10ième édition

Conférence sur

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

ASD 2016

ASD 2016

Actes de la 10ième édition

Conférence sur

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

Edités par

Mehdi NAFA, Abdelhalim BAAZIZ et Mohamed Amine FERRAG

14-16 mai 2016

Annaba, Algérie

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, Hammamet-Tunisie 2014, Tanger-Maroc 2015), ASD fait peau neuve et s'est convertie depuis sa 7ème édition en 2013 en ***Conférence Maghrébine sur les Avancées des Systèmes Décisionnels***. Cette nouvelle édition ASD 2016, la dixième de son rang, est accueillie cette année par l'Algérie.

ASD 2016 ambitionne de consolider les expériences conduites par les chercheurs, industriels et utilisateurs issus de communautés travaillant sur les systèmes décisionnels. L'objectif de cette dixième 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* au niveau du Maghreb.

Ces actes regroupent les articles acceptés et présentés à cette nouvelle édition. ASD 2016 a reçu 55 soumissions d'articles en provenance de différents pays (Algérie, France, Maroc, Tunisie). Après évaluation par les membres du comité scientifique, composé par 60 chercheurs-experts internationaux du domaine, 16 articles longs et 7 articles courts ont été retenus. Ces papiers couvrent différents thèmes de recherche et d'application sur les systèmes décisionnels.

ASD 2016 est organisée par l'université Badji Mokhtar d'Annaba, Algérie, et a reçu son soutien ainsi que celui de différentes institutions publiques d'enseignement et de recherche que nous tenons à remercier : Faculté des Sciences de l'Ingéniorat, Laboratoire de Recherche en Informatique (LRI), Laboratoire d'Ingénierie des Systèmes Complexes (LISCO), Laboratoire des Systèmes Embraqués (LASE), Laboratoire Réseaux et Systèmes (LRS), Laboratoire de Gestion des Document Electronique (LABGED) ; ainsi que des institutions internationales : l'Institut de la Communication (ICOM) et le Laboratoire ERIC de

l'Université Lyon 2 (France), l'Université HASSAN II Mohammedia-Casablanca (Maroc), la Faculté des Sciences et Techniques de Mohammedia (Maroc),, la Faculté des Sciences Economiques et de Gestion de Sfax (Tunisie), le Centre de Recherche en Informatique, Multimédia et Traitement Numérique des Données de Sfax (Tunisie), ainsi que 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
M. NAFA, A. BAAZIZ et Med A. FERRAG

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
- HARBI Nouria, ERIC, Université Lumière Lyon 2, France

Comité scientifique

- ABBACI Noudjoud Kahya, Université Badji Mokhtar, Annaba, Algérie
- 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
- ASFARI Ounas, Université Lyon2, France
- ATMANI Baghdad, Université d'Oran, Algérie
- AYACHI Sonia, ISG, Sousse, Tunisie
- AZIZI Nabiha, Université Badji Mokhtar, Annaba, Algérie
- BAAZIZ Abdelhalim, Université Badji Mokhtar, Annaba, Algérie
- BADACHE Nadjib, CERIST Alger, Algérie
- BADARD Thierry, Université Laval, Canada
- BADIR Hassan, ENSAT, Maroc
- BADRI Abdelmajid, Université Hassan II, Maroc
- BAHI Halima, Université Badji Mokhtar, Annaba, Algérie
- BELLAFKIH Mostafa, INPT Rabat, Maroc
- BELLATRECHE Ladjel, ENSMA Poitiers, France
- BENABES Farouk, Université Badji Mokhtar, Annaba, Algérie
- BEN ABDALLAH Hanene, Université de Sfax, Tunisie
- BENBLIDIA Nadjia, Université de Blida Algérie
- BENCHAALEL Amir, Université Badji Mokhtar, Annaba, Algérie
- BENHARKAT Nabila, INSA de Lyon, France
- BENSLIMANEl Djamel, Université de Lyon1, France
- BENTAYEB Fadila, Université Lumière Lyon 2, France
- BIMONTE Sandro, IRSTEA, Clermond-Ferrand, France
- BOUFAIDA Mahmoud, Université de Constantine, Algérie

