagiosité).On calcule le seuil épidémique : $R0 = \beta * k * D$,avec D : période de contagiosité. Si la valeur de $R_0 < 1$: l'épidémie décroit, si $R_0 > 1$: l'épidémie s'étend.Dans notre travail, on n'utilise pas k comme une valeur fixe mais plutôt comme une distribution de degré suivant une loi de puissance pour calculer la valeur de R_0.

4 Expérimentation

4.1 Jeux de données

La base de données utilisée dans cette étude concerne la maladie de la grippe saisonnière de l'année 2009 de la région d'Oran. C'est une base de données médicale/socio-économique obtenue de la Direction de la Santé et de la Population (DSP) d'Oran. Elle est composée de 5504 enregistrements (déclarations) pour les 26 communes de la wilaya d'Oran. Les données recueillies pour chaque patient inclus son identifiant, date de déclaration de la maladie (d'Août 09 à Décembre 09), l'âge (1 mois -93 ans), le sexe (F-M), la commune de résidence et le centre de soin sanitaire de déclaration de la maladie (14 centres : EPSP, EHS, EPH, EH).

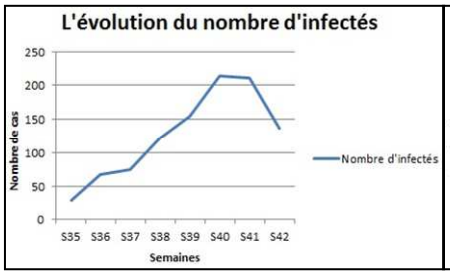


FIG. 2 – a : 1ère Vague d'épidémie

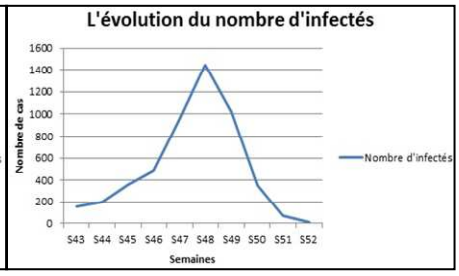


*FIG. 2 – **b** : 2ème Vague d'épidémie*

Pendant cette période la région oranaise a connu deux vagues de grippe saisonnière, la première était entre fin Août et fin d'Octobre 2009 (S35-S42) et la deuxième a commencé fin Octobre jusqu'au fin de décembre 2009 (S43-S52). La figure (FIG.2 : a-b) représente les deux vagues dans la population totale.

4.2 Protocole d'expérimentation

4.2.1 Estimation des paramètres

Plusieurs chercheurs ont fourni des estimations fiables de la durée moyenne des périodes de latence et d'infectiosité. Il n'existe pas de mesure directe permettant de disposer de la vitesse de transmission, et très peu d'estimations de la fraction des individus initialement susceptibles dans la population (Cauchemez et al., 2008). En effet, dans cette partie nous allons essayer d'estimer les paramètres non disponiblesdu modèle SEIR-SW, soit à partir des données de terrain (la base de l'épidémie de grippe survenue à Oran en 2009) soit par des méthodes statistiques comme la méthode maximum de la vraisemblance.

Présentation des paramètres du modèle SEIR-SW : Notre modèle dépend des huit paramètres suivants:

- *Relatifs à la population*: N, I_0

N = nombre total de la population supposée susceptible, I_0 = nombre initialement infecté

- *Relatifs au réseau* SW : k, p

k : degré d'un nœud, p : probabilité de recablage « rewiring »

- *Relatifs à la maladie* : α, β, γ

α : La probabilité infection, β : La probabilité de latente, γ : La probabilité de guérison

Pour mieux comprendre le phénomène de propagation d'épidémie, nous avons divisé la population en quatre tranches d'âges selon leurs activités. Dans le modèle SEIR-SW il y'a huit paramètres inconnus. Parmi ces paramètres, la durée de latence et la durée d'infectiosité peuvent être obtenues à partir des caractéristiques du virus de la grippe (Cori et al., 2008). D'autres paramètres doivent être estimés à partir des méthodes estimation ou à partir de la base de données réelle. Les valeurs estimées des paramètres pour notre modèle SEIR-SW sont donnés par le tableau suivant :

Paramètres	**Valeurs pour vague 1**				**Valeurs pour vague 2**				**source**
Age	0-5	6-17	18-59	60+	0-5	6-17	18-59	60+	Base de données
N	260	170	670	510	750	1850	3200	470	Estimé
I_0	2	1	1	1	52	21	75	25	Base de données
k	6 - 8	4-8	2- 8	2-8	2-6	2-8	2-8	2-8	Estimé
P	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	Estimé
α	+[1]	+	+	+	+	+	+	+	Estimé
Période de latence	2	1	1	2	2	1	1	2	Etude de Cori et al. (2012)
Période d'infectiosité	3	3	1-4	3	1-3	2-3	1-3	1-3	Etude de Cori et al.(2012)

TAB. 1 –Les valeurs des paramètres du modèle SEIR-SW

[1] + : Valeurs possible de α

4.3 **Résultats et interprétation**

Nous avons effectué un certain nombre de simulations avec les paramètres représentés dans le tableau ci-dessus, les résultats obtenus sont représentés dans la figure (FIG.3). Nous-savons présenté des graphes d'évolution de la grippe dans les deux vagues : la période entre S35 et S42 et la période entre S 43 et S 52 de l'année 2009, ces graphes représentent les données simulées et les données réelles. A partir de ces graphes nous remarquons que l'écart entre les deux graphes est petit, cela signifie que les paramètres choisis sont proches des paramètres réels et de la structure sociale d'individus.

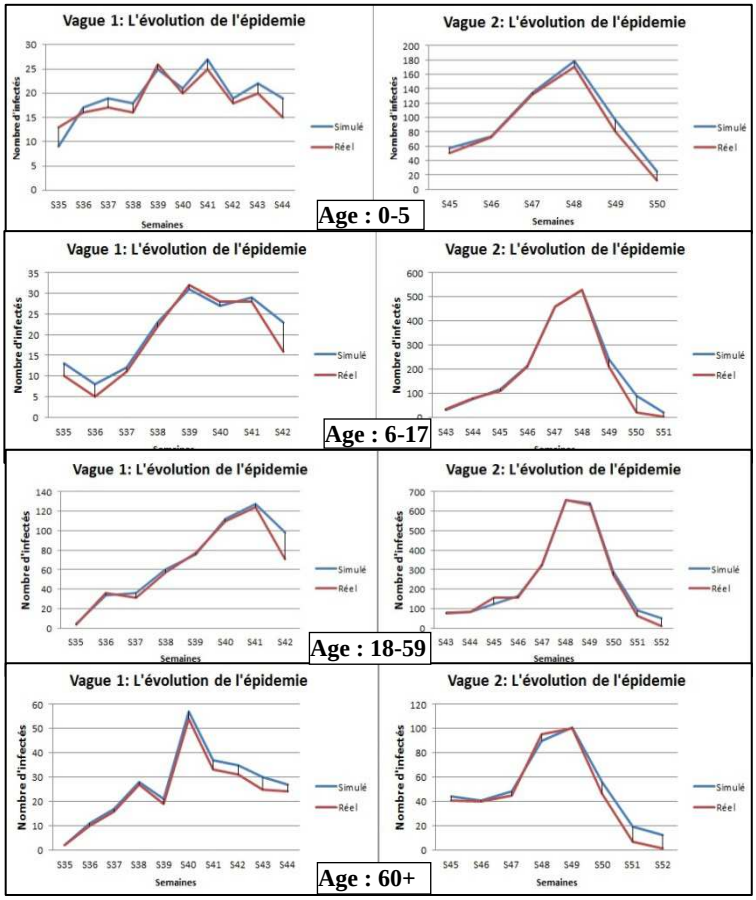


FIG. 3 – L'évolution de l'épidémie pour les données simulées et les données réelles pendant les deux vagues pour les 4 tranches d'âge.

Afin de mieux comprendre les facteurs qui favorisent la transmission et prévoir l'issue finale d'une épidémie, nous avons effectué quelques études sur le taux de transmission, le degré de contact entre individus (k) et la période de contagiosité par rapport aux tranches d'âges d'individus pour les deux vagues. La figure (FIG.4), représente la variation du taux de transmission par rapport à l'âge d'individus pour les deux vagues.

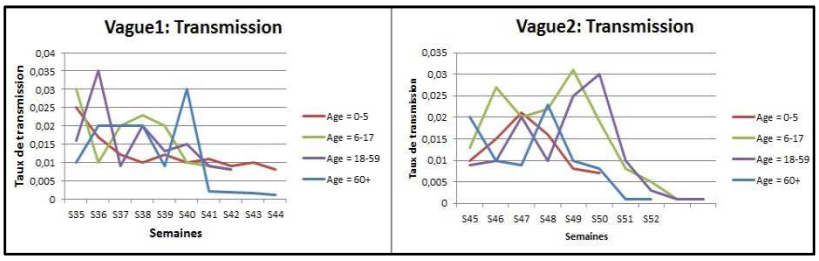


FIG. 4 – La variation du taux de transmission pendant les deux vagues pour les 4 tranches d'âge

À partir des résultats obtenus, nous constatons que le taux de transmission varie d'une tranche d'âge à l'autre et d'une vague à l'autre. Pour la vague 1, il existe moins de pics par rapport à la vague 2 ainsi qu'un taux de transmission plus faible. Dans la vague1, toutes les tranches d'âges ont presque les mêmes valeurs autour de 0.01 et 0.025. La transmission s'est arrêtée dans la semaine S 42 pour les tranches d'âges 6-17 et 18-59 et poursuit pour les tranches d'âges 0-5 et 60+ suite à leurs sensibilités devant la maladie (personnes fragiles). La transmission n'a pas beaucoup diminué ce qui indique une apparition future possible de la grippe**.** Dans la vague 2, la tranche d'âges 6-18 a eu un taux de transmission peu important par rapport aux autres. La transmission a diminué dans les quatre tranches d'âges ce qui indique la fin de l'épidémie. Un autre facteur qui peut aussi influencer la transmission de la maladie : le nombre de voisins pour chaque individu. La figure (FIG.5), représente la variation du nombre de degré k dans chaque tranche d'âges durant la période de la grippe.

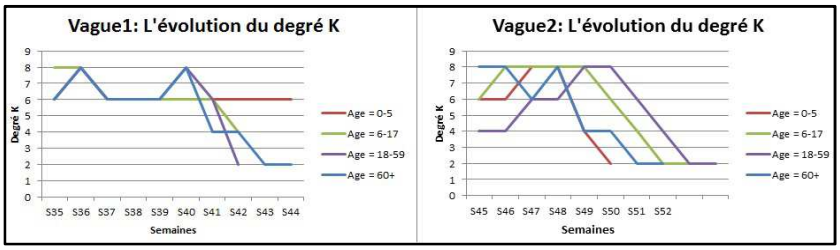


FIG. 5 – La variation du degré K pendant les deux vagues pour les 4 tranches d'âge

La différence dans la variation du nombre de contacts entre les deux graphes est très claire. Dans la 1ère vague : le nombre de contacts est fixe pour la tranche d'âge 0-5 ans, la majorité de cette tranche reste chez elle et n'exerce aucune activité. La tranche d'âge 6-18 ans, le nombre de contacts varie entre 4-8, cette tranche représente les élèves des écoles, du fondamental et des lyciens. Le nombre de contacts varie entre 2-8 pour les deux tranches d'âge 18-59 ans et 60+ ans. La première est la tranche la plus dynamique (employés, étudiants, privés, etc.) et la deuxième représente une catégorie de personnes d'où une partie exerce toujours son activité et une autre se rencontre dans les cafétérias ou dans les lieux publics. Pour la vague 2 : la variation dans le nombre de contacts reste presque comme la vague2, sauf à partir de la semaine S50, une diminution bien claire ce qui est expliqué par les vacances d'hiver. La période de contagiosité joue un rôle primordial dans la propagation d'épidémie, plus la durée est grande plus la maladie reste plus longtemps dans la population.

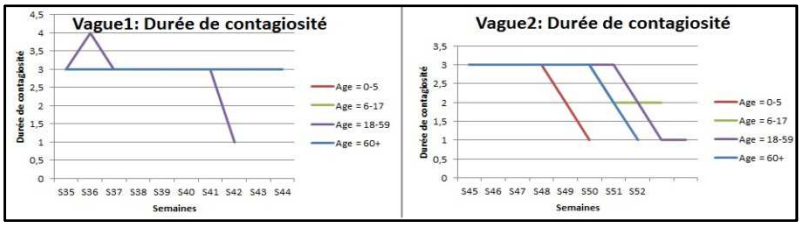


FIG. 6 – La variation de la durée de contagiosité pendant les deux vagues pour les 4 tranches d'âge

Selon les résultats obtenusreprésentés dans la figure (FIG.6), dans les deux vagues, la durée de contagiosité est presque stable pour les tranches d'âge 0-5 et 60+ car ce sont des personnes sensibles à la maladie et leurs anticorps sont plus faibles par rapport aux autres tranches d'âges. Les autres tranches d'âges 6-17 et 18-59 ans avaient une période de contagiosité de 3 jours en moyenne. Ce sont des personnes qui ont une certaine défense contre le virus, nous remarquons une diminution dans cette période dans les deux dernières semaines de l'année 2009, suite aux vacances d'hiver où la majorité des personnes reste chez elle.

Nous rappelons ici, que la période de contagiosité et le taux de transmission ont une relation directe avec le type de virus, sa structure et ses caractéristiques.

4.4 **Prédiction**

L'éradication d'une infection par la vaccination peut être comprise en termes de réduction des sujets susceptibles d'être infectés en dessous d'un seuil. Cet effet est appelé «immunité collective» puisque la population peut être protégée contre les épidémies, même si il y a quelques susceptibles dans la population. Ainsi, l'élimination est théoriquement possible avec un taux de vaccination important. Nous avons estimé l'indicateur du risque épidémique R_0. Ce calcul est effectué pour quelques semaines. Le tableau suivant récapitule les résultats obtenus pour l'ensemble de la population.

Semaines	**Nombre (k)**	**Période d'infection**	**Taux de transmission**	**R_0**
S46	6	3	0.056	0.79
S47	8	3	0.07	1.37
S 51	2	2	0.01	0.06

TAB. 2 –Représentation du taux de reproduction de base avec ses paramètres par rapport aux semaines

A partir des résultats du tableau (TAB.2), nous constatons que si le nombre de contacts augmente, le taux de transmission augmente aussi. De même, le taux de reproduction de base R_0a augmenté. Dans la S46 nous remarquons un risque d'épidémie (R_0=0.79) et dans S47 une présence d'épidémie (R0>1). Dans ce cas, nous pouvons prévenir une évolution future de l'épidémie de la grippe (voir FIG.3). Alors que dans la semaine S51, le taux de reproduction de base a diminué à 0.06. Cela suppose que c'est la fin de l'épidémie : ce qui est confirmé dans la figure (FIG.3).Comme mesure de prévention, les responsables sanitaires peuvent imaginer quelques scénarios : si une population est atteinte d'une maladie, la transmission est élevée et la période de contagiosité est élevée dans cette population il est recommandé

d'identifier les solutions de contrôle les plus efficaces (traitement, quarantaine, vaccinations, etc.) afin de réduire son évolution. Aussi dans le cas où le nombre de contacts est élevé de cette population, il est recommandé de fermer les écoles et les lieux de rencontre.

5. Conclusion

Les maladies infectieuses transmissibles telle que la grippe et la tuberculose sont de complexités multiples, d'une part la complexité de la maladie elle-même(structure et caractéristiques du virus, les modes de transmission de la maladie) et, d'autres part les structures sociales d'individus, les facteurs socio-économiques, démographiques, facilitant la transmission. En réponse à la problématique posée dans cette étude, nous avons tenté de décrire un modèle pour la propagation de la grippe au sein de la population oranaise.Ce modèle combine, principalement, deux modèles: le modèle mathématique d'épidémie SEIR et le modèle réseaux social petit monde. Les résultats fournis par le modèle SEIR-SW sont satisfaisants selon deux paramètres de performances les plus pertinents, à savoir: la compréhension du phénomène de propagation ainsi que la prévision et la représentativité des situations réelles.Dans les travaux futurs, il sera intéressant de créer un entrepôt de données médicale et d'intégrer le modèle de propagation SEIR-SW afin d'améliorer les fonctionnalités de l'analyse multidimensionnelle classique dans l'objectif de concevoir et d'élaborer un système d'information décisionnel dédié à la surveillance épidémiologique.

Références

Allard, A. (2008). *Modélisation mathématique en épidémiologie par réseaux de contacts*.Mémoire de maîtrise en Physique.L'Université Laval.

Anderson, R-M. et R-M. May (1991). *Infectious diseases of humans: dynamics and control*. 757 pages.Oxford University Press. New York.

Barthélemy, M., A. Barrat,R.Pastor-Satorraset A. Vespignani (2005). *Dynamical patterns of epidemic outbreaks in complex heterogeneous networks*. Journal of Theoretical Biology, 235: 275–288

Basileu, C., A.Bounekkar, N.Kabachi et M.Lamure (2010). *Vers un modèle de diffusion spatiale d'une pandémie*. 4*eme* Colloque international de Veille Stratégique Scientifique et Technologique (VSST 2010), Toulouse, France

Carrat, F., J.Luong, H.Lao, A.Salle, C. Lajaunieet H-A.Wackernagel (2006). *Small-world-like model for comparing interventions aimed at preventing and controlling influenza pandemics*. BMC Med 2006, 4-26.

Cauchemez, S., A-J. Valleron, P-Y. Boelle, A.Flahault et N-M. Ferguson (2008). *Estimating the impact of school closure on influenza transmissionfrom Sentinel data.*Nature 2008,754 :452-750.

Coria A.,A.J.Valleron, F.Carrat, G.Scalia Tomba, G. Thomas et P-Y Boëlle (2012). *Estimating influenza latency and infectious period durations using viral excretion data*.Epidemics, 4: 132–138

Degenne,A.,M.Forsé (1994).*Les réseaux sociaux*. Armand Colin, Paris.

Dorjee S., Z.Poljak, C-W Revie, J. Bridgland, B. McNab, E. Leger, J. Sanchez (2013). *A Review of Simulation Modelling Approaches Used for the Spread of Zoonotic Influenza Viruses in Animal and Human Populations*.Zoonoses and Public Health, 60:383–411

Hsu, C-I., et H-H. Shih (2010). *Transmission and control of an emerging influenza pandemic in a small-world airline network*.Accident Analysis and Prevention, 42 : 93–100.

INSP (2010). *Bilan de la saison (2009-2010) du réseau sentinelle algérien de surveillance de la grippe saisonnière*, Institut National de Santé Publique, Algérie

Kleinberg, J. (2000).*The Small-World Phenomenon : An Algorithmic Perspective*.In Proceedings of the 32nd ACM Symposium on Theory ofComputing (STOC), 163–170.

Lebhar, E. (2005). *Algorithmes de routage et modèles aléatoires pour les graphes petits mondes.* Thèse de doctorat, Ecole Normale Supérieure de Lyon.

May, R-M. (1979). *Population biology of infectious diseases: part II*. Nature 280: 455–461.

Milgram, S. (1967).*The small-word problem.* Psychology Today, 2: 60-67.

Miller, J-C. (2011). *A note on a paper by Erik Volz: SIR dynamics in random networks.* Journal of Mathematical Biology, 62:349–358.

Morens, D.M., G.K. Folkers, A.S. Fauci (2004).*The challenge of emerging and re-emerging infectious diseases*. Nature, 430: 242–249.

Pastor-Satorras, R. etA. Vespignani (2001). *Epidemic spreading in scale-free networks*. Physical Review Letters, 86:3200–3203

Rizzo C., A. Lunelli, A. Pugliese, A. Bella, P. Manfredi, G.Scalia-Tomba, M.Iannelli, M-L. CiofiDegliAtti(2008). *Scenarios of diffusion and control of an influenza pandemic in Italy, Epidemiol Infect*. 136(12): 1650–1657.

Sallet,G.(2010). R_0. EPICASA09 (INRIA & IRD)

Vazquez,A. (2006).*Spreading dynamics on small-world networks with connectivity fluctuations and correlations*, PHYSICAL REVIEW, 74, 056101.

Volz, E. (2008). *SIR dynamics in random networks with heterogeneous connectivity*. Journal of Mathematical Biology, 56: 293–310.

Warren C. P., L-M.Sander, I. Sokolov, C. Simon, et J. Koopman (2002). *Percolation on disordered networks as a model for epidemics*. Mathematical Biosciences,18:293–305.

Watts, D-J.etS-H. Strogatz (1998).*Collective dynamics of 'small-world' networks*. Nature,PubMed PMID: 9623998.

Yoneyama., T. et S. Krishnamoorthy (2012). *Simulating the spread of influenza pandemic of 2009 considering international traffic.* Simulation 88:437-449

Towards ontology building and updating from Big Data

Hanen Abbes, Faiez Gargouri
Miracl Laboratory
Higher Institute of Computer Science and Multimedia, Sfax University, Tunisia
Abbes.hanen@gmail.com, faiez.gargouri@isimsf.rnu.tn

Abstract. The main challenge of Big Data is to extract values from structured and unstructured data. Thus, we have to factorize data into knowledge. Ontologies represent knowledge as a formal description of a domain of interest. But we need an intermediate representation such as NOSQL database. As Big Data is always updated, updates must be reflected in the ontology.

1 Introduction

Handling the relevant data in Big Data scenarios is increasingly difficult due to their volume, variety, and velocity. The main challenge of Big Data is to benefit from structured and unstructured data. To deal with their complexity, we have to factorize data into knowledge. Ontologies seem to be a solution as a set of concepts about a domain, using a shared vocabulary to point out these concepts and their relationships. Thus, building ontology from Big Data is very interesting. Besides, Big Data are always updated due to the permanent arrival of new data. This must be passed to the ontology. In this paper we propose a methodology of building ontology from Big Data and we explain how to manage its updates. This paper is organized as follows. In section 2, we present Big Data and ontologies. Section 3 explains research context and related works. In section 4, we compare the different types of NOSQL databases. Our main ideas are proposed in section 5. Finally, section 6 concludes the paper and presents future works.

2 Big Data and Ontologies: State of Art

According to Gartner, "Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making"[1].

As for the McKinsey Global Institute, "Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze"[2].

[1]http://www.gartner.com/it-glossary/big-data/

IDC defines Big Data technologies as new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis3.

All definitions in the literature agree about Volume, Variety and Velocity. These are the three main characteristics of Big Data known as 3Vs.

According to Gruber (1993), ontology is a specification of a shared conceptualization of a domain. Its main purpose is to capture knowledge about a specific domain and to provide common accepted representation for re-use and share. In general, ontologies aim at adding semantics to raw data, solving heterogeneity problems, allowing reasoning and inferences.

In the context of Big Data, these purposes remain persistent with conjunction to other ones as mentioned in Hashemi and al. (2012):

-help people better understand and disentangle the complexity of big engineered systems and their social, economic, and natural environment,
-enable integration among systems and data through semantic interoperability,
-improve models and modeling, their adaptability and reuse, and resulting design,
-reduce development and operational costs,
-enhance decision support systems.

3 Motivations and Related Works

In the context of Big Data, Data comes mainly in two forms: structured and unstructured data. Structured data has semantic meaning attached to it whereas unstructured data has no latent meaning. The main challenge of Big Data is to benefit from both structured and unstructured data. Thus, the question is how to facilitate information retrieval from these data qualified as voluminous, varied and of high speed knowing that traditional tools of data management do not have required performances to deal with these data. So, we have to factorize these data into knowledge. Ontologies represent knowledge as a formal description of a domain of interest.

In the literature, we distinguish some research works that link ontologies to Big Data.

The Optique project Calvanese and al. (2013) advocates a next generation of the Ontology-Based Data Access (OBDA) approach to address the Big Data dimensions and in

[2]http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

[3]http://www.emc.com/leadership/digital-universe/iview/big-data-2020.htm

particular the data access problem. Authors propose an architecture to overcome the limitations of a classical OBDA system:

-*A query formulation component* aims at providing a friendly interface for nontechnical users.

-*An ontology and mapping management component* bootstraps an initial ontology and mappings and maintains the consistency between the evolving mappings and the evolving ontology.

-*A query answering component* is decomposed into several layers. The transformation layer transforms the formulated queries into executable and optimized queries with respect to the data sources. The planning and execution layers distribute queries to individual servers.

In Hoppe (2013), the author proposed an approach to automatic ontology based user profile learning from heterogeneous Web Resources in a Big Data Context. She developed, in collaboration with an industrial partner, an ontology that models the informational context that it works in. It captures entities directly touched by the analysis process, and concepts that stem from the commercial context of the enterprise

These two research works use domain ontologies for Big Data independently from data sources and structures. To the best of our knowledge, there isn't any work that deals with ontology building from Big Data. This idea seems to be innovative and motivates our thoughts. For this purpose, we judge important to represent structured and unstructured data in a same data store to facilitate the process. NOSQL databases are the suitable solution.

In the next section, we present different types of NOSQL databases.

4 NOSQL Databases

NOSQL[4] databases are next generation databases that are non-relational, distributed, open-source and horizontally scalable**.** The main reasons to use them are their ability to handle semi-structured and unstructured data and their horizontal scalability.

There are four types of NOSQL databases. In Moniruzzaman and al. (2013) and Nayak and al. (2013), the authors draw a comparison between these databases as follows:

-*Key/value stores* store items as alpha-numeric identifiers (keys) and associated values in simple, standalone tables. Requests can only be performed against keys, not values.

[4]http://nosql-database.org/

-*Column stores* do not store data in tables but in massively distributed architectures. In column stores, each key is associated with one or more attributes (columns). The data stored is based on the sort order of the column family.

-*Document databases* are designed to manage and store documents encoded in a standard data exchange format such as XML. Unlike the simple key-value stores, the value column contains semi-structured data – specifically attribute/value pairs. The number and type of recorded attributes can vary from row to row. Both keys and values are searchable.
Document databases are good for storing and managing Big Data-size collections of literal documents, like text documents, email messages, and XML documents.

-*Graph databases* replace relational tables with structured relational graphs of interconnected key-value pairings. The graphs are represented as an object-oriented network of conceptual objects (nodes), relationships (edges) and properties (object attributes expressed as key-value pairs).

We believe document oriented databases are suitable to initiate ontology building process as scalable, of high-performance and open source databases. They store structured, semi-structured and unstructured data. They are schema-less, very flexible and highly scalable.

5 How to build and update ontology from Big Data? Main ideas

As we mentioned above, the main objective of Big Data is to deal with structured and unstructured data in order to extract values. But dealing with huge volumes of data that are much diversified, complex and with high speed is not easy at all. This requires a pre-processing step to group these data into a store regrouping together heterogeneous data. This store is a NOSQL database. It will serve as a starter for the ontology building process and presents an intermediate representation between the raw data and the ontology. The next step is to build core ontology. It consists on defining rules to extract ontology's concepts, relations, roles, domains and ranges from NOSQL database specification. It will describe mappings of database components to ontology.

We notice that Big Data are dynamic by nature due to the permanent arrival of new data, so they are subject to different updates. New sources of data may also appear. These updates must be sent up to the ontology. Indeed, data about a domain may change in many different manners. On the one hand, new data may appear, and this leads to establish new concepts. These updates result on technological advances. On the other hand, some data may become obsolete, and so, some concepts must be removed from the ontology. Moreover, modeling a domain may necessitate concepts redefinition. New concepts that are more specific than the pre-existent ones can be defined if the domain needs to be more precise, or more abstract if

we want to simplify the domain and facilitate its comprehension. As an example, we consider the domain of alive species. An ontology that represents this domain may contain a concept “vegetable” and a concept “animal”. If we need to have a more precise description of the domain, we may add new concepts such as “mammal” and “bird” that specialize the concept “animal”. This example illustrates that the ontology may be updated whether the represented domain evolve or not. Indeed, the concepts “vegetable” and “bird” are part of the domain of alive species even before the ontology building Pruski (2009).

6 Conclusion and future Works

This paper presents our PhD project envisaged work. We introduced Big Data and ontologies and we presented different types of NOSQL databases. The intent is to develop a novel approach for ontology building from Big Data, based on NOSQL databases. We are convinced that Document Oriented Databases are the suitable ones.

As future works, we envisage to automatically feed NOSQL database, to define rules allowing to map NOSQL database components to ontology and to propose ontology’s update approach taking into account Big Data specificities.

References

Calvanese, D., M.Giese, P. Haase, I, Horrocks, T. Hubauer, Y. E. Ioannidis, E. Jiménez-Ruiz, E. Kharlamov, H. Kllapi, J. W. Kluwer, M. Koubarakis, S. Lamparter, R. Moller, C. Neuendtadt, T. Nordtveit, O. L. Ozçep, M. Rodriguez-Muro, M. Roshchin, D. F. Savo, M.Schmidt, A. Soylu, A. Waaler, and D. Zheleznyakov (2013). *Optique: OBDA Solution for Bid Data*. In *ESWC (Satellite Events)*, 2013, pp. 293-295.

Gruber, T. R. (1993).*A translation approach to portable ontology specifications.*Knowledge Acquisition. 5(2):199_220.

Hashemi, A. andT.Schneider (2012).*Ontology Summit 2012 Communique - v1.01 Ontology for Big Systems*.

Hoppe, A. (2013). *Automatic ontology-based User Profile Learning from heterogeneous Web Resources in a Big Data Context.* The 39th International Conference on Very Large Data Bases, August 26th 30th 2013, Riva del Garda, Trento, Italy. Proceedings of the VLDB Endowment, Vol. 6, No. 12.

Moniruzzaman,A. B. M., Hossain, S.A. (2013). *NoSQL Database: New Era of Databases for Big data Analytics - Classification, Characteristics and Comparison*. International Journal of Database Theory and Application Vol. 6, No. 4.2013.

Nayak, M., A. Poriya, D. Poojary (2013).*Type of NOSQL Databases and its Comparison with Relational Databases*. International Journal of Applied Information Systems (IJAIS) – ISSN: 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 5– No.4, March 2013 – www.ijais.org.

Pruski, C. (2009).*Une approche adapatative pour la recherche d'information sur le Web.* Phd thesis, April, 2009.

Résumé

Le principal défi des Big Data est d'exploiter les données structurées et non structurées pour faciliter la recherche d'information, sachant que les outils classiques de gestion de données n'ont pas les performances requises pour traiter ces volumes de données. D'où l'intérêt de construire une ontologie à partir de Big Data pour factoriser les données en connaissances. Ceci se fait en considérant une représentation intermédiaire qui est une base de données NOSQL. Dans ce papier, nous présentons les ontologies, les Big Data, et les bases de données NOSQL, nous proposons une démarche pour la construction d'une ontologie à partir de Big Data et nous expliquons la nécessité de la mettre à jour.

Towards ontology-based clustering of handicraft women

Rania Yangui *, Maha Maalej**
Achraf Mtibaa**, Ahlem Nabli***, Mohamed Mhiri***, Faiez Gargouri*

* Institute of Computer Science and Multimedia, Sfax, BP 1030 - Tunisia
yangui.rania@gmail.com, faiez.gargouri@isimsf.rnu.tn

** National School of Electronics and Telecommunications Sfax
maha.maalej@gmail.com, achrafmtibaa@gmail.com

***Faculty of Sciences, Sfax, BP 1171 – Tunisia
ahlem.nabli@fsegs.rnu.tn, med.mhiri@gmail.com

Abstract.
It is obviously that women participate in economic activities in whether developed or emerged countries. However, the majority of these women are confronted with different problems such as capital deficiency, price and marketing constraints, unavailability of raw materials, lack of knowledge on modern technologies, etc. These women can improve their socio-economic levels by using new technologies. Our aim, in this paper, is to represent knowledge about handicraft women through ontology. We proceed then, to cluster handicraft women file and production features.

1 Introduction

Women constitute more or less half of the population. Some of these women are involved in craft activities for survival. However, the majority of these women are confronted with different problems like capital deficiency, price and marketing constraints, unavailability of raw materials, lack of knowledge on modern technologies, etc. (Kartik et al., 2010). The aim of our work is to develop the socio-economic situation of handicraft woman. However, what is obvious for developed countries may not be so for emerging countries. For instance, technologies are not disseminated and technical infrastructures such as networks do not cover the whole country. Moreover, handicraft women can be illiterates. To enable the largest number of handicraft women to benefit from information technologies IT (Web, Social network, etc.) these technologies have to be adapted to the user's profile. Our principal aim is to cluster handicraft women in order to define for each cluster the affordable solution. The clustering is based on knowledge represented in an ontology. This knowledge is extracted from a set of interviews made on real cases of handicraft women from both Tunisian and Algerian countries.

In the literature, some clustering techniques have been used for the grouping of object described by knowledge contained in the ontology. For example, (Maedche et al., 2002) proposed an approach for grouping ontology instances to help users to work in cooperation. They used a set of similarity measures to compare ontology instances and then used these measures in a hierarchical clustering algorithm. In the same context, (Esposito et al., 2007) and (Mikroyannidi et al., 2011) proposed methods for grouping semantically annotated resources in order to facilitate the ontology management. These methods define similarity

measures which are totally dependent on the concepts semantic aspects in the ontology. The grouping of individuals is then realized by hierarchical clustering. In this paper, we will base on (Maedche et al., 2002) aproach. However, our contribution will appear in the selection of the relevant features and in the taking into account the artisan domain specificity.

The remainder of this paper is organized as follows. Section 2 presents the project aims. Section 3 introduces knowledge representation and section 4 exposes knowledge clustering. Finally, Section 5 gives a conclusion and future research directions.

2 Project aims

This work is involved in the BWEC[1] (Business for Women of Emerging Countries) project that treats women from rural and urban areas of Tunisian and Algerian countries. These women are involved in handicraft work. This project aims at improving the socio-economic situation of these women by bringing them information about supplier, customer and production process. To do that, we have made a questionnaire to collect information about these women and their productions. This questionnaire is made by a sociologist and covers five main topics: craft production nature, production process, the use of coordination tools, the latent needs and socio-demographic data.

The steps proposed to accomplish the purpose of this project are summarized in *FIG. 1*.

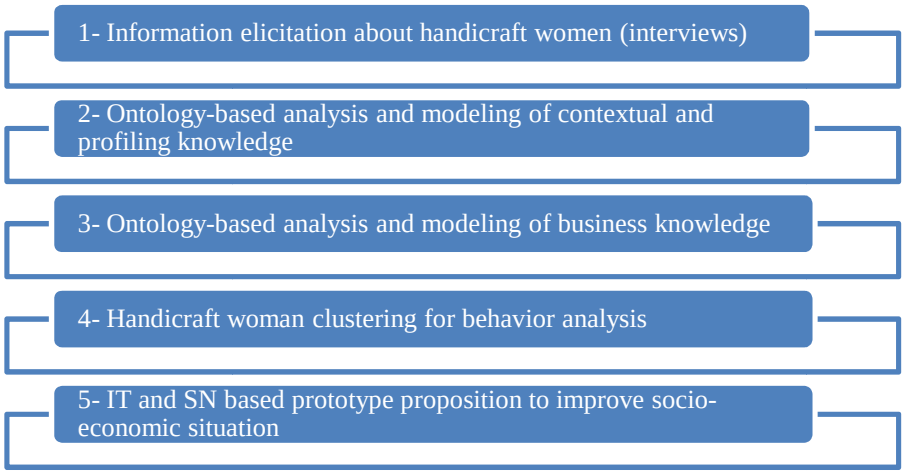


FIG.1 – Project aims

[1]Towards a new Manner to use Affordable Technologies and Social Networks to Improve Business for Women in Emerging Countries

First, a set of interviews based on the questionnaire are realized with handicraft women in many areas of Tunisian and Algerian countries. Besides, relevant concepts are extracted from interviews and represented through ontology. These concepts represent the profiling and contextual knowledge. Ontological analysis and modeling of business knowledge will be realized in the third step. Handicraft woman clustering for behavior analysis is concerned in the forth step. The final step consists on proposing innovative solutions to handicraft women in an appropriate manner to study the use of the affordable proposed solutions.

In this paper we present two important steps which are ontology-based knowledge representation which is part-of the second step in Fig.1 and knowledge clustering which is part-of step 4 in Fig 1. The first step consists on representing knowledge about handicraft women using ontology. In fact, ontology is used to organize knowledge in a structured manner. Besides, it allows reasoning with a high level of abstraction, offers interpretations related to the application domain and avoid ambiguities. The second step corresponds to knowledge clustering. In this step, we suggest to cluster handicraft women into homogeneous groups. The result of this clustering can imply the requirements of new kinds of affordable technologies fitted to each cluster. Other steps are current works and they are not mentioned in this paper. At the next sections we overview Knowledge representation and Knowledge clustering steps.

3 Knowledge representation

Interviews contain a lot of personal information about women (name, age, etc.) as well as information about their work (production tools, raw materials, etc.), their readiness to use new technologies, their interests, etc. After collecting information about handicraft women by means of interviews, we proceed to extract knowledge and represent it using ontology. An ontology is defined by Gruber as "A set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes, attributes, and relationships" (Gruber 2008).

In this step we present the relevant concepts representation. Then we determine the relationships between them. In the last step, we populate the ontology with instances already refined.

Ontology modeling: As the ontology is mainly used in representing knowledge, we create an ontology to structure the relevant concepts extracted from the interviews as shown in *FIG. 2*. Concepts represented in the ontology are ***Handicraft_woman, Production_tool, Profile, Product, Raw_material, Quality, Coordination_tool*** and ***Skills*** with sub-concepts ***Technical_knowledge*** and ***Business_knowledge***. The ***handicraft woman*** concept has the central role in this ontology. It has relationships with six concepts. The ***handicraft woman*** is related to ***Raw_material*** by “buy_raw_m” relationship. Besides this concept has relationship with ***Production_tool*** by “use_tool” relationship. “admit_skill” relationship links ***handicraft woman*** concept with ***Skills*** concept. Moreover, ***handicraft woman*** concept has relationships with ***Coordination_tool*** by “use_means_coord” relationship, with ***Product*** by means of “create” relationship and with ***Profile*** by means of “admit_profile” relationship. ***Production_tool*** is related to ***Quality*** by “have_quality” relationship. The concept ***Product*** has “need_means” relationship with ***Coordination_tool***. The concept ***Product*** has a reflexive relationship which is “composed_of_product”.

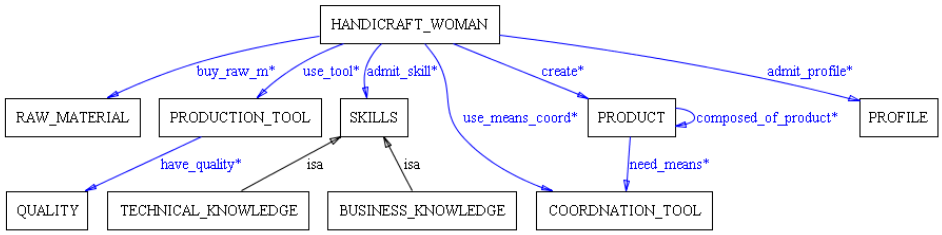


FIG. 2– Relationships between ontology concepts

Once concepts, relationships and axioms are defined, we pass to ontology populating. In order to correctly populate the ontology, there is a need for instances refinement.

Instances refinement and populating: As input to this sub-step, we have semi-structured interviews described in natural language. Based on these interviews, we notice that collected data are incomplete (i.e. there are missing values), noisy (i.e. there are redundancy) and not all understandable (i.e. some data are semantically ambiguous) which poses new challenges for knowledge extraction. To overcome some of these problems we proceed as follows:

- Detection of ambiguities and Determination of missing values: When the answer to a question is not understandable or missing, we propose, as solution, to use the links between concepts and instances in the ontology to detect ambiguous or missing data. For instance, we can detect the Production_Tool instances which are relative to a given Product. If the missing or ambiguous values are not found in the ontology, we can use the ONAT website by selecting the right tools name. For instance, we can use the ONAT website to detect which is the used feedstock in a production.

- Standardization of vocabulary: In fact, terms may give different interpretations within different user profiles. For example, terms “margoume” or “klim” having the same signification in Tunisian dialect lexicon are replaced by “traditional carpet”. Another example which is related to the age as “50 years old” and “born in 1963” are replaced by “50”. In this case, we can define semantic relationships (equivalence or synonymy) between the two instances in the ontology.

- Elimination of redundancy: When we find two questions that treat the same information, we proceed to eliminate one of them. For instance, the response of the question “Is your production rooted in a territorial or family tradition?” can be deduced from the response to the question “How did you acquire this skill?”. Another illustration of this case is represented by these two redundant questions: “What are the different stages of transformation?” and “Could you describe in detail each step of the process?”. In fact the ontology must not have redundant instances.

After refinement step, we have instantiated the ontology with instances from 80 treated interviews. This step consists mainly in filling the concepts individuals.

Once the ontology is populated, we proceed to cluster the women based on relevant features. This step is detailed in the following section.

4 Knowledge clustering

Recall that our aim in this step is to cluster handcraft women based on knowledge contained in the ontology. To do that, we propose to use the clustering data mining technique. Clustering aims at gathering the data set objects into relatively homogeneous groups called clusters. Objects in each cluster tend to be similar to each other and dissimilar to objects in the other clusters. It is appropriate for applications where labeled data are hard to find. Many algorithms can be used for clustering. In this paper, we will start by experimenting hierarchical and K-means algorithms then we proceed for other kinds of algorithm. We start by selecting features then we generate clusters according to selected features and we achieve by giving an analysis for the result.

Feature selection: Feature selection consists on the choice of the relevant knowledge contained in the ontology, as the basis for grouping handicraft women:

- ***WWW-Use****:* this criterion means if women use already new technologies. In fact, this criterion is important to detect if women need the use of the appropriate technology or need other type of solution.
- ***Intellectual level***: this criterion represents the women school level. In fact, it plays an important role in the selection of the adequate technology.
- ***Age***: this criterion corresponds to the exact or approximate age of women. It reflects the women's capacity of understanding. In fact, more women are older more it is difficult to teach them how to use new technologies.

In the next sub-section, we will apply the two chosen clustering algorithms based on the selected features.

Clusters generation: Clusters generated by "Hierarchical Clustering" algorithm are presented in the following table *TAB.1*:

	Hierarchical Clustering			
	Nombre d'instances	WWW-Use	Intellectual Level	Age
Cluster 0	16 (20%)	No	Illiterate	40-60
Cluster 1	33 (41%)	Yes	University	<40
Cluster 2	31 (39%)	No	Secondary	<40

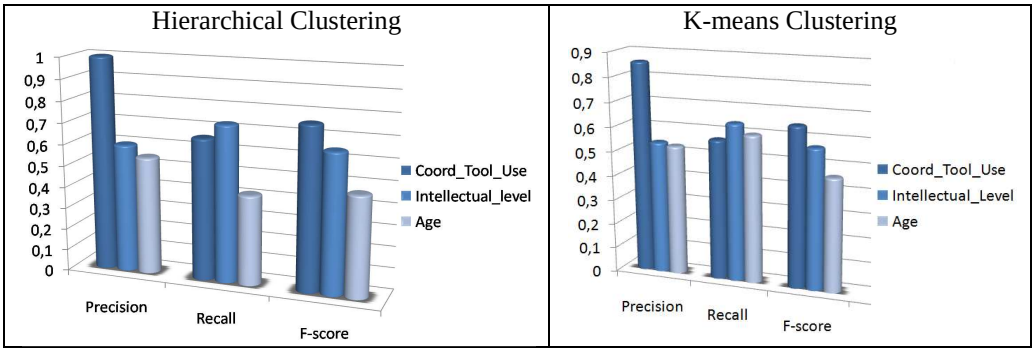
TAB.1– Hierarchical Clustering results

Clusters generated by "K-means" algorithm are presented in the next table *TAB.2*:

	K-means Clustering			
	Nombre d'instances	WWW-Use	Intellectual Level	Age
Cluster 0	33 (46%)	Yes	University	<40
Cluster 1	16 (22%)	No	Illiterate	>60
Cluster 2	23 (32%)	No	Secondary	40-60

TAB.2– K-means Clustering results

In order to evaluate the performance of the used clustering algorithms, we calculated three measures of quality widely used: the precision (P) the recall (R) and the F-score (F). Averages of each of these three measures are presented.



TAB.3. Hierarchical and K-means clustering effectiveness.

Based on F-sore values, we can conclude that, in our case study, the Hierarchical clustering is more effective than the K-means method (65%>56%). Thanks to ontology use, we have obtained these approved results. In fact, result efficiency is due to the ontology use for the validation of relevant concepts and features selection. As future work, we must improve obtained clusters by proposing new similarity measures that take into account ontology inferences and semantic relationships.

Results analysis: According to the obtained clusters by hierarchical clustering, we can infer the following notes:

- Women on cluster 0 require a specific type of media that support text to speech services. These women need also training on how to use these media.
- Women belonging to cluster 1 need to keep their customers and to promote one to one relationship to increase their business.
- Concerning women on cluster 2, if they are given the means, they can possibly use new technologies to improve their business and so to occupy a more powerful position within their family, their community and their country.

As for the next step, we are aiming to improve the women situation by offering them an IT interface adapted to their profile based on the analysis of the obtained clusters.

5 Conclusion

In this paper, we performed knowledge representation about handicraft women through ontology. Then, we achieved the knowledge clustering into groups by hierarchical and k-means algorithms. We obtained a better result through the hierarchical algorithm. The efficiency of the obtained results is justified by the ontology use. In terms of perspectives, we will improve these results by proposing similarity measures based on ontology inferences. After that we intend to offer IT solution interface adapted to handicraft women profiles.

Résumé

Il est évident que les femmes participent à des activités économiques qu'ils soient dans des pays développées ou émergées. Cependant, la majorité de ces femmes sont confrontées à différents problèmes tels que l'insuffisance du capital, les contraintes de prix et de commercialisation, l'indisponibilité des matières premières, le manque de connaissances sur les technologies modernes, etc. Pour cela, nous sommes impliqués dans un projet de recherche qui vise à aider les femmes artisanes, des deux pays la Tunisie et l'Algérie, à utiliser les nouvelles technologies. Dans cet article, notre objectif est de regrouper les femmes artisanes afin de leurs proposer de nouvelles technologies adaptées à leurs profils dans le but d'améliorer leurs niveaux socio-économique. Le regroupement est basé sur des connaissances représentées dans une ontologie. Ces connaissances sont extraites à partir d'un ensemble d'interviews effectué sur un cas réel de femmes artisanes des deux pays.

Acknowledgements

We are very thankful to the Algerian Tunisian Project dealing with the improvement of handicraft women business in emerging countries through affordable technologies and social networks.

References

Esposito F., Fanizzi N., and d'Amato C.. "Conceptual Clustering Applied to Ontologies by means of Semantic Discernability", Proceedings of the third ECML/PKDD international workshop on Mining Complex Data, 2007.

Gruber T., 2008, Ontology. Entry in the Encyclopedia of Database Systems, Ling Liu and M. Tamer Özsu (Eds.), Springer-Verlag, to appear in 2008.

Kartik D. and Karunesh S. (2010) "Problems and Prospects for Marketing of Rural Products: An Empirical Study of Tribal Region of Rajasthan (India)" Oxford Business & Economics Conference Program ISBN : 978-0-9742114-1-9.

Maedche A. and Zacharias V. "Clustering Ontology-based Metadata in the Semantic Web", PKDD 2002, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases.

Mikroyannidi E., Iannone L., Stevens R., Rector A.. (2011) "Inspecting regularities in ontology design using clustering". International Semantic Web Conference : Germany.

Une approche de conception multidimensionnelle d'entrepôt de données en utilisant les ontologies

Manel Zekri*, Atid Gharbi**
Abdelaziz Abdellatif ***

University of Tunis El Manar, Faculty of Sciences of Tunis
Department of Computer Science, El Manar II 2092, Tunisia
* manel.zekri@planet.tn
** gharbiatid@gmail.com
*** abdelaziz.abdellatif@fst.rnu.tn

Résumé. La construction d'un entrepôt de données est une tâche complexe qui vise à satisfaire les besoins des décideurs. Un des points clés de la réussite d'un projet d'entreposage est la conception du schéma multidimensionnel. De nombreuses recherches ont montré que l'utilisation de l'ontologie dans la conception de système d'information est prometteuse. Dans cet article, nous proposons une méthode pour la conception multidimensionnelle à partir de la source des données opérationnelle, en utilisant des ontologies. Nous introduisons le concept de l'ontologie multidimensionnel comme un outil pour la spécification des connaissances multidimensionnelle. Nous proposons une méthode basé sur l'ontologie pour la modélisation du schéma des données, et éventuellement couvrir les différentes phases du cycle de vie de l'entrepôt de données.

1 Introduction

De nos jours, les entrepôts de données sont devenus une technologie puissante qui a un intérêt croissant chez la communauté scientifique, et un déploiement à grande échelle auprès des décideurs dans différents domaines. Ils représentent l'un des aspects les plus importants des systèmes d'information décisionnels et des bases de données. La conception d'un ED est maintenant reconnue comme une tâche cruciale pour le succès d'un système d'information décisionnel (Golfarelli et al., 2009). Plusieurs méthodologies et approches ont été présentées dans la littérature. Ces approches peuvent être classées en trois catégories. La première a été introduite par (Phippset al., 2002). Elle est dite «guidée par les besoins des utilisateurs » car elle part des besoins de l'utilisateur final. Cependant, le fait de ne pas considérer les sources à partir desquelles l'ED sera alimenté risque de l'échouer si les données requises par les utilisateurs ne sont pas disponibles. Pour pallier à cette insuffisance, des chercheurs ont proposé de commencer la conception de l'ED à partir des sources de données dite « guidée par les sources de données ». (Kimball, 1996) et (Maiz et al., 2006) se sont basés dans leur travaux sur cette nouvelle méthode. D'autres chercheurs, à savoir (Bonifati al., 2001) et (Giorgini et al., 2008) ont pensé que la disponibilité des informations sources ne signifierait pas forcément que l'utilisateur est concerné par ces informations. Afin de profiter des avantages

des deux approches précédentes, des travaux comme (Phippset al., 2002) et (Soussi et al., 2005) proposent des approches mixtes partant des données sources et des besoins des utilisateurs. Toutes ces différentes méthodologies et approches sont généralement effectuées manuellement par des experts. Au cours des dernières années, quelques efforts de recherche ont tenté d'automatiser la conception de bases de données multidimensionnelles afin d'éviter que cette tâche soit (complètement) effectué par un expert, et de faciliter le processus. Pour ce faire, ces approches commencent à partir d'une analyse détaillée des sources de données pour déterminer les concepts multidimensionnels dans un processus de réingénierie, visant à identifier les connaissances multidimensionnelle pertinentes contenues dans les sources de données opérationnelles.

Le reste du papier est organisé comme suit. Nous commençons, dans la section 2 par présenter l'état de l'art. Ensuite, dans la section 3, nous présentons l'ontologie multidimensionnelle et la démarche derrière sa construction. Dans la section 4, nous discutons notre vision pour étendre l'ontologie multidimensionnelle. Enfin, nous finissons par une conclusion.