- BOUFAIDA Zizette, Université de Constantine, Algérie
- BOUFARES Faouzi, LIPN Paris France
- BOUKHALFA Kamel, USTHB, Alger, Algérie
- BOUKRAA Doulkifli, Univesrité de Jijel, Algérie
- BOULMALKOUL Azedine, Université Hassan II, Maroc
- BOURAMAOUL Ramzi Abdelkrim, Université de Constantine, Algérie
- BOUSSAID Omar, Université Lumière Lyon 2, France
- DARMONT Jérôme, Université Lumière Lyon 2, France
- DERDOUR Makhlouf , Université Cheikh Larbi Tbessi de Tebessa, Algérie
- DERRAR Hacene , Université de Blida, Algérie
- DIB Lynda, Université Badji Mokhtar, Annaba, Algérie
- EL HEBIL Farid, INPT Rabat, Maroc
- EL-MOUADiB Faraj, FIT, Benghazi, Lybie
- ELAMMARI Mohamed, FIT, Benghazi, Lybie
- FAVRE Cécile, Université Lumière Lyon 2, France
- FEKKI Jamel, Université de Sfax, Tunisie
- FERRAG Mohamed Amine, LRS, Université du 8 mai 1945, gUELMA, Algérie
- GARGOURI Faiez, Université de Sfax, Tunisie
- GHOUALMI Nassira, Université Badji Mokhtar, Annaba, Algérie
- GHOZZI Faiza, Université de Sfax, Tunisie
- HACHAICHI Yasser, Université de Sfax, Tunisie
- HAFIDI Mohamed, Université Badji Mokhtar, Annaba, Algérie
- HARBI Nouria, Université Lumière Lyon 2, France
- HIDOUCI Walid, ESI Alger, Algérie
- IDRISSI Abdellah, Université Mohammed V, Rabat, Maroc
- JERBI Houssem, University College Dublin, Ireland
- KABACHI Nadia, Université Lyon1, France
- KAZAR Okba, Université Biskra, Algérie
- KHADIR Mohamed Tarek, Université Badji Mokhtar, Annaba, Algérie
- LEMIRE Daniel, Université du Québec à Montréal, Canada
- MAHDAOUIi Latifa, USTHB, Alger, Algérie
- MAHNANE Lamia, Université Badji Mokhtar, Annaba, Algérie
- MALKI Mimoune, Université de Sidi Bel Abbes, Algérie
- MARGHOUBI Rabia, Université Hassan II, Maroc
- MELIT Ali, Université de Jijel, Algérie
- MEFTOUH Karima, Université Badji Mokhtar, Annaba, Algérie
- MEROUANI Hayet Farida, Université Badji Mokhtar, Annaba, Algérie
- MEZIANE Abdelkrim , CERIST, Algérie
- MOUSSA Rim, Université de Carthage, Tunisie

- MOUSSAOUI Abdelouaheb, Université de Sétif, Algérie
- NABLI Ahlem, Université de Sfax, Tunisie
- NAFAA Mehdi, Université Badji Mokhtar, Annaba, Algérie
- NOUAOURIA Nabila, Université Badji Mokhtar, Annaba, Algérie
- OMARY Fouzia, Université Mohammed V, Rabat, Maroc
- OUKID Saliha, Université de Blida, Algérie
- OULAD HAJ THAMI Rachid, ENSIAS Rabat, Maroc
- RAVAT Frank, Université de Toulouse, France
- REGUIEG F Zohra, Université de Blida, Algérie
- SEKHRI Larbi, Université d'Oran, Algérie
- SELMANE Sid Ali , Université de Blida Algérie
- SERIDI Hassina, Université Badji Mokhtar, Annaba, Algérie
- SIDHOM Sahbi, Université de Nancy, France
- SLIMANI Yahya, FS, Tunis, Tunisie
- TALEB Nora, Université Badji Mokhtar, Annaba, Algérie
- TERRISSA Labib, Université Med Khider, Bskra, Algérie
- TESTE Olivier, Université de Toulouse, France
- ZAROUR Nasreddine, Université de Constantine, Algérie
- ZEGOUR Djamel Eddine, ESI Alger, Algérie
-

Comité d'organisation

- BAAZIZ Abdelhalim (LABGED, Université Badji Mokhtar Annaba, Algérie)
- DJEMAM Youcef (LRI, Université Badji Mokhtar Annaba, Algérie)
- NAFA Mehdi (LRS, Université Badji Mokhtar Annaba, Algérie)
- FERRAG Mohamed Amine (LRS, Université du 8 mai 1945, Guelma, Algérie)
- AHMIM Ahmed (LRS, Université Larbi Tbessi Tebessa, Algérie)
- BENABES Farouk (LABGED, Université Badji Mokhtar Annaba, Algérie)
- MELOUAH Ahlem (LRS, Université Badji Mokhtar Annaba, Algérie)
- GHANEMI Salim (LASE, Université Badji Mokhtar Annaba, Algérie)
- BENCHALEL Amir (LRS, Université Badji Mokhtar Annaba, Algérie)
- BENTAYEB Fadila (ERIC, Université Lumière Lyon 2, France)
- HARBI Nouria (ERIC, Université Lumière Lyon 2, France)
- MOALLA Mohamed Sahbi, ISET Sfax - Tunisie

ASD'2016

Conférence sur les Avancées des Systèmes Décisionnels

14-16 mai 2016, Annaba, Algérie

Sommaire

Prototype of decisional system based Cloud for evaluation of urban transport projects
Imene Benatia, Mohamed Ridda Laouar, Sean b Eom, Hakim Bendjenna…….. *001*

Etude de la dynamique des phénomènes géographiques à travers une modélisation sémantique à base d'ontologie des objets spatio-temporels
Fethi Ghazouani, Wassim Messaoudi, Imed Riadh Farah…… *015*

Amélioration de l'effectivité et du temps de réponse dans la recherche d'information
Imene Mansouria Zemani, Lougmiri Zekri, Mohammed Senouci…….. *027*

Applications et enjeux des Big Data dans le contexte des défis mondiaux
Saouli Hamza, Kazar Okba, Kassimi Dounya…… *041*

Data mining pour la construction de communautés d'utilisateur
Nour El Houda Boulkrinat, Habiba Drias, Hakima Mellah, Hassina Khellouf, Aida Bouchabou…….. *053*

Elaboration d'un modèle artificielle de filtrage de SPAM basé sur les fonctions rénales humaines
Mohamed Amine Boudia, Reda Mohamed Hamou, Abdelmalek Amine…… *065*

A new approach based on the power saves of social bees for automatic summaries of texts by extraction
Mohamed Amine Boudia, Reda Mohamed Hamou, Abdelmalek Amine, Mohamed Elhadi Rahmani…… *077*

Unsupervised Classification of Unstructured Data using Filters Combination by Workers Social Bees with 3D Visualisation
Hadj Ahmed Bouarara, Reda Mohamed Hamou, Abdelmalek Amine…….. *089*

Indexing-based link discovery in Linked Data
Khayra Bencherif, Mimoun Malki, Soumia Berrahal…… *101*

Extraction de connaissances à partir des séries temporelles d'images satellites pour l'interprétation des changements aléatoires
Ali Ben Abbes, Imed Riadh Farah…… *113*

Méthode Déterministe pour la Fragmentation Horizontale des Entrepôts de Données
Mohamed Barr, Kamel Boukhalfa, Karima Hocine…….. *125*

New data selection approach for faster SVMs
Sonia Chaibi…….. *137*

Vers une Solution de Protection des Données Entreposées dans le Cloud basée sur les Systèmes Multi Agents *151*
Sara Rhazlane, Nouria Harbi, Nadia Kabachi, Hassan Badir……..

A New Approach for Privacy preserving in Big Data through Access Control
Amine Rahmani, Abdelmalek Amine, Reda Mohamed Hamou, Mohamed Elhadi Rahmani, Mohamed Amine Boudia…….....…….. *165*

Smart Approach for the solar irradiation estimation based on Multi-Agent System
Mohamed Amir Abbas, Nadjia Benblidia, Nour El Islam Bachari……. *179*

Un environnement sémantique à base d'agents pour la formation à distance (E-Learning)
Samir Bourekkache, Okba Kazar, Laid Kahloul, Faiez Gargouri, Aicha-Nabila Benharkat…… *191*

Médiation Sémantique dans MedPeer : Un Système d'Intégration de Sources de Données Hétérogènes Basé sur les Ontologies
Naima Souad Ougouti, Haféda Belbachir, Youssef Amghar…… *203*

Ontologie générique des concepts des Ahadiths El Nabawia
El Charifa Meftah Dahmouni, Hassina Aliane, Kamel Boukhalfa…….. *211*

Context aware recommender system for access to the adapted web information system
Fatiha Belkhir, Fatiha Rezoug ……..