2 Etat de l'art

La conception d'un ED est maintenant reconnue comme une tâche cruciale pour la réussite d'un projet de stockage. Plusieurs approches ont été proposées et celles qui sont basées sur des sources de données peuvent être classées en deux catégories: (i) les approches guidées par les sources de données, et (ii) les approches mixtes. Les approches relevant de la première catégorie sont basées sur les sources de données pour dériver des schémas multidimensionnels. Elles sont également basées sur le modèle d'affaires des données et profitent des relations entre les données en vue d'élaborer un schéma multidimensionnel d'une manière structurée. Ce type d'approche a été adopté par (Kimball, 1996) et (Maiz et al., 2006). Cependant, ces travaux ne peuvent pas être automatisés. En effet, malgré la dérivation d'un schéma multidimensionnel basé sur l'identification des faits, ils ne précisent pas un critère formel pour l'identification des faits à partir du modèle de données. Ils sont satisfaits par l'identification manuelle de ces faits. En outre, ces approches ne tiennent pas compte des besoins des utilisateurs. Ils assurent la disponibilité de l'information, mais ceci ne garantit pas que l'utilisateur sera satisfait par l'ED.

Les approches mixtes considèrent les sources de données et les besoins des utilisateurs afin de s'assurer que l'utilisateur trouve l'information qui l'intéresse et que cette information est disponible. Dans (Phippset al., 2002) une constellation a été dérivée à partir des sources de données et validée par les besoins des utilisateurs. En outre, (Mazón et al,. 2009), (Giorgini et al., 2008) et, (Soussi et al., 2005) ont dérivé un ensemble d'étoiles à partir des besoins des utilisateurs afin de valider les sources de données. Pour l'identification des faits, les premières approches considèrent «entités et associations n-aires avec au moins un attribut numérique" comme un fait. (Soussi et al., 2005) estime que cette dernière hypothèse génère un grand nombre de résultats dont la majorité d'entre eux ne correspond pas aux faits valides. Ceci s'applique aux entités et associations avec des attributs clés numériques ou des attributs qui ne correspondent pas à des mesures telles que "code postal". Pour s'attaquer à ces déficiences, ils proposent une nouvelle hypothèse : «entité ou association avec des attributs numériques non-clés".

3 Ontologie multidimensionnelle

Dans la conception d'ED, différentes techniques de modélisation sont utilisées pour représenter les concepts multidimensionnels extraits à partir des sources de données, ainsi que les sources elles-mêmes. Il peut être un schéma EA, un diagramme UML ou un graphe, etc. Contrairement aux ontologies, qui sont prêts pour faire des traitements, ces techniques sont des formalisations conceptuelles destinés à représenter graphiquement le domaine et non pas à requêter et raisonner. Ce travail est la continuation d'une recherche précédente (Zekri et al., 2011) et il vise à intégrer les ontologies dans le processus de conception. Donc, notre point de départ était notre méta-modèle du schéma d'ED (Zekri et al., 2011). Un ED est un ensemble de tables. Ces tables sont des tables de dimension ou des tables de faits. Une dimension est une perspective d'analyse. Un sujet est mesuré selon plusieurs perspectives. Ces perspectives sont des attributs qui caractérisent les mesures de l'objet analysé. Chaque table de dimension a un ou plusieurs niveaux de granularité, et forme une hiérarchie de dimensions utilisées pour définir la granularité d'une dimension. En fait, les attributs d'une dimension sont organisées à l'aide d'une relation de « est plus fin que » pour restreindre ou augmenter le niveau de détail de l'analyse. Par exemple, (jour-mois-année) est une hiérarchie de la dimension de fréquence. Chaque niveau peut être généré par un ensemble de règles d'agrégation, contient un ensemble d'attributs, soit une clé primaire. Ainsi, chaque niveau correspond soit à un ensemble d'attributs explicites ou aux règles attributs générés. En outre, les tables de fait ont une ou plusieurs mesures, et une clé primaire qui est une composition des clés étrangères (clé primaire de certaines tables de dimensions). Un fait est constitué de mesures correspondant à l'information de l'activité analysée.

L'ontologie multidimensionnelle est une représentation de connaissances dédiée aux systèmes décisionnels. Elle spécifie les concepts multidimensionnels et leurs relations sémantiques et multidimensionnelles. Diverses approches ont été proposées pour guider la création d'ontologies, nous avons surtout basé notre processus de construction d'ontologies multidimensionnel sur l'approche proposée dans (Maiz et al., 2006).

3.1 Concepts multidimensionnels

Ici, nous examinons plus en détails les concepts multidimensionnels de notre ontologie. En se basant sur notre méta-modèle du schéma d'ED (Zekri et al., 2011), nous avons défini les concepts suivants: Fait, Fait_ID, Mesure, Dimension, Dimension_ID, Hiérarchie, Niveau et Attribut. Ces concepts sont les éléments qui constituent un modèle multidimensionnel, dans notre cas, le schéma en flocon de neige (Kimball, 1996).

- **Fait**: Il représente un sujet analysé par le système décisionnel, et il est noté Fait (F).
- **Fait_ID**: Il s'agit de la clé primaire d'une table de fait qui est une composition des clés étrangères (clé primaire de certains tables de dimensions), et est notée Fait_ID (FID).
- **Mesure**: Chaque fait est caractérisé par un ou plusieurs attributs. Ces attributs sont généralement numériques et ils représentent des indicateurs d'analyse. Le concept mesure est noté Mesure (M).
- **Dimension**: est une perspective d'analyse d'un fait, et est noté Dimension (D).
- **Dimension_ID**: Il s'agit de la clé primaire d'une table de dimension, et c'est noté Dimension_ID (DID).

- **Hiérarchie**: Les attributs d'une dimension sont organisées à l'aide d'une relation de « est plus fin que » pour restreindre ou augmenter le niveau de détail de l'analyse. Cela permet de former des hiérarchies qui suivent un ordre à partir de la granularité la plus fine, jusqu'à la granularité maximale. Nous pouvons trouver plus d'une hiérarchie dans la même dimension. Un concept de Hiérarchie est notée Hiérarchie (H).
- **Niveau**: Chaque niveau de hiérarchie est représenté par le concept de Niveau, et est noté Niveau (N).
- **Attribut**: Les attributs sont les paramètres des niveaux, et sont notés Attribut (A).

3.2 Relations multidimensionnelles

Après avoir défini les concepts, nous devons préciser les relations qui existent entre eux. Chaque relation est de la forme Relation (X, Y), où Relation est un prédicat binaire, et X et Y sont des concepts. Nous définissons les relations décrites ci-dessous. Les concepts et les relations multidimensionnelles sont représentés dans la FIG. 1.

- Est_Fait_ID (FID, F) où Fait_ID (FID), Fait (F) et FID est l'identifiant de F.
- Est_Mesure (M, F) où Mesure (M), Fait (F) et M est une mesure de F.
- Est_Dimension (D, F) où Dimension (D), Fait (F) et D est une dimension de F.
- Est_Dimension_ID (DID, D) où Dimension_ID (DID), Dimension (D) et DID est l'identifiant de D.
- Est_Hiérarchie (H, D) où Hiérarchie (H), Dimension (D) et H est une hiérarchie de D.
- Est_Niveau (N, H) où Niveau (N), Hiérarchie (H) et N est un niveau de H.
- Est_Attribut (A, N) où Attribut (A), Niveau (N) et A est un attribut de N.
- Est_Plus_Fin_Que (Ni ,Nj) où Niveau (Ni), Niveau (Nj), Ni et Nj sont de la même hiérarchie et Ni a une granularité plus fin que Nj.

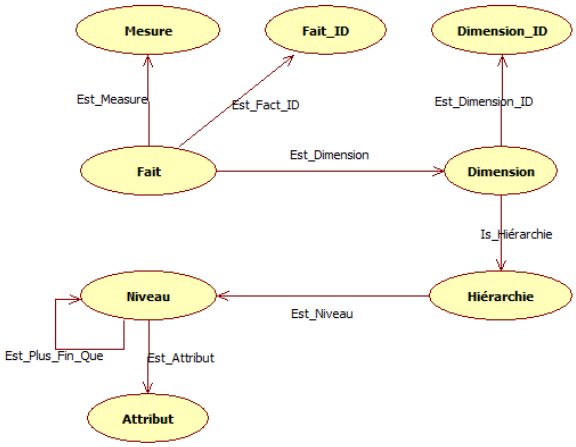


FIG. 1 – *Représentation graphique de l'ontologie multidimensionnelle.*

4 Etendre l'ontologie multidimensionnelle

Alors que d'autres techniques (mentionné ci-dessus) pour représenter les connaissances multidimensionnelles sont un choix approprié pour leurs approches respectives, leur utilisa-

tion se limite à la tâche de conception. En revanche, l'ontologie multidimensionnelle peut toujours être utile, car elle peut couvrir les différentes phases du cycle de vie d'un système d'information décisionnel, c'est-à-dire à partir de la spécification des besoins, la conception de l'ED, jusqu'à arriver aux phases d'exploitation. Ceci est pratique parce que les derniers SGBD permettent de stocker des ontologies parallèlement aux données de la même structure de base de données (Khouri et al., 2012). Cette base de données est appelée OBDB (ontology-based database ou base de données basé sur l'ontologie). Les ontologies sont évolutives et extensibles, et elles ont montré leur efficacité pour les systèmes d'information et la spécification des besoins (Calero et al., 2006). Plus concrètement, et de la même manière que les ontologies sont utilisées pour la clarification sémantique des sources de données, elles sont utilisés pour identifier et gérer les conflits sémantiques entre les concepts. Cela nous permet d'ajouter autant d'extensions que nécessaire pour l'ontologie multidimensionnelle pour couvrir les différentes phases du cycle de vie d'un système d'information décisionnel. Pour démontrer cet aspect, nous proposons une extension qui représente le schéma conceptuel de source de données opérationnelles, dans ce cas un diagramme EA. Comme l'illustre la FIG. 2, l'extension se compose de trois nouveaux concepts, qui sont Entité, Association et Lien. De cette façon, il est plus facile d'automatiser le processus d'extraction de connaissances multidimensionnelle à partir de la source.

L'heuristique utilisée dans (Zekri et al., 2011) pour déterminer les faits et les dimensions extrait les faits potentiels à partir des associations et les dimensions à partir des associations, d'où les relations Est_Fait_Potentiel, Est_Fait_Valide, Est_Dimension_Potentielle et Est_Dimension_Valide. L'utilisation de ces relations pour garder une trace de quel élément multidimensionnel correspond à quel élément source peut être utile dans les étapes ultérieures, par exemple la conception d'un ETL.

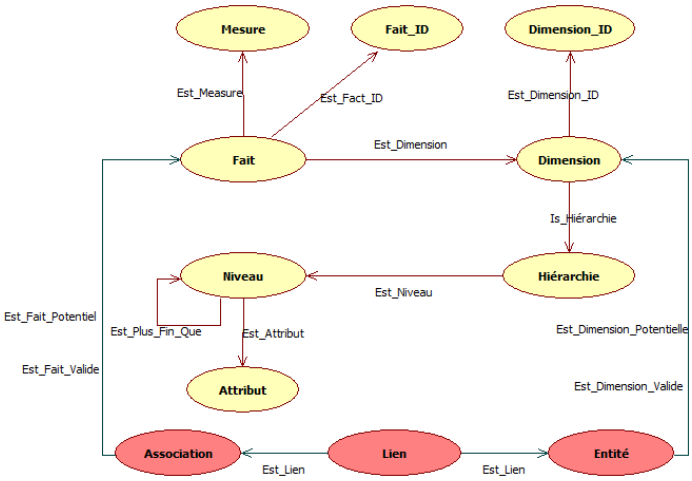


FIG. 2 – Représentation graphique de l'ontologie multidimensionnelle avec extension.

5 Conclusion

Dans ce papier, nous avons présenté une approche pour représenter le schéma d'ED basée sur une ontologie qui capture les connaissances multidimensionnelles. Nous avons discuté la

construction de l'ontologie et ses avantages par rapport à d'autres méthodes. Ensuite, nous avons discuté de notre vision pour étendre l'ontologie multidimensionnelle pour pourvoir éventuellement couvrir les différentes phases du cycle de vie de l'ED. Nous avons montré comment l'utilisation de l'ontologie multidimensionnelle combinée avec une extension peut être bénéfique. Dans l'avenir, nous avons l'intention de continuer à exploiter l'extensibilité des ontologies, en envisageant d'utiliser des ontologies de domaine comme des extensions pour tenter d'améliorer le schéma d'entrepôt de données qui en résulte.

Références

Bonifati, A., F. Cattaneo, S. Ceri, A. Fuggetta, et S. Paraboschi (2001). Designing data marts for data warehouses. ACM Trans. Softw. Eng. Methodol. 10(4), 452–483.

C. Calero, F. Ruiz, M. Piattini (2006). Ontologies for Software Engineering and Software Technology, Springer Verlag, Berlin, Heidelberg.

Giorgini, P., S. Rizzi, et M. Garzetti (2008). Grand : A goal-oriented approach to requirement analysis in data warehouses. Decision Support Systems 45(1), 4–21.

Golfarelli, M. et S. Rizzi (2009). A survey on temporal data warehousing. IJDWM 5,1–17.

Khouri Selma, Boukhari Ilye`s et al. (2012). Ontology-based structured web data warehouses for sustainable interoperability: requirement modeling, design methodology and tool. Computers in Industry 63 (2012) 799–812.

Kimball R. (1996). The Data Warehouse Toolkit, John Wiley and Sons, Inc., New York.

Maiz N., Boussaid O., Bentayeb F. (2006). Un système de édiation basé sur les ontologies pour l'entreposage virtuel, Hammamet-Tunisie, 31 May to 3 june 2006.

Manel Zekri, Imen Marsit et Abdellaziz abdellatif (2011). A new data warehouse approach using graph theory. IEEE International Conference on e-Business Engineering, Octobre 19-21, 2011 China.

Mazón, J.-N., J. Lechtenbörger, et J. Trujillo (2009). A survey on summarizability issues in multidimensional modeling. Data Knowl. Eng. 68(12), 1452–1469.

Phipps, C. et K. C. Davis (2002). Automating data warehouse conceptual schema design and evaluation. In DMDW, pp. 23–32.

Soussi, A. and F. Gargouri (2005). Génération et validation automatiques de schémas de magasins de données. In Tunisie :GEI'05.

Summary

Building a data warehouse is a complex task that aims to satisfy the needs of decision makers. One of the key points to the success of a data warehousing project is the design of the multidimensional schema. Many researches showed that the use of ontology in information system design is promising. In this paper, we propose a method for multidimensional design, using ontologies as a tool for specifying multidimensional knowledge.

Une vue d'ensemble des systèmes de traitement des requêtes sur des sources sémantiquement et/ou structurellement hétérogènes

Abderrafiaa ELKALAY*, Naoual MOUHNI**

*a.elkalay@uca.ma
**naoual.mouhni@edu.uca.ma

Résumé. Dans les systèmes classiques, l'extraction des données se faisait à partir des sources généralement structurées de façon homogène, ou même à partir d'une seule source centralisée. Au fil du temps, les choses ont changé, et la nature des sources de données devient de plus en plus hétérogène, dans certains cas les données sont physiquement réparties mais forment une même entité pour l'utilisateur final, qui lui envoie sa requête sans se soucier des hétérogénéités qui peuvent être d'ordre: structurel, sémantique ou même syntaxique.

Dans ce papier, nous allons mettre en lumière les différentes approches de traitement des requêtes sur des sources hétérogènes sous ses différentes formes.

1 Introduction

De nos jours, avec la grande expansion d'informations sur Internet et l'utilisation des sources multiples qui peuvent être hétérogènes et physiquement séparées, les vielles méthodes de traitement des requêtes utilisateur deviennent obsolètes, c'est pour cela les spécialistes en bases de données essayent de rechercher de nouvelles approches pour remplir cet écart.

Le premier pas était en proposant quelques méthodologies pour intégrer des sources de données hétérogènes, puisque dans une telle situation, chaque base de données indépendante a son propre schéma exprimé avec son propre modèle de données et a son propre langage d'interrogation (Mahendar et Zhou (1996)), donc une requête utilisateur doit passer par les étapes suivantes:

- Écrire la requête dans un langage de requêtage de départ.
- résoudre les incompatibilités entre les data sources en question en faisant le matching entre les types de données et aussi les noms d'attributs ;
- Décomposition de la requête initiale en plusieurs sous requêtes.
- réécriture des requêtes en langage compréhensible par les sources destination.
- appliquer une fonction d'intégration des résultats dans une réponse finale ne contenant pas de redondance et sans perte d'information.

Dans la section suivante nous allons discuter les problèmes d'intégration des données hétérogènes et quelques stratégies pour les résoudre. dans la section 3, nous allons nous focaliser sur le traitement des requêtes dans des environnements hétérogènes, non seulement dans

le cas des sources résultats d'une répartition horizontale, verticale ou même hybride qui peuvent être harmonisées en quelque sorte, mais plutôt dans le cas des data sources totalement indépendantes ce qui rend le processus d'intégration une mission compliquée. Puis, en section 4 nous allons discuter pour conclure les points forts et faibles des approches traitées dans ce papier.

2 Intégration des sources de données hétérogènes

Les techniques d'intégration de sources sont passées au niveau suivant, bien au delà de techniques traditionnelles comme JDBC (Java Data Base connectivity) ou encore ODBC (Object Data Base Connectivity) qui connectent des bases de données relationnelles entre elles (Nolen et McCauley(2004)).

Aujourd'hui les données peuvent être stockées dans des structures hétérogènes distribuées ou non, car on trouve dans un même domaine scientifique, économique ou autre, générant un très gros volume de données qui se multiplie dramatiquement jour après l'autre. Les données traitées dans un même champ de travail, peuvent être hétérogènes, non seulement au niveau structurel, mais aussi sur le plan sémantique. Par exemple, prenons les données du domaine médical distribuées sur des sites différents, on trouve qu'elles ne sont pas sémantiquement intégrées, on peut même trouver un effet secondaire exprimé différemment sur deux sites alors qu'il s'agit d'une même donnée. Si ce système était intégré, les requêtes client auront des résultats fiables.

Plusieurs recherches ont été faites pour améliorer les techniques d'intégration, dont on peut citer les techniques de data warehousing qui consistent à extraire les données de différentes sources, les nettoyer ensuite charger le résultat dans des entrepôts de données ou des data marts pour des fins d'analyse par la suite.

En effet, même en utilisant les techniques du data warehousing, et avec l'augmentation et l'expansion des données traitées dans certains champs, on peut faire face à des cas où un système décisionnel est composé de plusieurs entrepôts de données, créés dans des circonstances différentes, par des concepteurs différents, et qui finissent par se regrouper sous un même système décisionnel dit "fédéré", suite à une réunion de sociétés ou regroupement de petit Data warehouses existant sans avoir la nécessité de tout centrer dans un même entrepôt, dans ce cas , on peut faire appel à des technique de traitement des données sémantiquement hétérogènes telle que l'utilisation des ontologie comme proposé dans (MOUHNI et ELKALAY (2013)).

Donc on peut résumer les obstacles à l'interrogation des données dans:

- Hétérogénéité structurelle: différence au niveau des types de données et leurs structures.
- Hétérogénéité sémantique: différence au niveau des expressions utilisées pour représenter un attribut
- Hétérogénéité du modèle: différence au niveau du modèle et le type du schéma qui peut être par exemple, relationnel, objet,...

Ces éléments, nous conduit vers la description de deux processus utilisés pour identifier les hétérogénéités citées en dessus, à savoir le Mapping et le Matching.

Le processus de Matching est utilisé pour identifier si deux éléments sont reliés sémantiquement ou pas, si l'on prend par exemple deux schémas DB1.Client (ID, Name, Class) et

DB2.Customer (IDC, FName, LName, PointNbr), l'opération consiste à faire correspondre les éléments qui sont sémantiquement reliés mais différemment représentés, dans ce cas on peut citer ID dans DB1.Client et IDC dans DB2.Customer qui référencent tous les deux l'identifiant du client. Tandis que le processus de Mapping, est l'opération de transformation entre les éléments, par exemple, les attributs class et pointNbr sont utilisés pour classifier les clients, si les valeurs de l'attribut class sont représentées sous la forme : A, B, C et pour l'attribut pointNbr est un champs numérique, on peut alors faire le mapping entre les valeurs de ces deux attributs en assignant à chaque classe un intervalle de valeurs numérique par exemple, telle que (0-1000;C) (1000-5000;B).

Les systèmes d'intégration de données hétérogènes suivent des approches différentes même s'ils essayent de résoudre le même problème; pour cela nous allons citer quelques approches les plus répondues du domaine.

Une architecture dite central data integration, est un modèle des systèmes qui disposent d'un schéma global fournissant l'utilisateur final avec une interface commune pour accéder aux informations stockées dans des sites hétérogènes moyennant des requêtes formées de termes du schéma global (Xiao et I.F.C.a.H(2009), Amann et Scholl(2002), Juan (2010)).
Contrairement à l'architecture centralisée, les systèmes pairs-à-pairs d'intégration des données, chaque pairs ou source de données peut accepter des requêtes utilisateurs pour avoir accès dans d'autres pairs (Xiao et I.F.C.a.H(2009), Juan (2010)).

2.1 Les approches de Mapping et le processus de traitement des requêtes

Comme cité précédemment, pour assurer l'intégration de données, l'une des opérations primordiales est le processus de Mapping, que ce soit dans une architecture centralisée ou pairs-à-pairs.

Dans une architecture centralisée, le Mapping est fait entre le schéma global et les schémas locaux des sites. Alors que dans le cas des P2P, le Mapping est appliqué entre les pairs.

Il existe deux approches pour la création de ces types de Mapping; GLobal as View (GaV) et le Local as View (Lav) (Xiao et I.F.C.a.H(2009), Juan (2010), Nolen et McCauley(2004)).

La façon dont est établie la correspondance entre le schéma global et les schémas des sources de données à intégrer, L'approche GAV, qui provient du monde des bases de données fédérées, consiste à définir le schéma global en fonction des schémas des sources de données à intégrer. L'approche LAV est l'approche duale Selon l'approche LAV, il est très facile d'ajouter une source d'information, cela n'a aucun effet sur le schéma global. En revanche, la construction des réponses à des requêtes est complexe, contrairement à la construction de réponses dans un système adoptant une approche GAV qui consiste simplement à remplacer les prédicats du schéma global de la requête par leur définition.

Et puisque chaque système doit prendre en considération une stratégie de mise à jour, dans un système GaV, chaque fois qu'on change un élément au niveau des sites, il faut mettre à jour la vue global, alors que dans le cas des LaV, les changements affectant les schémas locaux n'affectent pas le schéma global.

Systèmes de traitement des requêtes sur des sources sémantiquement ou structurellement hétérogènes

L'approche Lav est une approche basée sur les vues, ce qui rend la construction de requête une mission compliquée, certes puisque l'utilisateur n'a pas accès aux relations des sources de données la seule information dont il dispose est via les vues.
Alors que dans le cas des approches GaV, les choses sont moins compliquées, puisque le processus de mapping spécifie directement pour chaque élément de la source son correspondant au niveau du schéma global.

2.2 Les approches basées sur l'ontologie

Plusieurs méthodologies ont été développées pour résoudre le problème d'intégration de données, parmi elles l'utilisation des ontologies; qui trouvent leurs origines dans le domaine de la philosophie. Une ontologie est la base du processus d'intégration sémantique de données
C'est un moyen d'utiliser une représentation conceptuelle de données et des relations entre elles afin d'éliminer les hétérogénéités. Elle est définie comme une spécification explicite d'une conceptualisation partagée (Xiao et I.F.C.a.H(2009))

Trois approches représentent l'utilisation des ontologies dans le processus d'intégration de données, elles sont représentées dans les figures ci-dessous:

Fig. 1. Single ontology approach

Fig. 2. Multiple ontology approach

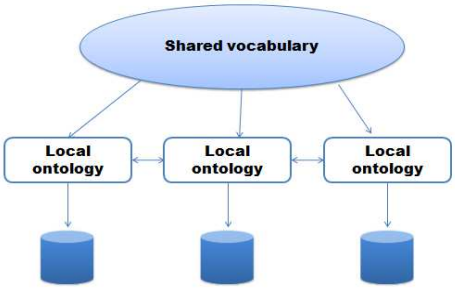


Fig. 3. Hybrid ontology approach

La première approche (Fig. 1), est basée sur l'utilisation d'une même ontologie partagée entre toutes les sources de données. Elle parait être l'approche la plus simple en comparaison avec les autres. Cependant, cette approche a un inconvénient dans sa simplicité, puisque dans certains cas, on a besoin d'une spécification détaillée au lieu d'une ontologie globale qui traite l'ensemble des sources comme étant une même entité intégrée.

L'approche suivante est l'approche basée sur des ontologies multiple (Fig. 2), elle est caractérisée par l'utilisation d'une ontologie pour chaque source de données. Chaque site a son propre ontologie et est intégrée en harmonie avec ses voisins. Le problème qui se pose pour cette approche, est qu'on risque en utilisant plusieurs ontologies locales de tomber dans le problème initial, à savoir l'hétérogénéité des ontologies si elles n'utilisent pas un vocabulaire partagé, et dans ce cas on aura besoin d'une ontologie pour résoudre le problème d'hétérogénéité des ontologies.

La dernière approche (Fig. 3), consiste à utiliser les concepts des deux premières approches, cette approche semble être la plus convenable au cas des grands projets, dans lequel les sources sont extrêmement différents, elle propose d'utiliser un vocabulaire partagé pour intégrer les ontologies locales définies sur chaque source.

3 Processus de traitement des requêtes dans des environnements hétérogènes

Le processus d'intégration de données est généralement basé sur des ensembles indépendants, hétérogènes de sources de données dans un même domaine d'intérêt.

La moelle épinière du processus d'intégration de données, est le traitement de requêtes, comme nous l'avons déjà cité, l'une des points qui doivent être présent dans systèmes d'intégration de données est un langage de requêtage commun, qui doit être définit malgré la présence d'une panoplie de langage de requête existante.

3.1 Modèle de requête (Query Model)

Chaque data source a son propre modèle de requête (query model), qui est le modèle de stockage des données et il doit être connu par l'utilisateur final afin de pouvoir rédiger des requêtes d'interrogation de la source. Le modèle de requête est caractérisé par quatre composants (Sujansky(2001)), le premier composant est le modèle abstrait qui donne une idée sur les types de structures de données pouvant être prises en charge par le nœud en question, par exemple, fichiers texte, des bases de données hiérarchiques des tables relationnelles. le deuxième composant est le schéma de données qui spécifie la représentation et la localisation des données sur la base de données, si la requête utilisateur porte sur deux données a et b , on doit savoir si a et b sont spécifier dans le même fichier? Dans le cas contraire faut-il joindre deux fichiers ?. Quel est le langage de requêtage qui sera utilisé pour interroger la base de données (e.g SQL)?

Le dernier composant du query model est le format de la base de données; pour expliquer ce point , faisons appel à l'exemple déjà cité un peu plus haut , l'exemple des deux tables DB1.Client (ID, Name, Class) et DB2.Customer(CID, FName, LName, PointNbr), pour ces deux sources de données (DB1 et DB2) les informations du client sont stockées différemment donc on a besoin avant de créer une requête d'interrogation comment ces données sont elles stockées.

Généralement, dans les systèmes d'intégration des sources hétérogènes, nous avons besoin de représenter ces éléments pour assurer le processus d'intégration.

Systèmes de traitement des requêtes sur des sources sémantiquement ou structurellement hétérogènes

Dans la figure (Fig. 4) nous décrivons un modèle standard d'architecture d'un système d'intégration.

.

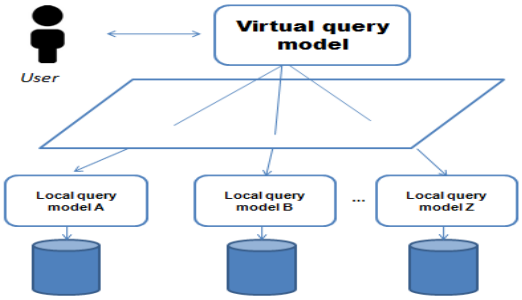


Fig. 4. Position of the Virtual query model in a heterogeneous data integration system

Ces systèmes offrent un query model virtuel permettant à l'utilisateur d'accéder aux sources de données sans se soucier de son query model local, tout le processus d'intégration doit être transparent pour l'utilisateur final.

4 Discussion et conclusion

Face au besoin croissant d'intégration de sources de données hétérogènes, qui peuvent être dispersées de part le monde , autonomes mais connectés afin de fournir une information complète, correcte pouvant représenter une base de prise de décision à l'utilisateur final;
Plusieurs applications ont vu le jour en vu de proposer une solution prenant en considération la variété des sources de données existantes aujourd'hui (sources XML, base de données relationnelles, entrepôts de données ou base de données orientée objet,...)
Ces applications offre généralement à l'utilisateur une interface lui épargnant de se soucier de la localisation des données recherchées, de leur structure, et des problèmes d'hétérogénéité et de redondance susceptibles.
Ces solutions, ont attaqué ce problème différemment, tantôt par utilisation des ontologies, afin de surmonter le problème d'hétérogénéité sémantique. D'autres présentent des médiateurs ou wrappers afin de traiter les requêtes utilisateur.

Cependant, le problème d'hétérogénéité persiste, vue à la nature dynamique des sources de données, car si l'en prend comme exemple la grande quantité de données insérée régulièrement dans les bases de données (e.g. les réseaux sociaux dans des bases de données noSQL et Big data) on s'est rendu compte que ces données font face aux problèmes des dirty data qui persiste parfois même après ces processus proposés.

Références

Bernd Amann , C.B., Irini Fundulaki, and Michel Scholl (2002), *Ontology-Based Integration of XML Web Resources.*

Bernd Amann, C.B., Irini Fundulaki, Michel Scholl (2002), *Querying XML Sources Using an Ontology-based Mediator.*

Juan E. (2010), *Ontology data integration for competitive decision making*

Lenzerini, M., *Data Integration: A Theoretical Perspective.*

Mahendar MacUaavaram, D.L.A., Ming Zhou (1996), *INTEGRATING HETEROGENEOUS DISTRIBUTED DATABASE SYSTEM* Computers ind. Engng . Vol. 31(No. 1/2): p. pp. 315 -318

Mike Nolen, L.M. (2004), *Integration of Heterogeneous Data Sources using Ontologies: An overview of current methodologies*, icai

MOUHNI N., EL KALAY A. (2013), *Ontology based data warehouses federation management system*. International Journal of Computer Science Issues. 10(4).

Sujansky, W.(2001), *Heterogeneous Database Integration in Biomedicine*. Journal of Biomedical Informatics. 34: p. 285–298.

Xiao, I.F.C.a.H.(2009), *Ontology Driven Data Integration in Heterogeneous Networks*. ADVIS Lab Department of Computer Science University of Illinois at Chicago, USA .

Summary

In the past, to answer a user query, we generally extract data from one centralized database or from multiple sources with the same structure. then things have been changed and we are facing the fact that in some cases, it is necessary to use a set of data sources to provide a complete information. These sources are physically separated, but they are logically seen as a single component to the final user. Besides the structure heterogeneity, there is another important point for what specialists are trying to find a solution which is the semantic heterogeneity of data sources. In this paper we are going to provide a list of different approaches that treated the query processing problem on heterogeneous data sources under different angles.

XACML et WS-policy pour la sécurité des données et des ETL dans un Webhouse

Nesrine Zaghdoud, Salma DAMMAK, Faiza Ghozzi

zaghdoud.nesrine@gmail.com
damak.salma@gmail.com
jedidi.Faiza@gmail.com
MIR@CL laboratory, University of Sfax, Tunisia

Résumé. L'intégration des sources Web dans les Webhouse a enrichi et amélioré l'aide à la prise de décision. Néanmoins, des risques infinis liés au web doivent être traités. L'intégration des aspects de sécurité dès la phase de conception devient nécessaire afin d'empêcher toute tentative d'intrusion ou d'altération des données. Dans cet article, nous visons à obtenir des politiques de sécurité relatives à notre Webhouse à partir d'un modèle indépendant de la plateforme (PIM) intégrant les contraintes de protection des données du Webhouse et des processus ETL. Aussi, nous détaillons les transformations basées sur le langage QVT permettant d'obtenir d'une part, la politique de sécurité XACML traduisant les contraintes et les règles permettant de protéger les données du Webhouse. Et, d'autre part, la politique WS-policy pour assurer la sécurité des processus ETL définis sous forme de services web.

1 Introduction

Le métissage entre les entrepôts de données et le web a augmenté les défis de ce dernier surtout ceux liés à la sécurité pour la nouvelle architecture nommée Webhouse Kimball et al, (2000). Cette nouvelle architecture intègre les données extraites du web tels que les click-streams représentant le comportement des internautes, les données des réseaux sociaux, ...etc. L'Extraction, la Transformation et le Chargement de ces données a mené à la redéfinition de la couche des processus ETL sous forme de Web Services Liu et al (2010), Marotta et al (2012). Ces services traitent les sources web et génèrent un Webhouse XML.

La complexité de l'architecture du Webhouse avec la variété des mécanismes et des techniques de sécurité nécessitent la définition préalable des besoins de sécurité afin de les traduire sous forme de solution physique. Ces besoins de sécurité couvrent la protection des données en répondant aux questions telle que : qui accède à quoi ?. Ils couvrent aussi la sécurisation des processus ETL devenus plus vulnérables par l'exposition au web. En effet, un incident d'intrusion causé par la perte de contrôle sur ces services peut permettre aux attaquants de rediriger les données lors de leur chargement dans le Webhouse.

2 Etat de l'art

Plusieurs travaux ont abordé l'intégration des aspects de sécurité dans la conception des entrepôts de donnés. Fernandez-Medina et al (2006) présentent leur approche fondée sur

MDA et définissent leurs modèles de sécurité. Ces modèles permettent au concepteur de spécifier les contraintes de contrôle d'accès et d'audit lors de la phase de modélisation.

Vela et al (2012) se basent sur le modèle PIM présenté par Fernandez-Medina et al (2006) et proposent un modèle PSM permettant de représenter l'entrepôt de données sécurisé sous forme d'un schéma XML. Les transformations du PIM vers PSM sont réalisées via le langage de transformation QVT.

Fugkeaw et al (2010) présentent un prototype d'authentification et d'autorisation pour les entrepôts de données. Ils proposent une modélisation des utilisateurs en suivant l'approche RBAC. Les auteurs proposent une approche pour implémenter la sécurité au niveau des données des entrepôts en utilisant XACML. kimball et al, (2000) fut le premier à parler du Webhouse et sa sécurité. La solution proposée par Kimball était totalement technique mais les points entamés représentent la clef d'une conception sécurisée : l'authentification à deux facteurs, une connexion sécurisée, une définition précise des rôles des utilisateurs. En 2012, Damak et al proposent une approche dirigée par les modèles pour la sécurisation d'un WeBhouse. Ils présentent un modèle d'exigence (CIM) intégrant les besoins de sécurité. Les auteurs proposent un méta-modèle de conception et d'analyse (PIM) qui se compose d'un modèle de sécurité des utilisateurs, un modèle de sécurité des données et un modèle de sécurité de communication. Le passage du CIM vers le PIM est réalisé via le langage QVT.
Le processus ETL est potentiellement l'une des tâches les plus importantes de la construction d'un entrepôt. Il est à noter que plusieurs travaux ont traité leur modélisation, mais nous constatons la rareté des travaux traitant le sujet de sécurité de ce processus. Mrunalini et al (2013) analyse le processus ETL selon deux mesures de sécurité: l'indice de vulnérabilité et l'indice de sécurité. Un cadre de sécurité de toutes les phases dans le processus ETL a été suggéré, ainsi qu'une méthode pour évaluer la sécurité du système. Un méta-modèle orientée objet pour modéliser les processus ETL en intégrant la préservation de la vie privée est proposée dans Kiran et al, (2012).

La plupart des travaux se concentre sur le contrôle d'accès aux éléments du Webhouse sans tenir compte du contrôle de la sécurité des processus ETL. Dans un contexte Web, ces processus sont définis sous forme de web services caractérisés par une plus grande portabilité et adaptabilité. La sécurisation des services Web peut être définie par la politique WS-policy et sa spécification WS-Security Gallino et al (2012). Dans le cadre d'une approche MDA, nous proposons de définir i) un méta-modèle de Webhouse intégrant les besoins de sécurité des données et des ETL au niveau PIM, ii) un ensemble de règles de transformation de ces besoins en PSM et iii) un méta-modèle PSM incluant les spécifications de sécurité des services ETL.

3 Sécurité des données et des ETL : PIM vers PSM

Notre approche de modélisation de la sécurité d'un Webhouse définit, en premier lieu, le méta-modèle SecWebhouse permettant de spécifier la sécurité des données et des processus ETL au niveau de la couche PIM dans le cadre d'une architecture orienté Modèle (MDA). Ensuite, nous modélisons un ensemble de règles de transformations sous le langage QVT qui traduisent ce modèle en deux modèles PSM. Ainsi, le développeur du Webhouse ne fera pas face à la complexité et la multitude des techniques et des outils de sécurité. Le premier modèle PSM respectant la politique XACML permet de couvrir la sécurité des données XML de notre Webhouse. Cette politique est utilisée pour décrire les exigences générales de contrôle

d'accès en termes de contraintes sur des attributs. Au niveau du deuxième PSM, nous proposons d'implémenter la sécurité au niveau ETL à travers la politique WS-policy et sa spécification WS-Security qui décrit la sécurité des services Web Gallino et al (2012).

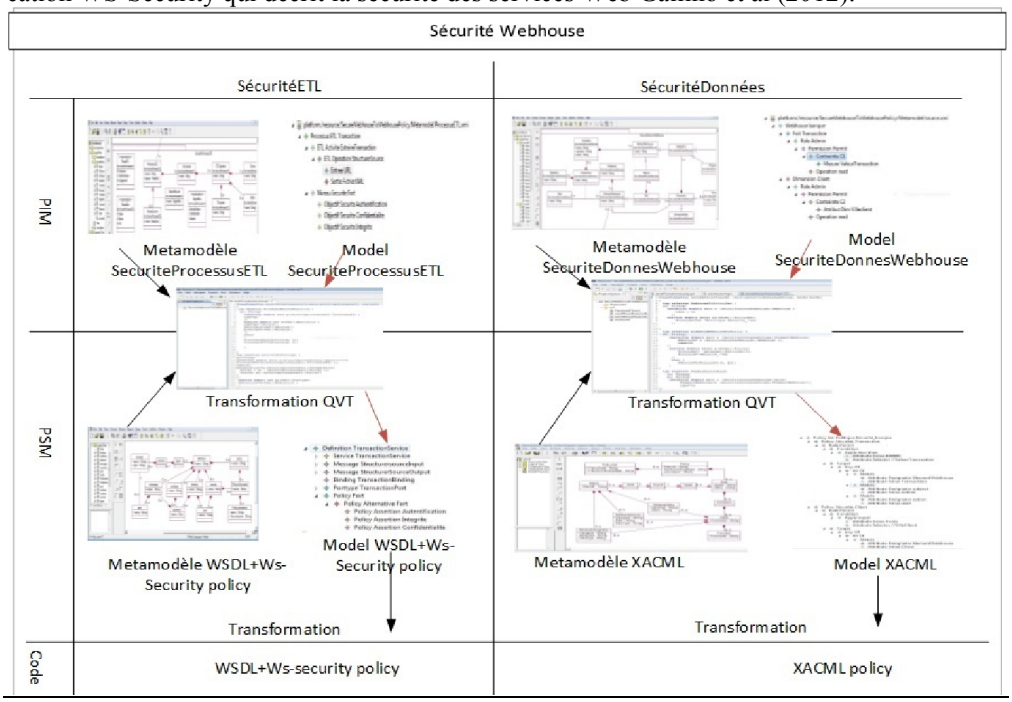


Figure 1: Approche de sécurité Webhouse

3.1 Le niveau PIM

Notre modèle PIM comprend deux packages : Un qui représente le modèle de sécurité des données et l'autre définit la sécurité des processus ETL. Le méta-modèle de la sécurité des données DWB permet de préciser notre stratégie de sécurité basée sur le modèle étendu du modèle de contrôle d'accès RBAC. Notre modèle permet de gérer les droits d'accès aux éléments du Webhouse (Fait, Dimension Web) en attribuant des permissions à des opérations en fonction des rôles. Les permissions ne seront exécutées que si les contraintes liées sont vérifiées. Les contraintes précisent la portée de la permission selon la valeur des attributs du Webhouse comparée avec l'opérateur et la valeur définie par la contrainte. Notre méta-modèle permet de modéliser, aussi, le processus ETL composé d'un ensemble d'activités regroupant des opérations à accomplir. Dans ce méta, un niveau de sécurité (NiveauSecurite) est associé aux processus ETL (ProcessusETL), aux activités ETL (ETLActivité), et aux opérations ETL (ETLOperation). Selon le niveau de sécurité, nous définissons les objectifs de sécurité à appliquer. Trois niveaux de Sécurité sont définis par notre modèle pour fixer le degré de sécurité à appliquer. Et, trois objectifs de sécurité sont proposés : l'authentification (pour assurer la vérification de l'identité de l'expéditeur), l'intégrité (vise à garantir que les données n'ont pas été modifiées d'une manière non autorisée) et la confidentialité (pour s'assurer que les données ne sont pas divulguées à des personnes non autorisées).

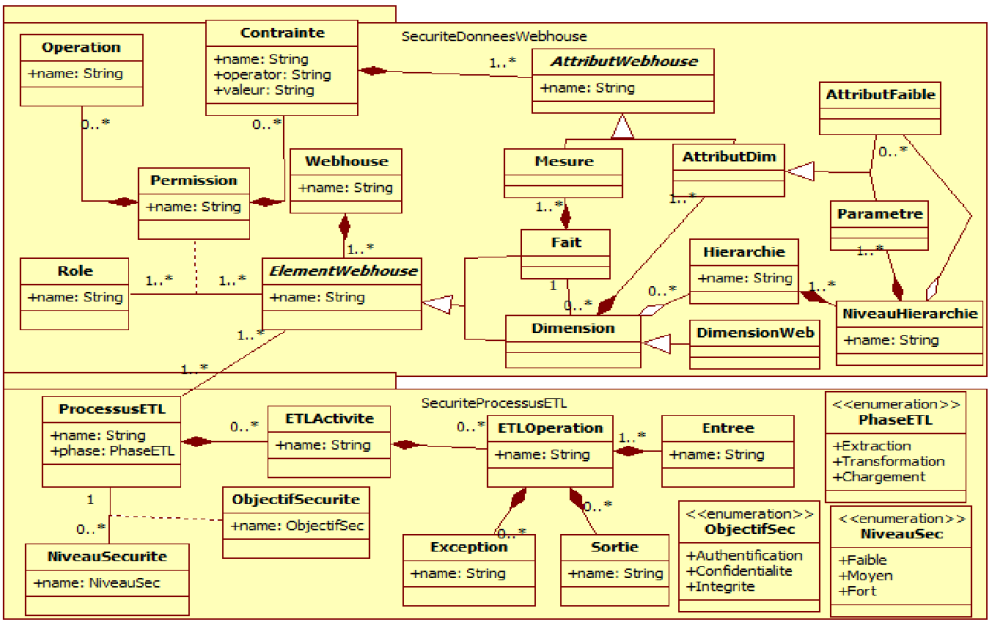


Figure 2 Méta-modèle PIM SecWebhouse

3.2 Le niveau PSM

Notre méta-modèle de sécurité webhouse au niveau PIM est traduit en une politique XACML conforme au modèle de XACML3 défini par l'OASIS Rissanen et al, (2013). Cette politique représente un ensemble de règles qui s'appliquent à des cibles représentées par des attributs formant les sujets, les ressources et les actions. En XACML, le « PolicySet » est le conteneur de différentes politiques. Chaque politique exprime un ensemble de règles permettant d'atteindre le contrôle d'accès voulu. Le « target » permet de contenir l'ensemble d'attributs traité par la règle de permission et qui contient le triplet (ElementsDWB, Role, Operation). Les contraintes, visant essentiellement à sécuriser des granularités du DWB, sont traduites en des conditions XACML exprimant des restrictions sur les attributs du Webhouse.

En outre, nous proposons la génération de la description des services ETL en utilisant la technologie WSDL. Cette description peut se référer aux descriptions WS-Policy et WS-Security Policy pour préciser les exigences liées à la sécurité. La recommandation WS-Security assure trois propriétés de sécurisation : l'authentification, l'intégrité et la confidentialité. Selon notre modèle, la politique de sécurité est composée de l'élément « Policy » qui contient l'élément « PolicyAlternative » qui contient une ou plusieurs assertions. Chaque assertion permet de réaliser un objectif de sécurité. L'attribut « SecurityLevel » permet de préciser le niveau de sécurité. En fonction de cette association entre l'objectif et le niveau de sécurité, nous générons le code de l'assertion au niveau code.

3.3 Les transformations QVT

Dans cette section, nous présentons l'ensemble des règles de transformation exprimées sous QVT-Relations (QVT-R) qui permet de spécifier de façon déclarative les transformations générant des modèles PSM à partir des modèles de sécurité du DWB.

3.3.1 Sécurité des données

Le modèle de sécurité de données est traduit en une politique XACML relative au Webhouse par la relation «WebhouseToPolicySet ». Des sous-politiques sont créés pour chaque élément du Webhouse à l'aide de la relation «ElementsWebhouseToPolicy».

```
top relation ElementsWebhouseToPolicy {fn: String; checkonly domain secw e
:secdonwebhouse::ElementWebhouse { Webhouse= w :securitedonneswebhouse::Webhouse {},
name=fn,Role=r :securitedonneswebhouse::Role{}};
enforce domain xacml p:xacml::Policy{PolicySet= ps:xacml::PolicySet{},
PolicyId='Sécurité_'+fn, Rule = pe :xacml::Rule{}};
```

L'ensemble des permissions associées à l'élément Webhouse est traduit en une règle XACML par la relation « PermissionToRule». Chaque règle comprend le triplet (ElementsDWB, Role, Operation) relatif à la correspondante permission. Ce triplet est traduit en des attributs XACML par les relations «ElementDWBToAttribut», «RoleToAttribut» et « OperationToAttribut». Chaque contrainte liée à une permission est traduite en une condition XACML par la relation «ContrainteToCondition».

```
relation ElementDWBToAttribut{en:String;
checkonly domain secw pr:secdonwebhouse::Permission{};
enforce domain xacml al:xacml::AllOf{Match= m :xacml::Match{ MatchId='string-equal'
AttributeDesignator=ade : xacml::AttributeDesignator{ Categorie='ressource', At-
tributeId='ElementWebhouse', DataType='string'}, AttributeValue = ave : xacml::AttributeValue
{Value=en, DataType='string'}}}; primitive domain prefix: String; where{en=prefix;}}

relation ContrainteToCondition{op,v:String; checkonly domain secw pr: secdon-
webhouse::Permission {Contrainte= c:secdonwebhouse::Contrainte{operateur=op, valeur=v}};
enforce domain xacml pe :xacml::Rule{ Condition= cx:xacml::Condition{ Apply=a:xacml::Apply
{FunctionId=op, AttributeValue=av:xacml::AttributeValue {Value=v, DataType='string' }}}};
where {AttributWebhouseToConditionAttribut(c, a);}}
```

3.3.2 Sécurité des services ETL

Pour chaque processus ETL, nous définissons un schéma WSDL. Les transformations QVT sont constituées de plusieurs relations, nous ne présentons que les relations liées à la génération des politiques de sécurité. A chaque processus ETL, un niveau de sécurité est associé, la relation «ProcessusETLToPolicy » permet de créer la Politique correspondante.

```
top relation ProcessusETLToPolicy { nsn:securiteprocessusetl::NiveauSec; pn:String;
checkonly domain sece ns :securiteprocessusetl::NiveauSecurite{
ProcessusETL=p:securiteprocessusetl::ProcessusETL{}, name=nsn};
enforce domain wss po:wsdl::Policy{ Definition=d:wsdl::Definition{}, name=pn,
PolicyAlternative=pa:wsdl::PolicyAlternative{ name=pn } };
when{ProcessusETLTodefinition(p,d);} where{pn=if (nsn=Faible) then 'Faible' else if (nsn=Moyen)
then 'Moyen' else 'Fort' endif endif; ObjectifSecuriteToPolicyAssertion(ns,pa, pn); } }
```

La relation « ObjectifSecuriteToPolicyAssertion » permet de définir les assertions de la politique ws-security policy. Chaque objectif de sécurité est transformé en une assertion instanciant l'attribut « SecurityLevel » précisant le niveau de sécurité associé à cet objectif.

```
relation ObjectifSecuriteToPolicyAssertion { osn:secprocessusetl::ObjectifSec; pasn:String;
checkonly domain sece ns :secprocessusetl::NiveauSecurite{ ObjectifSecurite=
os:secprocessusetl::ObjectifSecurite{name=osn} };
```

```
enforce domain wss pa:wsdl::PolicyAlternative{ PolicyAssertion = pas : wsdl::PolicyAssertion
{name=pasn, SecurityLevel=prefix}}; primitive domain prefix: String;
where{pasn=if (osn=Authentification) then 'Authentification' else if (osn=Confidentialite) then 'Con-
fidentialite' else 'Integrite' endif endif;}
```

Pour valider notre approche, nous avons développé un prototype qui implémente les transformations QVT sous medini-QVT. Ces dernières génèrent les politiques de sécurité XACML et Ws-security policy à partir du modèle SecWebhouse.

4 Conclusion

Dans cet article, nous avons présenté notre approche basée sur MDA pour la sécurité des Webhouse en traitant la sécurité au niveau des données et processus ETL. La définition des transformations QVT vise à faciliter la tâche du passage du niveau PIM vers le niveau PSM. Nous envisageons dans les travaux futurs d'étendre la sécurité des processus ETL et de terminer la phase de génération de code afin d'obtenir les politiques de sécurité.

Références

Dammak, S., Ghozzi, F., Approche de sécurisation des communications d'un webhouse, 6ème édition Atelier des Systèmes Décisionnels, ASD 2012.

Fernández-Medina, E., Trujillo, J., Villarroel, R., & Piattini, M. (2006). Access control and audit model for the multidimensional modeling of data warehouses. Decision Support Systems, 42(3), 1270-1289.

Fugkeaw, S., Mitrpanont, J. L., Manpanpanich, P., & Juntapremjitt, S. (2010). Developing Access Control Model of Web OLAP over Trusted and Collaborative Data Warehouses. In Emergent Web Intelligence: Advanced Information Retrieval. London, 2010. 393-413.

Gallino, J. P. S., de Miguel, M., Briones, J. F., & Alonso, A. (2012). Domain-Specific multi-modeling of security concerns in service-oriented architectures. In Web Services and Formal Methods. Springer Berlin Heidelberg, 2012. 128-142.

Kimball. R et R. Merz, (2000). Le DATA WEBHOUSE : analyser les comportements client sur le Web, Eyrolles Edition.

Kiran, P., Kumar, S. S., Kavya, N. P. (2012). Modelling Extraction Transformation Load embedding Privacy Preservation using UML. International Journal of Computer Applications.

Liu J., Hu.Ch. Ju,Y. HeJin (2010), Application of Web Services on The Real-time Data Warehouse Technology. Beijing, China: ICAEE, 335 – 338.

Marotta A., L. González, R. Ruggia (2012), A Quality Aware Service-oriented Web Warehouse Platform. Berlin,Germany: EDBT-ICDT, 29-32.

Mrunalini, M., Kumar, T. S., & Kanth, K. R. (2013) Secure ETL Process Model: An Assessment of Security in Different Phases of ETL. Int. J. of Software Engineering, IJSE 6(1).

Rissanen, E. (2013). extensible access control markup language (xacml) version 3.0. Retrieved August, 7, 2013.

Vela, B., Blanco, C., Fernández-Medina, E., & Marcos, E. (2012). A practical application of our MDD approach for modeling secure XML data warehouses. Decision Support Systems, 52(4), 899-925.