Une nouvelle approche basée sur la détection d'opinion par SentiWordNet pour les résumés automatiques de textes par extraction
Mohamed Amine Boudia, Reda Mohamed Hamou, Abdelmalek Amine, Ishak H.A Meddah…… *231*

Une nouvelle méthode pour le calcul du skyline basée sur le tri
Lougmiri Zekri…….. *239*

Une nouvelle stratégie pour le calcul du skyline sur GPU
Hadjer Belaicha, Lougmiri Zekri, Larbi Sekhri…….. *247*

Prototype of decisional system based Cloud for evaluation of urban transport projects

Imene Benatia*, Mohamed Ridda Laouar**
Sean b Eom**, Hakim Bendjenna****

*benatia.imene@yahoo.fr
**ridda_laouar@yahoo.fr
***beom@semo.edu
****hbendjenna@gmail.com

Abstract. Urban planning is the discipline of planning on the physical, social, economic and environmental development of metropolitan regions, municipalities and neighborhoods. It consists to develop land-use and building plans as well as local building and environmental regulations. Several problems in the field of the city planning require computerized solutions; one of these problems is the problem of the choice and the evaluation of the urban projects. In this article, we proposed a decisional system based Cloud to help the various decision makers of the city to select the best urban transport project.
Keywords. Decisional system, Cloud Computing, virtualization, urban transport project.

1 Introduction

Decision systems should take advantage of advanced technologies in computer science; Cloud Computing is a new technology that is useful for decisional systems. Cloud Computing is defined by US National Institute of Standards and Technology (NIST) a as a model for enabling ubiquitous, convenient, on-demand network access to a 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 (Mell and Grance, 2011). The combination of cloud computing and Decision System is used to solve different problems in different areas such as: city planning, medicine and other domains.
The core of city planning aims to provide a safe, organized, and enjoyable home and work life for residents of both new and established town (anonymous). Each urban city tries to solve the problems in the various domains (Habitat, Transport, Economy…) by proposing different urban projects to reach a sustainable development. The Projects represent a set of approaches to obtain an agreement between the different decision makers of the city. The urban projects are often characterized by contradictory criteria (Schutte & Brits, 2012) which generate a problem of conflict between the decision makers.
Our study proposes a Decisional System to effectively manage the urban projects that involves several criteria that are often conflicting in nature whose importance is not the same. In addition, typical urban projects involve several entities with contradictory interests. Consequently, the urban projects are viewed and evaluated differently according to their knowledge, objectives and concerns during the evaluation and the choice stages of the urban

project. The decision-making process is distributed between several entities with the conflict interests that affect the final decision. Our contribution consists in implementing a Decisional System deployed on a platform Cloud and managed by an infrastructure Cloud.
This article is structured as follows: a brief introduction is presented in Section 1. In Section 2, we will present an overview of Cloud Computing including definition, service models, and types of Cloud computing. We will present thereafter, related works in Section 3. The proposed decision models and the proposed architecture of Cloud computing for evaluation of urban transport projects are described in Section 4. The paper ends in with a conclusion.

2 Overview of Cloud Computing

2.1 Definition of Cloud Computing

Cloud computing can be defined in many different ways. Sun Microsystems (Sun Microsystems, 2009), for example, defined it as the ability:

- to rent a server or a thousand servers and run a geophysical modeling application.
- to rent a virtual server, load software on it, turn it on and off at will, or clone it ten times to meet a sudden workload demand.
- to store and secure immense amounts of data that is accessible only by authorized applications and users.
- to use applications on the Internet that store and protect data while providing services such as email, sales force automation and tax preparation.
- to use a storage cloud to hold application, business, and personal data.
- to use a handful of Web services to integrate photos, maps, and GPS information to create a mashup.

One of the widely accepted definitions of cloud computing is presented by the National Institute of Standards and Technology (NIST). Cloud computing has been defined as “a model for enabling convenient, on demand network access to a 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 cloud provider interaction” (Mell & Grance, 2011). The NIST cloud computing definitions specifically address three different aspects of cloud computing: deployment models, service models, and service attributes (Sosinsky, 2011).