Contrôle d'accès aux entrepôts de données fondé sur le profil utilisateur

Amina El ouazzani*
Nouria Harbi **, Hassan Badir*

* Université Abdelmalek Essaadi ENSA LabTIC 1818 Tanger MAROC
{ a.elouazzani2000,h.badir}@gmail.com,
http://www.ensat.ac.ma/
**Université Lumière Lyon 2 Laboratoire ERIC 69635 Lyon, Cedex FRANCE
nouria.harbi@univ-lyon2.fr

Résumé. Un entrepôt de données ou Data Warehouse (DW) peut être utilisé comme un mécanisme très puissant pour découvrir les informations cruciales de l'entreprise. Il est donc important d'appliquer des mesures de sécurité qui garantissent la confidentialité des données qu'il contient. Dans ce sens, plusieurs propositions ont été présentées, néanmoins, aucune n'est considérée comme un standard dans la gestion des accès aux entrepôts de données. Dans cet article, nous allons d'abord analyser le problème de sécurité au sein d'un entrepôt et ses enjeux, ainsi, nous présenterons une architecture d'un entrepôt sécurisé qui consiste à gérer les accès en fonction des profils des utilisateurs. Cette solution a l'avantage de fonctionner d'une manière indépendante de la plate-forme cible.

1 Introduction

Les entrepôts de données occupent une place importante dans les organisations, ils sont considérés comme étant le système de soutien et l'élément clé dans les processus de prise de décisions stratégiques. Cependant, les infractions en matière de sécurité et de confidentialité des entrepôts continuent à poser une menace en raison de la grande sensibilité de l'information qui peut y être découverte, tel que des informations sur la vie privée des individus.
Dans ce cadre, plusieurs gouvernements ont promulgué des lois pour la protection des informations sur la vie privée de leurs citoyens. Parmi ces lois, HIPAA (Health Insurance Portability and Accountability Act HHS (1996)) vise à protéger les données médicales des patients américains en obligeant les établissements du secteur des soins de la santé de suivre des règles strictes de sécurité, de même GLBA (Gramm Leach Bliley Act GPO 1999) oblige les organisations financières américaines à protéger les données de leurs clients. Par conséquent, étant donné que la sécurité des entrepôts de données présente une préoccupation urgente dans la plupart des entreprises compte tenu de l'importance de l'information dont ces entrepôts regorgent, il est essentiel de définir des mesures de confidentialité.
Dans ce papier, nous nous concentrons sur la gestion des accès à l'entrepôts de données en proposant une approche de sécurité basée sur le profil de l'utilisateur qui décrit les droits d'accès à ce dernier, en utilisant la politique de contrôle d'accès RBAC (Role based Access Control)

qui se focalise sur le regroupement des utilisateurs selon leurs métiers. Avec cette approche, nous sommes en mesure de restreindre davantage l'accès des utilisateurs aux données qui leur sont interdites.
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 notre proposition à savoir une architecture d'entrepôt sécurisé dont l'accès est basé sur les profils des utilisateurs. Enfin, la section 4 présente nos conclusions et perspectives.

2 Etat de l'art des travaux existants

Récemment, un certain nombre de solutions de sécurité des entrepôts de données ont été proposés, nous avons organisés les travaux selon deux approches, la première est celle de l'intégration de la sécurité dans le processus de la modélisation des entrepôts (niveau conception), et la deuxième présente les modèles de contrôle d'accès pour un entrepôt déjà mise en place (niveau exploitation).

2.1 La sécurité dans la modélisation des entrepôts (niveau conceptuel)

Parmi les travaux qui ont été développés sur l'intégration de la sécurité dans la modélisation des entrepôts, on trouve :

Fernandez-Medina et al. (2006) ont développé un modèle de contrôle d'accès et d'audit (ACA) spécifique aux entrepôts de données, qui repose sur deux politiques de gestion des accès : MAC et RBAC. en intégrant la notion de «profil utilisateur». Ce modèle reste un modèle purement théorique car aucune solution concernant son implémentation n'a encore été proposée.

Soler et al. (2008) ont utilisé des mécanismes d'extension fournis par le CWM (Common Warehouse Metamodel) pour étendre le package relationnel et construire un schéma en étoile, qui représente les règles de sécurité et de vérification.

Soler et al. (2009) ont développé une méthodologie comprenant quatre phases : analyse, modélisation, implémentation et validation, qui couvrent les cinq niveaux d'abstraction qui sont : analyse des besoins, niveau conceptuel, niveau logique, niveau physique et l'examen post-développement, ce dernier étant une nouvelle discipline introduite par Lujan et Trujillo (2004).

Rodriguez et al. (2011) présentent une extension d'UML 2.0 du diagramme d'activité. Cette proposition, libellée comme BPSec (Business Security Process), permet de définir un ensemble d'exigences de sécurité (contrôle d'accès, détection des risques d'attaques, non-répudiation, intégrité, confidentialité et vérification de la sécurité).

2.2 L'intégration de la sécurité dans les entrepôts existants (niveau exploitation)

Le serveur OLAP (Online Analytical Processing) est censé d'assurer des accès à l'entrepôt de données en fonction des habilitations de chaque utilisateur. Cependant, le serveur OLAP tout seul ne peut pas protéger l'accès aux données interdites. Des travaux ont été réalisés pour renforcer les droits d'accès/habilitations, pour interdire tout utilisateur malicieux d'inférer des données qui lui sont interdites.

Kirkgoze et al. (1997) ont défini un modèle qui consiste à élaborer un cube personnalisé possédant ses propres dimensions et hiérarchies. Ce modèle repose sur la politique de gestion AMAC. Il s'agit d'une extension du modèle MAC qui permet de spécifier les tâches que l'utilisateur peut exécuter selon son rôle au sein de l'organisation.

Priebe et Pernul (2001) ils ont créé un mécanisme de contrôle d'accès sous forme d'un langage exprimant, il s'agit d'un langage basé sur MDX « MulDimensionnale Xpression », celuici étant un langage de requête spécialisé dans l'interrogation et la manipulation des données multidimensionnelles.

Triki et al. (2011) ont proposé une approche basée sur les réseaux Bayésiens. Pour protéger un entrepôt de données contre les inférences, elle vise à interdire à un utilisateur d'inférer des données protégées à partir des données qui lui sont accessibles en utilisant les fonctions d'agrégations Min et Max.

Eavis et Althamimi (2012) ont présenté une approche intuitive et puissante pour l'authentification de base de données qui est uniquement adaptée au domaine OLAP. Il est orienté objet et utilise des règles de réécriture de requêtes afin d'assurer l'accès aux données cohérentes.

2.3 Synthèse des travaux existants

En générale, les mesures de sécurité pour les entrepôts sont définies dans la mise en œuvre finale au-dessus des systèmes commerciaux, car il n'y a pas une norme pour l'échange et l'interopérabilité des métadonnées. Bien que la proposition du méta modèle CWM (Common Warehouse Metamodel) basé sur trois standards, à savoir UML, MOF et XML est largement acceptée comme la norme pour l'échange et l'interopérabilité des métadonnées.
Les travaux présentés sont cohérents et complémentaires. D'abord E Soler et al (2006) ont étendu le méta modèle CWM pour représenter correctement toutes les règles de sécurité et d'audit définies dans la modélisation conceptuelle des entrepôts de données Dans la suite de leurs travaux, Soler et al. (2008) et Soler et al. (2009) se sont basés sur le modèle MDA pour définir des règles de passage formelles entre le modèle conceptuel de l'entrepôt de données et le modèle logique, en exploitant le query/view/transformations (QVT) proposé par le modèle MDA. Soler et al. (2009) ont fait usage des mécanismes d'extension fournis par le CWM pour étendre le paquet relationnel afin de construire un schéma en étoile, qui représente les règles de sécurité et de vérification capturées pendant la phase de modélisation conceptuelle de l'entrepôt.

Concernant les politiques d'accès, peu d'auteurs fondent leurs travaux sur la politique DAC, bien que celle-ci soit intéressante, l'application d'une politique DAC implique un faible contrôle du flot d'information.

3 Proposition d'une approche de sécurité des entrepôts

3.1 Profil utilisateur

Le profil de l'utilisateur est une représentation des préférences et les droits d'accès d'un utilisateur individuel. Il contient les informations nécessaires pour l'authentification, et le niveau de sécurité pour accéder aux données de l'entrepôt, alors que l'entrepôt peut utiliser les détails de catégorie des données pour déterminer le contrôle d'accès.
Le modèle de contrôle d'accès à base de rôles adopté est le RBAC étendu. Il s'agit d'une politique d'accès répondant mieux aux besoins des grandes entreprises, gérant beaucoup de permissions pour un plus grand nombre d'utilisateurs. En effet, il est possible, pour certaines personnes, de cumuler plusieurs fonctions et par conséquent, plusieurs rôles. L'objectif est alors d'associer des profils aux profils, voire des rôles au rôle concerné. Il se crée une hiérarchie entre les rôles et donc une relation d'héritage. En effet, le rôle descendant hérite des permissions et restrictions du rôle ascendant.

3.2 Architecture d'entrepôt sécurisé basée sur la gestion des accès à base de profil

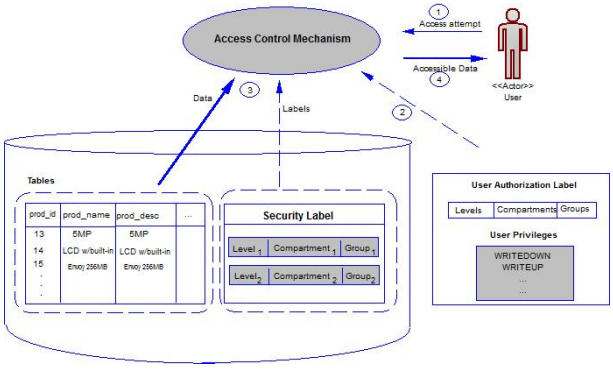


FIG. 1 – *scénario de gestion des accès à base de profil utilisateur.*

Dans l'architecture proposée, quelque soit la façon dont les utilisateurs se connectent à la table protégée (via une application, une interface Web ou SQL * Plus), le résultat est le même. Il n'y a pas de problème de sécurité des applications, puisque la politique d'accès est fixée à la table, et ne peut pas être contournée.

Pour atteindre les objectifs explicités dans l'architecture, on doit fixer les droits d'accès des utilisateurs de l'entrepôt. Comme dans les systèmes d'information, cette tâche de sécurité peut être traitée au niveau conceptuel ou au niveau logique, et les droits d'accès fixés sont appliqués par le serveur OLAP.

3.3 Le modèle d'authentification proposé

Le modèle conceptuel d'authentification proposé est chargé de vérifier les informations de l'identification de l'utilisateur. Il est constitué d'un ensemble des tables représentants les métadonnées nécessaires pour authentifier et autoriser l'utilisateur.
Par exemple, la table des utilisateurs stocke les informations de l'identification de base de l'utilisateur (login, mot de passe), tandis que la table des autorisations enregistre le fait qu'un utilisateur donné peut ou non accéder à certaines informations, tout en gardant la trace de chaque transaction effectuée par un utilisateur authentifié, durant une période autorisée indiquée dans la table des permissions d'accès.
La figure 2 illustre une version légèrement simplifiée du schéma d'autorisation de la base de données autorisation :

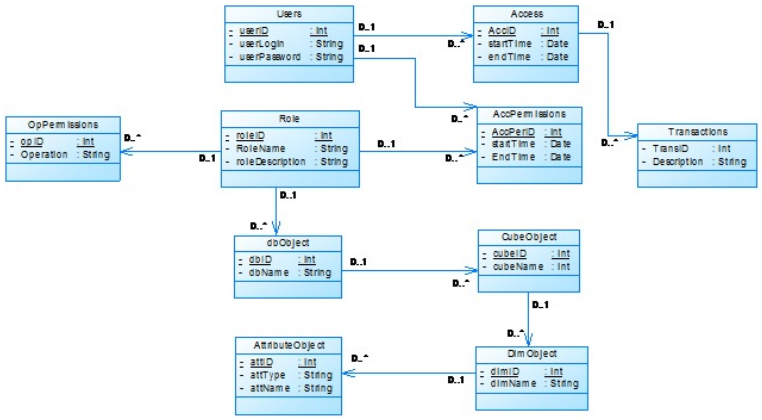


FIG. 2 – *Le modèle conceptuel d'authentification proposé.*

4 Conclusion et perspective

Dans cet article, nous avons essayés de présenter l'important rôle que jouent les entrepôts de données dans une organisation pour les prises de décisions stratégiques. Vue la sensibilité de leurs contenus, leur sécurité est une nécessité pour les entreprises. C'est pour cela que nous avons tenté de cerner les problèmes liés à la sécurité des entrepôts au niveau conceptuel et au niveau exploitation (OLAP) à travers une étude des travaux effectués dans la sécurisation des accès aux entrepôts de données.
En se basant sur ces travaux de recherche, nous avons proposé une architecture simplifiée et

fonctionnelle pour le contrôle d'accès aux entrepôts de données en adaptant la politique profil d'utilisateurs afin de garantir la confidentialité des données.
Notre travail est orienté vers l'avenir, car nous avons l'intention d'améliorer la solution proposée dans ce papier afin d'avoir une solution hybride combinant la sécurité liés aux Bases de données et surtout aux entrepôts de données et celles liées à l'environnement cloud computing.

Références

Eavis, T. et A. Althamimi (2012). Olap authentication and authorization via queryre-writing. *The Fourth International Conference on Advances in Databases, Knowledge, and Data Applications*, 130–139.

Fernandez-Medina, E., J. Trujillo, R. Villarroel, et M. Piattini (2006). Access control and audit model for the multidimensional modeling of dws. *Decision Support Systems*, 1270–1289.

Kirkgoze, R., N. Katic, M. Stolba, et A. Tjoa (1997). A security concept for olap. *Proceedings of the 8th International Workshop on Database and Expert System Applications (DEXA'97)*, 619–626.

Lujan, S. et J. Trujillo (2004). A data warehouse engineering process. *Proceedings of the 3rd International Conference on Advances in Information Systems (ADVIS'04)*, 20–22.

Priebe, T. et G. Pernul (2001). A pragmatic approach to conceptual modeling of olap security. *Proceedings of the 20th International Conference on Conceptual Modeling (ER'01) 2224*, 311–324.

Rodriguez, A., E. Fernandez-Medina, J. Trujillo, et M. Piattini (2011). Secure business process model specification through a uml 2.0 activity diagram profile.

Soler, E., J. Trujillo, C. Blanco, et E. Fernandez-Medina (2009). Designing secure data warehouse by using mda and qvt. *Journal of Universal Computer Science 8 15*, 1607–1641.

Soler, E., J. Trujillo, E. Fernandez-Medina, et M. Piattini (2008). Building a secure star schema in data warehouses by an extension of the relational package from cwm. *Computer Standards and Interfaces 30*, 341–350.

Triki, S., H. Ben-Abdallah, N. Harbi, et O. Boussaid (2011). Securing data warehouses: A semi-automatic approach for inference prevention at the design level. *1st International Conference on Model and Data Engineering, Lecture Notes in Computer Science (LNCS) by Springer-Verlag*, 311–324.

Summary

A data warehouse (DW) is a powerful mechanism that can be used to discover the critical business information, so it's important to specify security measures to ensure the confidentiality of data. In this sense, many proposals has been presented, however, any one is considered a standard in the management of access to data warehouse, in this article, we will firstly discuss confidentiality problems in data warehouse, and we present our secure architecture to manage access based on the user profile. This solution has the advantage of operating independently of the target platform , on any secure management system database.

Une approche logique de modélisation d'un moteur de règles de gestion hybride

Abdelfettah Idri* , Azedine Boulmakoul**

Laboratoire LAMIE ENCGC, Casablanca, Maroc
Département Informatique, Laboratoire Informatique de Mohammedia
Faculté des Sciences et Techniques de Mohammedia, Maroc
*abdelfattah_id@yahoo.com
**azedine.boulmakoul@gmail.com

Résumé. La complexité du processus métier ne cesse de croître s'accompagnant d'un besoin clair en termes de flexibilité et d'adaptation facile aux changements fréquents relatifs à la logique des activités métiers au sein des organismes. Le système d'information qui est sensé aider l'utilisateur à la prise de décision, nécessite de reposer sur une infrastructure logicielle lui permettant la gestion du dynamisme des règles de gestion métier. Plusieurs recherches ont été menées pour fournir des réponses à cette problématique en adoptant différents moteurs de règles et de workflow, mais une solution standard n'existe pas encore. On propose dans ce papier une approche hybride qui combine les techniques de la programmation conventionnelles et celles de la programmation logique pour offrir un socle qui s'apprête à implémenter avec un certain degré de simplicité les règles de gestion d'un processus métier qui peut être assez complexe. On présente également une étude de cas d'un organisme de santé pour illustrer le fonctionnement de notre architecture.

1 Introduction

Qu'il s'agisse de se positionner dans le marché et répondre aux critères de la compétitivité, d'offrir les services qui doivent accompagner la célérité du dynamisme des règles de gestion et du processus métier, ou de maintenir un système d'information existant, le système d'information reposant sur une programmation conventionnelle ne peut faire face à ces défis. Depuis plus d'une décennie, les chercheurs dans le domaine des systèmes d'information se sont intéressés aux approches de modélisation des règles de gestion adoptant les notions de moteurs de règles de gestion et les moteurs de workflow, Ross (1997), Ross et Lam (1998), Ross (2003), OMG (2008), Hay et Healy (2000).

D'une manière générale, on distingue deux types de règles de gestion : statiques et dynamiques. Une règle de gestion statique est une règle qui peut être vérifiée à n'importe quel moment et elle est modélisable par les structures statiques des outils de modélisation tels que l'UML et ERD (exemple de règle statique : l'utilisateur doit avoir un identifiant unique). Par contre, une règle de gestion dynamique ne peut être vérifiée qu'au moment du déclenchement d'un certain événement et elle est liée au comportement des objets métiers (exemple de règle dynamique : on ne peut arrêter un

moteur que lorsqu'il est en train de tourner). Une règle dynamique est vue comme une structure ayant une condition et une action qui se déclenche lorsqu'il y a satisfaction des conditions composant la règle dynamique. Parfois, on introduit même un déclencheur qui s'associe au contexte de la condition pour activer l'action, Terry (2005).

Plusieurs méthodes et approches ont été proposées pour modéliser et implémenter les règles de gestion métier. Quoiqu'on puisse constater l'abondance de ces outils de modélisation, la phase de spécification explicite et rigoureuse des règles de gestion reste incontournable Hebst (1996), herbst et al. (1994). Les problèmes identifiés durant la conception et l'exploitation des modèles adoptés pour les règles de gestion ont poussé les chercheurs de la communauté à fournir des solutions pour comprendre la nature des règles de gestion et de trouver des méthodes rationnelles pour les manipuler et maîtriser leurs comportements tout au long du cycle de vie d'un système d'information, Bajec et Krisper (2004). Parmi ces problèmes, on peut citer :

- La répartition des règles de gestion sur la logique applicative : il est difficile d'identifier et de tracer une règle de gestion et de déterminer son impact sur le reste des composantes du système.
- Souvent, il n'est pas évident de déterminer l'arborescence de dépendance des règles de gestions.
- Les définitions incomplètes des règles de gestion sont à la source des problèmes de divergence entre le besoin et l'implémentation de la règle de gestion.
- La plus part des organismes ne possèdent pas de documentations explicites sur les règles de gestion .
- Le changement au niveau des termes et de structure d'une règle de gestion.

L'objectif est devenu donc de trouver une représentation consistante et solide des règles de gestion qui constituera un modèle susceptible d'être implémenté et manipulé selon les paradigmes de programmation connus.

En partant du fait que le moteur des règles et celui du workflow sont destinés en fin de compte à des utilisateurs finaux qui ne sont pas sensés être expérimentés d'une part, et la spécification des règles de gestion qui peut être une tâche assez lourde d'autre part, le constat général sur la majorité des méthodes utilisées est la complexité de l'exploitation de ces outils. On peut citer à titre d'exemple la méthode de Ross qui connait même une structuration assez riche de la règle de gestion. Dans le même sens, la correspondance entre les règles de gestion et les outils adoptés offrant des interfaces graphiques n'est pas intuitive, Karami et Iijima (2010). Et comme indiqué ci-dessus, il n'existe pas encore de standard qui gère les règles de gestion et garantit la communication et l'intégration entre les différents systèmes.

L'objectif de ce papier est de proposer une approche hybride qui bénéficie des technologies existantes telles que J2EE et XML et offrir une couche d'abstraction pour modéliser les règles de gestion à l'égard des moteurs de règles basés sur la programmation logique, avec Prolog par exemple. Notre motivation pour cette approche est la difficulté recensée lors de la projection des concepts relatifs à la connaissance tels que ceux des règles de gestion originaires du monde réel et qui ne reposent pas sur une architecture conventionnelle, vers un modèle procédural de traitement d'information représenté par l'ordinateur (processeur, mémoire) et les couches logicielles (base de données, couches métier et présentation), Merritt (2004), The Why Engineer (2013). Ce document est structuré comme suit : le paragraphe 2 présente les composantes d'un moteur de règles. Le paragraphe 3 expose l'architecture du moteur

de règles de gestion proposé. Le fonctionnement du moteur de règles de gestion est traité dans le paragraphe 4. Le paragraphe 5 est réservé à l'étude de cas d'un organisme de santé. On conclut dans le paragraphe 6 avec nos suggestions et recommandations.

2 Composantes d'un moteur de règles

La figure ci-dessous montre les composantes principales d'un moteur de règles.

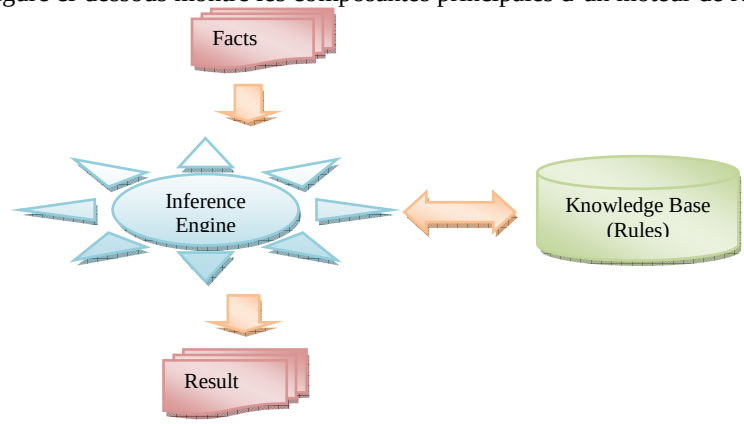


Figure 1 Architecture d'un moteur de règles

La base de connaissance englobe les définitions de l'ensemble des règles qui seront exécutées pour chaque scénario par le moteur d'inférence et qui constitueront une correspondance (*match*) pour les données formulées sous forme de faits. Le moteur suit un raisonnement déductif et son objectif est de déterminer qu'elles sont les règles à appliquer sur la base des faits disponibles. Souvent l'ensemble des règles identifiées par le moteur pour être exécutées sont insérées dans un agenda selon des critères prédéfinis. L'exécution des règles sélectionnées produira de nouveaux faits qui s'ajouteront à la base des faits initiale. Le raisonnement continuera jusqu'à ce que le résultat souhaité soit généré ou qu'aucun match n'est possible, cad, aucune des règles ne peut être sélectionnée pour exécution.

3 Architecture du moteur de règles de gestion proposé

L'objectif de notre architecture est d'exposer les règles de gestion métier sous une forme simple, flexible, fortement paramétrable et surtout directement exploitable par les utilisateurs finaux qui ne sont pas censés être des informaticiens. La structuration hiérarchique des règles de gestion comme expliqué ci-après, montre qu'à travers un simple fichier XML, l'utilisateur peut spécifier (programmer) ses règles de gestion avec une abstraction assez élevée sans se soucier du détail technique encapsulé dans l'implémentation du contrôle. Les composantes principales sur lesquelles se base la conception de l'architecture de notre moteur de règles de gestion se résument comme suit :

- Le modèle de règles de gestion : c'est une représentation hiérarchique comportant un niveau atomique inextensible qui reflète la structure d'une règle logique avec ses conditions et ses actions et que nous avons nommé « contrôle ». Au niveau de ce modèle, le contrôle est représenté juste par son interface (signature) qui définit ses arguments et ses propriétés. Le comportement du contrôle est séparé du modèle de la règle de gestion, ce qui garantit leur indépendance. En plus du niveau bas de cette hiérarchie, deux autres niveaux viennent compléter le concept de règles de gestion, notamment, le niveau des opérations simples qui s'exprime en termes de contrôles et le niveau des opérations composées qui, lui, s'exprime en termes des opérations simples ou composées. La récursivité de ces deux niveaux nous permet de spécifier une règle sur plusieurs niveaux hiérarchiques illimités mais qui en fin de compte converge obligatoirement vers le niveau de contrôles qui implémente la logique métier. Cette partie est spécifiée à l'aide d'un fichier XML dédié qui joue le rôle de la base de connaissance du système et qui devrait être défini en amont de l'exécution des opérations par le moteur des règles.
- Le modèle des faits : ce sont les événements interceptés par le système et qui sont exprimés en termes d'opérations composées décrites ci-dessus. Ces faits sont traités soit en temps réel, soit en mode différé à l'aide de scripts.
- Le moteur de règles : c'est un interpréteur de règles de gestion qui exécute chaque événement qui n'est autre qu'une opération composée équivalente à une instruction d'un programme ordinaire mais d'une manière unique du moment que la correspondance entre la règle concernée et l'opération est directe selon notre approche et aucun raisonnement sur les autres règles n'aura lieu dans ce contexte : le match effectué par le moteur est orienté opération et non pas fait.
- Le mécanisme d'exécution des règles de gestion : il s'agit de la logique dernière qui suit une démarche technique pour gérer toutes les composantes définies ci-dessus et garantir la bonne exécution des opérations.
- Puisqu'il s'agit ici d'un concept similaire à celui d'un langage de programmation procédurale dans son exécution et à un langage de programmation logique dans sa structure (contrôles), les opérations composées et simples sont enrichies par des propriétés sémantiques qui garantissent entre autres la synchronisation de l'exécutions des opérations et l'exécution itérative.
  La figure suivante illustre l'architecture décrite ci-dessus.

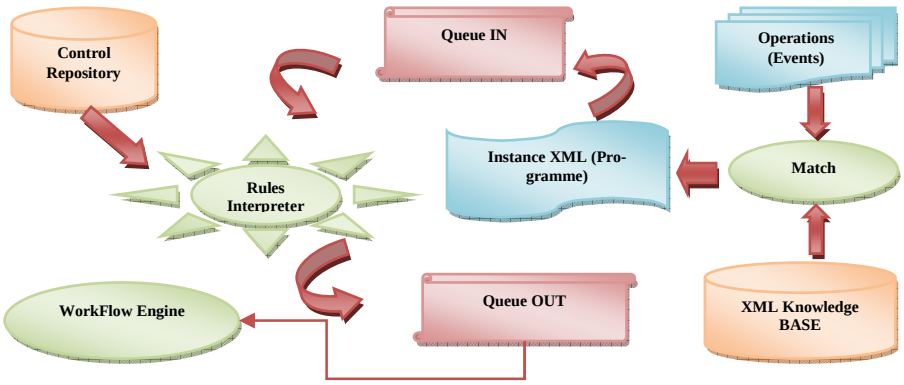


Figure 2 Architecture du moteur de règles de gestion proposé

4 Fonctionnement du moteur de règles de gestion

Selon l'architecture proposée, le moteur de règles est similaire à un interpréteur à haut niveau dont les instructions à exécuter sont constituées de l'ensemble des opérations composées encapsulées dans une instance du fichier XML. La syntaxe du langage de programmation macro est implicitement projetée sur la structure du fichier XML représenté par une instance du conteneur de la connaissance (voir la structure et la syntaxe des opérations composées dans les sections suivantes). Le fonctionnement du moteur est par conséquent décliné sur les phases suivantes décrites après l'introduction des définitions de base.

4.1 Définitions

On introduit les définitions suivantes des concepts adoptés dans notre architecture. La *Figure **3*** schématise l'hiérarchisation de ces concepts :

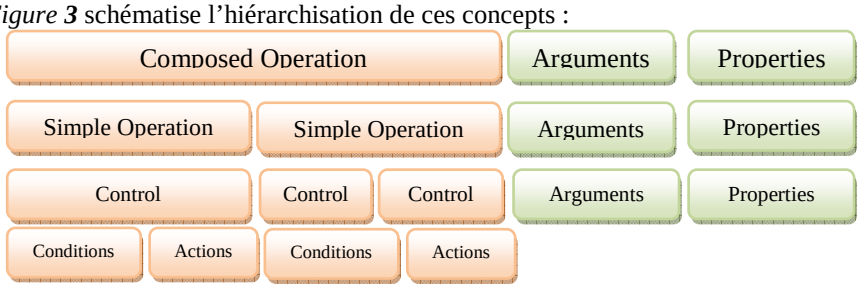


Figure 3 Hiérarchisation des opérations

- Argument : les arguments sont formés par les objets métiers regroupés en deux catégories : les entrées et les sorties connues en standard (exemple : user, card, name). Ces arguments proviennent du modèle de données spécifique au contexte du système à développer (exemple : «exploitable », « blocked », « delayed », etc.
- Condition : c'est la fonction qui s'occupe de vérifier les critères d'exécution d'une action. C'est une partie intégrante du contrôle et elle est réutilisable. La condition possède les mêmes caractéristiques que le contrôle, notamment « les arguments » et « les propriétés » et elle est implémentée en langage de programmation conventionnel.
- Action : similaire à la condition dans sa structure : elle implémente la fonction à exécuter « l'action » une fois la condition correspondante vérifiée. Elle possède des arguments et éventuellement des propriétés.
- Contrôle : représente la brique atomique possédant des arguments en entrée et en sortie, et qui implémente une fonctionnalité métier élémentaire qui est traduite vers un langage conventionnel tel que JAVA. La couche des contrôles est le socle des instructions de base de la logique métier qui est transparent pour l'utilisateur final et qui fournit une abstraction pour le niveau supérieur (opération simple). Un exemple de contrôle est celui de la création d'un utilisateur : « **create_user** (in_name, in_age, out_user) » et « **delete_user** (in_user) ». L'utilisateur final n'exploite que l'interface du contrôle pour construire sa logique métier basée sur les couches supérieures. Sa structure interne est composée de deux parties : conditions et actions. L'atomicité

permet de répondre à une situation complexe en combinant plusieurs briques élémentaires réutilisables : exemple : « If not **exist_user** (in_user)) then **add_user** (in_name, in_age, out_user) », où la première fonction est la condition et la deuxième est l'action ; ceci peut être la définition par exemple de «**create_user** (in_name, in_age, out_user) ». l'objectif du contrôle est d'implémenter les conditions et les actions et ne pas de les exécuter, c'est le moteur des règles qui se charge de cet aspect.

- Opération simple : elle est basée sur la couche des contrôles et elle est exposée aux utilisateurs finaux. L'objectif de cette couche est de traduire la complexité métier vers des combinaisons spécifiques de contrôles implémentant un processus métier. Les couches basses de cette hiérarchie alimentent cette couche soit pour implémenter définitivement le processus métier, soit partiellement dans le cas de décomposition d'un processus métier complexe en sous-processus moins complexes en se servant de la couche supérieure des opérations composées. L'opération simple possède naturellement des arguments qui sont, en fait, la réunion de tous les arguments des contrôles qui la constitue. Elle s'appuie sur les propriétés pour refléter la synchronisation entre contrôles et gérer leur flux d'exécution (exemple :**OS_Man_User** (in_name, in_age, out_user): «**create_user** (in_name, in_age, out_user) , delayed=true, exploitable=true ; **Update_user** (in_user), delayed=false, exploitable=false ; **Delete_User** (in_user), delayed=true, exploitable=true ».
- Opération composée : celles-ci est basée sur la couche des opérations simples pour fournir un niveau d'abstraction plus élevé. Elle permet de simplifier la complexité du processus métier et de gérer la récursivité du point de vue type (type récursif) des opérations, puisqu'elle peut aussi faire appel à une autre opération composée. L'opération composée possède par nature des arguments (réunion de tous les arguments des opérations simples qui la constituent) et c'est l'interface directe avec l'utilisateur final. Elle possède les mêmes caractéristiques que l'opération simple. Comme exemple, on peut citer l'affectation d'une carte à un utilisateur : **OC_Assign_Card_To_User** (in_name, in_age, out_user, in_card, out_card) : « **OS_Man_User** (in_name, in_age, out_user), delayed=true, exploitable=true ; **OS_Man_Card**(in_card, out_card), delayed=true, exploitable=true) », où l'opération simple **OS_Man_Card** peut être éclatée de la même façon que l'opération **OS_Man_User**.
- Propriétés : elles sont dédiées aux opérations simples et composées pour gérer la synchronisation, l'ordonnancement des opérations, les itérations d'exécution d'une opération, le statut d'une opération, etc. C'est le moyen adopté pour s'interfacer et servir le moteur de workflow qui n'entre pas dans le cadre de ce papier.

4.2 Chargement d'une opération composée

Deux modes sont proposés pour le chargement d'une opération composée : mode événementiel et mode batch.

- Mode événementiel : c'est la génération en temps réel d'une opération composée suite à un événement précis (action de l'utilisateur).
- Mode batch : c'est le cas d'un script contenant une liste d'opérations composées qui vont s'exécuter en arrière plan.

Le processus de chargement cible la préparation de l'opération composée pour qu'elle soit susceptible d'être exécutée par le moteur des règles. La tâche principale

est par conséquent le processus de confrontation entre la définition de l'opération composée se trouvant dans la base de connaissance et le fait ou événement représenté par l'instance de l'opération à exécuter. Cette instance est exprimée selon une convention adoptée mais qui doit contenir les éléments essentiels participant à la définition d'une opération composée. Tout ceci est dans l'objectif de générer l'instance de la définition (classe opération) de l'opération équivalente à l'instruction prête à être exécutée (objet opération). Du moment que nous avons adopté une structuration XML, ce match peut être réalisé par un outil dédié pour économiser l'effort de conception et de réalisation. Nous avons adopté à titre indicatif, l'outil JAXB qui transforme la définition de l'opération en un objet instancié prêt à l'exploitation.

Un modèle de base de connaissance XML est à voir dans la Figure 4.

```
<Operations>
    <OperationComposee Reference="OC_Assign_Card_To_User" Description=" affectation d'une carte à un utilisateur'
        Exploite="true" Succes="false">
        <ArgumentsIN>
        <ArgumentsOUT>
        <OperationsSimples>
            <OperationSimple Reference="OS_Man_User"
                Description="Création d'un nouveau utilisateur" Exploite="true"
                Succes="false">
            <OperationSimple Reference="OS_Man_Card"
                Description="gestion d'une carte" Exploite="true"
                Succes="false">
                <ArgumentsIN>
                <ArgumentsOUT>
                <Regles>
                    <Regle Reference="RG_01" Description="Contrôle 1"
                        Exploite="false" Nature="Non">
                </Regles>
            </OperationSimple>
        </OperationsSimples>
    </OperationComposee>
</Operations>
```

Figure 4 Modèle de base de connaissance XML

4.3 Exécution d'une opération par le moteur des règles de gestion

La phase d'exécution d'une opération prend en entrée une instance d'une opération composée (un objet et non pas la définition décrite dans le fichier XML) syntaxiquement correcte et procède à son traitement comme suit :

- Pour gérer le flux d'exécution et le synchroniser avec les ressources disponibles de la machine hôte (celle qui héberge le moteur), nous avons intégré dans l'architecture deux files d'attentes « Queue IN » et « Queue OUT ». La première file « Queue IN » est alimentée par le moteur de workflow à chaque fois qu'un match est vérifié entre l'opération à exécuter (l'événement) et la définition contenue dans le fichier XML (base de connaissance), et c'est le moteur de règles de gestion qui la consomme selon une priorité FIFO et le prince « Fetch, load, execute » connu au niveau du cycle d'exécution des instructions élémentaires d'un processeur dans le contexte d'une architecture matérielle. Une fois l'opération exécutée, le moteur de règle insère le résultat dans la file d'attente « Queue OUT » après avoir mis à jours tous les arguments en sortie afin de permettre au moteur de workflow à son tour de l'extraire de cette file et continuer le cycle de vie de l'opération.

- Puisqu'il s'agit d'une instruction à haut niveau, on peut se servir des techniques existantes telles que les interfaces et les classes abstraites. La décortication de l'opération devient assez simple, vu la séparation fournie au niveau contrôle entre les conditions et les actions. Le moteur commence par regrouper l'ensemble des conditions de toutes les règles de gestion (contrôles) dans une liste temporaire et de même pour les actions de telle façon à assurer un traitement transactionnel (commit / roll-back) et garantir l'exécution de l'opération composée toute entière ou son rejet intégral.
- Disposant de deux listes, l'une des conditions et l'autre des actions, le moteur procède à vérifier l'ensemble des conditions pour déduire l'indicateur global d'exécution de la liste des actions : si au moins une seule condition n'est pas vérifiée, l'opération composée est rejetée, sinon elle est exécutée.
- Dans le cas d'une validation globale des conditions, le moteur active l'exécution des actions selon les critères des propriétés adoptés par les définitions des opérations composées et simples au niveau de la base de connaissance XML.
- Dans le cas contraire, cad lorsque la non vérification d'au moins une condition empêche la validation globale de l'exécution, le moteur procède au traitement des rejets qui peuvent être définis au niveau de la base de connaissance. Ces aspects seront détaillés dans nos travaux futurs.
- Le résultat de l'exécution est soit une opération composée renseignée par les résultats calculés lors de l'exécution (arguments en sortie), soit une opération composée renseignée par les rejets générés lors de la vérification des conditions ou lors de l'exécution des actions. Dans les deux cas, l'opération résultante est insérée dans la file d'attente de sortie (Queue OUT).

5 Etude de cas : organisme de santé

Dans ce paragraphe, nous allons projeter les concepts de notre architecture et de la démarche de fonctionnement adoptée par le moteur des règles de gestion sur le cas d'un organisme de santé pour lequel un système d'information a été développé reposant sur le framework objet de ce papier.

5.1 Contexte global du projet

Il s'agit d'un système d'information qui gère la couverture médicale des adhérents. Chaque personne voulant bénéficier d'une assurance maladie, doit être inscrite au régime et en possession d'une carte d'adhérence. Dans cet exemple, nous allons nous restreindre aux règles de gestion décrites ci-dessus et on ne s'intéresse en aucun cas au traitement détaillé du métier de l'organisme.

Ce système d'information compte quelques centaines de règles de gestion et une dizaine d'opérations composées. Le mode adopté est le mode batch.

5.2 Spécifications détaillées de la base de connaissance XML

Structure du fichier batch des opérations (représentant les événements) : Le format de la description d'une opération ne pèse pas selon notre concept. Le plus

important c'est la présence des informations requises dans la base de connaissance XML. Un exemple de format du fichier en entrée (événement), qui peut bien être en format XML, est le suivant :

TypeOper	Nrdossier	Rang	Nom	Prenom	DateNaiss	LieuNaiss	Adresse	Sexe
N_OC	110	1	NOM1	PRENOM1	01/02/1970	110	ADRESSE1	M
N_OC	110	10	NOM2	PRENOM2	02/06/1990	110	ADRESSE1	M
N_OC	110	2	NOM3	PRENOM3	03/02/1972	110	ADRESSE1	F
N_OC	111	12	NOM4	PRENOM4	04/03/1980	110	ADRESSE2	M
N_OC	112	11	NOM5	PRENOM5	05/12/2000	110	ADRESSE3	M
N_OC	113	1	NOM6	PRENOM6	06/02/1970	110	ADRESSE4	M
N_OC	113	11	NOM7	PRENOM7	07/02/1970	110	ADRESSE4	M
N_OC	114	2	NOM8	PRENOM8	08/02/1970	110	ADRESSE5	F

Figure 5 Exemple de fichier des opérations (événements)

Il s'agit de l'exemple de l'opération d'une nouvelle adhésion référencée dans la figure par « N_OC ». Le reste des données, ce sont celles nécessaires pour le bon déroulement de ladite opération.

La base de connaissance XML relative à l'opération « N_OC » : Le fichier XML (voir *Figure* ***6***) suivant englobe la description globale de l'opération « N_OC » et l'infrastructure qui devrait l'accompagner : opérations simples, contrôles (règles de gestion).

```
<Operations>
    <OperationComposee Reference="N_OC" Description="Nouvelle adhesion"
        Exploite="true' Succes="false">
        <ArgumentsIN>
        </ArgumentsIN>
        <ArgumentsOUT>
        <OperationsSimples>
    </OperationComposee>
</Operations>
```

Figure 6 Structure globale XML de l'opération composée

Les arguments en entrée et en sortie sont décrits dans la *Figure* ***7***

```
<ArgumentsIN>
    <Argument>
        <Name>ALL_ARGUMENTS</Name>
        <Value></Value>
    </Argument>
    <Argument>
        <Name>Rang</Name>
        <Value></Value>
    </Argument>
    <Argument>
        <Name>Sexe</Name>
        <Value></Value>
    </Argument>
    <Argument>
        <Name>Nom</Name>
        <Value></Value>
    </Argument>
```

```
<ArgumentsOUT>
    <Argument>
        <Name>Immatriculation</Name>
        <Value></Value>
    </Argument>
    <Argument>
        <Name>Beneficiaire</Name>
        <Value></Value>
    </Argument>
    <Argument>
        <Name>DOSSIER</Name>
        <Value></Value>
    </Argument>
```

Figure 7 Structure des arguments IN et OUT

La *Figure* ***8*** montre les opérations simples composant l'opération composée « N_OC » : on trouve les deux opérations « VP_OS » pour valider une personne et « CF_OS » pour créer un foyer.

```
<OperationsSimples>
    <OperationSimple Reference="VP_OS"
        Description="Validité d'une personne" Exploite="true"
        Succes="false">
    <OperationSimple Reference="CF_OS"
        Description="Création d'un foyer" Exploite="true"
        Succes="false">
</OperationsSimples>
```

Figure 8 Opérations simples en XML

Une partie des règles de gestion (contrôles) de l'opération simple « VP_OS » est listée dans la *Figure* ***9***. (Dans la figure, « bénéficiaire veut dire une personne qui bénéficie de l'assurance maladie comme décrit au début dans le contexte du projet).

```
<OperationSimple Reference="VP_OS"
    Description="Validité d'une personne" Exploite="true"
    Succes="false">
    <ArgumentsIN>
    <ArgumentsOUT>
    <Regles>
        <Regle Reference="RG_053" Description="validité de la date de naissance"
            Exploite="true" Nature="Beneficiaire">
            <Controles>
        </Regle>
        <Regle Reference="RG_059" Description="validité de nom et prenom"
            Exploite="true" Nature="Beneficiaire">
        <Regle Reference="RG_041" Description="validité de rang et sexe"
            Exploite="true" Nature="Beneficiaire">
```

Figure 9 Structure des règles de gestion (contrôles) en XML

Le contrôle de la règle « RG_053 » référencé par « ctrl_B_Nom » est décrit dans la *Figure* ***10***.

```
<Regle Reference="RG_059" Description="validité de nom et prenom"
    Exploite="true" Nature="Beneficiaire">
    <Controles>
        <Controle Reference="Ctrl_B_Nom"
            Description="le nom est obligatoire"
            Nature="Nom_Prenom" Priorite="1" Bloquant="false" Exploite="true"
            Succes="false">
            <ArgumentsIN>
            <Conditions>
            <Actions >
            <RejetsPossibles>
        </Controle>
    </Controles>
</Regle>
```

Figure 10 Structure d'un contrôle en XML

Et finalement la structure des conditions et des actions est présentée dans la Figure 11. Nombreux sont les avantages qu'on peut tirer d'une telle architecture. Les plus importants sont listés au niveau de la conclusion.

6 Conclusions et perspectives

Nous avons introduit dans ce papier un framework hybride qui s'inspire des paradigmes de programmation conventionnelle et logique dans l'optique de proposer une

```
<Controle Reference="Ctrl_B_Nom"
    Description="le nom est obligatoire"
    Nature="Nom_Prenom" Priorite="1" Bloquant="false" Exploite="true"
    Succes="false">
    <ArgumentsIN>
    <Conditions>
        <Condition Reference="Cond_B_Nom"
            Description="le nom est obligatoire"
            Nature="Nom_Prenom" Priorite="1" Bloquant="false" Exploite="true"
            Succes="false">
            <Name>ALL_ARGUMENTS</Name>
            <Value></Value>
        </Condition>
    </Conditions>
    <Actions >
        <Action Reference="Act_B_Nom"
            Description="le nom est obligatoire"
            Nature="Nom_Prenom" Priorite="1" Bloquant="false" Exploite="true"
            Succes="false">
            <Name>ALL_ARGUMENTS</Name>
            <Value></Value>
        </Action>
    </Actions>
```

Figure 11 Structure des conditions et des actions en XML

alternative simple et pratique pour la réalisation des systèmes d'information dynamiques. Nous nous sommes focalisés dans notre travail particulièrement sur le moteur de règles de gestion. On peut citer les avantages suivants :

- La logique métier qui est normalement codée en dur dans la couche métier devient entièrement exposée aux utilisateurs finaux dans la base de connaissance représentée dans notre cas par le fichier XML lisible et bien structurée.
- La solution finale est fortement paramétrée et est susceptible d'être mise à jour à chaud sans l'intervention d'informaticiens.
- Le système est facilement évolutif et flexible aux changements
- La documentation précieuse requise par tout organisme est disponible gratuitement dans le fichier XML et peut être convertie vers n'importe quel format souhaité à l'aide de la panoplie d'outils disponibles
- Si plusieurs systèmes adoptent ce framework, la communication et l'échange d'informations peut évoluer vers l'échange d'expertise et de connaissance d'une manière très simple.
- La réutilisation des règles de gestion à tous les niveaux décrits dans ce papier devient possible.

Perspectives :

- Pour une vision plus complète, le moteur de workflow accompagnant cette architecture sera détaillé dans les futurs travaux
- La spécification de la base de connaissance XML peut être améliorée en introduisant une interface graphique supportant la syntaxe adoptée
- L'exploitation du framework dans un contexte connaissant des règles de gestion plus complexes.

Références

Ross, R. G. : The Business Rule Book: Classifying, Defining and Modeling Rules, 2nd edition, (Database Research Group. Boston. MA. 1997)

Ross, R.G. et Lam, G. (1998). Putting business rules to work: A tutorial and workshop on business rules, business tactics and policies. Chicago : Business Rule Forum, Technology Transfer Institute.

Ross, R.G. (2003) : About the Business Rules Manifesto. The Business Rule Message in a Nutshell," Business Rules Journal, Vol. 4, No. 1

OMG (2008) : Semantics of Business Vocabulary and Rules (SBVR)

Hay, D. et Healy KA. (2000) : Defining Business Rules - What are They Really? The Business Rules Group, Technical Report Revision 1.3

Terry, H. (2005) : Fact-orientation meets agent-orientation. The 6th International Bi-Conference Workshop (AOIS 2004), New York, USA, 97-109

Herbst, H. (1996) : Business Rules in Systems Analysis: A Meta-Model and Repository System. Information Systems, 21 (2), 147-166.

Herbst, H., Knolmayer, G., Myrach, T. et Schlesinger, M. : The Specification of Business Rules: A Comparison of Selected Methodologies. Methods and Associated Tools for the Information Systems Life Cycle (A. Verrijin and T. W. Olle, Ed), Amsterdam at al.: Elsevier (1994), 29-46

Bajec, M. et Krisper, M. (2004) : A methodology and tool-support for business rule management in organizations. Information Systems 30, 423–443.

Karami, N., Iijima, J. (2010) : A Logical Approach for Implementing Dynamic Business Rules, Contemporary Management Research. Pages 29-52, Vol. 6, No. 1

Merritt, D. (2004) : Best Practices for Rule-Based Application Development, , Microsoft Architect Journal

The Why Engineer (2013): *Business Rules Journal*, Vol. 14, No. 11, http://www.BRCommunity.com/a2013/b727.html

Summary

Several organizations and companies are in need of high performance information systems to deal with the complexity of their business rules especially when it becomes necessary to overcome the frequent changing requirements. Therefore, it is highly recommended to employ software components and architectures the dynamism of the business rules and processes. Nowadays, no standard solution is available for such a challenge although many alternatives are proposed by researchers and practitioners. This paper describes a hybrid approach that combines both conventional and logic programming paradigms to provide a framework that allows the implementation of complex business rules following simple steps. Moreover, the resulting knowledge base can easily adjusted even by end users. Furthermore Our business rules engine and its advantages are demonstrated with a case study of healthcare organism.

Identifying Relevant Contextual Parameters to Enhance Mobile Search Query

Sondess Missaoui*, Rim Faiz**

*LARODEC, ISG University of Tunis, Le Bardo, Tunisia
sondes.missaoui@yahoo.fr
**LARODEC, IHEC University of Carthage, Carthage Presidency, Tunisia
Rim.Faiz@ihec.rnu.tn

Abstract. Within Mobile Information Retrieval research, context information provides an important basis for identifying and understanding user's information needs. However, the challenge is how to define the best contextual information to be integrated in search process. In this paper, our intention is to build an approach to identify which contextual dimensions strongly influence the outcome of the retrieval process and should therefore be in the user's focus. We create a new measure in order to specify the relevance degree of each contextual dimensions. Our experiments show the interest of this measure to define the importance of user's current context to enhance the search process.

1 Introduction

We live in an information society and expanding technologies provide faster and broadband Internet connections which make users able to access information anywhere at any time in their daily lives. This has encouraged the use of mobile devices as one of the most important web search tools. Since, it is therefore natural to suggest new approaches of Information Retrieval (IR) in order to meet the special information needs of mobile users. Often, with mobile applications, some aspects of the user's context are available, and this context can affect what sort of information is relevant to the user. The context can include a wide range of dimensions that characterize the situation of the user. But the question is: What contextual dimensions reflect better the mobile user's need and lead to the appropriate search results? In this paper, we focus our research efforts on this area that has received less attention which is the context filtering. We have brought a new approach that has addressed this issue. How to define the relevant contextual dimension accurately and rapidly? In fact, our hypothesis is that an accurate and relevant contextual dimension is the one that provides an interesting improvement in both query profiles (Preferences and Content). Those dimensions can improve the quality of search by proposing to the user results tailored to the user's current context. The remainder of this paper is organized as follows. In section 2, we give an overview of related work which address Context-centered mobile web search. We describe in section 3, the Context Filtering approachl for mobile search. In Section 4, we discuss experiments and obtained results. Finally, section 5 concludes this paper and outlines future work.

2 Related work

The mobile users enter limited number of terms in a query. This creates a big challenge to the IR systems which called "query mismatch problem" Mario et al. (2010). So many studies integrate different context fields to enhance the query such as Bouidghaghen (2011), and especially to modelize context, allowing to identify information that can be usefully exploited to improve search results such as Tsai et al. (2010), Ahn et al. (2008), Tamine-Lechani et al. (2010), and Ingwersen and Jarvelin (2005). In fact, the Related work in the domain can be summarized in terms of two categories. Firstly, approaches which are using a set of contextual dimension to personalize all search queries. In this category, several research efforts are proposed in the literature to modelize the current user's context. Some approaches such as Chirita et al. (2006) and Gravano et al. (2003) have build models able to categorize queries according to their geographic intent. When Aréchiga et al. (2009) operate including Time and Location as main dimensions besides others to automatically infer the user's current context. The previous works propose to use a set of contextual dimensions for all queries and do not offer any context adaptation models to the specific goals of the users. In fact, only a subset of dimensions can be relevant and have the potential to influence the outcome search results Secondly, approaches that are performed to the aim of filtering the user's context and exploit only the relevant information to personalize the mobile search. This category of approaches such as Kessler (2012); Stefanidis et al. (2007) are proposed to identify the appropriate contextual information in order to tailor the search engine results to the specific user's context. In this category, our work has proceeded in terms of filtering the mobile context and identifying relevant contextual dimensions. We propose a new approach allows to define the most relevant and influential user's context dimensions (parameters) for each search situation. In the next section, we describe our definition of the context filtering problem.

3 Context filtering approach

Context filtering is the problem of identifying contextual dimensions who are eligible to encompass the user's preferences in order to enhance the the search process. The key notion of context may have multiple interpretations. In this paper, the context is the user's context which modeled through a finite set of special purpose attributes, called context parameters p_i, where $p_i \in C$ and C is the user's context defined by a set of n parameters $\{p_1, p_2..., p_n\}$. A parameter is a contextual dimension which is represented by a unique value.

3.1 Filtering features

To define the relevant contextual parameters, we will measure the query sensitivity to each one and we will select only the most influential dimensions considered as relevant parameters. Using two main features, we will measure the relevance of parameters in two ways: first, to enhance the capacity of document model in order to generate better the query (Content Profile). Second, to enhance the capacity of document model to generate the query with respect to the user's preferences (Preference Profile). In this section, we describe in details those two profiles as our filtering features. We investigate language modeling approach as described in Ponte and Croft (1998) to filter the context. We offer two types of features to classify parameters.

3.1.1 Content Query Profile

According to Diaz and Jones (2004), we infer that the best way to analyze a context parameter is to look at its effect on the query. So, its effect on the type of documents the query

retrieves (top N documents of an initial retrieval results). The context parameters can then be ranked by the probability that they "generated" the mobile query. This technics is called the "Content Query Profile". Lavrenko and Croft. (2001) define the document likelihood of having generated the query as the content query profile formally presented by the following equations:

$$P\left(Q\backslash D\right)=\prod_{w\in Q}P\left(w\backslash D\right)^{q_w} \tag{1}$$

Given a query Q and a document D, q_w is the number of times the word w occurs in query Q. According to Croft and Lafferty (2003), document language models $P\left(w\backslash D\right)$, are estimated using the words in the document. This ranking allows to build a query language model, $P\left(w\backslash Q\right)$, out of the top N documents:

$$P\left(w\backslash Q\right)=\sum_{D\in R}P\left(w\backslash D\right)\frac{P\left(Q\backslash D\right)}{\sum_{D\in R}P\left(Q\backslash D\right)} \tag{2}$$

Where R is the set of top N documents. This query language model is named “the Content Profile of the query Q”. By analogy to this content query profile, we create a new query feature in language model called "Preferences Query Profile" that helps us to define the effectiveness of the query to overcome user’s interests. This query feature will be explained as follows.

3.1.2 Preferences Query Profile

A relevant retrieved result is a ranking list which meets, in a better way, the individual user needs according to their preferences. E.g., Searching for "Music", the mobile search system must take into account the user’s preference "Jazz". Therefore, we build a preferences query profile, where the documents can be ranked by the probability that they have been generated depending on the user’s preferences which is initially defined as:

$$\hat{P}\left(Pre\backslash Q\right)=\sum_{D\in R}\hat{P}\left(Pre\backslash D\right)\frac{P\left(Q\backslash D\right)}{\sum_{D\in R}P\left(Q\backslash D\right)} \tag{3}$$

Where “Pre” is a term that describes a user preferences category from a data base containing all user’s preferences which are defined as “Pre”.

$$P\left(Pre\backslash D\right)=\{\begin{array}{cc}1 & if\,Pre\in Pre_D\\ 0 & Otherwise\end{array} \tag{4}$$

Where Pre_D is the set of categories names of interests contained in document D (e.g. Sport, Music, News, Cinema, Horoscope ...). A very helpful step is about smoothing maximum likelihood models such as $\hat{P}\left(Pre\backslash Q_{in}\right)$. We used Jelinek and Mercer process created by for smoothing. We use the distribution of the initial query Q_{in}(reference-model) over preferences as a background model. Such background smoothing is often helpful to handle potential irregularities in the collection distribution over interests. Also, it replaces zero probability events with a very small probability. Our aim is to assign a very small likelihood of a topic where we have no explicit evidence. This reference-model is defined by:

$$\hat{P}\left(Pre\backslash Q_{in}\right)=\frac{1}{|N|}\sum_{D}\hat{P}\left(Pre\backslash D\right) \tag{5}$$

Our estimation can then be linearly interpolated with this reference model such that:

$$P^{'}\left(Pre\backslash Q\right)=\lambda\hat{P}\left(Pre\backslash Q\right)+\left(1-\lambda\right)\hat{P}\left(Pre\backslash Q_{in}\right) \tag{6}$$

Given λ as a smoothing parameter. The assumption of the profile analysis is that relevant parameter provides query profiles with variance comparing to the initial query profile. Given that, this contextual parameter is important for a query and should be selected. Therefore, we chose Kullback-Leibler divergence as a probabilistic measure to define the influence of each parameter on the search results.

3.1.3 Parameter Enhancement

To compute the parameter effects on the retrieved results, we will comparing the preferences and content models yielded by the initial mobile query, to the models of the query enhanced by parameter, using Kullback-Leibler (KL) divergence. We will integrate the content background model $P\left(\backslash Q_{in}\right)$ and preferences background model $P\left(Pre\backslash Q_{in}\right)$ in a linearly interpolated score, which can be defined as the total dissimilarity between initial query and enhanced query. This score represents the query enhancement degree accomplished by the mobile contextual parameter. It can be represented as "Score KL":

$$ScoreKL\left(Q_p,Q_{in}\right)=D_{kl}\left(P\left(Pre\backslash Q_p\right),P\left(Pre\backslash Q_{in}\right)\right)$$
$$+\,D_{kl}\left(P\left(w\backslash Q_p\right),P\left(w\backslash Q_{in}\right)\right) \tag{7}$$

In order to measure the influence of contextual parameter on the mobile user's query, we build the "Parameter Relevance Metric" as described in the following.

3.2 Parameter Relevance Metric Rp

Indeed, there is no existing measurement method that allows the quantification of the mobile contextual parameter pertinence especially using a statistical property of retrieved result lists. Therefore, we propose a metric measure based on a set of components.

$$R_p\left(P,Q\right)=\frac{ScoreKL(Q_p,Q)*NbparameterTerms\left(P\right)}{NbqueryTerms\left(Q\right)} \tag{8}$$

Where Q is the mobile query, we denote the appearance of a parameter P in a mobile context C (a set of parameters) as P $\in${C}. **Score KL** is a shift between both initial and enhancing search results using contextual parameter. **NbqueryTerms** is the number of terms in mobile query which is inversely related to the Parameter Relevance measure. **Nbparameter-Terms** is the terms number per parameters. More a contextual parameter contains terms, more it is accurate and its relevance increase.

4 Experimental Evaluation

For the experiments reported in this work, we used a sample of real queries submitted to the America Online search engine AOL [1] in order to to evaluate the Rp metric. We randomly selected initial sets of 2000 queries for training, development and testing purposes. After a filtering step to eliminate duplicate and navigational queries, we obtained a set of 1300 queries. then, we simulate the user's current context for each query. Thence, we use three contextual

1. http://www.gregsadetsky.com/aol-data/

parameters (Time, Location and Nearby people). In this experiment, we evaluated the classification effectiveness of our approach comparatively to DIR approach developed by Kessler (2012). By using the DIR measure, contextual information is only classified as relevant or irrelevant using a threshold value δ. Whence, we compared the two approaches only on this basis. We implemented the DIR approach using the SVM classifier [2]. Relevant contextual information must have an impact that goes beyond a threshold value. Hence, we should obtain a high value of DIR measure to classify a context as relevant. Table 1 presents the precision, recall, F-measure and accuracy achieved by the SVM classifier according to the both approaches. The result of comparison show that, our approach gives higher classification performance than DIR approach with an improvement of 1% at accuracy. This improvement is mainly over Relevant context parameters with 1.3% at Recall.

TAB. 1 – *Classification performance on Relevant and Irrelevant parameters: comparison between our approach and DIR measure approach.*

Approach	DIR approach			Our approach Rp					
Class	Relevant	Irrelevant	Average	Relevant	Impro	Irrelevant	Impro	Avrege	Impro
Precision	1	0.968	**0.982**	1	**0%**	0.984	**1.7%**	**0.991**	**1%**
Recall	0.956	1	**0.981**	0.978	**2.3%**	1	**0%**	**0.99**	**1%**
F-measure	0.977	0.984	**0.981**	0.989	**1.3%**	0.992	**0.9%**	**0.99**	**1%**
Accuracy	**98%**			**99,5%**					**1,5%**

5 Conclusion

We proposed in this paper a filtering approachl for mobile user's context that evaluates the relevance of contextual dimensions using different features. This approach is based a new metric "Parameter Relevance measure" Rp, that allows to classify the contextual parameters according to their relevance to enhance the search results. Our experimental evaluation show the classification performance of our metric measure comparatively to a cognitively plausible dissimilarity measure namely DIR. For future work, we plan to exploit our proposed approach to personalize mobile Web search. We will customize the search results for queries by considering the determined user's contextual parameter classified as relevant.

References

Ahn, J., P. Brusilovsky, D. He, J. Grady, and Q. Li (WWW 2008). Personalized web exploration with task modles. In *Proceedings of the17th international conference on World Wide Web*, pp. 1–10.

Aréchiga, J. Vegas, and P. F. Redondo (2009). Mymose: Ontology supported personalized search for mobile devices. In *Proceedings of ONTOSE*.

Bouidghaghen, O. (2011). Nowledge based flexible query answering.

Chirita, P., C. Firan, and W. Nejdl (CIKM 2006). Summarizing local context to personalize global web search. In *Proceedings of the Annual International Conference on Information and Knowledge Management*, pp. 287–296.

2. http://www.cs.waikato.ac.nz/ml/weka/

Croft, W. B. and J. Lafferty (2003). *Language Modeling for Information Retrieval*. Kluwer Academic Publishers.

Diaz, F. and R. Jones (2004). Using temporal profiles of queries for precision prediction. *SIGIR'04 ACM*.

Gravano, L., V. Hatzivassiloglou, and R. Lichtenstein (2003). Categorizing web queries according to geographical locality. In *Proceedings of the twelfth international conference on Information and knowledge management*, pp. 325–333.

Ingwersen, P. and K. Jarvelin (2005). The turn: Integration of information seeking and retrieval in context. *Springer-Verlag Eds*.

Kessler, C. (2012). What is the difference? a cognitive dissimilarity measure for information retrieval result sets. *Knowledge and Information Systems*, 319–340.

Lavrenko, V. and W. B. Croft. (2001). Relevance-based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 120–127.

Mario, A., J. M. Cantera, P. Fuente, C. Llamas, and J. Vegas (2010). Knowledge-based thesaurus recommender system in mobile web search.

Ponte, J. M. and W. B. Croft (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in Information Retrieval*, pp. 275–281.

Stefanidis, K., E. Pitoura, and P. Vassiliadis (2007). Adding context to preferences. In *Proceedings of the 23rd International Conference on Data Engineering (ICDE)*.

Tamine-Lechani, L., M. Boughanem, and M. Daoud (2010). Evaluation ofcontextual information retrieval effectiveness: overview of issues and research. *Knowledge Information Systems*, 1–34.

Tsai, F. S., X. Xie, W. Lee, and Q.Yang (2010). Intro for mobile ir mobile information retrieval: A survey. *In European Journal of Scientific Research ISSN 55*, 394–400.

Résumé

"La Recherche d'Information contextuelle dans un environnement mobile" devient un domaine de recherche de plus en plus évolutif et prometteur. Ceci est dû à la popularité croissante des supports mobiles d'accès à l'information. Ces dispositifs permettent de spécifier des informations tels que ; localisation, temps, profils,... Qui constituent les paramètres d'un contexte multidimensionnel de recherche. Mais la mobilité de l'utilisateur fait de sorte que les valeurs de ces paramètres se changent rapidement et influencent les résultats de recherche différemment d'une requête à une autre. Notre objective dans ce travail est de préciser leurs pertinences et savoir lesquels à intégrer à fin de personnaliser la recherche. Notre approche se base sur la construction d'une mesure pour la classification des paramètres contextuels, nommé "Parameter Relevance Measure"(Rp), selon leur capacité à enrichir les résultats de la recherche pour une requête donnée.

Warehouse design approaches: A survey and a new insight based on business entities

Imen Jellali[1], Mounira Ben Abdallah[1], Nahla Z Haddar[1], HanêneBen-Abdallah[2]
[1]MIR@CL Laboratory, University of Sfax, Tunisia
{imen_ij, mounira_kriaa, nahla_haddar}@yahoo.fr
[2]King Abdulaziz University, Jeddah, KSA
hbenabdallah@kau.edu.sa

Abstract. Business intelligence applications, such data warehouse applications, transform raw data into useful information for improving decision-making and identifying new business opportunities. Current business intelligence application and most researches on enterprise data analysis do not exploit the strong relationship between the business data and the business processes of an enterprise. They consider exclusively either process warehouses or data warehouses. Only few researches propose the integration of business data and business processes as a solution to some decisional problems which cannot be solved by data or process warehouses separately. This integration is somehow problematic because business data and process data are stored separately. In this paper, we first present a survey of approaches to design data warehouses, process warehouses as well as the integration of business data and business processes. Secondly, we propose a new warehouse approach that integrates data and processes in a natural way based on business entities.

Keywords: data warehouse, process warehouse, integration, entity warehouse.

1 Introduction

The goal of business intelligence is to analyze an enterprise's performance, to control the achievement of its goals and to increase its competitiveness. Applications provided by business intelligence, such as data warehouses (DW), transform raw data into useful information for improving decision-making and identifying new business opportunities. These applications typically use one of two data types present within an enterprise: either the data issued from the information system or the data related to the business process of the enterprise. Despite the strong relationship between both data types, they are analyzed separately.

Indeed, for more than a decade, several approaches have been proposed to design a DW for data issued from information systems. These approaches provide for data analysis by organizing the data into some multi-dimensional form where the analytical data is integrated from different sources. However, the so-far proposed multi-dimensional models (Cabibbo et Torlone, 1998) (Romero et Abelló, 2007) have two weaknesses regarding to business process perspectives. First, because they are constructed regardless of the business goals, they are far from covering the enterprise goals. This weakness motivated some researches (Leal et al., 2013) (Bargui et al., 2011) to look into the construction of DW models based on business goals and the design of DWs from a business perspective in order to cover all the enterprise

goals. A second weakness of the data-based DW multi-dimensional models is their failure in providing sufficient information about business processes. This weakness motivated some researchers (Neumuth et al., 2008)(Mansmann et al., 2007) to propose the concept of *process warehouse* (PW) in order to obtain an adequate basis for a detailed analysis of business processes. Although these approaches provide sufficient information about organizational, functional and behavioral aspects, they neglect information about behavioral changes of data or processes during their execution.

Only few researches (Beltran et Ravat, 2009)(Stefanov, 2007) propose the integration of business data and business processes as a solution to some decisional problems that cannot be solved by data or process warehouses separately. It is worth noting that the integration of both data types is somewhat problematic because business data and process data are stored separately. Indeed, the separation causes information losses about organizational, behavioral, functional and/or informational aspects. However, because it is not always possible to control the achievement of enterprise goals, designing a warehouse that stores business "entities" covering all the perspectives in a natural way is a promising solution. This motivated us to propose a new concept of *warehouse based on such entities which we call entity warehouse*. More specifically, we present the new concept of entity warehouse and we outline a modeling approach for it.

To highlight the need for these contributions, we start by setting up a framework to evaluate existing DW and PW approaches in terms of business perspectives; the framework consists of the five perspectives: functional, behavioral, organizational, informational, and goal perspectives. Afterwards, we overview researches that integrate business data and business processes. Finally, we define the concept of business entities and we sketch an approach for the design of a business entity warehouse.

2 Warehouse analysis framework

In this section, we present a warehouse analysis framework that is used in the next sections for the evaluation of warehouse design approaches. Our analysis framework, inspired from (Curtis, 1992), consists of the four perspectives used when modeling business processes: functional, behavioral, organizational and informational perspectives.

The functional perspective considers the elements that compose a process including activities, flows, sub-process… It includes three parameters (List and Korherr, 2006):*Activity analysis(AA)* which evaluates performance and execution of activities; *Flows of informational entities(FA)* which gives information about flows between elements of a process and elements involved in the flows; and *Sub process analysis (SA)* which informs about the collaboration between the sub-processes of the main process to achieve a goal. To these parameters, we add a new parameter called *process analysis (PS)* to evaluate approaches that analyze processes as subjects regardless of their details.

The informational perspective considers the resources manipulated in a process. This perspective includes three parameters (List and Korherr, 2006); *Input analysis (IA)* to assess the information required to trigger a process; *Consumption analysis (CA)* to appraise the resources consumed during a process execution; and *Output analysis (OA)* to evaluate the number of times a process was successfully executed. Because it is interesting to highlight the structure of data in the analysis, we add a new parameter to this perspective called *data structure (DS).*

The organizational perspective considers the responsible for executing process activities. This perspective includes four parameters (List and Korherr, 2006)*: Resource analysis (R)* to answer questions about resources available and consumed; *Organizational unit analysis (OU)* to consider the associated unit with the processes; *Participant Analysis* to assess participants associated with a process and the number of processes associated with a participant; and *Software or service analysis (SSS) to estimate* the role of each software associated with a process.

The behavioral perspective considers the flows of data and activities within a process. It includes five parameters (List and Korherr, 2006): *Execution order analysis (EO)*to evaluate the execution order of elements; *Cycle time analysis (CT)* to assess period of time consumed by each process, start time and stop time of a process; *Anomalous behavior analysis (AB)* to consider anomalies in execution of a process; *Path analysis (P)* to appraise path followed in parallel flows against an event; and *Deadline analysis (D/E)* to evaluate number of deadlocks occurred during process execution.

To the above four process modeling perspectives, we add a fifth perspective called the *goal perspective*. It represents the extent to which the goal of a model is achieved. It is useful to evaluate the satisfaction of enterprise goals by a warehouse.

3 Data warehouse design approaches

The DW concept was formalized for the first time in 1990 by Bill Inmon (Inmon, 2002). It is a database built for analytical processing whose primary objective is to maintain and analyze historical data. A *DW* is an integrated and time-varying collection of data derived from operational data and primarily used in strategic decision making by means of online analytical processing OLAP (OnLine Analytical Processing) techniques. Several approaches are proposed to design DWs. We can classify the studied approaches into two categories: pure DW approaches and DW approaches based on goals.

3.1 Pure data warehouse approaches

This category of DW approaches focuses purely on transactional data stored in the warehouse regardless of their manipulating processes; organizational units… Some approaches start from a detailed analysis of the data sources from which the designer must select relevant blocks of data to decision making. The selected blocks are organized in a DW model regardless of any analytical needs (Golfarelli et al., 1998) (Husemann et al., 2000) (Cabibbo et Torlone, 1998) (Romero et Abelló, 2007). Other approaches are based on requirements collected from decision makers (Kimball et Ross, 2002) (Mazón et al., 2005) (Giorgini et al., 2005). Finally, some mixed approaches consider both data sources and requirements collected from decision makers (Bonifati et al., 2001) (Giorgini et al., 2008).

All pure DW approaches fail to meet business objectives. Indeed, they produce multi-dimensional models regardless of the needs of the business goals. Consequently, they do not provide information about the business processes which manipulate the data. In addition, they do not address the satisfaction of the enterprise's goals.

3.2 Data warehouse approaches based on goals

This category of approaches adds new information to the multi-dimensional models constructed by the first category: the goals that an enterprise wants to achieve on the one hand, and the processes manipulating the data taken as a dark box, on the other hand. Some researches propose to design a DW from a business perspective and to derive the DW model based on business goals (Leal et al., 2013) (Bargui et al., 2011). These works start from business goals to be sure that the designed model covers all the goals of the enterprise. Although these approaches take into account business processes, the produced models are informational and do not provide sufficient information about business processes. Indeed, they try to analyze processes as subjects regardless of its details (sub-processes, activities, flows…). Consequently, the functional perspective is only partially covered.

All the approaches studied in this category lack information about actors, resources, organizational units, services…As a result, it is not possible to analyze data with respect to the resources manipulated by processes, the actors, organizational units and services associated to processes. For example, it is not possible to detect a shipping delay caused by an actor when analyzing the performance of a sales process. Consequently, the organizational perspective is not covered. In addition, these approaches do not provide either information about the lifecycle of business objects such as execution order, cycle time, behaviors...Therefore, it is difficult to recognize the presence of an anomalous behavior or a deadlock, to determine the execution order of activities and the cycle time associated to each one. In other words, the behavioral perspective is not covered by these approaches.

Furthermore, all studied approaches deal with data analysis by providing details on the data structures to analyze them. Despite the efforts provided to design data structure, these approaches neglect information about organizational and behavioral aspects. In addition, they do not provide an adequate basis for a detailed analysis of business processes, because they do not capture information about ongoing activities from a process perspective (Casati et al., 2007)(Shahzad, 2012). For example, information about activities, their starting time, ending time and status of activities is not captured in a DW is missing. As a consequence, judicious decisions cannot be taken on potential process improvements using data-oriented warehouses as information sources (Casati et al., 2007)(Sayal et al., 2002)(Shahzad, 2012).Moreover, these approaches do not control the achievement of the enterprise goals.

4 Process warehouse approaches

The *process oriented data warehouse* concept, or simply *PW*, was introduced as an alternative to traditional DW (List et al., 2000) to satisfy the needs of decision makers. A PW is defined as a separate read-only analytical database that stores data about executed business processes including information about actors, executed activities, execution time and frequency of these activities. It is used for analysis and improvement of business processes (Casati et al., 2007) (List et al., 2000).

The main difference between a DW and a process warehouse resides at two levels: the *structural level* (Connolly et Begg, 2005) and the *architectural level* (Wingenious, 2005). Indeed, a DW is designed to support subject-oriented analysis of an enterprise and it is populated from operational sources. But a process warehouse is designed to support analysis

of process performance and it is populated with business process execution data and process related information (Schiefer et al., 2003).

A great number of studies were proposed to design process warehouses. Several researches carry out a top–down modeling of the process warehouse by starting from an operational database schema which represents the recording schema of a process (activities, flows….). This schema may be an E/R model (Neumuth et al., 2008), a UML class diagram (Mansmann et al., 2007), a workflow meta-model (Eder et al., 2002) or a workflow audit trial data (Pau et al., 2007). Other approaches propose a bottom up approach by starting with user analytical needs. For example, Niedrite et al propose an approach based on GQ(I)M (Goal Question Indicator Measurement) to define goals and questions to identify indicators (Park et al., 1996). The authors start from business goals and analyze them to produce sub-goals and measurement goals. Questions (which can identify achievement of goals) are developed for measurement goals. Later, these questions are used to identify indicators (Niedrite et al., 2007)**.**These indicators are expressed in OCL[1](Object Constraint Language) using entities and attributes of the model. Then, based on the structure of the OCL expressions, potential facts and dimensions of the process warehouse are identified. Some other works don't propose a formal methodology for PW design (Kueng et al., 2001), (Schiefer et al.,2003), (Casati et al., 2007). In (Kueng et al., 2001) *for example*, the authors developed a performance PW with the aim to facilitate BP improvement. In (Schiefer et al.,2003), the authors introduced architecture for BP monitoring based on a process data store. In (Casati et al., 2007), the authors develop a generic warehouse for BP models. The authors in (Beltran et Ravat, 2009) propose a simplified view of the process, called context warehouse, based on business data and BPs. The warehouse stores information about processes in their production context. Indeed, they consider that the production context of a process is more interesting for BP analysis than too detailed process description. They argue that the latter would be superfluous for decisional analysts. However, the approach used to design the context warehouse is ambiguous and the resulting warehouse does not store decisional data and does not provide any information about organizational or behavioral aspects. In fact, the main role of the context warehouse is to explain how the business data are produced by relating this context warehouse to an existing DW. So, the context warehouse is dependent on the DW.

In (Shahzad, 2012), the author proposes a PW consisting of two parts: a stable and a case-specific part. The stable part stores information about goal structure and its relationship with PW structure, such as satisfaction conditions, indicators, and goal related dimensions. The case-specific part captures the dimensions and facts about a process that are essential for performance analysis of processes. This part is dynamic in the sense that a data model is developed for each process i.e. the dimensions and facts identified for a process can be different from those of another process.

Most of the studied approaches operate on more or less detailed information about functional, behavioral and organizational BP perspectives. For example, some works fully cover the functional perspective (Mansmann et al., 2007) (Schiefer et al., 2003). Others lack information about sub-processes (Neumuth et al., 2008)(Pau et al., 2007)(Niedrite et al., 2007)(Casati et al., 2007).Concerning the behavioral perspective, we find a full coverage of it in (Casati et al., 2007). Indeed, this approach presents enough information about execution order, cycle time, deadlocks … Other approaches partially cover this perspective. For example, (Neumuth et al., 2008) considers execution order and cycle time. Also, most of

[1] OCL 2.0 Specification http://www.omg.org/spec/OCL/2.0/

mentioned approaches provide sufficient information concerning the organizational aspect including participants, organizational units, resources and services (Mansmann et al., 2007), (Eder et al., 2002), (Pau et al., 2007) and (Kueng et al., 2001).

On the other hand, most of the approaches cited above consider the resources manipulated by a process including input, output and consumed resources. But all the studied works do not consider the source data structure. As a result, it is not possible to identify the influence of processes on data because they do not keep track of how process activities change data during their execution. Moreover, only some works consider the satisfaction of the enterprise's goals (Niedrite et al., 2007)(Kueng et al., 2001).

To summarize, all the proposed studies to design process warehouses provide an adequate basis for a detailed analysis of business processes. They also provide information about the organizational, behavioral and/or informational aspects especially information about the resources manipulated in the processes. However, this information is not integrated and all perspectives are not fully covered by one approach. Furthermore, proposed approaches lack information about the source data structure.

5 Integration of DW and business processes

DWs and PWs, taken separately, are considered as insufficient for decision making because they could not meet the expectations of decision analysts. Indeed, each of them focuses on one part of the business in isolation. However, decision makers need a global view on the enterprise in order to analyze the influence of the process on data and vice versa. In this context, some researches propose to integrate business data and business processes as a solution to some decisional problems that cannot be solved by data or process warehouses separately (Beltran et Ravat, 2009)(Stefanov, 2007)(Chowdhary et al., 2006).

In (Beltran et Ravat, 2009), the authors propose the integration when analyzing the enterprise performance. They use a set of contextual rules to connect a DW that stores decisional data to a context warehouse containing production contexts. This integration allows the analysis of business data with their production context. However, this approach lacks information about the business processes. In fact, no information is provided concerning the organizational aspects, data behavior changes, etc.

In (Stefanov, 2007), Stefanov et al. introduce an integrated view of DW and enterprise models using model weaving (Bézivin et al., 2005). The model weaving approach was also used in this work to make the relationship between the DW and the enterprise goals more visible and accessible. This work focuses on providing this information for users within an analysis tool to support and improve data interpretation. However, the details about what extensions should be made to DW to capture information about goals and their relationship are not discussed. The drawback of this approach is that the integration is proposed when analyzing performance, not at design level. So, if the data structure of a DW or business models is changed, then some links can become invalid.

In (Chowdhary et al., 2006), the authors present a Model Driven Data Warehousing approach in the area of Business Performance Management with the aim to bridge the gap between BP models and DW models. This approach aligns the DW with business processes. Hence, it enables the adaptation to changes in the business environment. However, it does not provide information about organizational units, actors, behavior changes of business processes and business data.

All the works studied above in this section propose the integration of business processes and business data, which are stored separately, when analyzing enterprise performance. Most of them use bindings to link them. These links cause several problems. Indeed, they do not guarantee that all the relevant details about organizational, functional, behavioral, informational aspects are fully covered. In addition, most of them do not control the satisfaction of the enterprise goals.

Table TAB. 1 summarizes the comparison of warehouse approaches studied above according our warehouse analysis framework.

	Perspectives	Functional				Behavioral					Organizational				Information				Goal
	Approaches	AA	FA	SA	PS	EO	CT	AB	P	D/E	R	PA	OU	SSS	IA	CA	OA	*DS*	
DW approaches	(Golfarelli et al., 1998)																	×	
	(Husemann et al., 2000)																	×	
	(Cabibbo et Torlone, 1998)																	×	
	(Romero et Abelló, 2007)																	×	
	(Kimball et Ross, 2002)																	×	
	(Mazón et al., 2005)																	×	
	(Giorgini et al., 2005)																	×	
	(Bonifati et al., 2001)																	×	
	(Giorgini et al., 2008)																	×	
	(Leal et al., 2013)	×									×				×		×	×	×
	(Bargui et al., 2011)				×													×	×
PW approaches	(Neumuth et al., 2008)	×	×			×	×				×	×	×		×				
	(Mansmann et al., 2007)	×	×	×		×	×		×		×	×	×	×	×	×	×		
	(Eder et al., 2002)	×		×		×	×	×				×	×	×			×		
	(Pau et al., 2007)	×				×	×			×		×				×	×		
	(Niedrite et al., 2007)	×	×				×				×	×			×		×		×
	(Kueng et al., 2001)	×		×			×			×		×	×						×
	(Schiefer et al., 2003)	×	×	×		×	×	×	×		×	×	×				×		
	(Casati et al., 2007)	×	×			×	×	×	×	×	×	×	-	-	×	×	×		
	(Beltran et Ravat, 2009)	×				×											×		
	(Shahzad, 2012)	×	×	×	×		×				×	×				×	×		×

Integration	(Beltran et Ravat, 2009)	×				×											×	×	
	(Chowdhary et al., 2006)	×	×	×	×					×	×	×	×					×	
	(Stefanov, 2007)	×	×	×		×					×	×	×	×			×	×	×

TAB. 1- *Evaluation of data and process warehousing approaches.*
AA : *Activities Analysis* | FA : *Flows of informational entities Analysis* | SA : *Sub process Analysis* | PS : *Process Analysis* | EO : *Execution Order* | CT : *Cycle Time* | AB : *Anomalous Behavior* | P : *Path* | D/E : *Deadlock / Exception* | R : *Resource* | Pa : *Participant* | OU : *Organizational Unit* | SSS : *Software System & Services* | IA : *Input Analysis* | CA : *Consumption Analysis* | OA : *Output Analysis* | DS : *Data Analysis* (see section 2)

6 Towards a business entity warehouse

Separate DW and PW analysis could not meet the expectations of decision analysts because it does not provide consistent information about facts and the process that generates them. As discussed in the previous section, some researches propose the integration of business data and business processes as a solution to some decisional problems which cannot be solved by data-or process warehouses separately. This integration is somehow problematic because business data and process data are stored separately and relationships must be established between them.

For example, in the field of business management, decision makers need all information relating to orders, invoices including details on business objects, processes that manipulate them, organizational units that handle them, …, in order to detect anomalous behavior, delayed activities and to find out the responsible of the delays. Therefore, designing a warehouse that stores Business Entities (BEs) covering all the perspectives in a natural way is promising. Recently, a new approach to BP modeling, called *Entity-centric business process modeling* (Nandi, et al., 2010),was proposed following the emergence of the need for reconciliation between business goals, business operations and business data. This approach combines BEs, their life cycle, their informational model as well as business activities that manage this model. Indeed, business entity includes both an information model for data about the business objects during their lifetime, and a lifecycle model, which describes the possible ways and timings that tasks can be invoked and performed on these objects. An example of a BE type is Courier Shipment, whose information model would include attributes for package ID, sender, recipient, shipping method, arrival times, delivery time, and billing information. The lifecycle model would include the multiple ways that the package could be delivered and paid for, and would be used in tracking each instance of the Courier Shipment BE type. Other examples of BEs are Claim in an Insurance Claims process, going through the states of Filed, Approved, Fulfilled, and so on; Trouble Ticket for a Services Delivery process, going through the lifecycle states of Opened, Assigned, Rejected; financial Deal in a loan-giving organization, going through the lifecycle states of Draft, Offered, Signed, Active, and so on.

In much of the literature to date on BEs, the lifecycle models are specified using finite state machines, where each state of the machine corresponds intuitively to a business-relevant milestone, or operational objective, that might be achieved by a BE instance. BEs define a useful way to understand and track business operations, such as the locations that the package has passed through and its arrival times, and the distribution of timings (for

example, how many two-day air shipments took longer than two days in the last week) and ways of handling (for example, what percentage of cash-on-delivery shipments required more than one delivery attempt), which are useful for monitoring, dashboards, and more broadly, business intelligence. More generally, BE types can provide a unifying basis for understanding many aspects around the operations of an enterprise, including requirements gathering, business rules, compliance, and process user interactions (Casati et al., 2007), (Bargui et al., 2011), (Stefanov, 2007). Because business operations models based on BEs can be implemented in a reliable manner (for example, (Bonifati et al., 2001), (Chowdhary et al., 2006)) this approach opens the door to the development of a variety of BPM (Business Process Management) tools that are focused on the common basis of BEs.

The entity warehouse approach allows analyzing both business and transactional information during execution of the system to meet some analytical purposes. Consequently, designing a warehouse storing such entities guarantees the coverage of all BP perspectives (behavioral, functional, organizational and informational).

In this paper, we propose a new approach to design a *warehouse based on such BEs which we called entity warehouse*. Since operational systems are used as input when designing warehouses and most of enterprises' operational systems do not provide the BE concept, so it is necessary to go through a first step to design BEs. Indeed, we assume the existence of some models such as information system models, BPM models, BP logs, a business motivation model. These models usually exist since the enterprise creation. For such enterprises which do not use the concept of BE, we start from these models to design these entities as shown in figure 1. In the second step of the approach, we exploit the entities produced in the first step to design a model of the *entity warehouse* as shown in FIG. 1.

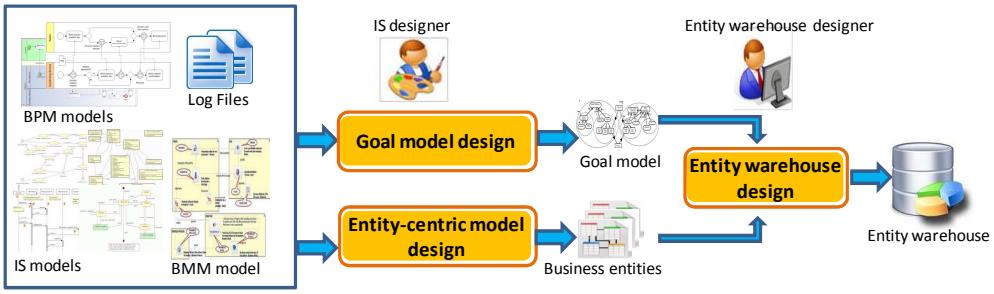


FIG. 1 – *Entity warehouse modeling approach*.

7 Conclusion

In this paper, we presented a warehouse analysis framework consisting of five perspectives: functional, behavioral, organizational, informational, and goal perspectives. Then, we studied warehouse approaches according to this framework: DWs, PWs and the integration of business data and business processes. Finally, we presented a new approach of warehouse which integrates data and processes in a natural way based on BEs.

Regarding future work, we are currently detailing the BE design process of enterprise models. Subsequently, we intend to detail the modeling approach of an entity warehouse.

Reference

Bargui, F., Ben-Abdallah, H., & Feki, J. (2011). A domain ontology based approach for analytical requirements elicitation. *proceeding of: 10th International Conference on Information and Knowledge Engineering (IKE)* , 29-35.

Beltran, T., & Ravat, F. (2009). Architecture pour la prise en compte des contextes dans les entrepôts de données. *Actes du XXVII° congrès INFORSID* .

Bézivin, J., Jouault, F., Rosenthal, P., & Valduriez, P. (2005). Modeling in the Large and Modeling in the Small. (U. Aßmann, M. Aksit, & A. Rensink, Eds.) *Model Driven Architecture, 3599*, 33-46.

Bonifati, A., Cattaneo, F., Ceri, S., Fuggetta, A., & Paraboschi, S. (2001). Designing Data Marts for Data warehouses. *ACM Transactions on Software Engineering and Methodology (TOSEM), 10* (4), 452 - 483.

Cabibbo, L., & Torlone, R. (1998). A Logical Approach to Multidimensional Databases. *VIth International Conference on Extending Database Technology (EDBT 98), 1377*, 183–197.

Casati, F., Castellanos, M., Dayal, U., & Salazar, N. (2007). A generic solution for warehousing business process data. *Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB'07)* , 1128-1137.

Casati, F., Castellanos, M., Dayal, U., & Salazar, N. (2007). A generic solution for warehousing business process data. *Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB'07)* , 1128-1137.

Chowdhary, P., Mihaila, G., & Lei, H. (2006). Model Driven Data Warehousing for Business Performance Management. *IEEE International Conference on e-Business Engineering ICEBE '06* , 483 - 487.

Connolly, T. M., & Begg, C. E. (2005). *Database Systems: A Practical Approach to Design, Implementation and Management* (4th Edition ed.). USA: Addison Wesley.

Curtis, B., Kellner, M. I., & Over, J. (1992). Process Modeling. *Communications of ACM, 35* (9), 75-90.

Eder, J., Olivotto, G. E., & Gruber, W. (2002). A data warehouse for workflow logs. *EDCIS '02 Proceedings of the First International Conference on Engineering and Deployment of Cooperative Information Systems* , 1-15.

Giorgini, P., Rizzi, S., & Garzetti, M. (2008). A goal-oriented approach to requirement analysis in data warehouses. *Decision Support Systems (DSS) jounal, 45* (1), 4-21.

Giorgini, P., Rizzi, S., & Garzetti, M. (2005). Goal-oriented requirement analysis for data warehouse design. *Proceedings of 8th ACM International Workshop on Data warehousing and OLAP (DOLAP'05)* , 47 - 56.

Golfarelli, M., Maio, D., & Rizzi, S. (1998). Conceptual Design of Data Warehouses from E/R Schemes. *Thirty-first Hawaii International Conference On System Sciences, 7*, 334 - 343.

Husemann, B., Lechtenborger, J., & Vossen, G. (2000). Conceptual Data Warehouse Design. *Design and Management of Data Warehouses* .

Kimball, R., & Ross, M. (2002). *The Data Warehouse Toolkit* (Second Edition ed.). New York: John Wiley & Sons, Inc.

Kueng, P., Wettstein, T., & List, B. (2001). A holistic process performance analysis through a performance data warehouse. *7th Americas conference on information systems (AMCIS'01)* , 349-356.

Leal, A. C., Mazón, J.-N., & Trujillo, J. (2013). A business-oriented approach to data warehouse development. *Ingeniería e Investigación Journal, 33* (1), 59-65.

List, B., & Korherr, B. (2006). An evaluation of conceptual business process modelling languages. *Proceedings of ACM Symposium on Applied Computing (SAC'06)* , 1532-1539.

List, B., & Korherr, B. (2006). An evaluation of conceptual business process modelling languages. *Proceedings of ACM Symposium on Applied Computing (SAC'06)* , 1532-1539.

List, B., Schiefer, J., AM, T., & Quirchmayr, G. (2000). The process warehouse: a data warehouse approach for business process. (W. Abramowicz, & J. Zurada, Eds.) *Selected Aspects of Knowledge Discovery for Business Information Systems* .

Mansmann, S., Neumuth, T., & Scholl, M. H. (2007). Multidimensional data modeling for business process analysis. *Proceedings of 26th International Conference on the Entity Relationship Approach (ER'07), 4801*, 23-38.

Mazón, J.-N., Trujillo, J., Serrano, M., & Piattini, M. (2005). Designing Data Warehouses: From Business Requirement Analysis to Multidimensional Modeling. (K. Cox, E. Dubois, Y. Pigneur, S. J. Bleistein, J. Verner, A. M. Davis, et al., Eds.) *1ST INTERNATIONAL WORKSHOP ON REQUIREMENTS ENGINEERING FOR BUSINESS NEED AND IT ALIGNMENT* .

Nandi, P., König, D., Klicnik, V., Claussen, S., Moser, S., Kloppmann, M., et al. (2010). *Data4BPM, Part 1: Introducing Business Entities and the Business Entity Definition Language (BEDL).* Retrieved from http://www.ibm.com/ developerworks/websphere/library/.

Neumuth, T., Mansmann, S., Scholl, M. H., & Burgert, O. (2008). Data warehouse technology for surgical workflow analysis. *21st IEEE International Symposium on Computer-based Medical Systems* , 230 - 235.

Niedrite, L., Solodovnikova, D., Treimanis, M., & Niedritis, A. (2007). Goal-Driven Design of a Data Warehouse-Based Business Process Analysis System. *Proceedings of the 6th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Database, 6*, 243-249.

Park, R. E., Goethert, W. B., & Florac, W. A. (1996). *Goal-driven software measurement – a guidebook.*Technical report, Carnegie Mellon University.

Pau, K.-C., Si, Y.-W., & Dumas, M. (2007). Data warehouse model for audit trial analysis in workflows. *Proceedings of IEEE International Conference on e-Business Engineering (ICEBE'07)* .

Romero, O., & Abelló, A. (2007). Automating Multidimensional Design from Ontologies. *Proceedings of the ACM tenth international workshop on Data warehousing and OLAP DOLAP '07* , 1-8.

Sayal, M., Casati, F., Dayal, U., & Shan, M.-C. (2002). Business process cockpit. *VLDB '02 Proceedings of the 28th international conference on Very Large Data Bases* , 880 - 883.

Schiefer, J., List, B., & Bruckner, R. M. (2003). Process data store: A real-time data store for monitoring business processes. (V. Mařík, W. Retschitzegger, & O. Štěpánková, Eds.) *Proceedings of Database and Expert Systems Applications (DEXA'03), 2736*, 760-770.

Schiefer, J., List, B., & Bruckner, R. M. (2003). Process data store: A real-time data store for monitoring business processes. *Proceedings of Database and Expert Systems Applications (DEXA'03), 2736*, 760-770.

Shahzad, M. K. (2012). *Improving Business Processes using Process oriented Data Warehouse.* Doctoral Dissertation, Royal Institute of Technology, Stockholm, Sweden.

Stefanov, V. (2007). *Conceptual Models and Model-Based Business Metadata to Bridge the Gap between Data Warehouses and Organizations.* Ph.D. Thesis, Vienna University of Technology Institute of Software Technology and Interactive Systems, Vienna.

Wingenious. (2005). Database Architecture. (Fourth Edition). Retrieved from Wingenious.

Résumé

L'objectif de l'intelligence métier au sein d'une entreprise est d'analyser sa performance, de contrôler la réalisation de ses objectifs et d'accroître sa valeur. Les applications fournies par l'intelligence métier, telles que les applications d'entrepôt de données, transforment des données opérationnelles en informations utiles pour améliorer la prise de décision et identifier de nouvelles opportunités métier. Malgré la forte relation qui existe entre les données métier et les processus métier dans les entreprises, elles sont analysées séparément par les applications de l'intelligence métier. En outre, la plupart des recherches existantes sur l'analyse des données de l'entreprise font une séparation entre elles. Ils considèrent exclusivement soit des entrepôts de processus ou des entrepôts de données. Seules quelques recherches proposent l'intégration de données d'entreprise et de processus métier comme une solution à des problèmes décisionnels qui ne peuvent pas être résolus par des entrepôts de données ou de processus séparément. Cette intégration est en quelque sorte problématique parce que les données métier et celles de processus sont stockées séparément. Dans cet article, nous présentons une étude des approches de conception des entrepôts de données, des entrepôts de processus ainsi que l'intégration de données métier et des processus métier et nous mettons l'accent sur leurs limites. Ensuite, nous proposons une nouvelle approche de conception d'entrepôt qui intègre les données et les processus d'une manière naturelle basée sur les entités métiers.

Modèle multidimensionnel en toile d'araignée : Modélisation conceptuelle et logique

Omar Khrouf, Kaïs Khrouf, Jamel Feki

Laboratoire MIR@CL, Université de Sfax
Route de l'Aérodrome km 4, B.P. 1088, 3018 Sfax, Tunisie
Omar.Khrouf@yahoo.fr, Khrouf.Kais@isecs.rnu.tn, Jamel.Feki@fsegs.rnu.tn

Résumé.
Les systèmes OLAP (« On-Line Analytical Processing ») ont été proposés pour améliorer le processus de prise de décision par l'analyse de grandes masses de données. Ces données peuvent être issues des sources opérationnelles ou de multiples documents manipulés et échangés. Dans ce contexte, nous proposons un nouveau modèle multidimensionnel que nous appelons *modèle en toile d'araignée* permettant l'analyse OLAP de documents XML (eXtensible Markup Language). Le modèle proposé se base sur une combinaison des différentes facettes standards extraites des documents pour permettre une exploitation plus naturelle et une vision plus ciblée des données aux décideurs.

1 Introduction

La gestion du contenu des documents est un facteur essentiel permettant aux entreprises d'améliorer leurs processus de prise de décision et de participer à la continuité et le succès de leurs activités. En effet, les documents constituent une capitalisation des connaissances dans leurs systèmes d'information (Tseng et al., 2006). Pour les décideurs, l'analyse du contenu de ces documents représente un véritable défi.

Dans ce contexte, plusieurs travaux comme (Hernandez et al., 2008) et (Charhad et al., 2005) se sont intéressés à la multi-représentation des documents en utilisant un ensemble de facettes qui permet de décrire un aspect utile du document ; ces facettes ne prennent pas seulement en compte l'aspect sémantique d'un document, mais aussi d'autres éléments relatifs au contexte d'exploitation des documents. Ces facettes ont pour objectif de répondre de façon plus adaptée aux besoins des utilisateurs. Cependant, ces facettes varient selon le domaine d'application. Il serait intéressant alors de définir des facettes standards permettant la représentation des documents, indépendamment de leur domaine.

D'autres travaux de la littérature se sont intéressés à l'application des traitements analytiques en ligne (OLAP : On-line Analytical Processing) sur les documents. Certains travaux ont repris les modèles multidimensionnels classiques, comme les modèles en étoile, en flocon de neige, et en constellation (Hachaichi et al., 2010) et (Feki et al., 2013) pour les document centrés-données ; (Ravat et al., 2008) pour les documents centrés-documents). D'autres travaux ont proposé des modèles multidimensionnels spécifiques comme le modèle

en Galaxie (Ravat et al., 2008) et le modèle en diamant (Azabou et al., 2013). Cependant, ces études ne traitent pas l'hétérogénéité des structures et exigent la définition des paramètres et des hiérarchies des dimensions à l'avance.

Afin de permettre l'analyse OLAP de documents XML, nous proposons un nouveau modèle multidimensionnel intitulé *modèle en toile d'araignée* dédié à l'OLAP de documents XML ; ce modèle est basé sur un ensemble de facettes standards. Dans nos travaux, une facette permet à l'utilisateur d'exprimer plus naturellement ses besoins d'analyse ; c'est pour cette raison que nous transformons chaque facette en une dimension (ensemble de paramètres organisés sous forme de hiérarchie et d'attributs faibles). Le *modèle en toile d'araignée* que nous proposons se diffère des autres modèles proposés dans la littérature par les spécificités suivantes : Contrainte d'exclusion entre les dimensions, présence de paramètres réflexives, autorisation de dimension dupliquée et de dimensions corrélées

Cet article est organisé comme suit. La section 2 détaille les travaux de la littérature relatifs d'une part à la représentation et l'exploitation des facettes des documents et d'autre part à l'analyse OLAP des informations documentaires. La section 3 décrit les facettes standards de documents que nous retenons. Dans la section 4, nous décrivons le modèle multidimensionnel en toile d'araignée avec ses différentes spécificités. Enfin, nous présentons le modèle logique ainsi que les règles d'obtention de ce modèle.

2 Etat de l'art

Dans cette section, nous commençons par présenter les travaux concernant la représentation et l'exploitation des facettes extraites des documents. Par la suite, nous décrivons les travaux qui se sont intéressés à l'analyse OLAP de documents.

La notion de facette a été utilisée dans plusieurs domaines et pour différents types de documents.

Pour les documents textuels, (Evéquoz et al., 2008) ont proposé un système de navigation par facette dans une collection personnelle nommé *Weena* qui permet à l'utilisateur de gérer ses informations personnelles avec plus de flexibilité et rapidité. Dans ce cadre, ils ont créé une interface graphique composée de trois parties : Facettes, Fil d'Ariane et Panneau de résultat. Les facettes définies pour faciliter l'accès aux documents dans une collection d'informations personnelles : textual search, file explorer, type, social network, date accessed, date modified, et date created. Le Fil d'Ariane permet à l'utilisateur de consulter le chemin parcouru, le nombre d'éléments obtenu pour chaque facette après les opérations de filtrage. Le panneau de résultat affiche sous forme d'un tableau les documents qui correspondent aux différents types de recherche effectuée.

Pour le même type de documents et dans le cadre du domaine de l'apprentissage en ligne, (Hernandez et al., 2008) ont proposé un modèle basé sur une représentation multi-facette de documents permettant d'associer plusieurs facettes à un même document qui a un objet éducatif, chacune est utile pour l'accès au contenu du document. Ils ont défini deux types de facettes : une facette qui représente la sémantique du contenu et les autres facettes (description des théories éducatives, description par des métadonnées, la structure du document, etc.) regroupent les différents paramètres qui doivent être pris en compte pour améliorer les résultats d'une recherche de document comme la description des théories éducatives, la description par des métadonnées, etc.

Pour les documents vidéo, Le modèle EMIR² (Mechkour, 1995) est un modèle basé sur les graphes conceptuels. Il permet d'explorer une image fixe à travers un ensemble de facettes. (Charhad et al., 2005) ont proposé d'étendre ce modèle pour inclure les documents audiovisuels. Ils ont ajouté deux facettes : la facette temporelle et la facette événementielle. Ces deux facettes caractérisent l'aspect dynamique spécifique à ce genre de documents. Ce nouveau modèle permet la prise en compte synthétique et intégrée des éléments d'information concernant l'image, le texte et le son.

Pour les tweets, (Kumar et al., 2012) ont proposé un système de navigation par facette intitulé NIF-T « Navigating Information Facets on Twitter » basé sur trois facettes : La Facette *Geo* indiquant les emplacements des tweets dans une carte. La facette *Sujet* représente un nuage de mots indiquant les différentes thématiques échangées par les tweets. La facette *Temps* présente le nombre de tweets à une date donnée.

Concernant la représentation et l'exploitation des facettes, nous notons que certains travaux utilisent des facettes variables ; elles varient suivant le domaine d'application. Par contre, pour d'autres approches, les facettes sont fixes mais pour un domaine d'application spécifique.

Pour la modélisation multidimensionnelle des données factuelles (issues de sources opérationnelles de l'entreprise), trois principaux modèles ont été proposés : modèle en étoile, modèle en flocon de neige, et modèle en constellation (Kimball, 2003). Concernant l'analyse OLAP des documents, la plupart de travaux ont repris les trois modèles proposés et ont suggéré des approches ou fonctions pour l'analyse du contenu textuel.

Pour les travaux qui ont utilisé le modèle en étoile, nous citons : (Tseng et al., 2006) évaluent le contenu de documents (emails, articles, pages Web…) selon des dimensions construites à partir des métadonnées définies par le Dublin Core (Dublin Core, 2007). Par contre ces analyses restent limitées à un simple comptage de documents. (Zhang et al., 2009) ont proposé un nouveau modèle basé sur un schéma en étoile intitulé *Topic Cube* qui permet d'étendre le cube de données traditionnel en intégrant une hiérarchie de thèmes *Topics* comme étant une dimension d'analyse. Ce modèle propose deux mesures probabilistes : la distribution d'un mot dans un thème $p(w_i)$ et la couverture d'un thème par les documents $p(topic_j)$. Par contre, *Topic Cube* ne supporte qu'un ensemble prédéfini de thèmes.