2.2 Service Models of Cloud Computing

Sosinsky suggests thinking of cloud computing service model in terms of a hardware/software stack working together to process data to produce a result. This is known as the SPI (Software, Platform, and Infrastructure) model of cloud computing.

- **Software as a Service (SaaS):** Software as a Service provides the actual applications to end users via the Internet including e-mail, customer relationship management, etc. These services are accessible via different interfaces, web browser.
- **Platform as a Service (PaaS):** PaaS offer to developers an environment for developing, deploying and provisioning cloud applications, including programming languages, application programming interfaces (APIs), etc.

- **Infrastructure as a Service (IaaS):** IaaS allows the client to rent hardware (memory, computation, storage, network capability), virtualization tools (virtual machines, virtual storages, and virtual infrastructure), and co-location (facilities) services.

2.3 Deployment Models of Cloud Computing

The deployment models address the purpose and the location of the cloud with the following four types:

- **Public Cloud:** a public cloud is typically hosted off-premise, accessible and available over Internet to a wide public, multi-tenant sharing, and utility pricing.
- **Private Cloud:** a private cloud is hosted on-premise, accessible and available over private network, single-tenant sharing, and capacity pricing. It is dedicated to be used for a single organization. The private cloud can be managed by the organization itself (internal Private Cloud) or a third party (external Private Cloud).
- **Community Cloud:** in this type of Cloud, a.k.a. a vertical cloud, resources are shared by major industries or government organizations that have common interests and incentive to leverage common resources (Rhoton, 2011). It can be managed by the organizations themselves or by a third party.
- **Hybrid Cloud:** a hybrid cloud, a.k.a. multi-sourced models, is the use of multiple clouds: public, private or community.

3 Related works

In the context of the urban planning, several researchers concentrated on the resolution of the problems of planning in various domains of the city (habitat, transport, ecology, evaluation… etc.).

● **Transport:** Moreira de Oliveira and al. (2012) developed a decision support system for the public transportation, the decsional system is based on geographic information system (GIS) which contributes to a more transparent and efficient way of regulate, supervise and plan the bus lines.

Hasan (2010) proposed an intelligent decision support system for Traffic Congestion Management System. The IDSS suggested educes the dependability on the expertise and level of education of the transportation planners, transportation engineers, or any transportation decision makers.

Yang and Mou (1993) proposed a geographical information system (GIS)-based DSS in the planning of public transport to decrease urban traffic in China. Jun and Yikui (2009) developed an intelligent DSS for planning a road urban network. Fayech and al. (2002) presented a multi-agent DSS for planning the urban transport network. Longfei and Hong (2009) proposed an approach for resolving the problem of parking in the city using the negotiation process based on calculation of routes utility.

● **Planification du development durable:** Dur and al.(2009) proposed a decision support system based on GIS for helping decision maker in selecting policy options according to the economic, environmental and social impacts that will be introduced in urban sustainable development. Ahris and al. (2009) suggested a decision support system based on the use of GIS and Multi Criteria Decision Making that allows to plan and manage a sustainable development;

● **Enviroment:** The rapid growth of urban areas of the city will generate big open spaces, for this purpose Maktav, et al. (2011) developed a multicriteria spatial DSS to manage open spaces in the city. Shi and Li (2010) describe the development of the framework for an integrated DSS for assessing and controlling urban traffic-related air pollution

● **Evaluation:** Yujing, and al. (2011) constructed a DSS model with gray relational Decision-Making and Fuzzy AHP which helps government decision makers to improve the performance of urban projects. Alshawi and Dawood (2009) developed a DSS which allows decision makers to evaluate the construction projects of new cities to solve the problem of slums in the Islamic world.