(Boussaid et al., 2006) ont proposé une modélisation en *flocon* des données multidimensionnelles XML avec des méthodes de fouille de données. Plus précisément, ils ont défini une approche appelée *X-Warehousing* qui permet de décrire le modèle logique d'un cube XML. Ce modèle implique beaucoup de redondance dans les données des dimensions puisque pour chaque mesure du fait, il faut indiquer les valeurs des dimensions correspondantes, ce qui peut impliquer des difficultés de mise à jour et de maintenance. En plus, *X-Warehousing* est spécifique à des données complexes, mais il ne peut pas être adapté pour des analyses de données textuelles.

D'autres travaux ont proposé des modèles spécifiques, comme : (Ravat et al., 2008) ont proposé un modèle en *Galaxie* qui repose sur un seul concept, celui de dimension. La notion de fait est non explicitée. Il s'agit d'un regroupement de dimensions liées entre elles par un ou plusieurs nœuds centraux ; chaque nœud modélise les dimensions compatibles pour une même analyse. Enfin, (Azabou et al., 2013) proposent un modèle en diamant qui étend le modèle en étoile par une dimension centrale, appelée sémantique. Les paramètres de cette

dimension seront reliés aux paramètres des autres dimensions. L'inconvénient de ces deux travaux est le fait de proposer un modèle par collection de documents.

En conclusion, les travaux traitant l'analyse OLAP de documents présentés dans cette section proposent soit des analyses simples, soit des analyses pour des documents ayant des structures identiques ou similaires. Contrairement à ces travaux, nous proposons un nouveau modèle multidimensionnel permettant l'analyse des documents hétérogènes de point de vue structure et contenu. Nous intégrons ainsi la notion de facette dans notre modèle puisque qu'une facette (de structure, contenu, métadonnées, ou sémantique) peut être considérée comme un moyen d'expression des besoins de l'utilisateur. Cet article sera consacré à la modélisation conceptuelle et logique de notre modèle multidimensionnel.

3 Facettes standards de documents

Une facette décrit un aspect utile à l'exploitation d'un ou de plusieurs documents selon un point de vue, ce qui offre à l'utilisateur la possibilité de consulter le même document ou un ensemble de documents à partir de plusieurs vues (facettes) et d'avoir un accès plus ciblé à l'information selon ses besoins. Les facettes que nous proposons sont indépendantes de tout domaine d'application spécifique. Pour ce faire, nous définissons un ensemble de cinq facettes, chacune représente une vision du document (Khrouf et al., 2013).

- La facette *Contenu*: Cette facette permet l'accès au contenu (texte, image, vidéo…) proprement dit du document. Elle a pour rôle de mettre en évidence l'information véhiculée par le document, en éliminant tout ce qui concerne la structuration, la présentation, les commentaires, etc.
- La facette *Structurelle* : Cette facette permet de définir une vue globale sur la structure d'un document ou d'un ensemble de documents homogènes. Il s'agit plus précisément de l'arborescence du document. Cette facette à pour rôle de se focaliser sur des parties du document et non sur sa totalité.
- La facette *Métadonnée* : Elle représente un ensemble structuré des données décrivant un document. Dans nos travaux, nous utilisons les métadonnées définies par le Dublin Core (Dublin Core, 2007). Cette facette offre aux utilisateurs une description complète sur le document (comme : le titre, l'éditeur, le format, les droits, etc.).
- La facette *Sémantique* : Cette facette indique la sémantique véhiculée dans le contenu du document. Pour la détermination de cette sémantique, nous utilisons les travaux de (Ben Mefteh et al., 2013) qui proposent une structure sémantique par document, et ce en projetant le contenu du document sur une taxonomie sélectionnée[1].
- La facette *Mot-clé* : Cette facette représente les Mots-clés les plus pertinents qui décrivent le contenu d'un ou de plusieurs documents. Ces Mots-clés sont déterminés en utilisant les techniques d'indexation de la recherche d'information.

En se basant sur les facettes précédemment citées, nous présentons, dans ce qui suit, un nouveau modèle multidimensionnel dédié à l'OLAP de documents XML, que nous appelons

[1] Une taxonomie est une ressource sémantique permettant la représentation *hiérarchique* de ses concepts.

modèle en toile d'araignée. Nous commençons par la modélisation conceptuelle. Ensuite, nous décrivons la modélisation logique.

4 Modélisation conceptuelle

L'idée principale de notre *modèle en toile d'araignée* consiste à transformer chacune des facettes définies dans la section précédente en une dimension puisque ces facettes peuvent représenter un moyen d'expression de besoin pour les utilisateurs. En effet, les facettes regroupent les différentes informations et métadonnées concernant les documents à analyser. En plus, nous avons ajouté la dimension *Document* afin de relier les différentes informations issues des facettes et représentées par des dimensions à leurs documents. Nous considérons que le fait correspond à une observation sur les documents, il est décrit par des mots-clés, possède une sémantique, etc. Au moment de l'interrogation une dimension pourrait devenir un fait en appliquant la fonction d'agrégation sur ses valeurs. La figure 1 présente ce modèle conceptuel en toile d'araignée.

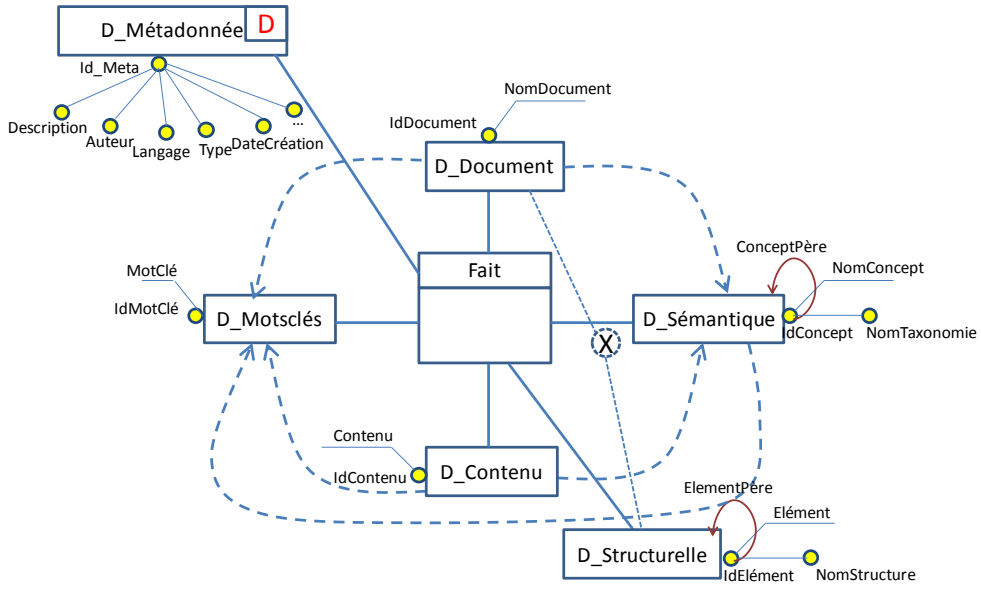


FIG. 1 – *Modèle en toile d'araignée*

Ce modèle en toile d'araignée est une extension du modèle en étoile, il se caractérise par les spécificités suivantes :

- Contrainte d'exclusion entre les dimensions : La contrainte d'exclusion exprime le fait que deux dimensions ne peuvent pas participer simultanément à une même analyse OLAP de documents. Dans ce cas, une seule de ces deux dimensions au plus peut être retenue au moment de l'analyse. Graphiquement, cette contrainte d'exclusion est exprimée par un cercle contenant **X** relié aux dimensions concernées, à savoir : Document et Structurelle. (cf. Figure 1)

- La dimension dupliquée : Dans la modélisation classique (telle que le modèle en étoile), une même dimension peut participer une seule fois dans une analyse OLAP. Pour remédier à ce problème, nous proposons un nouveau type de dimension intitulée dimension dupliquée. Cette dernière peut participer deux voire trois fois dans la même analyse OLAP. Graphiquement, une dimension dupliquée est schématisée par la lettre **D** dans la dimension concernée. Dans notre modèle multidimensionnel, nous disposons d'une seule dimension dupliquée, à savoir : Métadonnée. (cf. Figure 1)
- Paramètre récursif : Dans les modèles multidimensionnels classiques, les paramètres et les hiérarchies des dimensions sont connus à l'avance, ce qui n'est pas le cas pour nos travaux (Exemple : la structure des documents diffère d'une collection à une autre). Pour la représentation de telles dimensions, nous utilisons un nouveau type de paramètre, dit paramètre récursif car les documents et les taxonomies utilisées dans nos travaux sont représentés d'une manière hiérarchique. Graphiquement, un paramètre récursif est schématisé par une boucle. (cf. Figure 1)
- Dimension corrélée : Dans la modélisation multidimensionnelle classique, la substitution des dimensions n'est pas réalisable en raison de la contrainte d'orthogonalité des dimensions (Exemple : absence des relations inter-dimensionnels). Face à cette problématique, nous proposons le concept de la dimension corrélée qui permet de remplacer une dimension par une autre pour la même requête OLAP. Graphiquement, la corrélation possible entre les dimensions de notre modèle multidimensionnel est schématisée par des flèches entre les dimensions. Le passage d'une dimension à une autre est accepté tout en respectant le sens de la flèche. A titre d'exemple, il est possible de passer de la dimension Sémantique vers la dimension Mot-clé pour la même requête. (cf. Figure 1).

5 Modélisation logique

Cette section commence par présenter la nécessité d'une nouvelle modélisation logique pour notre *modèle en toile d'araignée*, à savoir ; la modélisation OROLAP (Object-Relational OLAP). Ensuite, nous décrivons les règles de passage du modèle conceptuel multidimensionnel au modèle logique OROLAP. Enfin, nous présentons le modèle logique obtenu.

- **Modèle OROLAP (Object-Relational OLAP)**

Dans la littérature, trois modèles logiques pour les systèmes OLAP ont été proposés : Les modèles ROLAP (Relational OLAP) qui se base sur un SGBD relationnel en transformant le fait et les dimensions en des tables relationnelles, les modèles MOLAP (Multidimensional OLAP) qui proposent une représentation multidimensionnelle des données (sous forme de cube) et les modèles HOLAP (Hybrid OLAP) qui combinent ROLAP et MOLAP à la fois. Nous notons que les modèles ROLAP sont les plus utilisés car ils sont associés au modèle relationnel qui est bien connu par les concepteurs de logiciels.

Cependant, les modèles ROLAP exigent une clé étrangère dans le fait qui référence chaque dimension liée au fait. Plus précisément, chaque valeur d'une clé étrangère correspond à une seule ligne de la dimension. Dans nos travaux, un fait peut concerner

plusieurs Mots-clés et non un seul, c'est la raison pour laquelle nous proposons l'OROLAP (Object-Relational OLAP). Ce modèle se caractérise par des liens monovalués (pour référencer une seule ligne) de la table de fait vers la table de dimension ainsi que des liens multivalués (pour référencer plusieurs lignes) de la table de fait vers d'autres tables de dimensions.

- **Règles de transformation**

Le modèle conceptuel multidimensionnel est transformé en modèle logique OROLAP selon les règles suivantes :

- Chaque dimension *d* est modélisée en une table relationnelle composée par un ensemble d'attributs qui représente les paramètres et les attributs faibles relatifs aux différentes hiérarchies composant cette dimension. La clé primaire de *d* est définie par l'attribut le plus fin au niveau de granularité.
- Les paramètres récursifs seront transformés par des liens monovalués récursifs.
- La corrélation entre les dimensions est transformée par des liens multivalués, si on référence la dimension Mot-clé et par des liens monovalués pour les autres cas.
- Le fait *f* est transformé en une table relationnelle composée par un lien multivalué vers la dimension Mot-clé et par des liens monovalués vers les autres dimensions. La clé primaire de *f* est multiple et elle est formée par la concaténation des attributs représentant les liens monovalués et multivalués.

- **Modèle logique OROLAP**

Le résultat de transformation de notre *modèle en toile d'araignée* en un modèle logique est présenté dans la figure 2.

Soit le document XML suivant :

```
<Article>
<Meta name="DC.Creator" content="Jamel FEKI" />
<Meta name="DC.Language" content="English" />
<Meta name="DC.Type" content="text" />
<Meta name="DC.Date" content="2009-04-16" />
<Title>OLAP Cube</Title>
<Author>Jamel FEKI</Author>
<Content>
<Section>When the cube...</Section>
<Section>The advantage of...</Section>
</Content>
</Article>
```

Modèle en toile d'araignée

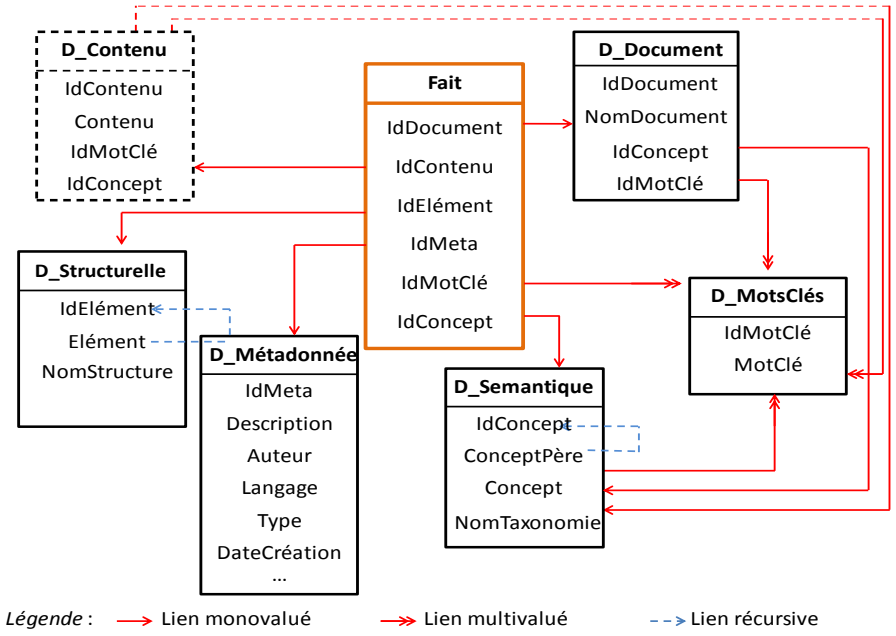


FIG. 2 – *Modèle logique OROLAP*

Pour ce document, nous aurons une ligne de fait par balise (une ligne pour *Title*, une ligne pour *Author*, etc.). La figure 3 décrit la ligne de fait *Title*, elle correspond au document *Doc1*, la structure *Article*, l'aspect *DW* (*Data Warehouse*) et elle contient les métadonnées suivantes (*Auteur*, *langue*, *type* et *date de publication*). Cette ligne de fait contient un lien multivalué vers les Mots-clés: *DW*, *FACT* et *OLAP*.

Afin de valider les propositions présentées dans cet article, nous avons implanté ce modèle logique sous Oracle 10g en utilisant les attributs de type REF pour les liens monovalués et les tables imbriquées (nested tables) pour les liens multivalués. Nous avons aussi instancié ce modèle et testé un ensemble de requêtes multidimensionnelles afin de vérifier et de valider notre modèle en toile d'araignée.

Exemple : Supposons que nous souhaitons analyser le nombre de documents par thématique (dimension sémantique), par auteur et année. Le système génère automatiquement la requête suivante :

```
SELECT    F.idConcept.Concept, F.idMeta.Auteur,
          TO_CHAR(F.idMeta.DateCreation, 'YYYY'),
          Count(Distinct F.idDocument)
FROM      Fait F
WHERE     F.idConcept.Concept IS NOT Null
GROUP BY  F.idConcept.Concept, F.idMeta.Author,
          TO_CHAR(F.idMeta.Date, 'YYYY')
```

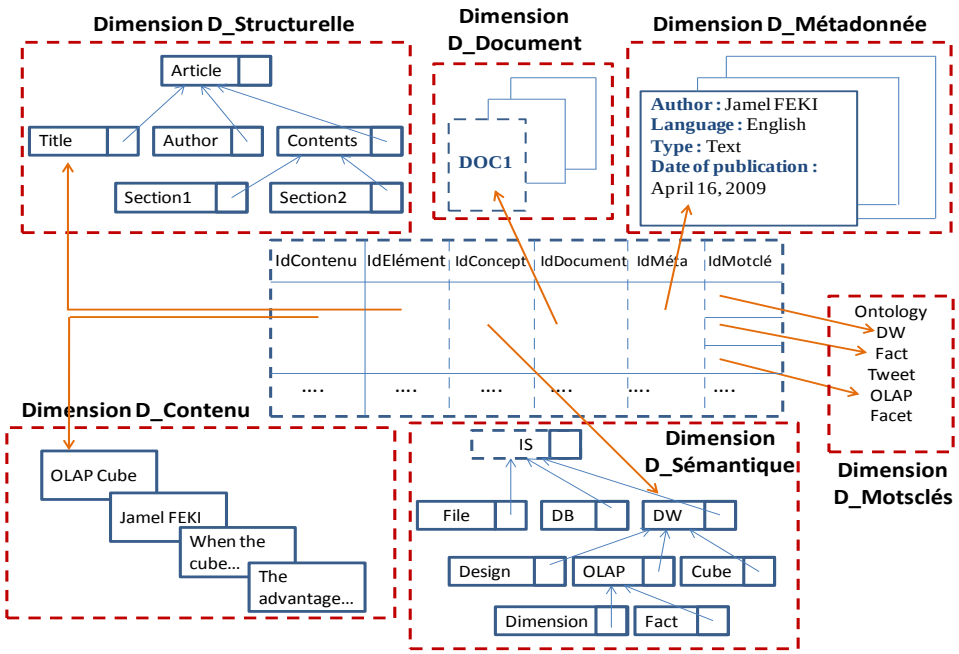


FIG. 3 – *Exemple d'instanciation du modèle logique OROLAP*

6 Conclusion

Le modèle en toile d'araignée proposé est dédié à l'OLAP de documents XML à base de facettes. Ce modèle est générique, c'est-à-dire non limité à un ensemble de documents prédéfinis ou de même structure. Le modèle proposé se distingue des autres modèles par les extensions suivantes : la contrainte d'exclusion entre les dimensions, les paramètres récursifs, la dimension dupliquée et les dimensions corrélées. Dans cet article, nous nous sommes focalisés sur la phase de modélisation conceptuelle et celle logique.

Plusieurs perspectives sont envisageables pour ce travail. Dans l'immédiat, nous développerons des interfaces graphiques conviviales permettant à l'utilisateur d'exprimer sa requête multidimensionnelle (spécification du paramètre d'analyse par dimension et la mesure pour le fait) ; Ainsi, le système qui se charge d'une part de générer automatiquement les scripts nécessaires pour le calcul et la récupération du résultat et, d'autre part, de l'afficher sous forme d'une table multidimensionnelle. Il est important aussi de proposer des opérateurs OLAP dédiées pour le modèle en toile d'araignée, à titre d'exemple un opérateur permettant la corrélation des dimensions. A long terme, nous souhaitons partager les analyses OLAP entre les différents utilisateurs d'une même organisation, par l'introduction de l'aspect collaboratif comme par exemple, un système de recommandation de requêtes et de collaboration entre des utilisateurs ayant des intérêts communs.

Références

Azabou M., Khrouf K., Feki J., Vallès N. et Soulé-Dupuy C. (2013). *Modèle multidimensionnel en diamant dédié à l'OLAP sémantique de documents*. Conf. Magrébine sur les Avancées des Systèmes Décisionnels (ASD), 25-27 March, Marrakech, Maroc.

Ben Mefteh S., Khrouf K., Feki J., Ben Kraiem M. et Soulé-Dupuy C. (2012). *Une approche d'extraction automatique de structures sémantiques de documents XML*. INFormatique des Organisations et Systèmes d'Information et de Décision (INFORSID 2012), p. 523-538, Montpellier, France.

Boussaid O., Ben Messaoud R., Choquet R. et Anthoard S. (2006). *Conception et construction d'entrepôts XML*. Journée francophone sur les Entrepôts de Données et l'Analyse en ligne (EDA'2006), p. 3–21, Versailles, France.

Charhad M. et Quénot G. (2004). *Semantic Video Content Indexing and Retrieval using Conceptual Graphs*. Edition JohnWiley, Damascus, Syria, 2004, p.19-23.

Dublin Core Metadata Initiative (DCMI) (2007): Dublin Core Metadata Element Set, Version 1.1, ISO Standard 15836, from http://dublincore.org/documents/dces/.

Evéquoz F., Thomet J. et Lalanne D. (2010). *Gérer son information personnelle au moyen de la navigation par facettes*. Conférence Internationale Francophone sur I 'Interaction Homme-Machine, 20-23 September, Luxembourg.

Feki J., Ben Messaoud I. et Zurfluh G. (2013). *Building an XML Document Warehouse*. Journal of Decision Systems (JDS), Edition Lavoisier, No.1/2013.

Hachaichi, Y., Feki, J. et Ben-Abdallah, H. (2010). *Modélisation multidimensionnelle de documents XML centrés-données*. Journal of Décision Systems (JDS), Edition Lavoisier, Vol 19/3, pp. 313-345.

Hernandez, N., Mothe, J., Ralalason, B., Ramamonjisoa, B., Stolf, P. (2008). A Model toRepresent the Facets of Learning Objects. Interdisciplinary Journal of E-Learning andLearning Objects, Informing Science Institute, Santa Rosa - USA, vol. 4, p. 65-82.

Khrouf O., Khrouf K. et Feki J. (2013). *A new multidimensional model for the OLAP of documents based on facets*. The International Arab Conference on Information Technology (ACIT2013), 17-19 December, Khartoum, Soudan, 2013

Kimball R. (2003). *The data warehouse toolkit*. Edition JohnWiley and Sons, 2nd edition, USA.

Kumar S., Morstatter F., Marshall G., Liu H. et Nambiar U. (2012). *Navigating Information Facets on Twitter (NIF-T)*. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 12-16 August, Beijing, China

Mechkour M. (1995). *A multifacet formal image model for information retrieval*. MIRO final workshop, Glasgow, UK, 1995, p. 18-20.

Ravat F., Teste O., Tournier R. et Zurfluh G. (2008). *Designing and Implementing OLAP Systems from XML Documents*. In : Annals of Information Systems, Springer, Special issue New Trends in Data Warehousing and Data Analysis, Vol. 3, pp. 295-315, November 2008.

Tseng F.S.C., Chou A.Y. (2006). *The concept of document warehousing for multidimensional modeling of textual-based business intelligence*. Decision Support System (DSS), vol. 42, (2), Elsevier, 2006, p. 727-744.

Zhang, D., C. Zhai. et J. Han (2009). *Topic cube: Topic modeling for olap on multidimensional text databases*. SDM '09: Proceedings of the 2009 SIAM International Conference on Data Mining, Sparks, NV, USA, 1124–1135.

Summary

Nowadays, the information systems of companies handle a large volume of data. The OLAP (On-line Analytical Processing) systems have been proposed to improve the processes of decision-making by providing to the decision-makers a space of multidimensional representation of the data. These data may be derived from operational sources of these companies or multiple manipulated and exchanged documents. In this context, we propose a new multidimensional model for the OLAP of XML documents, named CobWeb. This model is based on a combination of different standard facets extracted from documents in order to provide more opportunities for the expression of analytic queries and a vision more targeted of data to decision-makers compared to the classical multidimensional models. In this paper, we describe the conceptual and the logical modeling of our CobWeb model.

ETL-Web process modeling

Hana Mallek, Afef Walha
Faiza Ghozzi, Faiez Gargouri

hana027@gmail.com,
afef_walha@yahoo.fr
jedidifaiza@gmail.com
faiez.gargouri@isimsf.rnu.tn
MIR@CL Laboratory, University of Sfax, Tunisia

Abstract. Data integration in WeBhouse (DWB) is crucial to get valuable decisions in the web context. ETL processes dealing with Web data integration challenge complexities of data extraction from heterogeneous web sources, their transformation, and their loading in the DWB. Even though, some approaches in the literature recommend the modeling of these processes, web sources aren't supported yet. In this paper, we propose to model ETL-Web processes integrating several web sources, especially log files and websites, using UML Activity Diagram.

1 Introduction

Nowadays, the web is the largest source of information, which leads to the importance of using Data Warehouse (DW) and Business Intelligence (BI) applications in the web in order to integrate the web data from several sources and to make better decision. In the web context, the mega-volumes of data, the increasing number of users and the heterogeneity of requirements must be taken into account by the experts in order to fit business needs. That's why the extraction and analysis of useful information from the web (like net surfer behavior, social network, commercial or medical web sites,...) become a practical necessity. Furthermore, ETL (Extraction - Transformation - Loading) processes are responsible of data extraction from different sources, their transformation, and their loading into the DW. These processes achieve the synchronization between heterogeneous data sources and the target store. Therefore, the modeling of ETL processes in the web context (ETL-Web) is an efficient solution for integrating web data. Modeling ETL in the web context is intended to facilitate the various data treatments (cleaning, aggregation, ...) to obtain final homogeneous DWB. In this paper, ETL modeling is not only used to generalize the ETL processes but also to map different sources. We consider three main sources in the web which are: log files (clickstreams), business sources and websites. To summarize, our main objective is to free designers from having to deal with web data specificities (Web logs, web sites, clickstreams,...) and to overcome the complexity of ETL processes. So, we propose to model the ETL-Web processes using UML 2.0 Activity Diagram.
The remainder of this paper is organized as follows. In Section 2, we present a state of the

art dealing with ETL process design and web data integration. In Section 3, we depict the Meta-model of Web-ETL process. In Section 4, we detailed our ETL-web process modeling approach using activity diagram. In Section 5, we validate our approach by the extension of talend Open source. We finally conclude the paper and discuss future research issues.

2 Related works

In this section, we review and discuss the literature related to web data integration and ETL processes modeling. Web data integration approaches can be classified into 3 groups. The first one is based on ontology, in which web data are extracted through key words Embley et al. (1998) or glossaries Liu et al. (2009). The second group considers structural integration, in which web data are presented through XML files or class diagrams Darmont et al. (2007), Vangipuram et al. (2010). These approaches model various web sources without considering pertinent web data like web logs, web pages, ... In the third group of web data integration approaches, authors analyze web users behavior (Clickstreams) to multidimensional design. Andersen et al. (2000) propose web sessions modeling of log files. Also, Hernández et al. (2010) take into account the variety of log files formats and propose a unified meta-model of Clickstreams. These approaches are limited to web data integration without modeling ETL processes. In the other hand, ETL processes design is a crucial task in DW development due to its complexity and its time consuming. Many approaches are proposed in the literature to deal with it. They can be classified into two main groups: specific ETL modeling and standard ETL modeling. The first group offers specific notations and concepts to give rise for a new specialized modeling languages Vassiliadis et al. (2002), El-Sappagh et al. (2011). These authors propose a new conceptual or formal models for ETL processes in order to facilitate the modeling of complex ETL workflow. However, the standardization is an essential asset in modeling. The goal of the second group is to overcome this problem by using modeling languages like UML, BPMN, ... Wilkinson et al. (2010), Akkaoui et al. (2012) use BPMN standard where ETL processes can be a particular type of business. Trujillo and Luján-Mora (2003) use UML class diagram to represent ETL processes statically or dynamically by using UML activity diagram Muñoz et al. (2008). Even though these approaches are very interesting, they don't cover web data sources. After studying different ETL design approaches, several modeling languages are provided. In our context, UML 2.0 and BPMN 2.0 are the most popular standards. Despite that many works in the literature presented comparisons between these two modeling languages, most researchers have concluded that there is a significant similarity between them Russell et al. (2006). In this paper, we choose UML 2.0 (Activity Diagram) to describe the ETL-Web processes because of its understandability and its relevance. Muñoz et al. (2010) prove the efficiency of activity diagram in representing dynamic aspects of ETL processes through a set of validation experiments. So, we adopt the approach of Muñoz et al. (2008) and we enrich it by pertinent web data.

3 Meta-model of Web-ETL processes

To model our web-ETL process, we extend UML activity diagram Meta-model OMG (2007), by adding some classes presented with blue color in Fig.1. These classes are described

as follow:

- "Steps" inherits from "ActivityGroup" class. It models the extraction, transformation and loading processes. Each "Step" is composed of "Activities".
- "Sons" and "Parents" classes inherit respectively from the "inputPin" and the "outputPin" classes.
- "Xml Buffer", "Log file", "Website" and "Relational DB" inherit from the "data store Node" class are used as storage means.
- "XML Dimensional" is an "XML Buffer" class. It represents a dimensional structure (fact, dimension, attributes ...).

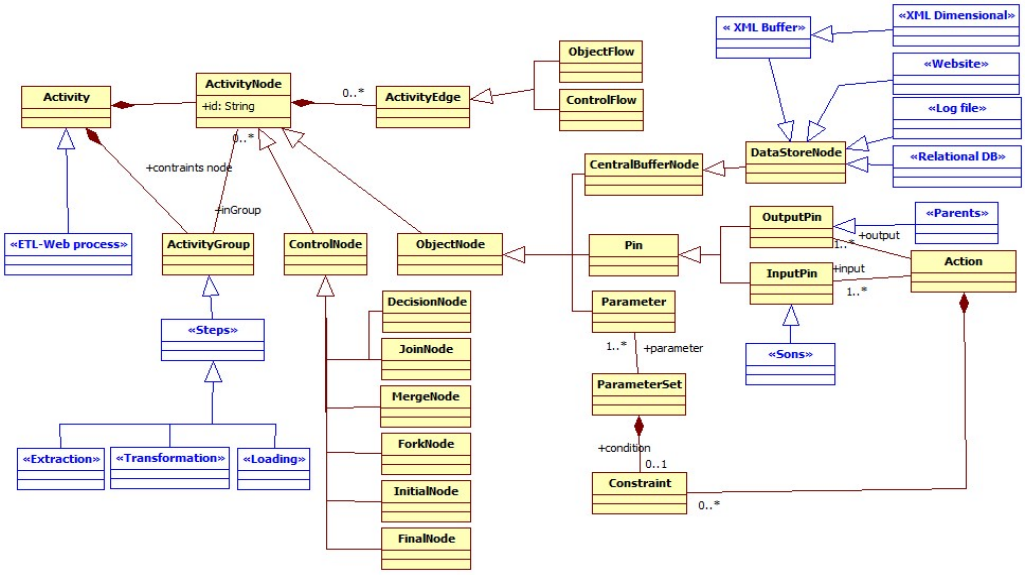


FIG. 1 – *Web-ETL processes meta-model*

4 ETL-Web process modeling

web data are heterogeneous, noisy and voluminous. We propose a modeling of ETL web processes in order to unify, clean and map web data (log files, websites and business data). In this section, we focus on modeling the first two steps of ETL processes: Extraction and Transformation.

4.1 Extraction Step

This step consists of extracting data from log files (stored in web servers), websites and business sources (relational DB, XML ...). The main objective of this step is to unify these

sources, into a single format to be used later . So, heterogeneous and unstructured sources are organized in XML format. We propose to model extraction step through many activities, which are described in the following sections.

"Log file structure" activity Log files exist in several formats like ELF, CLF, ECLF, ... [1]. Each one is composed of specific fields. Indeed, it's recommended to set these logs in a common format in order to simplify the data extraction step. In our approach, we normalized all these formats to ECLF (Extended Common Log File format) because it covers all needed fields: client's host name or IP adress, request date and time, operation type (GET, POST, ...), requested resource (URL), status code, size of requested page, referrer and user agent. Fig.2 presents "LogFileStructure" activity describing log files structuring process. This activity starts by verifying the validity of the input log file. A log file is valid if its structure respects a specific log file format. Then, it analyzes the structure of this file. Finally, it transforms this structure into an XML file represented as buffer store.

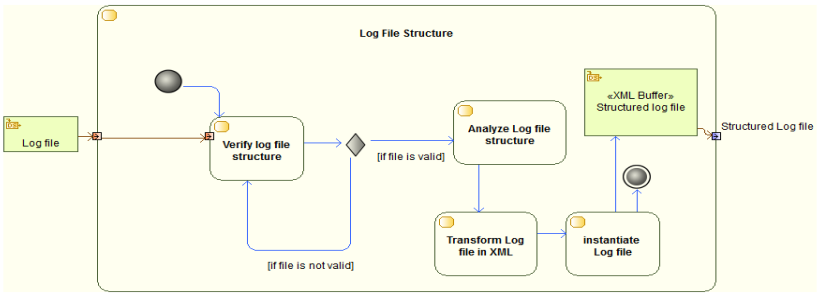


FIG. 2 – *"Log File Structure" activity diagram*

"Website Structure" activity Analyzing web page descriptions can provide pertinent data about net surfer interests. We propose to extract web site structure by using Sitemap generator [2]. This tool returns a list of web pages addresses (URL) with their descriptions. For example, the URL "http://www.yvesrocherusa.com/control/fragrance-gift-ideas/" corresponds to "performed gift ideas" description. We model website data extraction process through "Website structure" activity (Fig. 3). The first step consists in verifing website URL validation. Then, Sitemap generator provides website structure in XML format. The final step is the instantiation and storing data in XML buffer files.

1. http://www.w3.org
2. http://www.sitemaps.org

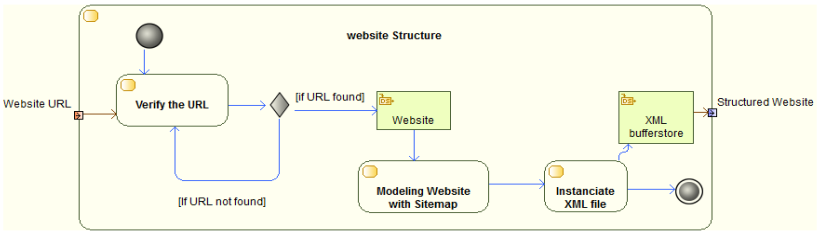


FIG. 3 – *"Website Structure" activity diagram*

"Convert business source to XML" activity Some previous works deal with log data integration in DW. These works omit business data sources which provide pertinent decisional information as traditional sources Hernández et al. (2010). Business sources can be in several formats (XML, relational, HTML, DB, ...) . To overcome this heterogeneity problem, we propose to convert all business sources into XML format.

4.2 Web Transformation Step

In ETL design process, transformation step is an important and complex task. In the literature, it assembles several classic operations such as join, conversion, filtering, aggregation and mapping between sources and target data. These operations are necessary in this step. However, they are insufficient in a web context. As a solution, we propose others web transformation activities specific to web sources shown in Fig. 4. This activity takes as input web data sources structured in previous step (section 4.1). First, it removes unnecessary and useless data from log files through "Log file cleaning" activity (for example removing not found web pages). Second, it identifies web surfer sessions and subdivides log file according to the different user sessions that shouldn't exceed 30 minutes Kimball et al. (2000) ("session identification" activity). In another hand, "Business-web mapping" activity maintains the correspondence between various DWB sources to maintain homogeneous file ("structure unification" activity). Finally, the different results are unified and stored in a temporary Buffer as a structured file. These activities will be more detailed in the following subsections.

"Log file cleaning" activity Log files can contain invalid data, empty fields, useless information, or wrong pages. As a solution, we propose "log file cleaning" activity in order to remove unnecessary, irrelevant records stored in these files, which are unnecessary in analysis phase. Log file cleaning activity is presented in Fig. 5. This activity takes as input a log files and designer requests. First, it analyzes requests and files content. Then, it removes useless records from log file according to designer requests. The result of this activity is a cleaned log file stored in "XML buffer".
In this activity, we use some cleaning operations presented in Cooley et al. (1999):

- Deleting records, where the extension suffixes of the requested files express an image or a multimedia object such as (.gif, .Jpg, .Png, .wav ...) Sumathi C.P. (2011).
- Eliminating of dynamic web pages that are considered as pages interpreted by the server.

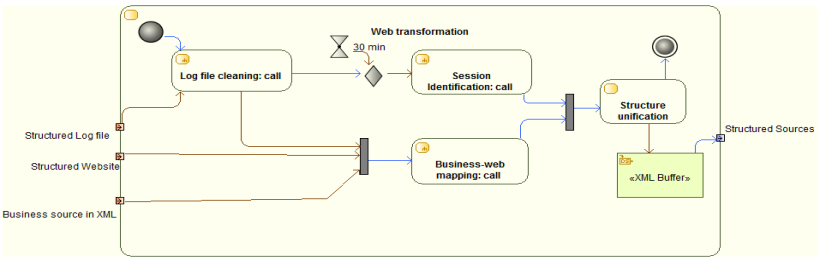


FIG. 4 – *"Web Transformation" activity diagram*

– Removing records with state code different to 200 Kimball et al. (2000). For example, we take a record from log file where the status code is "404". This code indicates that the requested page cannot be found. In our context, we can maintain the correct queries where its status code is "200".
– According to Charrad et al. (2010) a log file contains requests from web called "Web crawler" that help search engine to index web pages such as Google. To achieve well-structured log files, it is strongly recommended to remove queries generated by the robots to obtain requests web visitors.
– We distinguish the type of queries at the level of its HTTP method sent by the visitor such as (GET, POST, HEAD). In this case, we only select requests with the "GET" method Charrad et al. (2010).

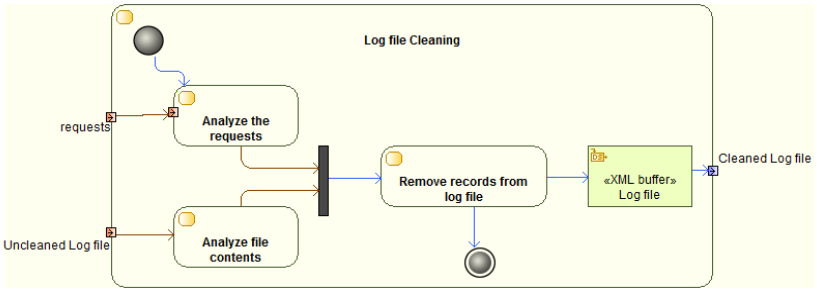


FIG. 5 – *"Log file cleaning" activity diagram*

"Session identification" activity The activity of identifying sessions (presented in Fig. 6) is an important step that leads to list the different sessions initialized at the entrance of Internet users. The aim of this activity is to combine different queries that have the same IP addresses for a period of 30 minutes. At the end, the structure of each session and its instantiation parameters will be saved in an XML buffer.

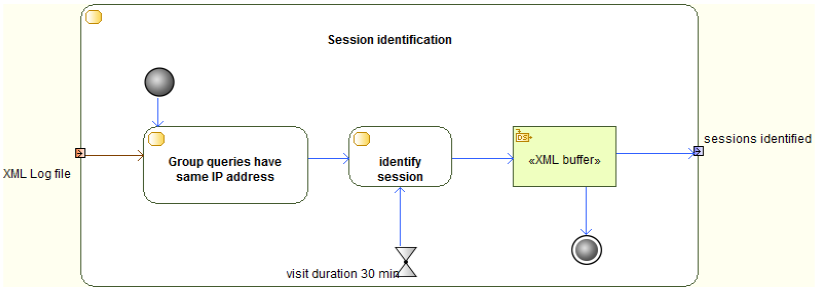


FIG. 6 – *"Identify session" activity diagram*

"Business-web mapping" activity This activity is the most important phase, in the transformation step. It tackles semantic mapping between log files, website and business sources. A log file stores queries without details of the requested page, a website provides URLs described through keywords, and relational databases presents only business data.
To make the mapping between these sources (Fig. 7) we start by integrating XML website and XML business sources ("Integrate sources by Key word" activity) in "XML buffer" mapping file. We follow by integrating this mapping file and XML log file ("integrate source for each row in log file" activity). The result of this activity is a "mapping web file". This mapping in-

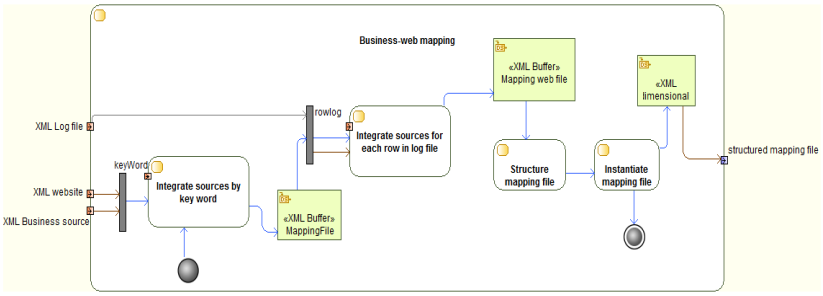


FIG. 7 – *"Business-web mapping" activity*

corporate a problem of how to match key words extracted from log files and their structure in business sources. Let's take an example, we extract "URL" of visited page from referrer in the log file. URL information can't be used alone to define web user needs. It must be matched to website structure to extract key words for this URL. Then, the meaning of these key words are completed by accessing business sources. For example, "Sublin Skin cream" can be found in business sources as a product's name belonging to "Body care" family.

5 Prototype

To validate our approach, we extend the ETL tool Talend Open Studio (TOS) [3]. It have as role to generate for each treatment of data integration, a specific code that can be written in Java or Perl. This tool has a palette of components called "base ETL library" which is composed of different operations (mapping, join, merge, etc...). It is an open source tool, for that we can add new components to the palette. We create a new library called "ETL-Web" that contains ETL operations to extract, transform and load data. These operations are defined according to our activities diagrams (section 4).

The Fig. 8 shows the architecture of TOS which consists of three layers. The first is the conceptual layer. It is composed of the base palette and our extended palette "ETL-Web". In this case, webhouse designers can choose the appropriate ETL-Web operation to extract and transform web data. The second is the logical layer, where we have our two libraries (the base ETL and ETL web). The third layer is the physical layer which we develop for each operation a java code, used by Talend to generate JavaJet executable component. The algorithm 1 depicts a method defined to implement "mapping between sources" operations. Operations wokflow forms an ETL process. This last is stored ETL-Web Library, and can be generated and executed to DWB.

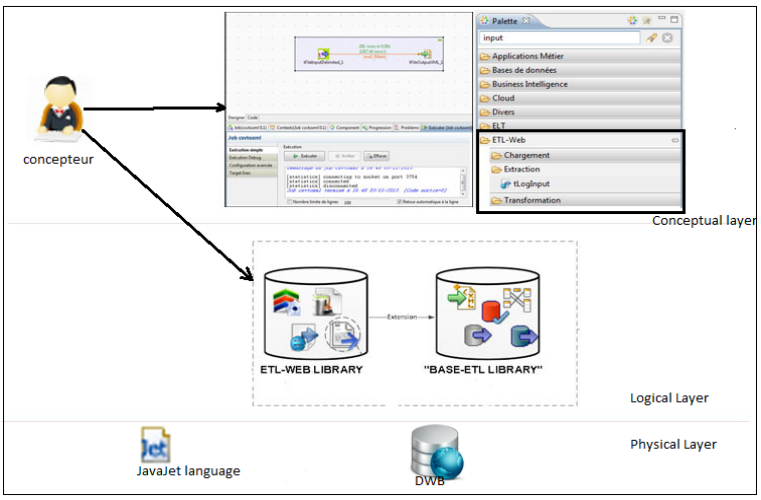


FIG. 8 – *Architecture of TOS*

3. http://www.talend.com/products/talend-open-studio

Algorithm 1 Mapping between sources

Input: The log file **FLOG**, the file Containing URL and description of each page **FwebSite**, the business source in Xml files **Fbusiness**

Output: **FileMappingweb**

```
1:  Begin
2:      for each metakeyWord in Fwebsite do
3:          Search (metatkeyword, Fbusiness)  ▷ search the description of each meakeyword
    in the data base
4:          if Search return true then
5:              CreateFileMapping (Fwebsite, Fbusiness)        ▷ writing In the XML file
    "FileMapping" the result of research
6:          end if
7:      end for
8:      for each rowLog in Flog do
9:          Search (wordLog, FileMapping)    ▷ search the description of the words in each
    row of the log file, in the "FileMapping" folder we created in step 4
10:         if Search return true then
11:             CreateFileMappingweb (FileMapping, Flog)       ▷ writing In the XML file
    "FileMappingweb" the result of research
12:         end if
13:     end for
14: End
```

Therefore, in our work we created a component "tlogInput" as presented in Fig. 9 that can read and unify the log file:

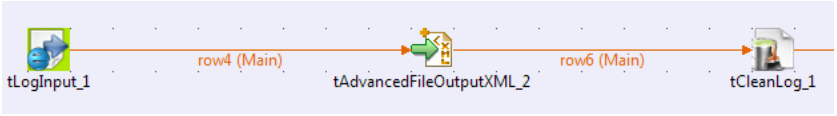


FIG. 9 – *Log file structuring and cleaning*

Furthermore, the component "tCleanLog" shown in Fig. 10 allows log file cleaning. We give the hand to the designer to choose what cleaning operations are required for its design. Thanks to our component, cleaning operation becomes easier. The user doesn't need to compose a complex query, he can simply check the derived operations.

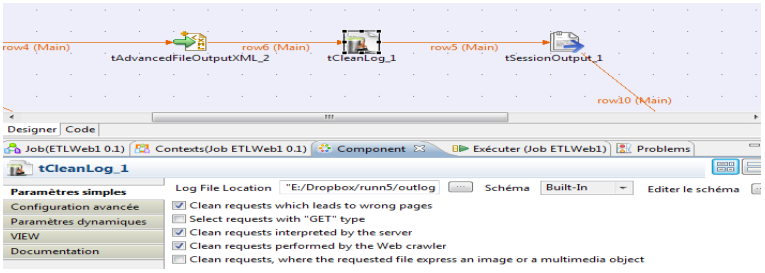


FIG. 10 – *"tCleanLog" component*

6 Conclusion

In this paper, we have presented a model of ETL-WEB processes with UML 2.0 Activity Diagram. This model minimizes the complexity enhanced by the integration of web data in the DWB. Thus, our approach consists in modeling each step in ETL-Web processes. In particular, at the extraction step, we dealt with the problem of log files heterogeneity where we have proposed to unify these files into unified format ECLF. Furthermore, at the Transformation step, we found that data derived from log files are incomplete and didn't identify members of different dimensions. This task requires performing a semantic mapping between the log file key words and the content of the different sources. Finally, we validated our approach with the modeling tool TOS. Our future work consist in extending this model to social networks like Facebook and Twitter. The Behavioral analysis is getting more and more interesting. It requires the use of preferences and particularities of internet users, which lead us to the concept of webhouse personalization. We intend then to use this concept in ETL processes.

References

Akkaoui, Z. E., J.-N. Mazon, A. A. Vaisman, and E. Zimanyi, E.nyi (2012). Bpmn-based conceptual modeling of etl processes. In A. Cuzzocrea and U. Dayal (Eds.), *DaWaK*, Volume 7448 of *Lecture Notes in Computer Science*, pp. 1–14. Springer.

Andersen, J., A. Giversen, A. H. Jensen, R. S. Larsen, T. B. Pedersen, and J. Skyt (2000). Analyzing clickstreams using subsessions. In R. Missaoui and I.-Y. Song (Eds.), *DOLAP*, pp. 25–32. ACM.

Charrad, M., Y. Lechevallier, M. B. Ahmed, and G. Saporta (2010). Wcum pour l'analyse d'un site web. In *EGC*, pp. 669–672.

Cooley, R., B. Mobasher, and J. Srivastava (1999). Data preparation for mining world wide web browsing patterns. *Knowledge and information systems 1*(1), 5–32.

Darmont, J., O. Boussaid, and F. Bentayeb (2007). Warehousing web data. In *ER*, Volume abs/0705.1456.

El-Sappagh, S. H. A., A. M. A. Hendawi, and A. H. E. Bastawissy (2011). A proposed model for data warehouse {ETL} processes. *Journal of King Saud University - Computer and Information Sciences 23*(2), 91 – 104.

Embley, D., D. Campbell, Y. Jiang, S. Liddle, Y.-K. Ng, D. Quass, and R. Smith (1998). A conceptual-modeling approach to extracting data from the web. In T.-W. Ling, S. Ram, and M. Lee (Eds.), *ER*, Volume 1507 of *Lecture Notes in Computer Science*, pp. 78–91. Springer Berlin Heidelberg.

Hernández, P., I. Garrigós, and J.-N. Mazón (2010). Model-driven development of multidimensional models from web log files. In *ER Workshops*, Volume 6413, pp. 170–179.

Kimball, R., R. Merz, and J.-M. Berthier (2000). *Le data webhouse : analyser les comportements client sur le web*. Solutions bases de données. Paris: Eyrolles. La couv. porte : Au sommaire : coupler les technologies data warehouse et web, commerce électronique et personnalisation client, analyse du clickstream et modélisation multidimensionnelle, guide de conduite de projets data warehouse.

Liu, Y., R. Chen, and H. Yang (2009). Web information extraction based on hierarchical model. In *Computational Intelligence and Software Engineering, 2009. CiSE 2009. International Conference on*, pp. 1–5.

Muñoz, L., J.-N. Mazón, J. Pardillo, and J. Trujillo (2008). Modelling etl processes of data warehouses with uml activity diagrams. In R. Meersman, Z. Tari, and P. Herrero (Eds.), *OTM Workshops*, Volume 5333 of *Lecture Notes in Computer Science*, pp. 44–53. Springer.

Muñoz, L., J.-N. Mazón, and J. Trujillo (2010). A family of experiments to validate measures for uml activity diagrams of etl processes in data warehouses. *Information & Software Technology 52*(11), 1188–1203.

OMG, O. M. (2007). Omg unified modeling language (omg uml), superstructure, v2.1.2.

Russell, N., W. M. P. van der Aalst, A. H. M. ter Hofstede, and P. Wohed (2006). On the suitability of uml 2.0 activity diagrams for business process modelling. In *APCCM '06: Proceedings of the 3rd Asia-Pacific conference on Conceptual modelling*, Darlinghurst, Australia, Australia, pp. 95–104. Australian Computer Society, Inc.

Sumathi C.P., Padmaja Valli R., S. T. (2011). An overview of preprocessing of web log files for web usage mining. *Journal of Theoretical and Applied Information Technology 34*, 88–95.

Trujillo, J. and S. Luján-Mora (2003). A uml based approach for modeling etl processes in data warehouses. In *ER*. Springer.

Vangipuram, R., V. Sreekanth, and B. Rangaswamy (2010). Implementation of web-etl transformation with pre-configured multi-source system connection and transformation mapping statistics report. In *Advanced Computer Theory and Engineering (ICACTE), 2010 3rd International Conference on*, Volume 2, pp. V2–317–V2–322.

Vassiliadis, P., A. Simitsis, and S. Skiadopoulos (2002). Conceptual modeling for etl processes. In *DOLAP '02: Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP*, New York, NY, USA, pp. 14–21. ACM.

Wilkinson, K., A. Simitsis, M. Castellanos, and U. Dayal (2010). Leveraging business process models for etl design. In *WorkShop ER*, Volume 6412, pp. 15–30.

ETL-Web process modeling

Résumé

Le changement rapide et le volume important des données web, génère une grande hétérogénéité au niveau de ces données. Alors, l'intégration de ces derniers au niveau du DWB devient plus complexe, ce qui exige le traitement, le nettoyage et l'importation de ces données à partir de différentes sources (fichiers log, sources métier, ...). Ce processus de traitement constitue le but des outils ETL (Extraction, Transformation, et Chargement). Afin de minimiser la complexité des traitements et d'augmenter la qualité des données web, la modélisation des processus ETL s'avère être une solution primordiale. Dans ce cadre, nous proposons de modélises les processus ETL adapté aux données web à travers le langage de modélisation conceptuelle UML 2.0.

Towards an approach for the development of a DWaaS with adaptable security requirements

Emna Guermazi[1], Hanêne Ben-Abdallah[2], Mounir Ben Ayed[3]
[1]MIR@CL Laboratory, University of Sfax, Tunisia
guermaziemna1@gmail.com
[2]King Abdulaziz University, Jeddah, KSA
hbenabdallah@kau.edu.sa
[3]REGIM Laboratory, University of Sfax, Tunisia
mounir.benayed@ieee.org

Summary. Cloud computing is increasingly been adopted by enterprises to deploy and manage their various business applications. Recently, enterprises started examining the Cloud for the deployment of their data warehouse systems. However, security remains a key challenge for a widespread adoption of the Cloud for data warehousing. In this paper, we first overview different approaches for the development of a secure data warehouse. Secondly, we propose an approach to develop a secure Data Warehouse as a Service in the Cloud. Finally, we outline an approach for adapting a security service in DWaaS to accommodate new security requirements.

1 Introduction

Despite the importance of data warehouses in the success of decision making processes, small and medium size enterprises are not able to afford them because of their prohibitively expensive cost (Herwig, 2013). With the emergence of Cloud computing, several enterprises could overcome the cost predicament and benefit from a data warehouse (DW) deployed as a service. Indeed, the Software as a Service (SaaS) model of the Cloud offers cost effective solutions (Liyang et al., 2011) by providing for reduced capital expenditures, lower operational costs and fast implementation times[1]. Recently, DW applications delivered via SaaS on the Cloud have proved to be the next generation in the BI market (Liyang et al., 2011). The resulting model, called Data Warehouse as a Service (DWaaS), can be used over the Internet on-demand and pay-per-use. Among the examples of DWaaS on the Cloud, we find Amazon Redshift[2]and Treasuredata[2].

When a data warehouse is deployed in the Cloud, security becomes a more pressing requirement. Moreover, DW security requirements can evolve due to two reasons: *(i)* the DW security model depends on its schema which can be modified after its implementation (Kimball, 1997) inducing changes on the security policy; and *(ii)* when a DW is deployed as software-as-a-service in the Cloud, the different tenants may have different security needs. Indeed, De-Capitani-di-Vimercati et al. (2007) affirm that when data is outsourced, selective authorization policies must be enforced and the support of policy updates must be addressed.

1. http://www.enterprisemanagement.com/research/asset-free.php/2102/pre/Cloud-Business-Intelligence-and-Data-Management-as-a-Service-pre
2. http://www.treasuredata.com

In addition, similar to any other secure system, a DW security solution must be modified after the detection of any vulnerability.

In this paper, we try to answer the following two questions: *(i)* what are the security requirements needed to have a secure DWaaS in the Cloud? and *(ii)* how to offer this solution as a service that meet pay-per-use model? To answer these questions, we start by comparing some approaches to secure a DW and analyze approaches to develop a DW using MDA approaches. Afterwards, we propose a solution for the development of a secure DWaaS and we outline an approach to meet changing security requirements. Our development approach has two merits. First, it produces a DW as a set of interacting services deployed in the Cloud; such architecture provides for a better scalability. Secondly, it treats security and privacy requirements as first-class citizens in the system design in order to cater to the security requirements of the inherently multi-tenant Cloud systems.

The remainder of the paper is organized as follows: Section 2 identifies DW security requirements. Section 3 overviews different approaches for the development of a secure DW. Section 4 first presents our approach to develop a secure Data Warehouse as a Service in the Cloud; secondly, it outlines an approach for adapting a security service in DWaaS to accommodate new security requirements.

2 DW security requirements in the Cloud

To examine existing works on secure data warehousing, we fixed a set of security requirements. We extended the standard types of security services proposed in the framework of Firesmith (2004) to account for the specificities of data warehouses and the Cloud computing environment.

The standard set of security services cover the following requirements:

- Access Control: defines the "degree to which the system limits access to its resources only to its authorized externals" Firesmith (2004). This security service covers identification, authentication and authorization services. In addition, it can be defined for direct access or indirect access where an authorized external infers partially or totally unauthorized data from data to which they have authorized access;
- Non-repudiation: is defined as "the degree to which a party to an interaction (e.g., message, transaction, transmission of data) is prevented from successfully repudiating (i.e., denying) any aspect of the interaction." Firesmith (2004);
- Security Auditing: is the degree to which security personnel are enabled to audit the status and the use of security mechanisms by analyzing security-related events Firesmith (2004);
- Privacy and Anonymity: represent the degree to which unauthorized parties are prevented from obtaining sensitive information Firesmith (2004);
- Recovery: describes the degree to which unintentional manipulated, corrupted or 'lost' data may be partially or fully recovered Höner (2013) ;
- Availability: represents the property of being accessible and usable upon demand by unauthorized entity (ISO 27001:2005).

To the above standard services, we added the following security criteria to meet specific characteristics of data warehouses and the Cloud computing context:

- Processing analytical queries over encrypted data: when a data warehouse is deployed in the Cloud, its sensitive data are encrypted. Such security mechanism induces a challenge for querying the encrypted data;
- Fine-grained access control on the Cloud: Allow the definition of access right in encrypted data through a fine management of decryption Keys Bhartiya et al(2012). An example Key-Policy Attribute Based Encryption (KP-ABE) Goyal, and al. (2006).
- Inference of the security policy from the source The Extract-Transform-Load operations must take into account the security model of the data source. They must handle any conflicting security requirements among the attributes extracted from the data source;
- Cloud Level: We choose to compare work also according to the level of Cloud in which the security solution is integrated (IaaS, PaaS or SaaS).
- Adaptable security policy: DW security requirement can evolve due to the:
  - Modification of the tenant DWaaS schema in the cloud: Since the security model depends on the Cloud DWaaS schema, the security model must therefore be modified when the tenant DWaaS schema changes;
  - Change of security needs of the different multi-tenant when a data warehouse is deployed as software-as-a-service (SaaS) in the Cloud: In fact, De-Capitani-di-Vimercati and al. (2007) affirm that when data is outsourced (to the Cloud), selective authorization policies must be enforced.

3 Related work

Cloud computing is a model for enabling convenient, on-demand network access, to shared pool of configurable computing resources, (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction (NIST, 2012). (d'Orazio, Bimonte; 2011) affirm that considering OLAP analysis and data warehousing in the cloud becomes a major issue.

An interesting solution on the market is Amazon Redshift which is an SQL data warehouse and uses industry standard ODBC and JDBC connections and Postgre SQL drivers. The security solution used in Amazon Redshift: all data written to disk will be encrypted using hardware-accelerated AES-256 and SSL. Amazon Redshift enables client to configure firewall rules to control network access to data warehouse cluster. Data warehouse can be isolated using cluster in a virtual network and connect it to existing IT infrastructure using industry-standard encrypted IPsec VPN. Keys are managed using hardware security modules (HSMs). All API calls, connection attempts, queries and changes to the cluster are logged and auditable[1].

(Essaidi M., 2010 a; 2010b, 2013) present ODBIS, an open source infrastructure to build and deliver On-Demand Business Intelligence Services. ODBIS supports services for DW projects management and models design based on a model-driven approach using the 2 Track Unified Process (2TUP) and the Model Driven Architecture (MDA).

[1]http://aws.amazon.com/

(Essaidi M., 2010 a; 2010b, 2013) present Model-Driven Data Warehouse as a service (MDDWS) represents an approach that aligns the development of data warehouse systems with a general model-driven development paradigm. MDDWS is a web-based model-driven data warehouse development environment and it's developed under the ODBIS platform (Essaidi, 2010 a). The Technical Architecture of ODBIS is based on Java Entreprise Edition (JEE) technologies using Spring framework. The technical architecture of ODBIS is based on Spring framework. At the data layer, authors use PostgreSQL. For the persistence layer, authors use Java Persistence API(JPA) to define the object-relational mapping using Java metadata annotation and Hibernate is used as persistence provider for JPA. The domain model is based on a java implementation of CWM metamodel which is defined using Java Metadata Interface (JMI) specification. JMI allow also metamodel and metedata interchange via XML by using the industry standard XML Metadata Interchange (XMI) specification. For the presentation layer authors use Java Server Faces (JSF) technology like Sun Mojarra, Apache Trinidad, and Spring Faces. All services run under Apache Tomcat web server. In terms of security, ODBIS uses a security framework for web applications a named spring security. The working principle of is based on a Spring Security filter chain. Each of these filters address a specific need related to the identification, authentication or authorization. Authentication can be realized for example through (BASIC method) X.509 certificate (servers CAS delegation)[1]. Access control can be realized by declaring security constraints using two methods: patterns to filter URL (*e.g.*, /admin/** -> ROLE_Director); and Application classes instrumentation by decorating methods with access rights (*e.g.*, @Secured("ROLE_Director") attached to a Boolean public method delete(MyResource resource))[2] .

In the Cloud, the deployment of analytical data at one Cloud service provider is not very secure. In fact, the provider may disappear, use analytical data in unauthorized manner, etc. In order to resolve this security problem, Karkouda et al. (2012) propose to share data among many providers through a Secret Sharing algorithm which ensures three security levels: 1) service availability through the use of secret sharing algorithm which is capable of data restitution based on a percentage of data less than 100% stored at each provider; 2) secure transaction between the client and provider because transiting data is incomplete and unusable; and 3) secure Cloud data storage because each provider has only partial data. To handle queries, this approach proposes to decompose each query into several queries according to the data sharing strategy across the providers; for example, if we choose three cloud service providers, the query "select CA from table where Month= parametres1 and Name=parameters2" would be decomposed into three queries: "select CA from table where Month= parameters11 and Name=parameters21", "select CA from table where Month= parameters12 and Name=parameters22" and "select CA from table where Month= parameters13 and Name=parameters23" to cover all three providers. Note that the higher level of security ensured by this approach comes with a higher storage cost (which will be proportional to the number of Cloud service providers).

---

[1]http://www.ippon.fr/documents/10226/26207/Ippon%20Technologies%20%20La%20s%C3%A9curit%C3%A9%20des%20applications%20Java%20EE.pdf
[2]http://ippon.developpez.com/tutoriels/java/gwt-spring-security

(Attasena et al., 2013) extend the work of (Karkouda et al., 2012) by using cryptography and hash function to maintain data integrity. They encrypted each data blocks before sharing it and in order to verify data correctness they propose two types of signatures: inner and outer which are created by hash function. The sharing process is composed of four steps: organizing data into blocks; creating signature in each block; encrypting data and signature; and creating a signature for each encrypted data. To reconstruct data, they apply five steps: selection of *t* out of *n* cloud service providers; verification of correctness of encrypted data; transfer of encrypted data to user; computation and signature of original data; and verification of data correctness. This proposition offers Security, Reliability, and Cost analysis.

Monomi (Tu et al., 2013) is a system for practical analytical query processing. Its goal is to provide a solution to two types of queries: sum of the multiplication (e.g.,SUM(cost * quantity)) and Inequality of summation of the multiplication (e.g.,SUM(cost * quantity) > 1000000). The technique used in this system relies on splitting a query and creating a query plan in order to find: 1) a query part to run on the server over the specific crypto systems; and 2) another query part to run at the client.

Table 1 synthesizes our evaluation of existing approaches for secure Cloud data warehouses. The notation used is the following: Direct access control (C11); indirect access control (C12); Non repudiation (C2); Auditing (C3); Anonymity (C4); Recovery (C5); Availability (C6); Processing analytical queries over encrypted data (C7); Fine-grained access control (C8); Inference of the security policy from the source (C9); Cloud level (C10); *Yes* for covered requirement, No for not covered, P for partially covered, NT for Not Mentioned, NI for not include in the comparisons (there are some security requirement which are related only to DW on the Cloud computing).

TAB. 1 – Evaluation of works for secure data warehouse as a service and works to deploy and query a secure DW in the Cloud

	Security requirements covered										
	$C1_1$	$C1_2$	C2	C3	C4	C5	C6	C7	C8	C9	C10
Essaidi M., 2010 a; 2010b	Y	No	NT	NT	No	No	No	No	No	No	NI
Karkouda, 2012; Attasena, 2013	Y	No	No	Yes	No	Y	Y	P	No	P	IaaS
Tu et al., 2013	Y	No	No	No	No	No	No	Y	No	No	IaaS

For the results shown in this table, we can conclude that few works address the development of DW as a service or securing a DWaaS. In addition, none of the DW security works supports the MDA-SOA architecture nor adaptable security policy security requirements.

In addition to the above presented works, we have studied the work of Atigui and al., 2012 on developping a DW using MDA oriented approach. The authors present a unified conceptual model that describes both the DW and its ETL process using the constellation model and the Object Constraint Language (OCL). We choose this approach for the

development of DWaaS but we plan to add the security and SOA perspective Also In (Mazón and Trujillo, 2008; Zepeda et al., 2008), MDA oriented approaches for the development of DWs are presented. And by applying the Model Driven Architecture (MDA) approach, Vela et al., 2013 proposed development approaches for secure data warehouses which incorporated security aspects into all stages of the DW development cycle. All these works do not take into account the software architecture design and especially the SOA architecture and they do not construct a secure DWaaS.

Table 2 synthesizes our evaluation of existing approaches for DW development using MDA. In fact, Service Oriented Architecture (SOA), is a recent software development approach. The SOA paradigm is based on 'separation of concerns' and the composition of services. Each service has a description interface. While externalizing security management and enforcement tasks from the application ensure that the security is not tightly coupled with the adopted platform.

TAB. 2 – Evaluation of works for DW development using MDA

	Service Oriented Architecture	Security aspect	Security externalization and security runtime verification
Atigui and al., 2012	No	No	No
Mazón and Trujillo, 2008	No	No	No
Zepeda and al 2008	No	No	No
Essaidi and Osmani 2011 b, 2013	No	No	No
Vela and al., 2013	No	Yes	No

4 DWaaS development approach and security requirement adaptation in a DWaaS

We suppose that the DW application is completely web-based and designed for sharing across multiple customers (via multi-tenancy). A Software as a Serivce (SaaS) is built from scratch using web-based architectures. With a SaaS, user data is housed within data centers controlled by the SaaS provider. The SaaS model exploits internal cost advantages from multi-tenancy and configurability. In fact, SaaS provides software application vendors a web-based delivery model to serve big amount of clients with multi-tenancy based infrastructure and application sharing architecture so as to get great benefits from the economy of scale (Guo and al. (2011)).

4.1. MDA and SOA approach for the development of secure DWaaS deployed in the Cloud

Through our approach, we aim to develop a data warehouse as a service oriented application and to implementing it in the Cloud. Our approach is based on a specific Service Oriented Architecture for model driven secure data warehouse development. It is adopted from Architecture-Centric Model-Driven Architecture (ACMDA) (Marcos and al. (2006)).

The CIM-DW contains the business model of analysis requirements, ETL process, and security requirements. At this level, the application domain is modeled, but because the architecture does not depend on the domain therefore it is not modeled.

The PIM-DW-SOA contains services involved in the architecture. More specifically, for this level, we identify three services: (1) MDM +ETL service, (2) Analysis service, and (3) Security services. Each of these services has a service description. At this level, we will determine the services and the behavior of services in terms of their composition, operations and interactions between different operations. In this paper, we focus only on presenting the services. This layer is typically realized with component-based middleware which utilizes the Service Oriented Architecture (SOA) pattern (Eggert and al. (2013)).

The PSM-DW-Cloud level contains the deployed services such as data base servers, ETL engine and analysis server. The PSM level can have different physical models (PSM) according to the deployment platform chosen by the designer. For example, the designer may choose Hive which is a data warehouse infrastructure that provides data summarization and ad hoc querying, he may also choose any other type of platform that contains a relational database like MySQL or PostgreSQL or any OLAP server.

The PSM should take into account how to place the security mechanisms within the architecture. At the PSM level, security patterns can be used to model security mechanisms. We identify three groups of security services to be modelled at this level:

1. The first service group (S1) contains two services: (S11) which generates a multi- dimensional model in terms UML profile or constellation model via model transformation; and service (S12) which generates OCL expressions that conceptually describe the extraction and multidimensional attributes transformation formulas (projection, conversion, selection and aggregation).
2. The second service group (S2), DW analysis service, allows to receive analysis query and to select a set of instances that satisfies the query and to apply aggregation functions such min max or to apply operator such as Cube and to display the analysis result. The query can be written using SQL, MDX or HiveQL.
3. The third service group (S3) contains four services: (S31) is a security service linked to the DW specificity (*i.e.* managing the security of multidimensional model and of ETL processes); (S32) is a security service related to the DW domain (e.g the medical field contains private data requiring the application of anonymity security requirements but the commercial domain does not requires this security requirements); (S33) is a security service linked to specific Service Oriented Architecture.; (S34) is a security service related to the specificity of deployment model which is in our case Cloud computing such cryptography

solution as an approach for Cloud data storage security ,or or fine grained access control solution

A secure DWaaS architecture is given in Figure 1 a metadata store information of tenant-functionality. Every event responsible caused by changing security requirements will modifies Meta-Data. If Meta-Data was modified then materialized view security mechanisms must be updated. All elements communicate via SOAP or REST messages and are outsourced in cloud platforms. With multi-tenancy, the hosted SaaS application runs a single code base for all customers, while ensuring that the data seen by each customer is specific to them. According to the needs of each client, a security service is elaborated to secure the data warehouse. Some security services are common to all tenants. But the client needs are variable, so the security service must adapt to these changes. If the schema source changes then the security service change to ensure that the DW is well secured, we test the security of the DW and if we detects a vulnerability then security service will change

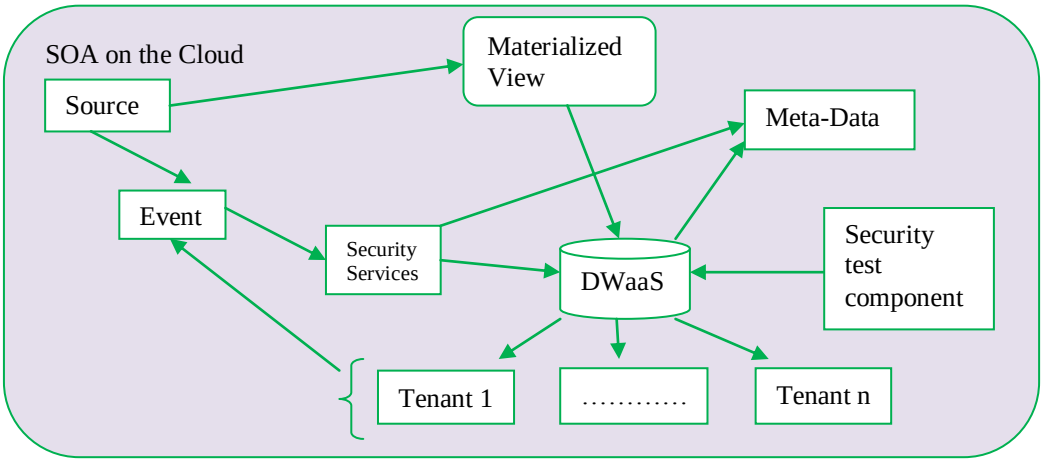


FIG. 1– **Secure DWaaS architecture**

4.2. DWaaS adaptation approach to meet changing security requirements

In this section, we present a solution that is based on the service oriented architecture to adapt Cloud DWaaS in order to meet changing security requirements. In fact, this on-demand solution is re-executed each time security requirements change. The result is the generation of a new code and the deployment of the new DW or a new virtual DW in the Cloud after each DWaaS modification.

When an event causes the change of a security requirement or security mechanism, the DWaaS will be adapted to take into account this change. This adaptability ensures security elasticity. If we need to change or to add a security service, we propose to regenerate security code using Aspect-oriented Programming (AOP) automatically to adapt to the

propose change. The security code will be generated using SOA-model driven approach (discussed section 4.1.).

The steps to adapt a DWaaS to meet changing security requirements are:

1. New security requirements flowing an event such as changing cloud DWaaS schema or changing security requirement of multi-tenants DWaaS web application.
2. Automatically generation of executable service deployed in the Cloud after adding new security service or modifying old security service.

Conclusion

In this paper, first we have presented a comprehensive comparison between DW security approaches based on a set of criteria fixed to cover the security requirements for data warehouses. Second we have presented an example of adapting security mechanism to meet changing security requirements in the Cloud DW. Third we have presented a DWaaS development approach and security mechanisms adaptation in a DWaaS. In the future we aim to detail the behavior between the identified services and also to detail the CIM-DW, PIM-SOA-DW and PSM-Cloud-DW level.

Acknowledgements

The presented research and innovation are performed in the context of a MOBIDOC thesis financed by the European Union within the program PASRI.

References

Agrawal D., El-Abbadi A., Wang S. (2013). Secure and Privacy-Preserving Database Services in the Cloud, Tutorial in ICDE. 29th international conference on data engineering, Bisbane, Australia 2013, april 8-11.

Atigui, F., Ravat F., Teste O., Zurfluh G. (2012) Modélisation conjointe des données et des processus pour l'implantation de schémas d'entrepôts. Journal of DecisionSystems, 21(1), 27-49

Attasena, V. Harbi N. Darmont J. (2013). Sharing-basedPrivacy and Availability of Cloud DataWarehouses.9èmes journées francophones sur les Entrepôts de Données et l'Analyse en ligne EDA 2013, vol. B-9, p.17-32.

Bhartiya A.S., Agrawal L.S., Gawande Y.V., Rapartiwar S.S; Achieving sheltered , Scalable and fine-grained data access control in Cloud computing ; World Research Journal of Engineering and Technology ; ISSN: 2278-8530 & E-ISSN: 2278-8549, Volume 1, Issue 1, 2012, pp.-04-07.

De-Capitani-di-Vimercati S., F. Jajodia, S. Paraboschi, P. Samarati (2007). Over-encryption: Management of Access Control Evolution on Outsourced Data. VLDB 123-134.

d'Orazio L., Bimonte S.(2011). Entreposage et analyse en ligne dans les nuages avec Pig. Ingénierie des Systèmesd'Information 16(6): 139-162 .

Dubé E., Badard T., Bédard Y. Une architecture orientée service web pour la constitution de minicubes SOLAP pour clients mobiles. Revue Internationale de Géomatique 19(2): 211-230 (2009)

Essaidi, M. (2010a) ODBIS: towards a platform for on-demand business intelligence services. EDBT/ICDT Workshops 2010

Essaidi, M. A. Osmani A (2010b) Model driven data warehouse using MDA and 2TUP. J. Comput. Meth.in Science and Engineering 10(3-6): 119-134.

Essaidi, M. A. Osmani (2011a) Business Intelligence-as-a-Service: Studying the Functional and the Technical Architectures. Business Intelligence Applications and the Web: Models, Systems and Technologies. (pp. 199-221). IGI Global, September.

Essaidi, M. "Model-Driven Data Warehouse And Its Automation Using Machine Learning Techniques"; thèse à L'UNIVERSITE PARIS XIII; 2013

Essaidi, M. A. Osmani (2011 b) Trnasforming Learning in the Context of Model-Driven Data Warehouse : An Experimental Design Based on Inductive Logic Programming. 23rd IEEE International Conference on Tools with Artificial Intelligence.

Firesmith, D. (2004), Specifying reusable security requirements. Journal of Object Technology, 3, 1, 61-75.

Goyal, V. O. Pandey, A. Sahai, B. Waters, Attribute-based encryption for fine-grained access control of encrypted data, in Proc. Of CCS'06, 2006.

Guo C., Wei Sun, Zhong-Bo Jiang, Ying Huang, Bo Gao, and Zhi-Hu Wang; Study of Software as a Service Support Platform for Small and Medium Businesses ; Information and Software as Services, LNBIP 74, pp. 1–30, 2011; © Springer-Verlag Berlin Heidelberg 2011

Herwig V . (2013) business intelligence as a service for Cloud based application; The 7th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications.

Höner, P. (2013). Cloud Computing Security Requirements and Solutions: a Systematic Literature Review, 19 Twente Student Conference on IT, June 24, 2013, Enschede, The Netherlands.

Jun L., Hu ChaoJu,Yuan HeJin ,Application of Web Services on The Real-time Data Warehouse Technology; 2010 International Conference on Advances in Energy Engineering

Karkouda, K. Harbi N. Darmont J., Gavin G. (2012). Confidentialité et disponibilté des données entreposées dans les nuages, 9ème atelier Fouille de données complexes, Bordeaux : France.

Kimball, R.(1997).The Data warehouse Toolkit, John wiley and sons, Inc.

Liyang,T. Zhiwe,iN. Zhangjun,W. Li W.(2011).A Conceptual Framework for Business Intelligenceas a Service (SaaSBI) Intelligent Computation Technology and Automation (ICICTA).

Marcos, E., Cesar J. Acuña, and Carlos E. Cuesta; Integrating Software Architecture into a MDA Framework. V. Gruhn and F. Oquendo (Eds.): EWSA 2006, LNCS 4344, pp. 127 –143, 2006. © Springer-Verlag Berlin Heidelberg 2006.

Mazón, J.-N. J. Trujillo (2008). An MDA approach for the development of data warehouses. Decis. Support Syst., 45(1):41–58.

NIST. (2012). National Institute of Standards and Technology, The NIST Definition of Cloud Computing, Information Technology Laboratory, 2009

Pintar D., Mihaela Vranić, Zoran Skočir. (2009) Metadata-driven SOA-based application for facilitation of real-time data warehousing

Sen, S. ; Datta, D. ; Chaki, N. An Architecture to Maintain Materialized View in Cloud Computing Environment for OLAP Processing; Computing Sciences (ICCS), 2012 International Conference on

Sharma, R, S. Manu, S. Divya, Modeling Cloud SaaS with SOA and MDA Advances in Computing and Communications Communications in Computer and Information Science Volume 190, 2011, pp 511-518

Shroff G. (2010) enterprise cloud computing technology architecture application.

Vela B., Mazónb J.N., Blanco C., Fernández-Medinad E., Trujillo J., Marcosa E. (2013) Development of Secure XML Data Warehouses with QVT. Information and Software Technology, Volume 55, Issue 9, September , Pages 1651–1677.

Zepeda, L. M. Celma, R. Zatarain (2008). A Mixed Approach for Data Warehouse Conceptual Design with MDA. In ICCSA, pages 1204–1217. Springer.

Résumé

Les entrepôts de données contiennent des informations très sensibles, de ce fait il est indispensable de leurs définir des mesures de sécurité. Dans ce papier, nous comparons différentes approches pour la sécurité des entrepôts de données dans les nuages et les entrepôts de données comme un service. De plus, nous proposons (1) une approche de développement d'ED sécurisé comme un service (2) une approche pour adapter un ED comme un service pour répondre aux nouvelles besoin de sécurité induites par le changement de schéma d'ED, ED avec plusieurs utilisateurs, de l'infrastructure.

Organizational structure assessment based on structural analysis: many-to-many Relations-Process Integration

Azedine BOULMAKOUL, Zineb BESRI

LIM/IDS Lab. Computer Sciences Department, Mohammedia Faculty of Sciences and Technology, B.P. 146 Mohammedia, Morocco
azedine.boulmakoul@gmail.com, z.besri@gmail.com

Abstract. To maintain and strengthen their positioning, organizations are registered in a continuous process of quality improvement. In this context, they regularly carry a critical review or redesign their processes for more effectiveness and efficiency. Such projects often supports organizational overhaul, the implementation of a new information system. To carry out and achieve the desired objectives, organizations must be able to analyze their current modes of operation, to compare best practices and engage all stakeholders to rethink and restructure the process mode. The redesign is to conduct changes of distribution of work, allocation of tasks to resources involved and change the roles of performers if necessary. The organizational structure is usually designed to follow the strategy of the company. A strength company depends on its organizational structure and performance of its staff. The paper proposes identifying current dysfunctions of business processes and the definition of areas for improvement with concrete and measurable impact. It proposes a method for assessment of organizational structure based on structural analysis. A case study is presented to assess organizational structure via many-to-many relationship-process. The diagnosis detects non-conformances of the business organization. It analyses the organizational structure through structural indicators such as eccentricity, complexity, betweenness-centrality and closeness-centrality. The paper takes an example of two business process from Telecommunication Company. It derives two relationships that are: collaborative relationship and transfer relationship work.

Keywords. q-analysis, business process, structural analysis, algebraic topology, organization redesign.

1 Introduction

Organization redesign is a powerful tool for improving organizational structure and individual performance. Reorganization is the most frequent response of management to a performance problem and how to design to strategies for implantation. The organization redesign is concerned with changing the assigned goals, roles, responsibilities and relationships within a given enterprise organization. Why an organization should be redesigned? To respond this question we should at first assess the enterprise organizational structure and get a complete diagnose about it.

The objective of organizational diagnosis is the analysis of the organizational structure, leading to identify the cause of that failure and non-compliance. It also allows evaluates and makes an overall judgment to highlight the potentials and weaknesses of a company and to identify competitiveness factors. In general diagnosis generates a report with indicators measured. The shape of the diagnosis depends on the nature of the planned objectives, the means and resources available. Several shapes are identified and can be divided into three types of diagnostic (Clayton 1980):

- **Global diagnostic (in-depth diagnostic)** the basic model that analyzes the activity from a global perspective across functions and organizational structure and leads to proposals for improvement;
- **Express diagnostic.** This diagnosis objectives to identify the origins of the problems, but also to formulate measures speedy overhaul and prioritize actions to be taken in order of importance;
- **Functional diagnostic.** This is a function of specialized diagnostic fragmentary.

This paper focuses on the diagnosis in depth as it is to identify and ask the real problems and how realistic solutions to address them. This diagnosis is a systemic analysis of the environment in which the company's market and competitive position. In addition, there is a detailed and comprehensive analysis of the various components of the organizational structure indeed, provides a means to evaluate the structure of the organization by mapping and analyzing the relationships between employees and organizational units.

The work focuses on a particular type of assessment of the organizational structure of the company. The various structural indicators are computed for organizational assessment. Structural indicators like eccentricity, complexity and centrality of the simplicial complex in the proposed case. The paper came through a series of concepts and methods using the structural analysis and methods of simplicial complexes introduced by Atkin (Atkin 1974, 1977). Not only for the representation and discover of the informal organizational structure, but also for the analysis of structure discovery which aims to measure the non-compliance with the formal organizational structure. The theoretical concepts and associated methods are based on algebraic topology for structural analysis. Discover knowledge representatives a collaborative relationship between a set of performers and a set of activities. Or transfer work relationship between a set of performers and a set of organizational units. These relationships are derived in part from a given set of business processes. The analysis results are presented in diagnosis map. It will help in making management decisions for any organizational change. Finally, the paper describes the involvement of activity-relationships performers-organizational units. So how does this impact on the structure of the organization and productivity and business performance.

The remainder of this paper is organized as follows: Section 2 presents related work and focuses on the process of mining the organizational perspective. Then Section 3 presents the structural analysis framework. It allows to analyze and assess proposed in Section 4 organizational structure of the case. Then in Section 5 describes the application of structural analysis and integration of the many-to-many relationship-process. Finally in Section 6 as a synthesis conclusions concerning this document and present our perspective of possible future work.

2 Related Works

The organization plays a crucial role in achieving a competitive advantage for them. Therefore, a system of information management must be designed for change and organizational redesigns. There are many fields of application and research need to be explored. Like the organizational culture (study of collective human behavior which are part of an organization) (Kirikova 2007), management skills of employees and the precise distribution of work and assignments based on information skills updated (Kirikova2007, Schein1996). The basic concepts of the organization are: According to (Mintzberg 1992) an organization is the rational coordination of a number of people to achieve some common goals, the division of

labor and function, and a hierarchy of authority and responsibility. (Weber 1947), it is a collective action in pursuit of the realization of a common mission. And (Wil et al 2011) is a set of constraints on the activities carried out by a group of agents collaborating. Therefore, we can say that the organization is the sum of the means of division of labor among its members and to coordinate the results of different tasks.

In general, business process management consists of components that allow model to identify, analyze and manage business process models. It is through the use of all the entities which are needed to describe the workflow. Then the components that support the performers to fill key roles for the execution, monitoring and invoking the business process model defined by the first component. The purpose of the extraction process is to extract useful information from event logs that record the activities performed in an organization (Aalst 1998). Process mining can extract information from different points of view. The process perspective focuses on flow control. The goal here is to find a meaningful characterization of all possible paths through the model or kneaded scoring model business processes (Aalst et al 2007, Song et al 2006). The organizational perspective is the impact of resources, where the performers are involved in the process model and how they are related. The main objectives are: structuring the organization to classify people into roles and organizational units and show the relationships between the performers.

Information systems become increasingly interrelated with business processes they support. Consequently, a multitude of events are recorded by information systems. However, organizations have problems with data extraction. (Ying et al 2009) Process mining can extract useful information on the studied process complements existing in the business process management approaches (BPM). BPM is the discipline that combines knowledge of information technology and knowledge of management science and applies it to business processes (Aalst 1998). The purpose of the extraction process is to use the event data to extract information related to the process, to discover, monitor and improve real processes, the processes are not assumed by knowledge extraction from event logs readily available in enterprise information systems. It is between the data extraction and modeling and analysis process. We establish links between the actual processes and their data on the one hand and process models, on the other hand (Song et al 2006).The proposed work is the diagnosis through a structural analysis of the organization and redesign of the organizational structure of the company.

3 Organizational diagnosis framework

Topological foundation and structural approach reflect and models the whole communication and business processes in an organization. Who is responsible to whom and for what? The following approach provides a way to determine bottlenecks and conformance degree with the real business processes by structural quantities and indicator measures.

The proposed methodology is applied to the following solution architecture (figure 1):

Capturing the enterprise world; organizational chart, business processes, roles, activities, performers… etc. Other enterprise components for diagnosis purpose such as event log and organizational structure investigations for process-mining are also taken into consideration. The paper use the meta-model proposed in (Boulmakoul et al 2012) for modeling enterprise architecture. We use the no-SQL system especially graph database. It is an alternative use of the traditional database management system. Considering the organization as professional network it is evident to choose a graph database. With a graph database, the focus is on the connections between data. Telling the database in advance that things are connected

and how, and representing those relationships physically, as opposed to storing them in tables and relating them through indexes. The graph database used in the case study is neo4J.

Then extract information's need for a redesign i.e. Business processes that have bottlenecks and need to be analyzed and assessed. By giving specific queries to ask the graph database, we derive the responsible relationships of those bottlenecks in the selected business process set. Or get a process mining before extracting relationships to overview bottlenecks and breakdowns. The extracted relationship can be represented by an adjacency matrix. It is the input of the next step, structural analysis.
We extract meaningful relationships from graph database using specific query language, Cypher. Query result can be exported as CSV, XML or JSON file. We derive the responsible relationship as an adjacency matrix. It is the input of the structural analysis framework

Topological structure analysis; using simplicial complex method, the analysis technique provides a diagnosis map of the whole organizational structure under the chosen relationships set. The diagnosis map summaries all structural quantities and indicator measures. Help top management to make decisions and chose the adequate actions and re-designs' operators for developing an alternative organization for more effectiveness. The new organizational structure is also a target for a new structural analysis until having a stable organizational structure with more compliance. Decision maker chose the right actions in order to be aligned with the goal and enterprise strategies.

In analysis phase the framework takes the shared-faced matrix of a chosen simplicial complex KP (A, λ) or KA (P, λ) to get co-occurrence information's in order to get structured vector "Q" as q-analysis result. Q-analysis result leads to measure structural quantities like eccentricities, complexity and other indicator like centrality. These measures defined the diagnosis map: it includes structural quantities and indicators like eccentricity, structural complexity and centrality. Those indicators and measures are to characterize term co-occurrence information about a dataset.

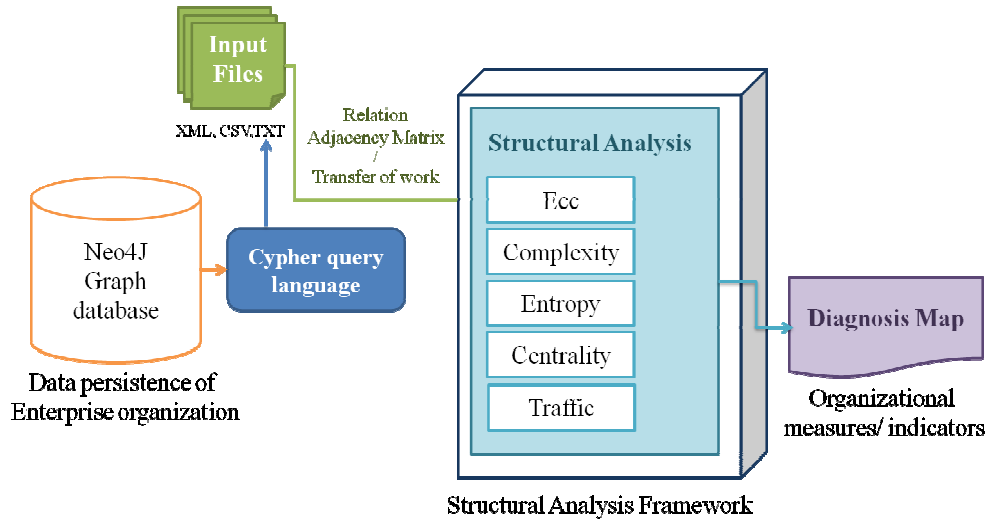


FIG. 1 – *Structural analysis framework*

4 Structural analysis for the diagnosis of the company structure's

We propose a case study of a telecommunication company. The paper focuses on the re-organization of the above company. Taking into consideration an organization structure and a set of business processes that have bottlenecks, we chose to make the diagnosis through two relationships from two business processes. The evaluation of this organizational structure is through the structural analysis. In addition we calculate various indicators measures to assess non-compliance of the organization.

In this document the use of several tools to redesign the organizational structure of the company. We used the ARIS model to capture the organizational structure and the different business processes via the notation of business process model (BPMN).

Figure 2 outlines the organizational structure of the company taken to end testing and analysis. The company consists of 5 directions; Branch, technical management, commercial management and marketing, finance and administrative management and human resources management. Each direction in turn consists of organizational units. Technical management includes the operations department, infrastructure department and after-sales service, IT and R&D department. The business management and marketing consists of two sales and marketing departments. Administrative and finance management includes accounting department, finance department, legal department and procurement department and logistics. Human resources management consists of personnel departments, department assistance and social work services.

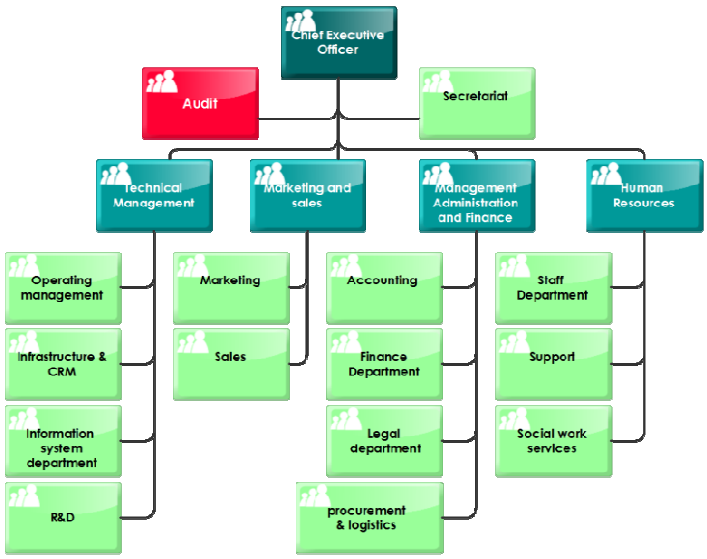


FIG. 2 – *Telecommunication's services company chart diagram*

4.1 Capturing business processes

To diagnose the organizational structure must firstly extract business processes that generate non-compliance and inefficiencies. From these processes derive intervening relationships of the above problems. In this case study we chose two business processes related to

changes in providing telecommunications services. Figures 3 and 4 illustrate the captured business processes.

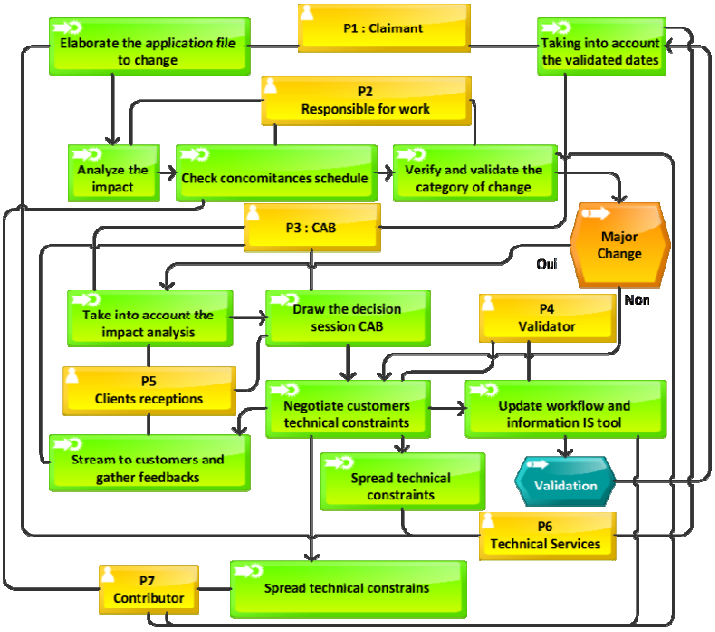


FIG. 3 – *Business process 1*

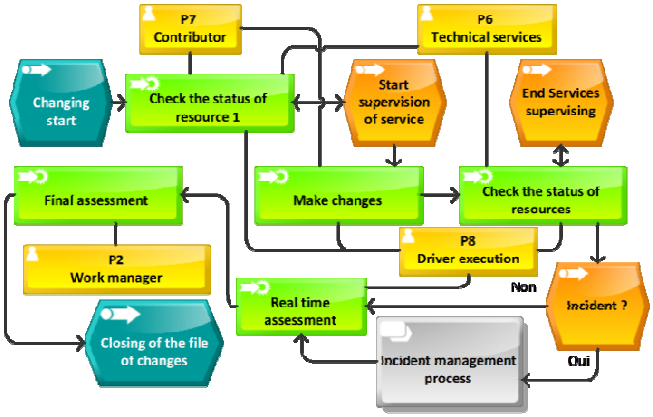


FIG. 4 – *Business process 2*

4.2 Deriving relationships set from the chosen business processes

There are many types of relationships, such as collaborative relationship, transfer of work relationship, hierarchical relationship, etc. From these relations we evaluate the organizational structure through simplicial complexes and structural analysis. After capturing business processes, we record all relationships between employees and activities of these processes and the relationships between organizational units and resources. Data persistence is via the database Neo4j graph. This is a database system of the Not-SQL. Several relationships can be studied. Resources/activities, activity/business processes, organizational units/resources,

etc. The two derived relations are: The first relationship λ_1, is the collaborative relationship. It represents the relationship between performers and activities of the business process. This type of relationship is shown schematically through an adjacency matrix λ= (P x A). Tables I and II represent the results of querying the database graph Neo4j data. We use Cypher, specific query language.

BUSINESS PROCESS' ADJACENCY MATRIX

λ_1	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}
P_1	1	1	0	0	0	0	0	0	0	0	0	0
P_2	0	0	1	1	1	0	0	0	0	0	0	0
P_3	0	1	0	0	0	1	1	0	0	1	0	0
P_4	0	0	0	0	0	0	0	1	1	0	0	0
P_5	0	0	0	0	0	1	1	0	0	1	0	0
P_6	1	1	0	0	0	0	0	0	0	0	1	0
P_7	0	0	0	1	1	0	0	0	1	0	0	1

TAB. 1 – Adjacency matrix of cooperation relationship between activity and performers from the business process 1

BUSINESS PROCESS'ADJACENCY MATRIX

λ_2	A_{13}	A_{14}	A_{15}	A_{16}	A_{17}
P_2	0	0	0	0	1
P_6	1	0	1	0	0
P_7	1	1	0	0	0
P_8	1	1	1	0	0

TAB. 2 – Adjacency matrix of cooperation relationship between activity and performers from the business process 2

The second relation λ_2, is the transfer-of-work is that organizational units are related if there is a transfer-of-work from an organizational unit to another. It connects between organizational units and performers. This relationship is represented through a graph whose nodes are the branches or departments. Vertices are shared among employees' two organizational units. Nodes are disc-shaped variable diameter Figures 5 and 6. These two relationships will be subject to structural analysis.

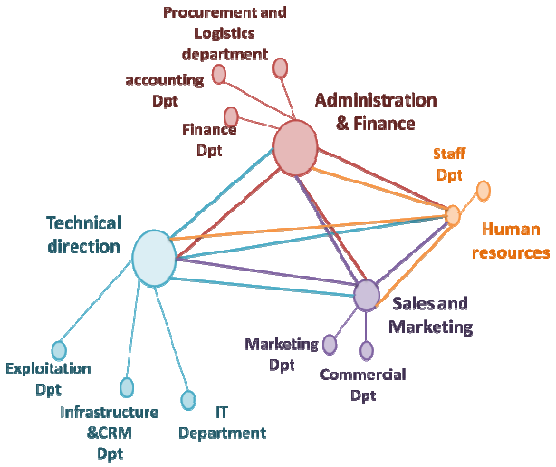


FIG. 5 – *Transfer of work related to business processes 1*

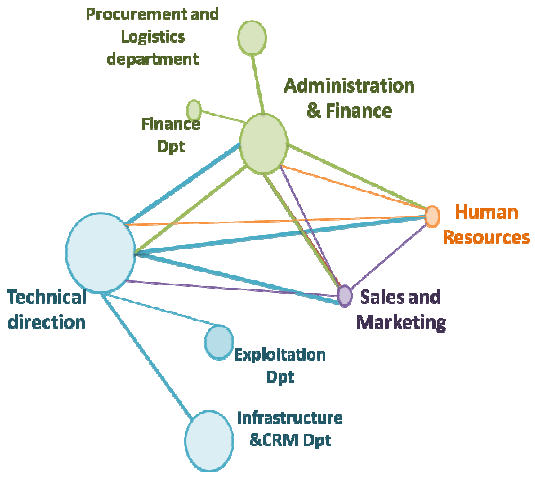


FIG. 6 – *Transfer of work related to business processes 2*

5 Structural analysis results

Structural analysis allow us to

- Compare formal organizational structure with that emerging from the analysis;
- Identify resources or activities eccentric organizational unit and process;
- Measure the complexity of the system;
- Measure indicators such as closeness centrality and betweenness centrality;
- Help managers to make an organizational redesign;
- Study the impact of the redesign of the organization on the organizational structure of the company.

5.1 Analysis of the relationship λ_1 of process1 and process2

Activity view: Table 3 represents the structural analysis of activities by simplicial complexes. We used the Q-analysis method. It reveals the chain connectivity (Boulmakoul et al 2012). Structured vector Q summarizes the results of Q-analysis. It shows the different equivalence classes with their dimensionality. This vector will be used subsequently for the calculation of structural indicators and the communication rate of this complex.

Q-analysis of shared face matrix for KA (P; λ)	***Q-analysis of shared face matrix for KA (P; λ)***
At **q = 2** we have $Q_2 = 1$; $\{A_2\}$ At **q = 1** we have $Q_1 = 4$; $\{A_1, A_2\}\{A_4, A_5\}\{A_6, A_7, A_{10}\}\{A_9\}$ At **q =0** we have $Q_0 = 1$; $\{A_1, \ldots A_{12}\}$ $Q = \{^{2\ 1\ 0}_{1\ 4\ 1}\}$	At **q = 2** we have $Q_2 = 1$; $\{A_{13}\}$ At **q = 1** we have $Q_1 = 1$; $\{A_{13}, A_{14}, A_{15}\}$ At **q =0** we have $Q_0 = 1$; $\{A_{13}, A_{14}, A_{15}, A_{17}\}$ $Q = \{^{2\ 1\ 0}_{1\ 1\ 1}\}$

TAB. 3 – Structural analysis of the relationship λ_1 extracted the first BP (column1), second BP(column 2). –Activity view-

Performer view: Table 4 represents the structural analysis of performers by simplicial complexes. We used the simplicial complex method that reveals the chain connectivity. Structured vector "Q" summarizes the results of Q-analysis. It shows the different equivalence classes with their dimensionality. This vector will be used later for the calculation of structural indicators and assess the communication within the complex

Q-analysis of shared face matrix for KP (A; λ)	*Q-analysis of shared face matrix for KP (A; λ)*
At q = 3 we have Q_3 = 2; {P_3}{P_7} At **q = 2** we have Q_2 = 4; {P_3, P_5} {P_2} {P_6}{P_7} At **q = 1** we have Q_1 = 5; {P_3, P_5} {P_2, P_7} {P_1}{ P_4}{ P_6} At **q =0** we have Q_0 = 1; all ; {P_1, P_2, P_3, P_4, P_5, P_6, P_7} $Q = \{^{3\ 2\ 1\ 0}_{2\ 4\ 5\ 1}\}$	At **q = 2** we have Q_2 = 1; {P_8} At **q = 1** we have Q_1 = 2; {P_8, P_7}{ P_6} At **q =0** we have Q_0 = 1; all ; {P_2, P_6, P_7, P_8} $Q = \{^{2\ 1\ 0}_{1\ 2\ 1}\}$

TAB. 4 – Structural analysis of the relationship λ_1 extracted the first BP (column1), second BP(column 2). –Performer view-

5.2 Structural indicators measure

Q-analysis provides a set of structural vectors that reflect the relationship of incidence of Vectors and chain structure of higher dimension, and also provides a new way of viewing organizational complexity through the analysis of organizational processes. Several structural indicators can be computered to characterize the overall structure (complexity) and the local structure (concentricity, eccentricity) of a complex.

5.2.1 Eccentricity

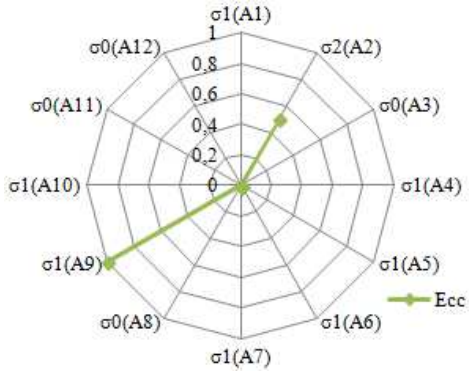
The eccentricity measures the degree of proximity of a component among all components of a complex. It defines the status of a simplex within the entire complex. By indicating the degree of integration of a specific complex with respect to the simplex complex, (Jiang et al 2006, Duckstein1997, Casti et al 1979, Dandan) defining a conventional measurement of the eccentricity noted ECC (Equation 1):

$$ecc(\sigma) = \frac{\hat{q} - \check{q}}{\check{q} + 1} \qquad (1)$$

With top-q "q̂" rank the level where the simplex appears first in the complex. "q̌" This is the rank of level where the simplex appears connected with another simplex in the complex. A simplex is said eccentric, if not properly integrated in the complex.