	Domain	Tools					
		GIS	MCDM	ES	AI	MAS	Negotiation
Yang et Chen (1993)	Transport	X					
Fayech & al. (2002)	Transport					X	
Jun & Yikui (2009)	Transport				X		
Longfei et Hong (2009)	Transport						X
Dur and al.(2009)	Economic Social Environmental	X					
Hasan (2010)	Transport				X		
Alshawi & Dawood (2009)	Evaluation Housing		X				
Ahris et al. (2009)	Economic Social Environmental	X	X				
Shi & Li (2010)	Environmental			X			
Maktav & al. (2011)	Environmental	X	X				
Yujing & al. (2011)	Evaluation		X				
Moreira de Oliveira & al. (2012)	Transport	X					
Takahashi & al. (2013)	Transport						X

TAB. 1 – *Different research in urban planning domains*

These studies showed that Decisional System is an effective tool for city planning in different sectors, using a wide range of DSS tools such as artificial intelligence(AI), multi-agent systems(MAS) , negotiation, multiple criteria decision making (MCDM) , Geographic Information Systems(GIS), etc. Examining works on transport and evaluation field, there is no contribution which deals evalutation of urban transport projects, so we proposed a DSS based on the use of multi-criteria decision making and negotiation and we propose an architecture Cloud to integrate our decisional system in the new technology Cloud Computing in order to ameliorate our DSS and to enjoy the benefits of cloud in terms of accessibility of decision makers,decrease time of development and deployment of our decisional system and data sharing.

4 proposed model for evaluation of urban transport projects

The thrust of urbanization in China has grown considerably, causing the growth of motor vehicles. This generates urban environment and urban traffic problems. For these reasons, the Chinese authorities underlined to give priority to public transport in order to increase the proportion of public circulation in the whole of the traffic and to reduce the carbon emissions in China. In this context, LIU [8] proposed a multi-criteria decision analysis in order to evaluate and rank Chinese urban transport projects. Thus, in order to validate our decisional system, we used urban transport projects proposed by LIU [8]. Figure 1 describes the eight urban transport projects and their criteria.

		P1	P2	P3	P5	P6	P7	P8
Criteria	Sub-Criteria	Taiyuan	Xi'an	Xiangyang	Jiaozuo	Wuhan	Changzhi	Xining
Economic	IRR (%)	13.5	14	17.1	17.39	15.3	19.2	16.1%
	NPV ($ million)	45.7	296	613	142.04	138.4	154	67.90
	Benefit Cost Ratio	n/a	n/a	n/a	1.59	n/a	n/a	1.28
	Traveling Time Saving (min) [1]	12	n/a	n/a	5.7	7	n/a	3.8
	Unit Saving in Fuel Efficiency (liter/¥)	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Engineering	Annual Average Daily Traffic (pcu/h)	8155	4208	n/a	2233	4124[6]	3028	2650
	Designed Life Span of Road (year)	15	15	15	15	15	13.75	13
	Adoption of Sustainable Material & renewable resource	n/a	No	No	Yes	No	No	n/a
	Improved Road Network Density [5]	n/a		50%	n/a	566.67%	n/a	420%
Environmental	Expected Air Pollution (mg/m3) [2]	0.305	0.140	0.1815	0.164	0.679	0.5517	0.607
	Expected Noise Level (db) [3]	71.9	77.9	67.2	67.4	80.2	64.2	65.9
	% of Green Area	14.73	16.7	n/a	n/a	25	17897.021/	20.2
	% of Investment in Environmental Protection	2.02	1.0	5.96	1.45	3.34	1.87	1.75
Social	# of Displaced Family	2193	527	272	38	391	0	121
	# of Displaced People	7752	2385	1498	120	1486	0	536
	Public perception of satisfaction with the project[4]	5.0	4.0	4.16	4.8	3.46	4.7	3.75
	Reduction of Traffic Accident Fatalities [6]	0.15	n/a	0.33	1.7	6	1.35	0.5
	Improved Access for Disadvantaged Groups to Jobs, Education and Healthcare	No	Yes	Yes	Yes	Yes	No	Yes
Risk	Project External Risk	n/a	Substantial	Low	Moderate	Moderate	Low	Moderate
	Project Design Risk	Substantial	Substantial	Moderate	Substantial	Moderate	Moderate	Moderate
	Project Implementation Risk	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate

2, Max Nitrogen dioxide (NO2) value estimated by Caline4 model
3. Average of estimated max noise level at the affected residential site in day time and in night time, estimated by CadnaA
4, Public perception of satisfaction with the project in the affected region
5, Increased number of intersection with traffic signals
6, Average of expected AADT of Heping Road and Jiefang Road; Percentage of decrease in the number of traffic accident fatalities reduction per 10,000 motorized vehicles

FIG. 1 – *Chinese urban transport projects.*

4.1 Decisional system

The suggested decisional system is composed of two models, the first model consists in performing a multi-criteria analysis to elaborate the rankings of the urban transport projects using the Promethee II method and the second executes the Hare method of negotiation process in order to determinate the final urban transport project to decision makers of the city. Our suggested Decisional System is illustrated in Figure 2.

FIG. 2 – *Proposed decisional System.*

In the first level of our decisional model, each decision maker should specify weights of each characteristic of urban transport projects according to his choice as represented by figure 3.

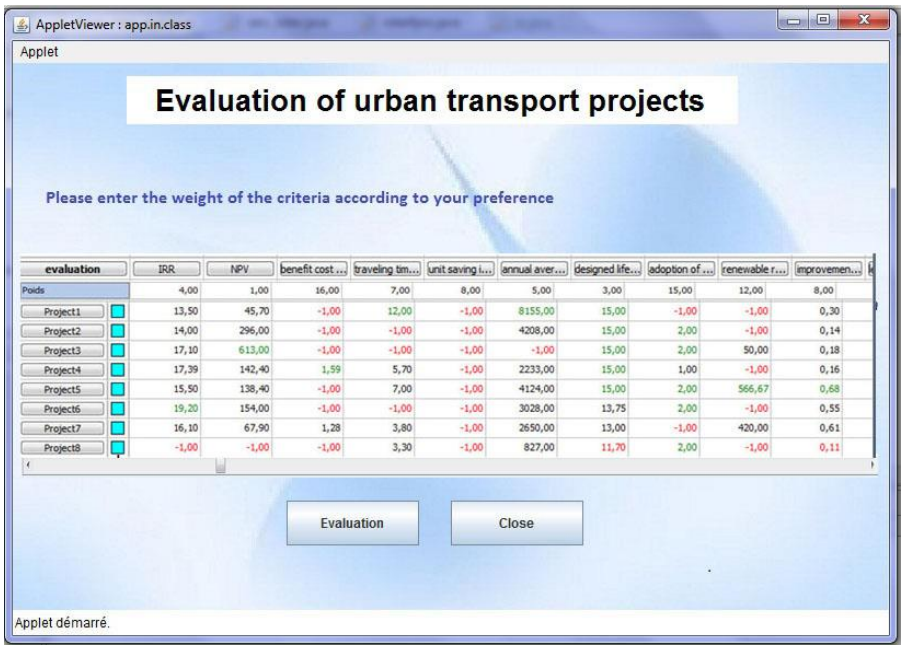
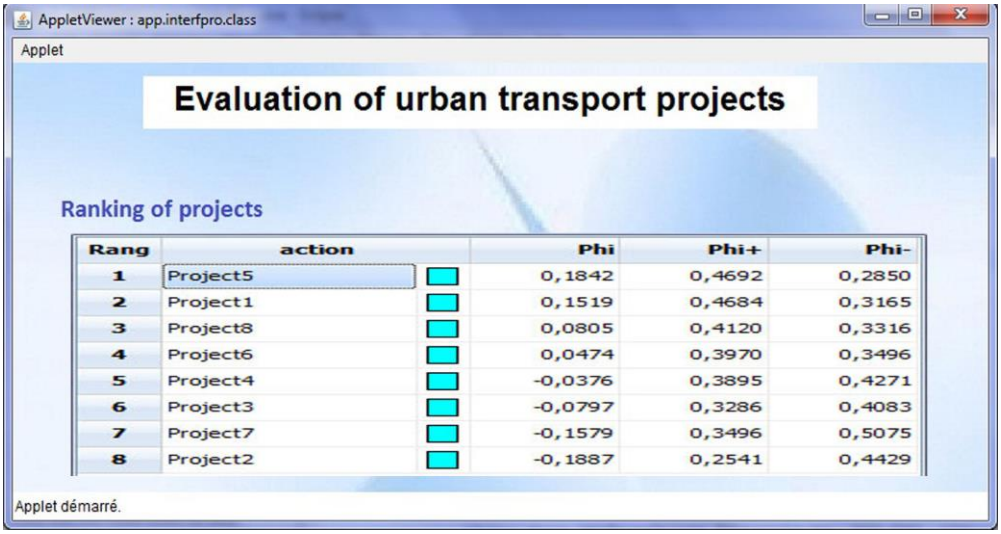