Activity view: Table 5 gives the values of the indicator eccentricity of the activity complex relationship λ_1 of business process 1

Simplex	Ecc
$\sigma_1(A_1)$	0
$\sigma_2(A_2)$	0.5
$\sigma_0(A_3)$	0
$\sigma_1(A_4)$	0
$\sigma_1(A_5)$	0
$\sigma_1(A_6)$	0
$\sigma_1(A_7)$	0
$\sigma_0(A_8)$	0
$\sigma_1(A_9)$	1
$\sigma_1(A_{10})$	0
$\sigma_0(A_{11})$	0
$\sigma_0(A_{12})$	0

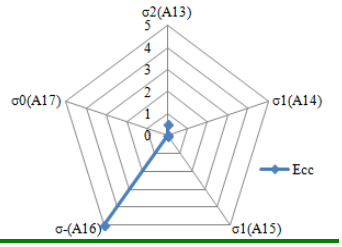


TAB. 5 – Eccentricity of each activity relationship λ_1 of business process 1

We observe that the simplex $\sigma_1(A_9) = 1$. It is eccentric relative to the other simplex complex KA (P; λ_1). As regards simplex $\sigma_2(A_2) = 0.5$. It is relatively with respect to the eccentric assembly. The rest of simplexes are concentric. It indicates that the correct definition and allocation of tasks to the activities of business process 1.

Table 6 gives the values of the indicator eccentricity activity of the complex relationship $\lambda_1$of business process 2.

Simplex	Ecc
$\sigma_2(A_{13})$	0.5
$\sigma_1(A_{14})$	0
$\sigma_1(A_{15})$	0
$\sigma_{-}(A_{16})$	∞
$\sigma_0(A_{17})$	0

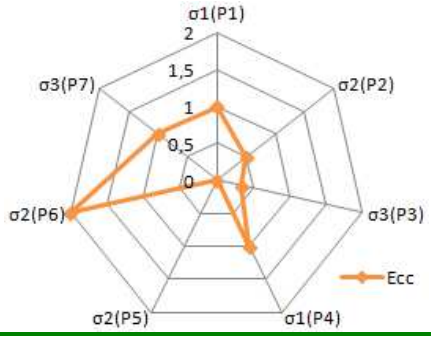


TAB. 6 – Eccentricity of each activity relationship λ_1 of business process 2

The analysis of the relationship λ_1 business process 2, we give the results in table 6: simplex $\sigma_{-}(A_{16}) = \infty$. It is most eccentric with respect to the other simplex complex KA (P ; λ_1). As regards simplex $\sigma_2(A_{13}) = 0.5$. It is relatively with respect to the eccentric assembly. The rest of simplexes are concentric. Indicating that the correct definition and allocation of tasks to the activities of business process 1

Performer view: Table 7 gives the values of the eccentricity indicator complex performer λ_1 of the business process 1.

SIMPLEX	**ECC**
$\sigma_1(P_1)$	1
$\sigma_2(P_2)$	0.5
$\sigma_3(P_3)$	0.33
$\sigma_1(P_4)$	1
$\sigma_2(P_5)$	0
$\sigma_2(P_6)$	2
$\sigma_3(P_7)$	1

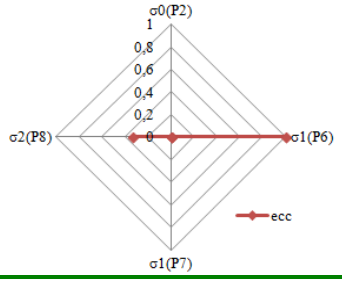


TAB. 7 – Eccentricity of each performer relationship λ_1 of business process 1

We note that the simplex $\sigma_2(P_6) = 2$. It is most eccentric with respect to the other simplex complex KP(A ; λ_1). As regards simplex $\sigma_1(P_1)$, $\sigma_2(P_2)$, $\sigma_3(P_3)$, $\sigma_1(P_4)$ and $\sigma_3(P_7)$, they are relatively eccentric. The simplex $\sigma_2(P_5)$ is concentric. Indicate that the correct definition and allocation of tasks to the activities of business process 1. We can conclude that the eccentric simplexes share tasks and activities in other organizational units as defining their roles

The table 8 gives the values of the indicator eccentricity complex performer λ_1 relationship in business process 2.

SIMPLEX	ECC
$\sigma_0(P_2)$	0
$\sigma_1(P_6)$	1
$\sigma_1(P_7)$	0
$\sigma_2(P_8)$	0.33



TAB. 8 – Eccentricity of each performer relationship λ_1 of business process 2

We note that the simplex $\sigma_2(P_6) = 1$. It is most eccentric with respect to the other complex $KP(A ; \lambda_1)$. As regards simplex $\sigma_0(P_2)$, $\sigma_1(P_7)$, they are concentric in the complex. This indicates good allocation of activities in relation to their roles in the organizational unit they belong eccentric.

5.2.2 Centrality

Characterizing networks who are most central? Which nodes are most 'central'? Definition of 'central' varies by context/purpose. Centrality defines who's important based on their network position. There are different kind of centrality measures; local measure with the degree indicator. Closeness, betweenness, eigenvector (Bonacich power centrality), etc are for expressing relative to rest of network. (Sabidussi 1966).

The paper proposes the use of tow centrality measure; betweenness and closeness.

Betweenness: The betweenness centrality is a bond for describing the linkage status of a simplex within a complex. This measure computes the influence that a node has over the spread of information through the network. A node (i.e., Performer) with high betweenness centrality value means that it performs a crucial role in the enterprise organization.. If this node is the only bridge linking these two groups and for some reason this node is no longer available, the change of information and knowledge between these two groups would be impossible. It identifies the "relay individuals". Thereby determine the betweenness centrality of a given simplex by the number of times a node is on the path of all the other pairs of nodes (Freeman 1979). (See equation 2).

$$B(\sigma_i)_q = \sum_i \sum_j P_{ikj} / P_{ij} \qquad \textbf{(2)}$$

Where P_{ij} is the number of shortest paths between simplicies σ_i and σ_j, and P_{ijk} is the number of shortest path from σ_i to σ_j through transational simplex σ_k.

Closeness: Closeness examines how a node is integrated within an organization structure. A lower closeness centrality value indicates a more central position in the graph organization structure. This means that a performer with a higher closeness centrality value, the number of individuals by which the performer must pass to get in touch with other performers in the system. One the other hand, another performer with a lower closeness centrality value is able to contact the same performer with fewer steps. (Freeman1979, Beauchamp 1966). (see equation 3).

$$C(\sigma_i)_q = n - 1 \Big/ \sum_{k=1}^{n} d(\sigma_i, \sigma_k) \qquad \textbf{(3)}$$

Where d is the shortest distance from a given simplex σ_i to every other simplex, n is the total number of simplicies within a complex. In the proposed case study, we compute the betweenness and closeness centrality indicator for both relations between the two processes. Table 9 shows the result of estimate betweenness and closeness centrality indicator in the performer viewpoint.

Organizational unit	Betweenness-Centrality	Closeness-Centrality
Operating management	0.115	0.052
Infrastructure & CRM	0.0451	0.036
Information system department	0.1093	0.106
R&D	**0.005**	**0.003**
Marketing	0.0682	0.041
Sales management	0.0791	0.057
Accounting	0.102	0.092
Finance department	0.0852	0.039
Legal department	0.0329	0.0412
procurement and logistics	**0.2094**	**0.305**
Staff department	0.1641	0.104
Support	0.0064	0.007
Social work services	0.0527	0.037

TAB. 9 – Betweenness and closeness centrality analysis result.

Regardless of high value of betweenness centrality, we notice that the performers in procurement and logistic are prominent in the enterprise organization. If we check the business process, it confirms that all activities pass through this organization unit and its performer. Than we have in second position staff unit. And the last one is the R&D with the lowest value. In other hand, closeness centrality shows up that information system team are more important than staff department.

5.2.3 Complexity

The complexity of the system structure of the case study can be described by measuring Ψ (K) proposed by (Casti et al 1979). In the case study proposed, the complexity of the system is measured as follows:

Complexity results of λ_l from BP1		**Complexity results of λ_l from BP2**	
Activity	Complexity	Activity	Complexity
$Q = \{^{2\ 1\ 0}_{1\ 4\ 1}\}$	Ψ(KA(P, λ)) = 2	$Q = \{^{2\ 1\ 0}_{1\ 1\ 1}\}$	Ψ(KA(P, λ)) = 1
Performers		Performers	
$Q = \{^{3\ 2\ 1\ 0}_{2\ 4\ 5\ 1}\}$	Ψ(KP(A, λ)) = 3.1	$Q = \{^{2\ 1\ 0}_{1\ 2\ 1}\}$	Ψ(KP(A, λ)) = 1.33

TAB. 10 – Betweenness and closeness centrality analysis result.

From the above analysis, we can conclude that the effectiveness of the topological structure analysis based Q-analysis method is affected largely by structural quantities and complexity indicator of a given relationship in the chosen business process. After providing the indicator report or the map diagnosis, it is the turn of the top management to take a decision on the choice of re-design to establish in order to reduce the complexity of the system while keeping the efficiency and alignment with the strategy of the company.

6 Conclusions &Future work

In this work, we have proposed a methodology for analysis and diagnosis of enterprise organizations' structure. The method is based on simplicial complex that overall structure analysis. To quantify structural properties of simplicies, we define a set of structural meas-

ures by taking into consideration chains of connectivity. Major contribution of this paper is to diagnose a case study of an organizational structure from telecommunication company by integrating many-to-many-relations-process. First we capture the enterprise repository. Then from a selected business processes that have non-compliance process, we derive relations responsible for these problems. Then we evaluate structural indicators eccentricity and complexity to measure the degree of non-compliance and to identify the point of impact on the effectiveness of the company. We extend the scope by two centrality-based measures. Betweenness-centrality and Closeness-centrality are introduced for characterizing structural properties of individual simplicies within organization system. Thus it is more appropriate to analyze the flows of information and activities within a system. We plan to extend our work by the integration of many-to-many relations-process within social network analysis techniques.

References

P. A. Clayton. The Methodology of Organizational Diagnosis. Vol. 11, No. 3 June (1980).

R. Atkin, Combinatorial Connectivities in Social Systems. Basel, Birkhäuser Verlag. (1977)

A. Boulmakoul, Z. Besri. Performing Enterprise Organizational Structure Redesign through Structural Analysis and Simplicial Complexes Framework. The Open Operational Research Journal, 2013, 7.pp- 11-24. (2013)

F.Luthans. Organizational Behavior. McGraw-Hill/Irwin, ISBN-0073404950, 9780073404950 (2006)

M. Kirikova. Flexibility of Organizational Structures for Flexible Business Process. Available at http://lams.epfl.ch/conference/bpmds05/program/Kirikova_10.pdf. (2007)

E. H. Schein. Organizational Learning: What is New? MIT Sloan School of Management. Working paper n°3912 July 1996. (1996)

H. Mintzberg. Structure in fives: Designing effective organizations. Upper Saddle River, NJ: Prentice Hall. (1992)

M. Weber. The theory of social and economic organization. (trans. T. Parsons). New York, NY: Oxford University Press. (1947)

M.P Wil, D. A. Van. Process Mining Discovery, Conformance and Enhancement of Business Processes. Springer Heidelberg Dordrecht London New York. ISBN 978-3-642-19344-6. (2011)

W.V.D Aalst. The Application of Petri Nets to Workflow Management. The Journal of Circuits, Systems and Computers, 8 (1): 21–66, 1998 (1998)

W.V.D Aalst, H. Reijers, A. Weijters, B. Dongen , A.M. Alves , M. Song and H. Verbeek."Business process mining: an industrial application," Information Systems, vol. 32, no. 1, pp. 713 – 732. (2007).

M. Song. Mining Organizational relations from business Process Data, Ph.D. Thesis, 2006

L. Ying Tat, J. Bockstedt. Structural Analysis of a business Entreprise. Service Science 1 (3), pp. 169-188 (2009)

P. K. Kwanghoon. Discovering Activity-Performer Affiliation Knowledge on ICN-based Workflow Models. Journal Of Information Science And Engineering 29, 79-97. Pp. 79-97 (2013).

J.L. Casti, Connectivity, Complexity and Catastrophe in Large-Scale Systems. Wiley, New York. (1979)

B. Jiang, and Omer I. Spatial Topology and its Structural Analysis based on the Concept of Simplicial Complex, 9th AGILE Conference on Geographic Information Science, Visegrád, Hungary, pp. 204-212, 2006 (2006)

L. Duckstein, A. Steven, Nobe. Q-analysis for modelling and decision making. European journal of operational Research 103 411-425. (1997)

L. Dandan and C. P. Kwong. Understanding Latent Semantic Indexing A Topological Structure Analysis Using Q-Analysis Method.

L.C. Freeman. Centrality in networks: I. Conceptual clarification. Social Networks 1, 215–239. 1979.

G.Sabidussi. The centrality index of a graph, Psychometrika, 31, 581-603. 1966.

Decision Evaluation System within Adaptive Business Intelligence

Abdelkerim Rezgui

Carl von Ossietzky University of Oldenburg
Ammerländer Heerstr. 114-118, 26129 Oldenburg Germany
abdelkerim.rezgui@uni-oldenburg.de

Abstract. Nowadays, there is an approved concept called Business Intelligence (BI) that supports the decision making process. By extending BI a new concept called Adaptive Business Intelligence (ABI) has emerged. The current state in ABI is that decisions are not evaluated in a periodic manner and the inappropriate decisions of the past might occur again. This hinders companies from benefiting from their historical pitfalls to enhance the decisions quality.
The enhancement of decision quality is one of the major outputs behind this paper. The evaluation of past decisions makes it helpful to take future complex decisions based on the uncertainty or confusion of historical decisions.
The adaptability behind the proposed solution is achieved through the evaluation, tracking, simulation and recommendation of decisions in any BI system. This paper presents a reference architecture for a new approach called ***decision evaluation system within adaptive business intelligence*** that can enrich the ABI applications.

1 Introduction

Today's Companies are faced with a continuous increase of both the number of data sources and the overall data volumes. These data have not only to be stored, but also needs to be collected, managed, filtered and analyzed to provide the best possible image of the business situation. In order to ensure correct strategic and operational decisions, the underlying knowledge must be gained and/or derived from the available data. These challenges prompted the development of Business Intelligence (BI) concepts, which support decision makers in their activities.

BI systems are known since many years of evolution in their architectures with the integration of adaptability module. This kind of extension is responding to real need to adapt the system based on very quickly changing environment. But with this important step, the decision itself is still not considered as a main component in this approach.

The evaluation of past decisions makes it helpful to take future complex decisions based on the uncertainty or confusion of historical decisions. For this purpose, the present situation has to be categorized in order to predict appropriate decisions and/or actions. In other words, the evaluation of past decisions is essential to make future decision making more efficient.

This paper presents a new adaptability module for BI as innovative approach called decision evaluation system within adaptive business intelligence that can enrich the traditional adaptive business intelligence applications.

The organization of this paper is as follows: this section presents a short introduction about the new concept; section two gives the main background information about business intelligence and self-adaptive systems. In section three, the process of the decision evaluation system will be explained. Section four shows the reference architecture of the decision evaluation system within adaptive business intelligence with its characteristics and components. The paper then concludes in section five with a brief summary regarding the contribution of this content and gives an outlook to the future directions.

2 Adaptive Business Intelligence

2.1 Business Intelligence Definition and Classification

Nowadays, there is in both academia and industry an approved and mature concept called Business Intelligence (BI) which supports decision-making process from extracting heterogeneous (internal/external, structured/unstructured) data to loading them into a central multidimensional data warehouse up to knowledge presentation and sharing.

In 1989, Howard Dressner from the Gartner Group introduced the term Business Intelligence as an umbrella term to describe concepts and methods to improve business decision-making by using fact-based support systems. Few years after, the definition was updated that (Kemper, Mehanna, and Unger 2006) identified seven different definitions, for instance BI is equal to data warehouse, alerting system, advanced management information system, etc.

However, it is important not to reduce BI as if it was only an extension of a data warehouse with the aim to transfer information from an operative system via online transaction processing (OLTP) into a systems supporting OLAP to report company's information (Gómez, Rautenstrauch, and Cissek 2008).

According to (Turban, Sharda, and Delen 2011) and (Rezaie et al. 2011) BI system is defined as "An umbrella term that encompasses tools, architectures, databases, data warehouses, performance management, methodologies, and so forth, all of which are integrated into a unified software suite".

BI is considered in this work as an organizational methodology with the objective to support decision-making process. It includes many methods that are based on technologies and tools for instance extract-transform-load (ETL), data quality, data warehousing, master data management, web, and portals. The BI methodology is useful for each actor in any organization regardless of their positions whether they work in the human resources, sales or marketing departments, etc. It helps them to have the appropriate knowledge about the factors affecting their business in order to support them by making decision.

Figure1 illustrates the actual architecture of BI systems.

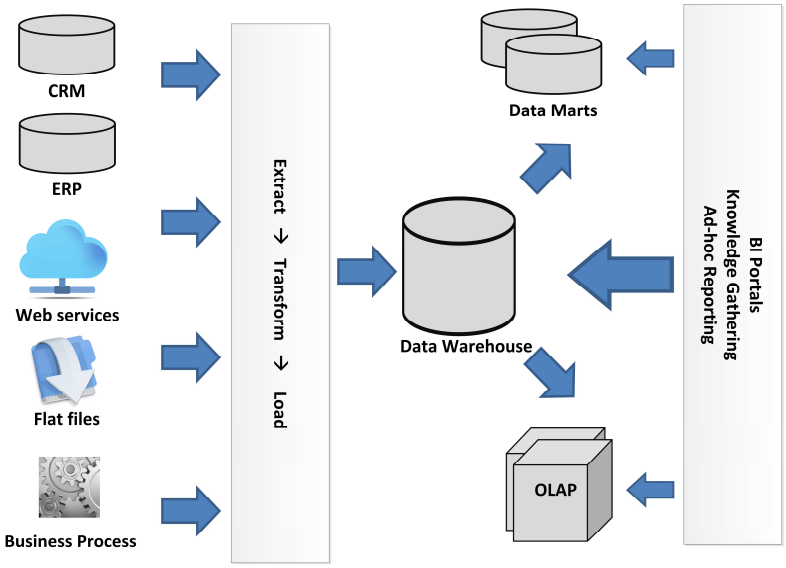


FIG. 1 – *Business Intelligence Architecture*

To analyze data within the data warehouse, and discover useful knowledge to deliver to business managers, Codd et al. introduced the On-Line Analytical Processing (OLAP) and defined that with twelve rules (Codd, Codd, and Salley 1993) . OLAP offers the possibility to explore, aggregate, and visualize data using operators. Nevertheless, upstream the on-line analysis processing comes a whole phase of extracting data, transforming, and then loading the data warehouse. In BI projects, being able to capture data necessary to provide the appropriate information, and integrate it into the data warehouse, is the most expensive phase in terms of time and resources. The BI project manager has to detect first what data could be useful to stakeholders for the decision-making process. After that, he usually shall extract these data from heterogeneous sources (DBMS, ERP, Excel files, Text files, etc.). The second stage of ETL process is the transformation. It includes all activities to transform the operational data into data that can be interpreted in terms of business and economy. It is composed of several sub-processes, i.e. filtering (eliminating redundancies and outliers), harmonization, aggregation and enrichment (Kemper, Mehanna, and Unger 2006). Finally, he shall load pre-processed data in the central multidimensional database and the data warehouse (DWH).

In the context of DWH, we are speaking about dimensions, facts, aggregations and hierarchies. Depending on business needs, the data model may be configured as a star, snowflake, or galaxy scheme. The data are grouped into data marts that are defined in functional terms. They include each subject-specific, previously aggregated, historical, current and planned data (Naana and Rezgui 2010). In order to ultimately bring the data into human-readable form, BI portals play a major role. They help user to have different views based on the analysis need, a pie chart for illustrating numerical proportion or a bar chart to show comparisons among categories. For this reason (Gluchowski, Gabriel, and Dittmar 2008) said “Finally, the value of the information [...] not only depend on the offered content, but also on the chosen form of presentation.” For both the preparation and presentation of re-

ports, a large number of software solutions is available. To improve the user's performance, many of these BI tools offer a graphical interface and drag-and-drop techniques for creating ad-hoc reports. Those tools are particularly suitable for employees being in lack of profound IT knowledge, thus enabling them to perform evaluations without needing to forward their requests to IT specialists. Standard reports and dashboards are prebuilt reports, which contain pre-calculated indicators and serve as a basis for decision-making. Another instance of information presentation is Balanced Scorecards (BSC) that describes varied tasks of the activity planning, communication and inspection of KPIs and contains typical possibilities for data analysis and reporting.

So far, we have spoken only about classical data sources, but the advent of the Web has launched a new challenge to the BI process. The quantity of data available and easily accessibility is increasing exponentially. Data volumes are growing, and corporations are demanding ever-more sophisticated BI and analytics to deal with that data. In addition to its big amount available, data are becoming more and more complex and heterogeneous and its integration from different sources requires more than a transformation of data into a single representation. The data sources are not necessarily all structured in the form of databases, but can be, for instance a corpus of documents, or come from the Web with recurring refresh. With this type of data, ETL process cannot be held traditionally. The data warehouse facilitates complex data analyses without placing a burden on the operational source systems that run the day-to-day business. In order to catch up with data changes in the operational sources, the data warehouse is refreshed in a periodic manner (usually on a daily, weekly, or even monthly basis). ETL process is a resources consumer. It lowers the performance of the source system that cannot support many selects of data on the running time. Based on that, data warehouse refreshment is typically scheduled for off-peak hours where both, the operational sources and the data warehouse experience low load conditions, e.g. at nighttime (Jörg and Dessloch 2010). However, if enterprises want to be more competitive on today's global economy, they are forced to satisfy the customer's constantly changing needs. Additionally, the speed and dynamic nature of business often negates the time required for long term planning and time consuming implementations in order to stay on top. Because of this, organizations must implement solutions that can be deployed quickly and in a cost-effective manner (Zicker 1998).

2.2 Self-Adaptive Systems

The complexity of software systems and the quick changing environments had led the software engineering community to look for a concept that allows systems to adapt themselves or their behaviors based on user profile or requirements. This forms the motivation of the initiation of self-adaptive systems.

Among several existing definitions for self-adaptive software, one of those definition is introduced by (Oreizy et al. 1999) and says that: "Self-adaptive software modifies its own behavior in response to changes in its operating environment. Here operating environment means anything observable by the software system, such as end-user input, external hardware devices and sensors, or program instrumentation." Three years ago Villegas add to this definition that "Such dynamic systems adapt in response to changes in their environments, either to ensure the continuous satisfaction of their functional and non-functional requirements, or to provide ubiquitous and context-dependent smart services" (Villegas Machado, Müller, and Tamura Morimitsu 2011).

Being adaptive as an inevitable requirement has also impacted Business Intelligence researchers to change applied BI concepts towards adaptability. The next section gives more details about the new generation of BI.

2.3 Adaptive Business Intelligence

The extension of BI by using adaptability based on prediction and optimization methods and techniques for forecasting and decision supporting is called Adaptive Business Intelligence (ABI) and was firstly introduced to public by (Michalewicz et al. 2006) and enriched by (Nenortaitė and Butleris 2009), (Fabac 2010) ,(Burmester 2011), (Lau et al. 2012) and (Kim et al. 2013).

Michalewicz defined the term Adaptive Business Intelligence as "the discipline of using prediction and optimization techniques to build self-learning 'decisionning' systems" (Michalewicz et al. 2006 p.5)

Adaptive Business Intelligence has been investigated within the different researchers from different points of view, adaptability in user interface, adaptability in models, automatic decision making, adaptive knowledge presentation; however, most of these initiatives have been isolated. From our perspective all of them ignored an important point: the adaptability in the content of knowledge (the decision itself), the human implication in decision-making acts, and in the recommendation of decisions.

The current state in ABI is that decisions are not evaluated in a periodic manner and the inappropriate decisions of the past might occur again and again over time. This hinders the companies from benefiting from their historical pitfalls to enhance the quality of their decisions over time. The same applies for archiving the perfect adequate decisions that are taken over time and recommendation of the optimal decision to a specific issue.

The management of BI decisions over time is one of the major outputs behind this paper. The evaluation of past decisions makes it helpful to take future complex decisions based on the uncertain or confusing historical decisions. For this purpose, we need to categorize our present situation in order to predict appropriate decisions and/or actions. In other words, the evaluation of past decisions is essential to make future decision making more efficient.

Such management includes storing, evaluating, and ranking of BI decisions, which will be stored in a central repository that serves as a core of a new adaptive BI. The adaptability behind the proposed solution is achieved through the decision evaluation of the already taken decisions in any BI system. Such evaluation will rebuild the harvested advanced knowledge in a way that any company will see its decisions in form of decision dashboard in which each single decision taken in the past can be seen with its reputation and number of occurrences over time.

2.4 Decision Making Process

In the literature, there have been many researches addressing the decision-making process issue. Simon consider that decision making includes four principal phases: finding occasions for making a decision, finding possible courses of action, choosing among courses of action, and evaluating past choices (Simon, 1977). Figure 2 illustrates his approach.

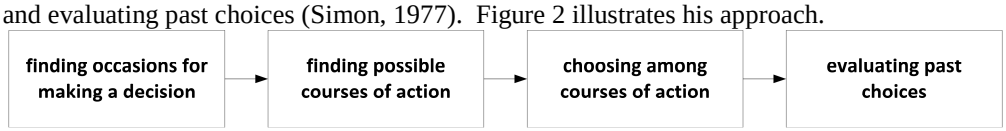


FIG. 2 – *Decision-Making Process, based on (Simon, 1977)*

He used the terminology, intelligence activity for the first step to take before taking a decision. The second phase, called design activity, involves inventing, developing, and analyzing possible courses of action. After that, comes the third phase, called choice activity, which consists on choosing a particular course of action from the available courses found while performing the design activity. Finally, evaluating the choices already made, called the review activity.

For others, it still remains an invisible process, only thoughts and operations inside the minds of managers. It is very difficult to understand document or improve a decision process. For many organizations, a managerial decision is treated as a 'black box', subject to neither explanation nor review. Human decision-making process is invisible (March, J. (1987), Davenport et. al 2001, Kahneman, J. (2003), p.131)

According to Turban et al., it is a process of choosing among two or more alternative courses of action for the purpose of attaining one or more goals (Turban, Sharda, and Delen 2011, p.42)

Business Intelligence Systems tried to assist decision-makers in their everyday tasks. The traditional decision making process is defined as the management of flows of data, to transform it into information, and then into useful knowledge. So far, this process deals only with the first phase (Simon, 1977) the intelligence activity. All three phases, which comes after, are taken only by humans, without any kind of assistance from computers. The machine automates the knowledge generation and the human takes this knowledge and makes decision. The decision itself is not supported by the BI system, only making knowledge available is supported.

Adaptive Business Intelligence Systems (ABIS) made some steps more and cover the first (finding occasions for making a decision), second (finding possible courses of action) and third (choosing among courses of action) phase of Simon's approach. However they cover only partially this phases and supporting the fourth step is missed.

The main characteristics of actual ABIS are:

- Fully-automated approaches, decision-making is done through machine, no human contribution is provided
- Poor decision evaluation
- Limited decision impact simulation:
  - only economical aspect
  - one dimensional impact (measurement of the impact on single KPI[1] without relationship to other KPIs)
- Anonymous decision responsibility
- Missed knowledge:
  - KPI/KPI impact relationships
  - KPI/decision impact relationship
  - Decision reputation

All these motivated the idea of treating the evaluation of decisions over the time and through all phases.

[1] KPI is an abbreviation of Key Performance Indicator.

3 Decision Evaluation Process

The decision evaluation process includes storing, simulation, recommendation, evaluating and ranking of BI decisions. These decisions will be stored in a central repository that serves as a core of the new adaptive BI system. The adaptability behind the proposed solution is achieved through all functionalities. The proposed system adapts the decisions raking and recommendation based on their evaluation, from other hand it adapts the decision recommendation based on user responsibility. Decisions are becoming then domain specific categorized (sales/presales decisions, marketing decisions…).

The evaluation will rebuild the harvested advanced knowledge in a way that any company will see its decisions in form of decision dashboard in which each single decision taken in the past can be seen with its maker, impact, reputation and number of occurrence over time.

In our approach, we make a difference between 'first-level' knowledge and 'advanced' knowledge. The first-level knowledge is used widely by the traditional BI/ABIS which transform data into knowledge and present it to users to make decisions. It is basically data manufactory with the goal to help user to get the right knowledge. This knowledge does not include the decision itself. The advanced knowledge includes first-level knowledge and knowledge about the decision.

Our conception of the adaptive business intelligence system distributes the decision between two main actors: ABI users and the ABI system itself. Human machine interaction is very valuable for the effectiveness of decisions. The most important added value within the proposed ABI system is the decision database component. This will relieve the loss of knowledge acquired from past decisions.

The next cross-functional flow chart in figure 3 represents the process of the proposed ABI system and the interaction between them and the users. Let us first explain the schema. Every successful business is built over measurable objectives, called KPI's. The user is the only one capable of selecting, describing and identifying them in order to monitor them. From this description, the system will generate a KPI matrix, and store it into a decision database to be used later. This matrix gives responses to the following questions:

- What is the level of this KPI (strategic, tactic, operational)?
- What is the domain of this KPI (sales, marketing, production…)?
- What are the dependency and the influence rate of this KPI with/on other KPI's?

So far there is no decision to be taken, the user follow up his metrics and performs the first phase of decision-making process, which is finding occasion for making a decision. If this occasion shows up (e.g. KPI under the defined threshold), the system adapts the user interface to declare the need of decision-making. The user will select the KPI in question and the system will search for previous taken decisions in relation with the selected KPI from the decision database with their evaluations. The system will display a decision matrix dashboard to the user who will either pick the one with the best evaluation, or define a new decision, which has never been taken before. If the user chose an old decision, the system will help him simulating and seeing the impact of this decision on his KPI matrix, and store the simulation results into the decision database. Otherwise, he should introduce the characteristics of his new decision and store it into the decision database, create a simulation scenario and perform the simulation. Only then, the user will be capable of taking the right action.

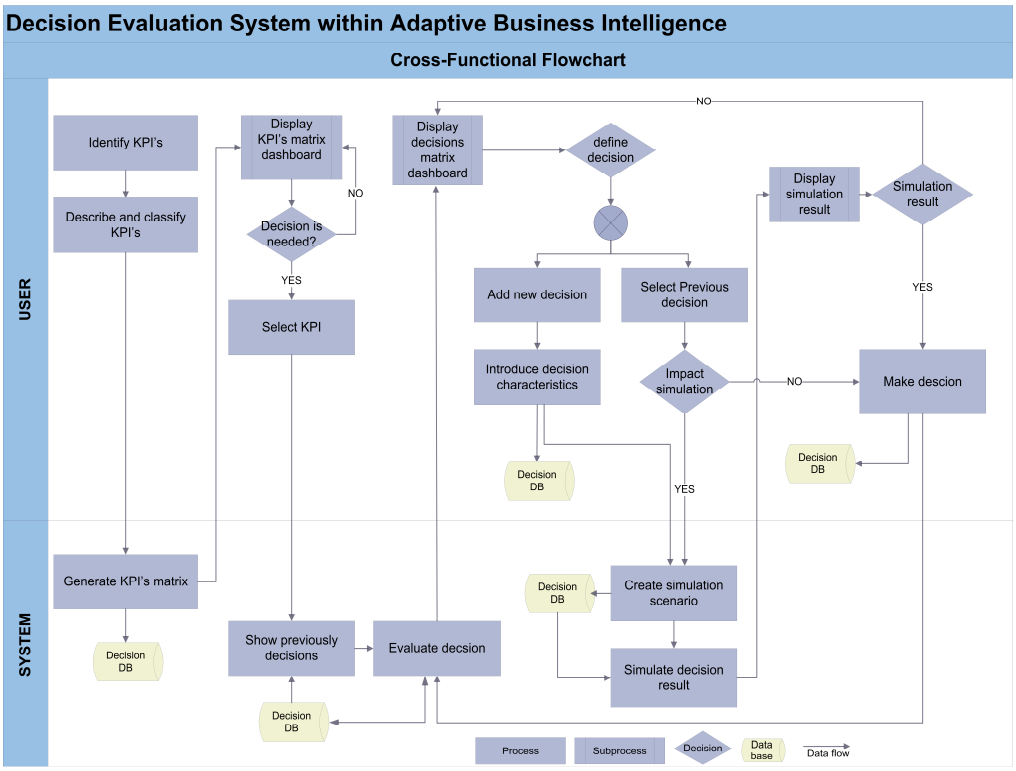


FIG. 3 – Cross-functional flow chart for decision evaluation system

4 Decision Evaluation System Reference Architecture

In this paragraph, we will explain the decision evaluation system (DES) reference architecture, its components and the interactions between them. The proposed solution assures interactivity between the user and the system while making decision and it is fully adaptive.

The main outcomes of this new concept are addressing the lack in decision-making process with ABI through enablement of:

- Human-machine interactivity in the decision making process
- Building of decision/KPI relationship
- Decision impact simulation and demonstration
- Decisions tracking, evaluation and recommendation

Figure 4 illustrates the DES reference architecture. The external component data warehouse (DWH) is used to get the KPIs and their values and changes (positive or negative). The role of the first component, the KPI & Decision Tuner, is assuring the relationship between KPI/KPI and KPI/decision and storing KPI values in the Decision DB. The main subcomponents are the KPI Generator, KPI Classifier, KPI Monitor and KPI Matrix Generator.

The communication between the KPI & Decision Tuner and the two databases (the DWH and the Decision DB) is assured through two Data Access Objects (DAOs).

The second component, the Decision Engine is composed of six subcomponents and enables the interactivity between user and system while making decision (1), building the relationship between actions and KPIs (2), simulation of decision impact (3), classifying decisions (4) and generation of decision matrix (5). These subcomponents are, Decision Generator, Decision/KPI Combiner, Decision Simulator, Decision Matrix Generator, Decision Maker and DAO.

During the activity of decision-making as a human act, errors may occur, it is hard for the individual to identify and determine these errors since they are presented as truth (Karlsson 2013). For this reason we need to evaluate decisions. The Decision Evaluator component assures the evaluation of already taken decisions. This evaluation is based on a set of different criteria defined as entry parameters. For now and for presentation purposes, we will not dive through the details of each class of criteria, rather we will explain the main strategy adopted in evaluating the BI decisions. The evaluation criteria are classified based on the classification aspects. Each criterion is going to be evaluated individually based on the impact of the decision on one or many KPIs. The Impact Monitor subcomponent is responsible for the measurement of the decision impact and gets this information through an internal DAO, which communicates with the Decision DB. The average of these criteria will be then evaluated on a higher level that is the set of each criterion.

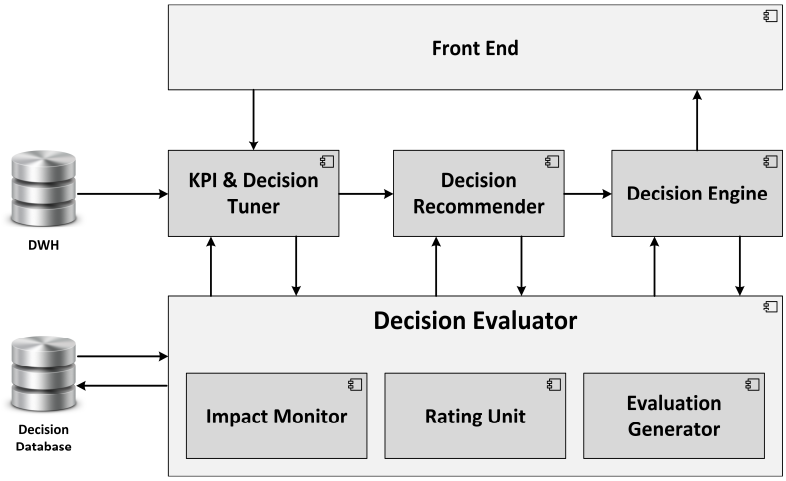


FIG. 4 – *Decision Evaluation System Reference Architecture*

To be able to make decision the user should know which possible choices are available and what is their potential success, based on their reputation. Based to (Etzioni 1988) "The term choice should be used to encompass the sorting out of options, whether conscious or unconscious. Deliberate choices are to be referred to as decisions". The Decision Recommender component gives the user all possible decisions (related to his profile) to a specific situation. The proposed options are sorted based on their evaluation. Each choice has it's own characteristics. This component adapts the recommended decision on user domain (sales, production, marketing…etc.) and responsibility (division manager, team lead, em-

ployee…etc.). Then it's not possible that each user can take all type of decisions, for example a sale consultant may make a decision to change his customer visits plan to prioritize important customers (knowledge gained from a customer ranking dashboard), but can't change his product portfolio or start a new marking campaign.

To illustrate more clearly the proposed concept, we will give an example. Let us consider a manufacturing company that is very involved within the environmental cause. So the board created a dashboard to monitor several KPIs such as CO2 emissions, chemical product emissions to water, energy use, etc. Assume that one day, a BI user opens the dashboard and found that the CO2 emissions exceeded the established threshold.

This event will trigger a decision making process. The proposed decision evaluation system within ABI will show previous decisions taken with an impact (direct or indirect) on this KPI from the decision database. The user will be able then, to choose a decision already defined, or define a new one. This depends on the evaluations of the already taken decisions but with relation to the new context. The suggested decision list does not include yet a decision with a direct impact on CO2 emissions, but it includes the historical decision: changing raw material that impacted the CO2 emission. This action was taken because of the increasing cost of the old used raw material and with to goal to decrease production coast and so the profit rate. It turned out that this new material generates much more pollution.

To overcome this issue, the stakeholder calls for a meeting to look out for new alternatives, and the technical staff came with a list of proposed new materials. The system will assist the stakeholder through his mission; by simulating the decisions results made by the decision evaluator and give him the best alternative given the characteristics and the impact on the KPI matrix. The best alternative could be a new raw material that maintain the profit rate at the same level but decrease the CO2 emissions to an acceptable level.

5 Conclusion and Outlook

Firstly a brief description of the different understandings of the term 'Adaptive Business Intelligence' gives a short motivation about the relevance of the combination of BI and Self-Adaptive Systems.

We addressed in this paper the lack in decision-making process within ABI to enable more efficient decision. Our concept proposed a decision evaluation system within adaptive business intelligence to form a bridge between decision preparation and decision-making and covered the lack of decision adaptability in business intelligence. The main point of interest in our approach is the decision evaluation and the simulation of its impact.

Future directions of this on-going research are the implementation of the proposed concept. On this basis, different technologies will be examined and a planned prototypical implementation will be conducted. Therefore a specific scenario can be applied coming from existing partners from the industry for example. This offers the possibility to evaluate the approach in the context of a real case and also underline the benefits companies can gain from our approach.

Acknowledgment

This work was financed by BI4YOU: www.biforyou.com

References

Burmester, L. (2011). Adaptive Business-Intelligence-Systeme. Springer. Codd, E. F., S. B. Codd, and C. T.

Salley (1993). Providing OLAP (on-line analytical processing) to user-analysts: An IT mandate. Codd and Date 32.

Davenport, T. H., J. G. Harris, D. W. De Long, and A. L. Jacobson (2001). Data to knowledge to results: Building an analytic capability. California Management Review 43(2).

Etzioni, A. (1988). The moral dimension: Toward a new economics, 1988. Cerca con Google, New York: Free Press., 150.

Fabac, R. (2010). Complexity in organizations and environment-adaptive changes and adaptive decision-making. Interdisciplinary Description of Complex Systems 8(1), 34–48.

Gluchowski, P., R. Gabriel, and P. Chamoni (2005). Management Support Systeme und Business Intelligence: Computergestützte Informationssysteme für Fach-und Führungskräfte. Springer-Verlag New York, Inc.

Gómez, J. M., C. Rautenstrauch, and P. Cissek (2008). Einführung in Business Intelligence mit SAP NetWeaver 7.0. Springer.

Jörg, T. and S. Dessloch (2010). Near real-time data warehousing using state-of-the-art ETL tools. In Enabling Real-Time Business Intelligence, pp. 100–117. Springer.

Karlsson, R. (2013). Data as intelligence: A study of business intelligence as decision support.

Kemper, H.-G., W. Mehanna, and C. Unger (2006). Business Intelligence - Grundlagen Und Praktische Anwendungen: Eine Einführung in Die It-Basierte Managementunterstützung. Springer DE.

Kim, J., M. Hwang, D.-H. Jeong, S.-K. Song, and H. Jung (2013). Business Intelligence Service based on adaptive user modeling and groups. Journal of Computer Science 9(10), 1396.

Lau, R. Y., S. S. Liao, K.-F. Wong, and D. K. Chiu (2012). Web 2.0 environmental scanning and adaptive decision support for business mergers and acquisitions. MIS Quarterly 36(4).

March, J. G. (1987). Ambiguity and accounting: The elusive link between information and decision making. Accounting, Organizations and Society 12(2), 153–168.

Michalewicz, Z., M. Schmidt, M. Michalewicz, and C. Chiriac (2006). Adaptive business intelligence. Springer.

Nenortaite˙, J. and R. Butleris (2009). Improving business rules management through the application of adaptive business intelligence technique. Information Technology and Control 38(1), 21–28.

Oreizy, P., D. Heimbigner, G. Johnson, M. M. Gorlick, R. N. Taylor, A. L. Wolf, N. Medvidovic, D. S. Rosenblum, and A. Quilici (1999). An architecture-based approach to self-adaptive software. IEEE Intelligent systems 14(3), 54–62.

Rezaie, K., A. Ansarinejad, A. Haeri, A. Nazari-Shirkouhi, and S. Nazari-Shirkouhi (2011). Evaluating the business intelligence systems performance criteria using group fuzzy ahp approach. In Computer Modelling and Simulation (UKSim), 2011 UkSim 13th International Conference on, pp. 360–364. IEEE.

Rezgui, A. and M. Naana (2010). Improving of environmental management accounting system for support the environmental information management. Shaker Verlag.

Simon, H. A. (1977). The logic of heuristic decision making. In Models of discovery, pp. 154–175. Springer.

Turban, E., R. Sharda, and D. Delen (2011). Decision support and business intelligence systems. Pearson Education India.

Villegas Machado, N. M., H. A. Müller, and G. Tamura Morimitsu (2011). On designing self-adaptive software systems. Sistemas y Telemática.

Zicker, J. (1998). Real-time data warehousing. DM Review March.

Résumé

Aujourd'hui la Business Intelligence (BI) est un concept bien mature et approuvé. Le rôle de la BI est de soutenir le processus de la prise de décision. Dans le cadre de tentative d'extension et d'amélioration de la BI, un nouveau concept appelé Adaptive Business Intelligence (ABI) a emergé. L'état actuel de l'ABI ne permet pas l'évaluation périodique des décisions prises, et les mauvaises décisions du passé peuvent encore se reproduire. Cela empêche les entreprises de tirer profit des erreurs commises. L'amélioration de la qualité de la décision est l'un des principaux avantages du concept présenté dans ce papier. L'évaluation des décisions prises devient utile afin d'améliorer les future décisions.

L'adaptabilité dans la solution proposée est obtenue par le suivi des évaluations, la simulation et la recommandation des décisions. Cet article présente une architecture de référence pour une nouvelle approche appelée système d'évaluation de la décision au sein de l'adaptive Business Intelligence qui permet d'enrichir les applications ABI existantes.

Desicion Support Systems: What is the next step?

Raji Ben Maaouia*,***, Abdelkerim Rezgui*,**, Faiez Gargouri***

*BI4YOU, Pôle des technologies de la communication, 2010 Manouba, Tunisia
r.benmaaouia,a.rezgui@biforyou.com
http://www.biforyou.com
**University of Oldenburg, Ammerländer Heerstraße 114-118, 26129 Oldenburg, Germany
abdelkerim.rezgui@uni-oldenburg.de
http://www.uni-oldenburg.de
***MIRACL, Pôle Technologique de Sfax, 3021, Sakiet Ezzit Sfax. Tunisia.
faiez.gargouri@isimsf.rnu.tn
http://www.miracl.rnu.tn

Abstract. "We are drowning in information, while starving for wisdom. The world henceforth will be run by synthesizers, people able to put together the right information at the right time, think critically about it, and make important choices wisely" Says Edward Osborne Wilson in his book "Consilience: The Unity of Knowledge"(Wilson, 1999). According to the above, Decision Support Systems should allow the management of data flows, first, by identifying and synthesizing the relevant information, and then, processing this information into a condensed and useful knowledge and intelligence.

1 Introduction

The amount of data exchanged is more and more important every day. And the cost of storing it into local hard drive or into distant servers is getting cheaper. This data can be generated from multiple sources and in different ways (Web pages, emails exchange, customer relationship management tools, enterprise resource planning, posts on blogs, status updates on social networks, etc.). Most people does not realize that this data is a very powerful weapon, and very few, know how to use it efficiently. In fact, data is a real gold mine. It takes only just a bit of reorganization and cleansing to discover how much actionable knowledge we can extract from this huge and non structured amount of data. And if we have the necessary techniques and the technologies, and dig a little bit deeper, we will be surprised with new and very useful knowledge "hidden" inside of it. This was among the central reasons behind the rise of Business Intelligence as a very promising field of research. In fact, in a very competitive and fast-changing world economy context, the acquisition, analysis, and exploitation of information has become essential for enterprises to be leaders in their fields. Hal Ronald Varian, founding dean of the School of Information, University of California, and chief economist at Google, pointed out that the opportunities for the future generations of data scientists and practitioners will be very big. He said "If you are looking for a career where your services will

be in high demand, [. . .] my recommendation is to take lots of courses about how to manipulate and analyze data: databases, machine learning, econometrics, statistics, visualization, and so on" (Varian, 2008). Note that every single discipline mentioned in his quotation is a fundamental element of Business Intelligence as is it is nowadays.

In this paper we will introduce a proposal of a collaborative decision support system based on users' contributions and evaluations.The collaboration needs to be introduced at the first stage when collecting complex and unstructured information from multiple sources, but it should be able to assist stakeholders all along the decision making process. In the first section, we will present a historical study of the evolution of decision support systems, and the impact of new mobile and web technologies on the DSS's capabilities. Then we will present the most important trends and application domains with some of the used techniques and research areas. The third section will be dedicated to the presentation of collaborative systems and the benefits of the integration of collaboration within decision support systems.

2 Business Intelligence

The term Business Intelligence was used for the very first time by Hans Peter Luhn, computer scientist for IBM, in his paper "A Business Intelligence System" back in 1958. He defined it as "a collection of activities carried on for whatever purpose, be it science, technology, commerce, industry, law, government, defense, et cetera. The communication facility serving the conduct of a business (in the broad sense) may be referred to as an intelligence system. The notion of intelligence is also defined here, in a more general sense, as the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal"(Luhn, 1958).

It took almost 30 years for Howard Dresner, an analyst for Gartner Research, who is also considered as the father of Business Intelligence, to come up with a new definition of Business Intelligence as "a broad category of software and solutions for gathering, consolidating, analyzing and providing access to data in a way that lets enterprise users make better business decisions" (Power, 2007). Since then, many researches were published, and even more definitions of the term Business Intelligence were proposed. Never the less, the major part of them turn around the extraction and the analysis of data. Throughout a business intelligence process, that data takes many forms.The most famous representation of the data's evolution until today is the knowledge pyramid (also known as the DIKW pyramid). It suggests the use of a hierarchical model with data at its base and wisdom as a top level.

Nevertheless, Jonathan Hey (Hey, 2004) criticized the fuzzy distinction between each stage of this presentation. He said that the problem is the fogginess that surrounds the transition between these concepts and how we shall proceed in order to "climb" the DIKW pyramid and reach wisdom. To overcome this conception, Clark, D. (Clark, 2010) introduced the DIKW chain with further details about how we shall define the different parts, and the transitions between every part of it.

According to the above, we can assume that the purpose of Business Intelligence is the management of data flows by first identifying and then processing the information into condensed and useful managerial knowledge and intelligence. As such, the BI task includes new topics, addresses very old managerial problems and it is one of the basic tasks among the

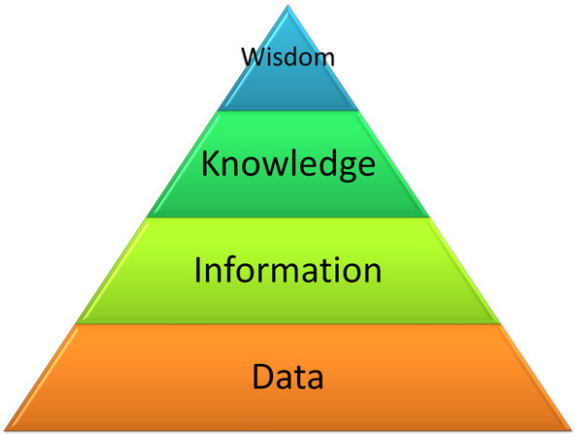


FIG. 1 – *DIWK Pyramid.*

many management tools: analyzing the complex business environment in order to make better decisions (Rouhani et al., 2012).

2.1 Data warehousing approach

The first approach used to implement Business Intelligence systems was basically a data-driven approach. It relies essentially on relational databases introduced by the IBM researcher Edgar Frank Codd (Codd, 1970). As the aim from RDBMS was operational in the first place, an enormous amount of redundancy was required for decision-makers to be able to handle multiple decision-support environments. Thus, a new kind of databases was being designed to let stakeholders use their data into decision support systems. That is why other IBM researchers, Barry Devlin and Paul Murphy, used the term "information warehouse" and began building experimental "data warehouses". But the concept of a data warehouse did not become "mature" until Bill Inmon and Chuck Kelley wrote the famous list of twelve rules defining it (Inmon and Kelley, 1994) and until the work of Ralph Kimball on its architecture (Kimball et al., 2002). Thus, data management and warehousing is considered as the first brick in the foundation of the classic decision-support systems. Since then, a wide range of Business Intelligence techniques was being developed. It goes from data marts design, tools for extracting data transforming, and loading it (ETL), reporting, dashboards and Online-Analytical-processing (OLAP), to predictive modeling and data mining. The major IT vendors including Oracle, IBM, Microsoft, and SAP were able to incorporate all of these technologies in their business intelligence platforms (Sallam et al., 2011). Nevertheless, researchers are still working very actively on some of these technologies such as data mining, in-memory database management systems and real-

decision support systems: what is the next step?

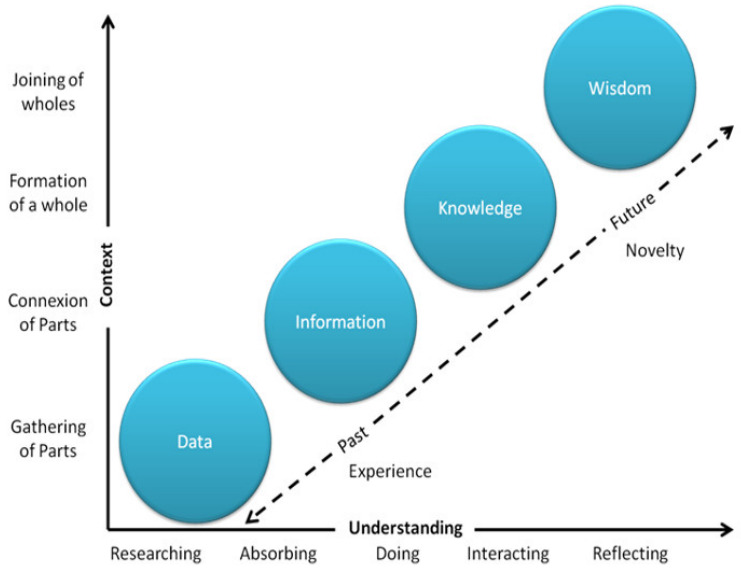


FIG. 2 – *DIWK Chain.*

time decision tools (Bitterer, 2011). Let us note that every single technique cited so far, relies on well-structured data extracted from enterprise-specific data.

2.2 Web-Based approach

Since the early 2000s, Business Intelligence has been directly impacted by the high rate of internet penetration everywhere worldwide. While the classic decision support systems relies on data warehousing, and multi-dimensional databases, the new generation of business intelligence systems found a priceless mine of information online. Indeed, a very big amount of information about customers, products, industries can be found on the Web. And through a range of text mining and web mining techniques, actionable knowledge could be discovered and used to take strategic, tactic, or operational decisions with big impact on business. Up to there, the scope of a business intelligence project was defined inside an economic context seeking for growth. The main objectives were increase revenue, cut costs, and save time. With new Web technologies (Web 2.0), everyone was able to share content with everyone. The high level of interaction allowed between the site and the connected users implies a very active contribution by users in the process of data creation. Many marketing researchers believe that social media analytics presents a unique opportunity for businesses to deal with the market as an exchange between businesses and customers instead of the traditional business-to-customer, one way "marketing" (Lusch et al., 2010).

This Web revolution makes it possible for BI applications to came out from the corporation world and reaches a new level into a socio-political and more global context. This user-

generated data, usually heterogeneous (text, images, videos, etc...), can make business intelligence more powerful and efficient. In addition, we can gather instant feed-backs and opinions on products, videos, songs, etc. The challenging aspect of this new era was how to organize non-structured or semi-structured data retrieved from the internet and use it as a decision support input. This requires the integration of techniques such as text-mining, sentiment analysis, social network analysis, web-mining, etc.

2.3 Sensor-Based approach

The mobile revolution reached a new milestone when the time spent on the internet from mobile devices (smart phones, tablets, PDAs, etc.) surpassed the time of personal computers used online in the United States (477 billion minutes vs. 481 billion) (Goodman, 2010). Other sensor-based Internet-enabled devices equipped with RFID, barcodes, and radio tags (the Internet of Things) are opening up exciting new steams of innovative applications (Chen et al., 2012). Wireless sensor networks are really going to provide data for the scientific community, citizen-driven activism, or organizations (Ganeriwal et al., 2008). Sensors are distributed across the globe leading to an avalanche of data about our environment. It is possible today to utilize networks of sensors to detect and identify a multitude of observations, from simple phenomena to complex events and situations (Sheth et al., 2008). This makes the decision support systems confronted more than ever to big challenges and opportunities in making scientific and societal impacts from this new kind of data. Thus, new perspectives for Business Intelligence and Analytics are open.

3 Trends and application domains

It is no doubt that since the beginning of 2013, big data and analytics have been among the most addressed themes for the information systems community. In contrast with classic business intelligence systems, in which we use data from corporate information systems, combined with calculation rules, and descriptive statistics, to show trends, and measure KPIs, big data is used with statistical inference to detect novelties that are cannot be discovered if we had a smaller volume of data. The emerging research opportunities from this context are priceless. And the enthusiasm about analyzing semi-structured and unstructured content hides many challenges.

As the most important portion of unstructured data is in a textual format, text analytics became a very important field of research. If we assume that the understanding is the key to reach wisdom, the analysis of text contents, will definitely help summarize and contextualize information. Text analytics techniques have been actively developed for a range of areas, including information retrieval (IR), opinion mining, sentiment analysis, multilingual analysis, and automatic summarization. Web 2.0 and social media content have created abundant and exciting opportunities for understanding the opinions of the general public and consumers regarding social events, political movements, company strategies, marketing campaigns, and product preferences (Pang and Lee, 2008). Community detection for example is also a very promising field of study. It is closely related to web analytics, as it can help apprehend the behavior of customers in a marketing context, or voter in a political context.

decision support systems: what is the next step?

4 Towards a collaborative decision support system

Despite their spectacular and fast evolving capabilities, DSS must focus on helping a human decision maker not replacing him. Indeed, the complexity of a decision making process cannot be simplified by solving mathematical equations and optimization problems. The need of an expert's know-how and experiences is essential, especially when the decision requires compromises, pragmatism and intuition. This is why the man judgment is more reliable than the machine's when confronted to complex situations. Computers should be able to assist stakeholders by playing at least two principal roles. The first consists on providing them with knowledge gathered from internal and external sources so every collaborator can apprehend the context, and the elements surrounding the problem to solve. The second role for DSS to play, is maintaining an open free-opinion-space for the involved persons to interact and suggest solutions.

4.1 General architecture for collaborative DSS

While several schools suggest that information gathering engines could be a very efficient alternative to traditional ETL approaches, others pointed out that ETL is a very mature process that has shown a high level of effectiveness as the basis for creating successful DSS projects.(Mahmoud et al., 2012). Our point of view reaches the second reflection. We consider that we should keep the existing BI IT infrastructure and try to integrate more components that can provide us with contextualized information and decision suggestions. The proposed architecture relies on the classic decision support system architecture. However, we will add some enhancements to enable human-human and human-machine collaboration.

Almost all DSS have three layers in common. Data sources, ETL, data warehouse, reports and dashboards. We will not illustrate these components and the interaction between them because it has been done over and over again in the literature (Chaudhuri et al., 2011; Duan and Da Xu, 2012; Grothe et al., 2012); however we will explain the new proposed components and the benefits of including them into a DSS.

4.2 Information gathering engine

The proposed information gathering engine can retrieve knowledge online to provide the DSS with up-to-date insight on the problem being treated. For instance, if an airline sales manager needs to know the weather changes and the actual flight position and understand in real time that people will miss their connecting flights, and plan alternatives for them on the one hand, and sell their seats to other people on the other hand. This example represents the rigidity of the actual architectural design, which does not consider the need of data from external sources. The restriction of this design can be also seen in the prerequisites of the data warehouse.

The problem with this situation is that the engine can retrieve a very large number of relevant as well as irrelevant information. To overcome this, the engine should be able to represent the context in a way that can interact with web very naturally. So it should be a semantic-based information retrieval engine. This enables a semantic filtering of information and will definitely help contextualize the extracted online information. Besides, the frequency of update will be much faster than if we use ETL process. Online information will be integrated in the

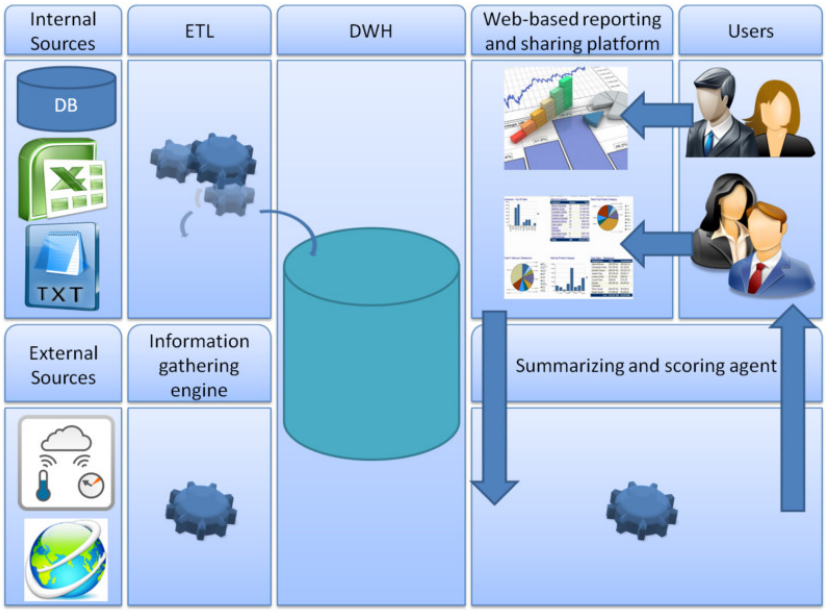


FIG. 3 – *general architecture for collaborative Decision Support System.*

decision support process as soon as it is uploaded, and taken into account by decision-makers right away. To extract valuable knowledge from online sources, various data and text mining techniques have been proposed to collect and analyze web contents(Agarwal et al., 2010; Kumar et al., 2010; Tsai, 2011). Also various classification-based and lexicon-based techniques have been proposed for finding opinions and sentiments relevant to a given topic using online content (Zhang et al., 2009; Lee et al., 2008; He et al., 2008). A technological and methodological support may lead to domain ontologies, intended as the definition of logical concepts for the definition of a semantic Web. They provide support to the activity of identification of rules and other specialized operations for interpretation of Web documents, through the construction of a network of relations and connections between documents(Buttarazzi et al., 2010).

4.3 Summarizing and scoring agent

It is more than clear that all DSS only provide a description of the actual state by presenting graphs, dashboards, alerts, etc. And according to the above seen on how human are more capable of getting the right decision in complex situations than machines, the summarizing and scoring agent is intended to get users feed-backs and suggestions on how we can solve the actual problems or concerns. For instance, if a simple analyst in a company could post "a comment" about the sales dashboard saying that "if we open a new store in the region X, sales will raise by 5%". This analyst has a background knowledge that his manager does not, because he comes originally from region X and he knows that the marketplace there needs

much more of the products his company is selling. The future system should be able to present a summarized view of the most relevant contributions to be presented to the board of the company. The system will filter feed-backs and opinions, and through a scoring system it will suggest the most accurate solutions. In the literature we have found related works which turned to be very useful to help us define better the component. For instance, Brooks et al, developed a system that automatically tagged blog posts based on the TF-IDF weighting of the top three terms extracted from the post (Brooks and Montanez, 2006). Also, Gilad Mishne has developed a collaborative filtering system. This system finds similar tagged posts and suggests some set of the associated tags to a user for selection (Mishne, 2006). Our system will most likely use a similar technique to present an insight on users' opinions weighted by the opinion frequency and evaluations by other users.

5 Conclusion

This paper presents a state of the art in the domain of decision support systems and demonstrates the absence of collaboration in such systems. We proposed a general architecture of new collaborative decision support system. The main difference between the presented architecture and related works (Luhn, 1958; Sallam et al., 2011; Rouhani et al., 2012) is that we introduced two new components. The first one is the information gathering engine, which provides users with external insight while making decision. And the second one is an agent capable of retrieving users' feed-backs and presenting suggestions based on the most relevant contributions. We argued that the proposed architecture will enhance the capabilities of DSS and emphasizes the role of the human in the decision-making process. Future directions of this on going work are the implementation of a detailed reference architecture with all components, sub-components and how they interact with each other. After the validation of this architecture, a prototypical implementation will be conducted. An industrial partner will examine the evaluation of the prototype through real case scenario.

Acknowledgment

This work was supported by the EU-funded project PASRI within a MOBIDOC Ph.D. thesis. The program is directed by ANPR and co-financed by the company BI4YOU.

References

Agarwal, N., M. Galan, H. Liu, and S. Subramanya (2010). Wiscoll: Collective wisdom based blog clustering. *Information Sciences 180*(1), 39–61.

Bitterer, A. (2011). Hype cycle for business intelligence. *Gartner, Inc., Stamford, CT*.

Brooks, C. H. and N. Montanez (2006). Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *Proceedings of the 15th international conference on World Wide Web*, pp. 625–632. ACM.

Buttarazzi, B., M. Mechilli, and L. Polinari (2010). How web 2.0 improves business intelligence: showcase of emerging technologies.

Chaudhuri, S., U. Dayal, and V. Narasayya (2011). An overview of business intelligence technology. *Communications of the ACM 54*(8), 88–98.

Chen, H., R. H. Chiang, and V. C. Storey (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly 36*(4).

Clark, D. (2010). Understanding and performance.

Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM 13*(6), 377–387.

Duan, L. and L. Da Xu (2012). Business intelligence for enterprise systems: a survey. *Industrial Informatics, IEEE Transactions on 8*(3), 679–687.

Ganeriwal, S., L. K. Balzano, and M. B. Srivastava (2008). Reputation-based framework for high integrity sensor networks. *ACM Transactions on Sensor Networks (TOSN) 4*(3), 15.

Goodman, E. (2010). There's a 20 billion dollars pot of gold at the end of the mobile advertising rainbow.

Grothe, M., U. Schäffer, et al. (2012). *Business intelligence*. John Wiley & Sons.

He, B., C. Macdonald, and I. Ounis (2008). Ranking opinionated blog posts using opinionfinder. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 727–728. ACM.

Hey, J. (2004). The data, information, knowledge, wisdom chain: the metaphorical link. *Intergovernmental Oceanographic Commission*.

Inmon, W. and C. Kelley (1994). The 12 rules of data warehouse for a client/server world. *Data Management Review 4*(5), 6–16.

Kimball, R., M. Ross, et al. (2002). The data warehouse toolkit: the complete guide to dimensional modelling. *Nachdr.]. New York [ua]: Wiley*.

Kumar, S., R. Zafarani, M. A. Abbasi, G. Barbier, and H. Liu (2010). Convergence of influential bloggers for topic discovery in the blogosphere. In *Advances in Social Computing*, pp. 406–412. Springer.

Lee, Y., S.-H. Na, J. Kim, S.-H. Nam, H.-y. Jng, and J.-H. Lee (2008). Kle at trec 2008 blog track: Blog post and feed retrieval. Technical report, DTIC Document.

Luhn, H. P. (1958). A business intelligence system. *IBM Journal of Research and Development 2*(4), 314–319.

Lusch, R. F., Y. Liu, and Y. Chen (2010). The phase transition of markets and organizations: the new intelligence and entrepreneurial frontier.

Mahmoud, T., J. Marx Gómez, A. Rezgui, D. Peters, and A. Solsbach (2012). Enhanced bi systems with on-demand data based on semantic-enabled enterprise soa.

Mishne, G. (2006). Autotag: a collaborative approach to automated tag assignment for weblog posts. In *Proceedings of the 15th international conference on World Wide Web*, pp. 953–954. ACM.

Pang, B. and L. Lee (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval 2*(1-2), 1–135.

Power, D. J. (2007). A brief history of decision support systems. *DSSResources. COM, World Wide Web, http://DSSResources. COM/history/dsshistory. html, version 4*.

decision support systems: what is the next step?

Rouhani, S., S. Asgari, and S. V. Mirhosseini (2012). Review study: business intelligence concepts and approaches. *American Journal of Scientific Research 50*, 62–75.

Sallam, R. L., J. Richardson, J. Hagerty, and B. Hostmann (2011). Magic quadrant for business intelligence platforms. *Gartner Group, Stamford, CT*.

Sheth, A., C. Henson, and S. S. Sahoo (2008). Semantic sensor web. *Internet Computing, IEEE 12*(4), 78–83.

Tsai, F. S. (2011). A tag-topic model for blog mining. *Expert Systems with Applications 38*(5), 5330–5335.

Varian, H. R. (2008). Hal varian answers your questions.

Wilson, E. O. (1999). *Consilience: The unity of knowledge*, Volume 31. Random House LLC.

Zhang, X., Z. Zhou, and M. Wu (2009). Positive, negative, or mixed? mining blogs for opinions. *ADCS 2009*, 139.

Résumé

« Nous nous noyons dans l'information, tout en étant affamés de sagesse. Le monde sera désormais géré par des synthétiseurs, des gens capables de mettre en place la bonne information au bon moment, penser de façon critique, et faire des choix importants à bon escient », explique Edward Osborne Wilson dans son livre « Consilience: The Unity of Knowledge »(Wilson, 1999). Selon ce qui précède, les systèmes d'aide à la décision devrait permettre la gestion des flux de données, d'abord, par l'identification et la synthèse de l'information pertinente, et ensuite, le traitement de cette information à la connaissance et l'intelligence condensée et utile.

Vers un Méta-Modèle de Système d'Aide à la Décision: application au domaine médical

Ali Ayadi*, Salma Sassi*
Anis Tissaoui*

*FSJEG, Université de Jendouba, Tunisie
aliayadi1@gmail.com,sassisalma@yahoo.fr,tissaouianis@yahoo.fr

Résumé. Les systèmes d'aide à la décision connaissent un essor important en raison de leur capacité à supporter efficacement les analyses sur les données disponibles dans les organisations. Cependant, les systèmes actuels ne suivent pas les besoins d'analyses et ne s'intéressent qu'à l'analyse partielle des données (données numériques ou textuelles). Nous proposons dans cet article un Méta-Modèle de système intelligent d'aide à la décision qui peut être implémenté dans n'importe quel domaine d'activité, permettant une analyse qualitative et quantitative de tous les documents complexes rendant ainsi possible l'évolution des contextes d'analyses. Cette approche place l'agrégation des données et leurs représentations au centre du dispositif et remplace l'analyse unidimensionnelle de la donnée par une approche multidimensionnelle en se basant conjointement sur le processus d'entreposage (OLAP), l'ontologie et la théorie de l'aide à la décision.

Mots clefs: Système décisionnel, OLAP, entrepôt de métadonnées, agrégation des données, ontologie

1 Introduction

Un Système d'Aide à la Décision (SAD) est l'ensemble des outils informatiques (matériels et logiciels) qui permettent l'analyse des données opérationnelles issues du système d'information des organisations. Ces SADs sont présents dans de nombreux domaines et ont pour objectif d'aider le décideur dans sa tâche en lui fournissant tous les éléments pertinents pour la prise de décision. Cette dernière est devenue récemment un thème actif et important dans l'extraction et la gestion des connaissances.

Cependant, les modèles de décision traditionnels sont confrontés à d'énormes problèmes tels que l'hétérogénéité des données, leur analyse sémantique, leur représentation multidimensionnelle, le problème de personnalisation et celui de la visualisation. Pour résoudre ces problèmes, il serait intéressant de proposer une approche traitant tous les types de données, permettant leurs analyses sémantique et leurs représentation multidimensionnelle. Ces données répondant aux besoins du décideur, sont transformées en une vision orientée décideur puis présentées au moyen d'une interface graphique significative basée sur une technique métaphorique 3D.

Dans le domaine médical, les Systèmes d'Aide à la Décision Médicale (SADM) sont "des applications informatiques dont le but est de fournir aux cliniciens en temps et lieux utiles les informations décrivant la situation clinique d'un patient ainsi que les connaissances appropriées à cette situation, correctement filtrées et présentées afin d'améliorer la qualité des soins et la santé des patients"(Lobach et al., 2013).

Les documents médicaux sont le cœur des systèmes décisionnels dans la médecine. En effet, ces ressources ne cessent d'évoluer. Cette évolution croissante engendre, ainsi, des difficultés prévenant de trois problèmes : (i) les volumes des données médicales sont difficiles à manipuler, mais également elles sont de nature hétérogène. En effet, ces ressources sont difficilement stockées et réutilisées car elles n'ont pas été conçues à cet effet. (ii) Les outils et les méthodes actuels utilisés dans la représentation multidimensionnelle et l'analyse sémantique des données telle que la technologie OLAP (OnLine Analytical Processing), sont adaptés à la gestion des données numériques qui ne représentent que 20% des données d'un SI. Les données restantes sont hors de la portée du système d'analyse multidimensionnelle. Ceci mène sûrement à l'omission d'informations pertinentes, ce qui nuit par la suite à la prise de décision médicale vu l'absence de l'analyse sémantique. (iii) Les modèles actuels ne sont orientés qu'aux besoins d'analyse et ne concernent pas le contexte. En effet, ces modèles classiques ayant une flexibilité limitée, ne permettent pas l'évolution des contextes d'analyse et ne représentent pas les connaissances du domaine nécessaires aux décideurs.

C'est dans ce contexte que nos travaux de recherche visent à développer des infrastructures sémantiques de communications adaptées aux systèmes d'informations initialement hétérogènes mais aussi aux besoins des utilisateurs. Ces infrastructures forment un méta-modèle sémantique générique et instanciable qui permet de résoudre le problème d'hétérogénéité dans la présentation et dans la sémantique des données ainsi que la représentation multidimensionnelle des données. Ce Méta-Modèle d'Aide à la Décision (M^2SAD) est basé sur l'approche des entrepôts de données, de la technologie OLAP ainsi que l'ontologie. On a développé ces thèmes en les illustrant sur un terrain d'applications qui présente de nombreux facteurs de complexité : les systèmes d'information dans le domaine médical.

Dans cet article, nous présenterons un court état de l'art sur quelques travaux liés aux problèmes déjà cités ainsi que les solutions existantes. Nous détaillerons ensuite notre méthodologie M^2SAD proposée. Puis nous montrerons quelques aspects d'expérimentations menées. Un résumé des apports de notre contribution ainsi que les travaux futurs concluront cet article.

2 Les systèmes d'aide à la décision et le domaine médical : un bref état de l'art

Depuis plusieurs années, les travaux de recherche sur le thème des systèmes d'aide à la décision (SAD) se sont multipliés. En effet, ils visent à étudier et améliorer l'ensemble des techniques permettant, pour une personne donnée, d'opter pour la meilleure prise de décision possible. Ces SADs représentent un énorme potentiel en médecine. Par conséquent, nous présentons dans cette section les principaux travaux effectués dans ce domaine.

CAMD (Conforti, 1999) permet le traitement d'images basées sur la morphométrie cellulaire. Ce SADM est composé d'un classificateur automatique basé sur un outil de programmation mathématique. La phase de classification constitue le cœur du processus de son diagnostic

médical. Ce système aide le médecin à distinguer entre les cellules bénignes et malignes pour la détection prématurée du cancer du sein. Ce SADM n'a l'avantage que de classifier les images d'une mammographie et d'y gérer la maladie infectée statiquement. Il est spécifique au domaine de cancer du sein, ne gère que des données structurées. Ces données sont analysées mathématiquement et affichées sous forme de texte.

SAPHIRE (Laleci et al., 2008) est un système basé sur la notion des systèmes multi agents. Son principal but est de soutenir la définition, le déploiement et l'exécution des directives cliniques à un patient. Il est, en fait, composé d'un ensemble de composants (agents) collaborateurs s'exécutant dans un environnement distribué hétérogène. SAPHIRE consiste à extraire les données cliniques des dossiers de santé électroniques du patient et effectuer des analyses statistiques afin d'aboutir à une décision. Ce système traite uniquement les données structurées en leur analysant statistiquement. Il n'est pas spécifique à un domaine bien précis. SAPHIRE représente les données d'une manière unidimensionnelle et les expose au médecin sous forme textuelle.

IHMDS (Chang et Lu, 2009) qui est basé sur la notion des services web en collaboration avec les arbres de décision. Ce système, basé sur le langage XML, permet la manipulation des données structurées. L'avantage de ce système réside dans le fait qu'il incorpore l'interopérabilité et l'extensibilité de XML. Ce système permet au médecin de faire des diagnostics sur les symptômes et de générer uniquement les maladies correspondant à la thyroïde. En effet ce système effectue des analyses sémantiques d'une manière unidimensionnelle, les affichant sous forme de texte et de formulaire.

CNS (Couturier et al., 2010) est également basé sur la notion de systèmes multi agents, favorisant un meilleur diagnostic médical du patient. En effet, CNS vérifie plusieurs conditions qui pourraient être ignorées par les humains grâce à son analyse sémantique et statistique leur permettant d'améliorer les décisions pouvant être prise par des médecins. Ce système de diagnostic médical coopératif ne s'intéresse qu'aux domaines de l'endocrinologie. Il ne permet pas de traiter les données complexes non structurées, ni la représentation multidimensionnelle. Il utilise l'affichage sous forme textuelle et de formulaire.

D'autres travaux s'avèrent aussi intéressants dans ce domaine à savoir : le système FCM (Peter, 2012) utilise l'approche de cartes cognitives floues comme méthode de modélisation informatique des données médicales. En effet, ce système combine le processus d'aide à la décision médicale et la représentation de ces connaissances sur une carte cognitive. Il ne traite que les données médicales structurées en langage XML. FCM est générique dans le domaine médical, il affiche ses connaissances sous forme textuelle.

Après ce survol de l'existant, nous élaborons une étude comparative des différents travaux déjà décrits. Pour ce faire, nous dressons un tableau dans lequel des critères de comparaison sont pris en compte. Ces critères sont : (i) *la spécialité du SAD* est le secteur couvert par son activité. On distingue deux catégories : le SAD générique (gén) qui concerne tous les secteurs du domaine et le SAD spécifique (spé) qui concerne un secteur bien précis du domaine tel que le secteur cardiologique du domaine médical. (ii) *la nature des données* (données), deux types de données sont pris en compte. Les données structurées (st) disposées de façon à être traitées automatiquement par le système, mais non nécessairement par un humain. Les données non structurées (nn) qui s'adressent à l'humain et dont il n'est pas possible de prédéfinir leur structure. (iii) *le type d'analyse*, soit une analyse statistique (sta) sur des données simples, soit une analyse sémantique (sém) sur des données textuelles complexes en prenant en compte ainsi

leurs contexte.(iv) *la représentation des données*, unidimensionnelle (uni) où le sujet à étudier est analysé en une seule dimension, multidimensionnelle (mult) qui consiste à analyser un sujet selon plusieurs dimensions (axes d'analyses).(v) *la technique de visualisation* qui montre la manière dont les informations sont affichées à l'utilisateur. On distingue la visualisation textuelle (tex), la visualisation formulaire (form) et enfin la visualisation graphique (graph).

	Spécificité		Données		Analyse		Resprésentation		Visualisation		
	Spé.	Gén.	St.	Nn	Sta.	Sém.	Uni.	Mult.	Tex.	Form.	Graph.
CAMD	*		*		*		*		*		
SAPHIRE		*	*		*		*		*		
IHMDS	*		*			*	*		*	*	
CNS		*	*		*	*	*		*	*	
FCM	*		*			*	*		*		

TAB. 1 – *Tableau Comparatif des SADMs.*

Il est également évident que ces systèmes se distinguent par la différence entre eux en certains points et se ressemblent dans d'autres.

De la perspective de critique, nous arrivons à relever les constatations suivantes :

- La plupart des SADMs sont spécifiques à des sous-domaines bien précis.
- Ces SADMs étudiés manipulent essentiellement des données très structurées.
- Les analyses des données sont généralement statistiques.
- L'analyse d'un sujet n'est faite que par rapport à un seul et unique axe d'analyse.
- Les traitements sont visualisés soit sous forme textuelle soit sous forme de formulaire.
- Aucun système ne permet de :
  - Traiter toutes les spécialités d'un domaine spécifique.
  - Traiter des données structurées et non structurées.
  - Analyser statistiquement et Sémantiquement les données médicales.
  - Représenter les sujets d'analyse d'une façon multidimensionnelle.
  - Visualiser les résultats d'analyse sous plusieurs formes de visualisation selon le besoin de l'utilisateur : forme textuelle, formulaire et graphique.

En prenant en compte ces points, nous allons proposer un nouveau SADM sûr et fiable répondant à tous ces critères.

3 Méthodologie M²SAD : vers la prise en compte de la spécialisation des domaines

Dans le contexte de la prise de décision, il semble pertinent de prendre en compte la notion de la spécialité dans le domaine. Dans ce contexte, nous présentons le système M²SAD qui apparaît donc comme un méta-modèle à partir duquel seront définies des primitives permettant l'analyse sémantique des données, leurs représentations multidimensionnelles, leurs personnalisations et leurs visualisations sous différentes formes. Ces primitives sont adaptées à un domaine et à un type d'activité particulière.

Notre méta-modèle est conçu dans le but de traiter des données hétérogènes provenant de différentes sources et d'être utile pour des profils d'utilisateurs aussi bien différents. Pour ce

faire, une architecture à trois niveaux est définie. La figure 1, propose une architecture à trois niveaux de construction d'une structure partagée et collective pour guider l'utilisateur dans sa prise de décision prenant en compte la notion de spécialisation dans le domaine étudié.

- Le niveau 1 constitué par le méta-modèle qui définit les primitives du système à savoir : l'entrepôt des données, l'entrepôt des règles, l'ontologie et la technologie OLAP.
- Le niveau 2 constitué par l'instanciation des primitives déjà citées permettant ainsi de structurer un secteur d'activité bien déterminé et représenter les différents objets professionnels du domaine sous une forme unifiée et prenant en compte les différentes spécialités de ce domaine. Par exemple, dans le domaine médical ce niveau servira à définir un entrepôt de métadonnées, un entrepôt de règles, une ontologie de structure médicale et des primitives d'OLAP adaptées au domaine médical. Selon sa spécialité, le professionnel de santé crée les objets médicaux qui correspondent à son statut.
- Le niveau 3 est l'instanciation des éléments du niveau 2 qui permet de construire ; effectivement ; le système d'aide à la décision de l'entreprise.

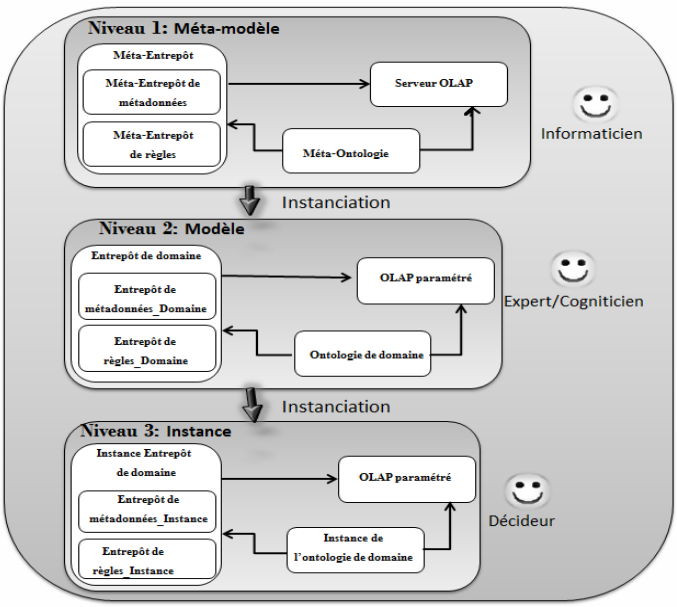


FIG. 1 – *Méta-modélisation de M²SAD.*

Le niveau méta-modèle est constitué par le méta-entrepôt qui contient le méta-entrepôt de métadonnées et le méta-entrepôt de règles, la méta-ontologie et le serveur OLAP. Le méta-modèle définit la description de la structure et de la sémantique de documents hétérogènes. A ce niveau, c'est l'informaticien qui intervient pour créer les concepts du méta-modèle de SAD. Les composants sont génériques pour les SADs de n'importe quel secteur d'activité. Ils serviront de patrons de modélisation de ces modèles de SAD.

Le niveau modèle définit le modèle de SAD, issu de l'instanciation de Méta-modèle, pour un domaine d'activité particulier. Le modèle de données définit l'entrepôt de métadonnées,

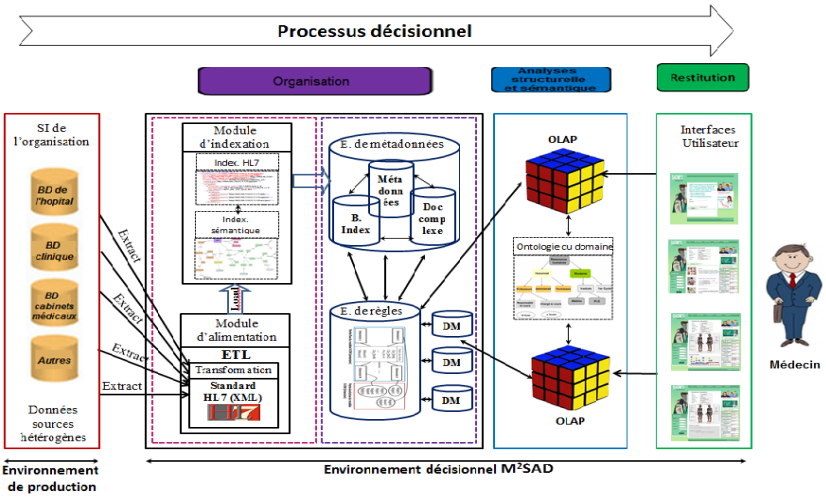


FIG. 2 – *Architecture de M²SAD.*

l'entrepôt de règles, l'ontologie de domaine du secteur d'activité et le serveur OLAP paramétré. Ici, c'est l'expert du domaine qui intervient et qui instancie le méta-modèle pour créer un modèle de SAD du domaine. L'instanciation correspond à limiter le méta-modèle de données aux concepts du méta-modèle de SAD autorisés par la loi du domaine d'activité.

Le niveau instance correspond à l'instanciation, par l'utilisateur final, des éléments du modèle de SAD décrits dans le niveau modèle de SAD. Cette instanciation lui permet de manipuler graphiquement, sur une interface dédiée, les différentes fonctionnalités du SAD selon ses droits et son profil. En effet, un médecin généraliste n'aura pas les mêmes fonctionnalités qu'un radiologue.

3.1 Architecture de M²SAD

L'architecture M²SAD est basée sur une organisation à trois phases, chacune d'elles étant basée sur un module de données particulier. Par exemple, le module de l'entrepôt gère l'évolution et l'indexation des données complexes alors que le module OLAP organise les données de manière multidimensionnelle pour optimiser l'analyse sémantique. La figure 2, montre de façon schématique l'architecture du Méta-Modèle de SAD proposé. Nous allons présenter, dans cette section, ces différentes phases une par une.

3.1.1 La phase d'organisation : une fusion d'entrepôt de métadonnées et d'entrepôt de règles

Cette phase a pour objectif d'assurer l'interopérabilité sémantique des données. En fait, les données provenant des sources de données hétérogènes sont restructurées, annotées, décrites, unifiées et contextualisées selon les standards de communication d'un domaine bien défini (dans le domaine médical, HL7 est utilisé pour standardiser les documents médicaux) afin de

résoudre le problème d'hétérogénéité. Cette phase permet également à travers l'entrepôt de métadonnées, de stocker de grandes masses de données et de les analyser.

Cette phase est composée de trois modules :

- Le module d'alimentation qui permet de récupérer les informations hétérogènes et pertinentes de diverses sources de données. Ce mécanisme prend en compte deux sortes d'hétérogénéité : structurelle et sémantique. L'hétérogénéité structurelle des sources se matérialise par des documents plus ou moins structurés et reposant ou non sur des standards. L'hétérogénéité sémantique est liée au fait que les documents peuvent concerner des contextes très divers (sujets, domaines, etc.). Ce module fournit des documents hétérogènes fortement structurés (standardisés par XML) en passant par deux étapes : une première étape d'extraction de structure et de contenu et une deuxième de standardisation.
- Le module d'indexation qui permet d'associer à un document un ensemble d'informations le caractérisant. Ces informations peuvent être des mots clés ou des Meta-informations. L'objectif du processus d'indexation consiste à décomposer les documents en entités informationnelles afin de les interroger et de les manipuler. Ce module se base sur trois étapes : la première étant l'étape de décomposition sert à segmenter le document en divers entités documentaires (objets ou granules). La deuxième étape est l'indexation, nous proposons à ce niveau, une indexation structurelle et sémantique. Finalement L'insertion du contenu qui consiste à insérer le contenu du document dans la base multidimensionnelle en rattachant chaque granule à l'élément de structure correspondant ainsi qu'aux métadonnées lui correspondant.
- Le module entreposage : le M^{2}SAD repose sur deux méta-entrepôts, le premier un méta-entrepôt de métadonnées qui rend facilement accessibles, exploitables et réutilisables les données. Le deuxième méta-entrepôt étant le méta-entrepôt de règles qui rend l'analyse des données plus flexible et offre ainsi au décideur la possibilité de définir ses propres règles pour déterminer de nouvelles hiérarchies de dimensions :

(a) Méta-entrepôt de métadonnées : Un entrepôt de métadonnées est une collection de données : intégrée, orientée sujet, non volatile, historiée, résumées et disponibles pour l'interrogation et l'analyse (Inmon, 2002). L'entrepôt intègre les données pertinentes, nécessaires à la prise de décision des systèmes d'informations de l'organisation. L'objectif de ce méta-entrepôt est de stocker toutes les informations issues des documents sources (contenus, structures et métadonnées) pour effectuer des analyses décisionnelles de tous types : recherche d'information, interrogation, analyse multidimensionnelle sémantique. L'entrepôt de métadonnées repose sur une base de données contenant des métadonnées, une autre contenant des hyperliens entre les objets des documents et des liens inter-documents, et enfin une base multidimensionnelle contenant les documents complexes ainsi que leurs différentes entités. Cet entrepôt propose une richesse sémantique grâce à la méta-ontologie.

(b) Méta-entrepôt de règles : Un critère essentiel qui distingue les SADs est la flexibilité. Bien évidemment les modules classiques apportent une certaine flexibilité temporelle au modèle par la mise à jour du schéma apportant une réponse à l'évolution des données. Cependant, ils n'apportent pas de solution à l'émergence de nouveaux besoins d'analyse qui sont orientés, non pas par l'évolution des données, mais par l'expression de nouvelles connaissances. L'en-

trepôt de règles pourra ainsi contribuer à la résolution de la problématique et rendre l'analyse plus flexible. En effet, il permet de créer des hiérarchies de dimension d'une façon dynamique en renforçant l'interaction entre l'utilisateur et le SAD. Cette génération de hiérarchie à la demande permet de faire évoluer les contextes d'analyses.

Le méta-entrepôt de règles est composé de deux parties : une fixe définie en extension, comprenant une table de fait et des dimensions du premier niveau. Une partie évolutive, définie en extension par des règles de type "Si Alors", qui déterminent les niveaux de granularité dans les hiérarchies de dimensions basées sur les dimensions existantes et de nouvelles connaissances de l'utilisateur. Les règles permettent d'établir des liens sémantiques entre les données (Favre et al., 2006). Ce méta-entrepôt garantie l'évolution, la flexibilité, la performance et l'optimisation des analyses.

3.1.2 La phase d'analyse : modélisation multidimensionnelle, sémantique de données complexes

Le modèle de données multidimensionnel OLAP est le responsable de l'organisation des données dans le système. Un modèle multidimensionnel contient les informations pertinentes et les présentent dans un format plus approprié pour faire leurs analyses. OLAP a pour but d'organiser les données à analyser par domaine ou thème et d'en ressortir des résultats pertinents pour le décideur. Pour garantir ce processus d'interrogation, nous avons besoin d'outils performants et conviviaux pour l'accès et l'analyse de l'information pertinente. Les systèmes OLAP fournissent des méthodes et des outils puissants permettant l'analyse de données transactionnelles (Sullivan, 2001) et (Tseng et Chou, 2006). Cependant, seul 20% des données d'un système d'information sont des données transactionnelles et peuvent être traitées par OLAP (Lioni et al., 2006). Les 80% restant des données représentent les documents électroniques (Sullivan, 2001) et (Tseng et Chou, 2006).

Par exemple, dans le domaine médical, les objets médicaux sont généralement annotés, c'est-à-dire qu'ils sont principalement constitués de texte, d'images, de sons, de vidéo, etc. Ces documents restent hors de la portée des SADs faute de manque d'outils de traitement et d'analyse.

Afin de remédier à cette problématique, nous visons à ajouter des fonctions d'agrégation textuelles. En effet, l'agrégation des données textuelles permet de résumer le volume de données à visualiser lors d'une même analyse. Ainsi, en réduisant le volume de données par les méthodes de synthèse, le décideur peut avoir une vision globale du domaine qu'il analyse d'où une analyse sémantique des données. Les fonctions textuelles utilisées sont détaillées dans (Ravat et al., 2008) et (Tseng et Chou, 2006) :

- TopKeyword : fonction qui extrait les n mots-clefs jugés les plus pertinents d'un fragment de texte.
- Topic : fonction qui extrait le sujet d'un fragment de texte.
- SUMMARY : génère un résumé automatique d'un fragment de texte.

Pour créer ces fonctions, il est nécessaire de disposer d'un moyen permettant de calculer et de trouver les mots-clefs. La méta-ontologie de domaine est employée à cette fin. Bien évidemment, les indicateurs sémantiques sont utiles pour la prise de décision, car ils résument l'information et permettent d'appréhender des systèmes complexes. Nous avons intégré l'ontologie vue que de plus en plus les décideurs souhaitent pouvoir tenir compte aussi des contextes sémantiques et des connaissances lors de leurs prises de décision (l'analyse numérique seule-

ment n'est plus suffisante). Pour cette raison, il serait indispensable d'intégrer un mécanisme performant en représentation des connaissances et permettant un raisonnement sémantique qui serait la méta-ontologie. Outre, cette ontologie de domaine permettrait l'interaction des "experts" qui seront responsables d'enrichir le processus de prise de décision tel qu'ils le perçoivent par leurs connaissances. De plus, l'ontologie permet à l'entrepôt de métadonnées de révéler tout un vocabulaire lié à l'utilisation et l'organisation de la base décisionnelle documentaire des SADs, d'où la garantie de l'aspect sémantique et la consommation en temps pour les interrogations de l'utilisateur. La connaissance associée à un domaine peut être représentée de façon plus formelle au travers d'une ontologie. Pour un utilisateur, accéder à la connaissance, grâce à une ontologie, peut lui permettre de spécifier son besoin et les lacunes de sa connaissance par rapport à l'information qui lui est disponible. D'autre part, la représentation des granules d'information à partir d'une ontologie peut définir un vocabulaire contrôlé (termes et concepts) à partir duquel l'utilisateur spécifiera son besoin. La description du besoin correspond, dans ce cas-là, aux caractéristiques des granules car elles ont été indexées à partir des mêmes ressources.

3.1.3 La phase de restitution : l'interface de visualisation

Cette interface permet de mieux appréhender le résultat de l'analyse. L'utilisateur final, n'étant pas forcément un informaticien, aura plus de facilités dans des interfaces simples, basées sur la visualisation et la perception de l'œil du décideur que d'aller requêter directement dans le serveur d'analyse. Pour atteindre cet objectif, nous proposons une interface graphique utilisant des diagrammes, des courbes statiques, une visualisation métaphorique et une technique de visualisation 3D.

4 Prototype et évaluation

Afin de valider notre approche, nous avons développé un prototype d'aide à la décision dans le domaine médical. Cet outil facilite la tâche du médecin dans sa prise de décision. Il est composé essentiellement d'une interface utilisateur et d'une base décisionnelle multidimensionnelle, sémantique.

Nous détaillerons ici quelques interfaces utilisateurs (voir figure 3) graphiques permettant au médecin de (i) consulter les données hétérogènes historiées du patient, (ii) effectuer des analyses sur un sujet bien déterminer selon diverses dimensions.(iii) répondre aux requêtes du médecin de manière analytique et sémantique. (iv) effectuer les diverses analyses pour obtenir un diagnostic lié au choix du symptôme et des résultats d'analyses. L'application se charge ensuite de répondre au médecin par la pathologie associée aux symptômes et aux analyses ainsi que la présentation des traitements nécessaires. La figure 3, présente des interfaces utilisateurs représentant respectivement les principales fonctionnalités du SAD : (a) consultation du dossier médical du patient, (b) aide à la décision, (c)aide au diagnostic et (d) aide à la prescription médicamenteuse. Elles montrent aussi les divers objectifs de notre méthodologie telles que la gestion des données structurées ou non (exploitation textuelle des comptes-rendus, exploitation de vidéo d'une échographie, etc.) ainsi que leur analyse statistique et sémantique (calcul de glycémie, exploitation de connaissances, etc.).

FIG. 3 – *Les principales fonctionnalités du prototype.*

Pour évaluer notre SADM, nous avons décidé de confronter les professionnels de santé aux fonctionnalités de ce prototype pour la pratique de prise de décision clinique et de tirer des premières recommandations sur l'utilisabilité de ses fonctionnalités. Dans cette évaluation, nous avons utilisé un corpus qui contient 23 dossiers médicaux de patient de l'hôpital régional de Jendouba (Tunisie). Nous avons mené une évaluation avec 5 cliniciens. Une étape préliminaire de l'évaluation a permis de présenter aux participants notre système. L'objectif de cette présentation est de fournir à tous les participants les connaissances nécessaires pour porter des jugements de l'utilité des services proposés par le SADM, des objectifs et de la nature concrète de notre recherche. Cette étape a permis aussi de familiariser les participants avec notre système ainsi qu'avec les notions utilisées lors de l'évaluation, afin de leur permettre de se concentrer davantage sur le test de l'utilité de notre SADM. Suite à cette démonstration, nous avons demandé à chaque participant de répondre à trois questions cliniques afin de tester sa compréhension et sa maîtrise des différentes fonctionnalités du système. Lorsque le participant pourrait répondre correctement aux questions, il a été considéré comme prêt à effectuer l'évaluation.

Cette même enquête qualitative est menée pour obtenir l'appréciation donnée par les praticiens sur l'impact perçu de notre SADM positif ou négatif par rapport à d'autres SADMs utilisés actuellement. Les métriques d'évaluation des SADMs étudiés concernent à la fois la validité des connaissances, la rapidité de la décision, la validité des solutions proposées et leurs impacts sur la qualité des soins, la satisfaction sur l'utilisateur ainsi que les technologies

déployées.

Les résultats ont démontré la faisabilité du système proposé par rapport aux autres existants. Les résultats des évaluations obtenus sont plutôt encourageants. Les cliniciens ont trouvé toutes les fonctionnalités utiles et l'interface très facile à manipuler, ils ont jugé également performant cet outil décisionnel en termes de temps de réponse, d'interopérabilité et surtout de qualité de soin. Bien que 70% des cliniciens disent ne pas avoir l'habitude de ce type d'outils, ils ont trouvé l'utilisation générale du système plutôt très simple à 90%. Seuls 10% des cliniciens ont dit avoir dû vraiment chercher des fonctionnalités (les autres les ont trouvées plus facile) et seuls 20% des cliniciens ont avoué avoir commis des erreurs de manipulation. La consultation, l'analyse, la manipulation et la visualisation des données dans son ensemble semble donc être plutôt très facile à appréhender. Ce qui était le plus encourageant, c'est que les cliniciens ont tous affirmé que ce prototype les aide énormément à élaborer des plans de soins corrects et à communiquer le plus facilement avec les autres professionnels de santés à travers les fonctionnalités du SADM.

5 Conclusion et perspectives

Dans cet article, nous avons proposé un nouveau méta-modèle générique pour l'aide à la décision. Ce méta modèle permet de trouver une syntaxe commune et partageable entre les utilisateurs du système pour décrire d'une façon identique les documents de structure différente. Ce méta-modèle assure également leur analyse structurelle et sémantique, leur représentation multidimensionnelle et leur visualisation sous différentes formes. Ce méta-modèle se base principalement sur un *méta-entrepôt de métadonnées* stockant de grosses masses de données et de différentes entités composant les documents hétérogènes grâce à la base d'index comportant les liens joignant ces entités à leurs métadonnées. Le *méta-entrepôt de règles* assure la personnalisation du SAD, procure plus d'agrégation de données et permet une analyse encore plus détaillée grâce à la multitude de dimension.

Finalement, la *base multidimensionnelle OLAP* couplée à la méta-ontologie de domaine permet d'analyser sémantiquement, contextuellement et multidimentionnellement les données décisionnelles. Enfin, nous avons validé ce méta-modèle en implémentant et évaluant un système d'aide à la Décision appliqué au domaine médical.

Bien que les résultats obtenus soient encourageants, plusieurs perspectives de travaux futurs peuvent être envisagées. Il serait intéressant de développer un formalisme qui pourra faciliter la tâche du concepteur. Nous envisageons également d'améliorer la complexité de notre approche et d'introduire d'autres mécanismes tels que le datamining qui permettra de ré-extraire et enrichir les connaissances des utilisateurs.

Ainsi, nous trouvons intéressant de développer et d'expanser les domaines d'application de notre système, par exemple l'adaptation de ce dernier dans le domaine économique, agricole, politique, etc.

Références

Chang, C. et H.-M. Lu (2009). Integration of heterogeneous medical decision support systems based on web services. In *Bioinformatics and BioEngineering, 2009. BIBE '09. Ninth IEEE*

International Conference on, pp. 415–422.

Conforti, M. (1999). Decision support systems for medical diagnosis. *Information Technology Applications in Biomedicine, 1999. ITIS-ITAB '99. 1999 IEEE EMBS International Conference on*, 25–26.

Couturier, V., M.-P. Huget, et D. Télisson (2010). Engineering agent-based information systems - a case study of automatic contract net systems. pp. 242–248. SciTePress.

Favre, C., F. Bentayeb, et O. Boussaid (2006). A rule-based data warehouse model. *4042*, 274–277.

Inmon, W. H. (2002). *Building the Data Warehouse,3rd Edition*. New York, NY, USA: John Wiley & Sons, Inc.

Laleci, G. B., A. Dogac, M. Olduz, I. Tasyurt, M. Yuksel, et A. Okcan (2008). Saphire: A multi-agent system for remote healthcare monitoring through computerized clinical guidelines agent technology and e-health. pp. 25–44.

Lioni, A., C. Sauwens, G. Theraulaz, et J.-L. Deneubourg (2006). A rule-based data warehouse model. In *BNCOD*, pp. 274–277.

Lobach, D. F., K. Kawamoto, K. J. Anstrom, G. M. Silvey, J. M. Willis, F. S. Johnson, R. Edwards, J. Simo, P. Phillips, D. R. Crosslin, et E. L. Eisenstein (2013). A randomized trial of population-based clinical decision support to manage health and resource use for medicaid beneficiaries. *J. Medical Systems 37*.

Peter, P. (2012). The challenge of modeling decision support systems for medical problems using fuzzy cognitive maps: An overview. In *BIBE*, pp. 132–138.

Ravat, F., O. Teste, R. Tournier, et G. Zurfluh (2008). Top_keyword: an aggregation function for textual document olap. In *Data Warehousing and Knowledge Discovery*, pp. 55–64. Springer Berlin Heidelberg.

Sullivan, D. (2001). *Document Warehousing and Text Mining: Techniques for Improving Business Operations, Marketing, and Sales*. John Wiley & Sons.

Tseng, F. S. et A. Y. Chou (2006). The concept of document warehousing for multi-dimensional modeling of textual-based business intelligence. *Decision Support Systems 42*(2), 727 – 744.

Summary

The decision support systems know an important upheavel due to their capacity to support effectively the data analysis available in the organization. However, the actual systems don't follaw the organization's analysis needs evolution and are only interested in the partial data analysis(digital or textual data). We propose in this paper, a new metamodel of decision support system generic which can be implemented in any field of activity, allowing quantitative and qualitative analysis for all complex documents, makes this possible the contexts of analysis evolution. This approach puts the aggregating data and their visualization in the center of the system. It replaces the one-dimentional approach by a new multidimentional one,wich is based on datawarehouse process, ontology and the decision support theory.

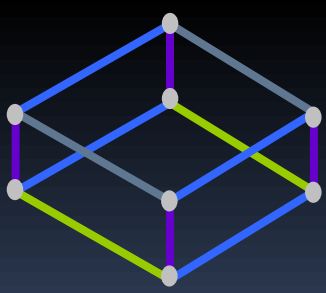
Keywords: *decision-making system, OLAP, data warehouse, data aggregation, ontology*

Index des auteurs

8ème édition de la conférence sur les Avancées des Systèmes Décisionnels

29-31 mai 2014, Hammamet — Tunisie

L'importance accordée par la communauté scientifique et par les industriels à l'informatique décisionnelle (ou Business Intelligence) ne cesse d'augmenter, comme en témoigne le nombre de travaux théoriques et d'outils mis sur le marché. En effet, les systèmes décisionnels permettent à l'entreprise de prendre des décisions à différents niveaux hiérarchiques en analysant son existant et son passé pour mieux prédire le futur. Sur le plan de la recherche, la conférence sur les Avancées des Systèmes Décisionnels (ASD) est dédiée aux systèmes décisionnels permettant de consolider les efforts des chercheurs d'une part, et d'autre part, de répondre aux aspirations des professionnels. Les contributions que cette édition accueille portent en particulier sur les thèmes suivants: architecture, organisation et conception des entrepôts de données, modélisation multidimensionnelle, sémantique et ontologies décisionnelles, big data et entrepôts de données, entrepôts de données complexes, business intelligence et cloud computing, les systèmes d'information décisionnels et applications industrielles …



www.asd-conf.net

Prix: 30 DT