FIG. 3 – *Specifying weights by a decision maker*

All weights introduced by decision makers will be exploited by the multi criteria analysis using Promethee II method to rank urban transport projects from the best to the worst. Promethee II is an effective multi-criteria decision-making tool often applied to deal with complex problems. Promethee II generates complete ranking for a final set of urban transport projects. It is based on pair -wise comparisons of transport projects . Promethee II method requires two additional types of information for each characteristic: a weight and a

preference function[1]. It is a method is based on a comparison pair per pair of possible decisions along each criterion. Possible decisions are evaluated according to different characteristics; the step-wise procedure of Promethee II starts by determining deviations based on pair-wise comparisons. The next step involves the calculation of preference function evaluation and relative weight. then, we calculate global preference index which is used to calculate positive and negative outranking flows and in the final step, we calculate the net out ranking flow. Figure4 represent the ranking of the eight urban transport projects for a decision maker.



Rang	action	Phi	Phi+	Phi-
1	Project5	0,1842	0,4692	0,2850
2	Project1	0,1519	0,4684	0,3165
3	Project8	0,0805	0,4120	0,3316
4	Project6	0,0474	0,3970	0,3496
5	Project4	-0,0376	0,3895	0,4271
6	Project3	-0,0797	0,3286	0,4083
7	Project7	-0,1579	0,3496	0,5075
8	Project2	-0,1887	0,2541	0,4429

FIG. 4 – *Rankings of urban transport projects.*

We propose that we have 7 decision makers; when all urban transport projects will be ranked by the multi criteria analysis according to each decision maker, the negotiation method will be executed to negotiate and give the best urban transport project. The negotiation method used in our decisional model is Hare method which is a voting method. Hare method is a method that tries to achieve the majority for some option in rounds. In each round each decision maker has one vote given by Promethee II method. In the first round, if one transport project receives the majority of first places, then it is elected, otherwise the transport project with the fewest first places is eliminated, and the process is repeated until we get a majority. Figure 5 shows steps of negotiation by Hare method.

[1] **The preference function** characterizes the difference for a criterion between the evaluations obtained by two possible decisions into a preference degree ranging from 0 to 1. six basic preference functions have been proposed in (Brans & Mareschal , 2005)

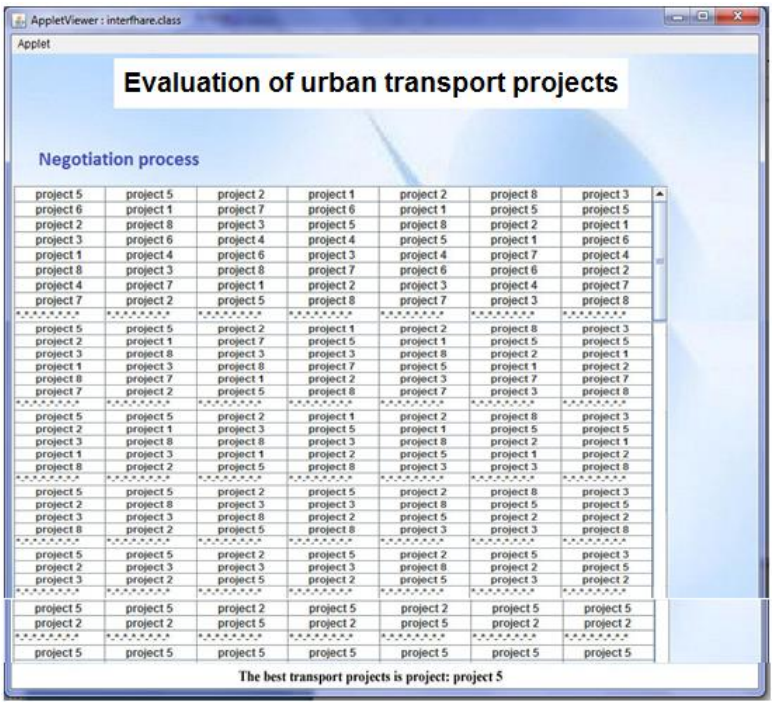


FIG. 5 – *Hare method steps*

4.2 Decisional system based Cloud

In this section, we will propose architecture to integrate our decisional system for evaluation of urban transport projects in private Cloud. Figure 6 illustrates our proposed architecture.

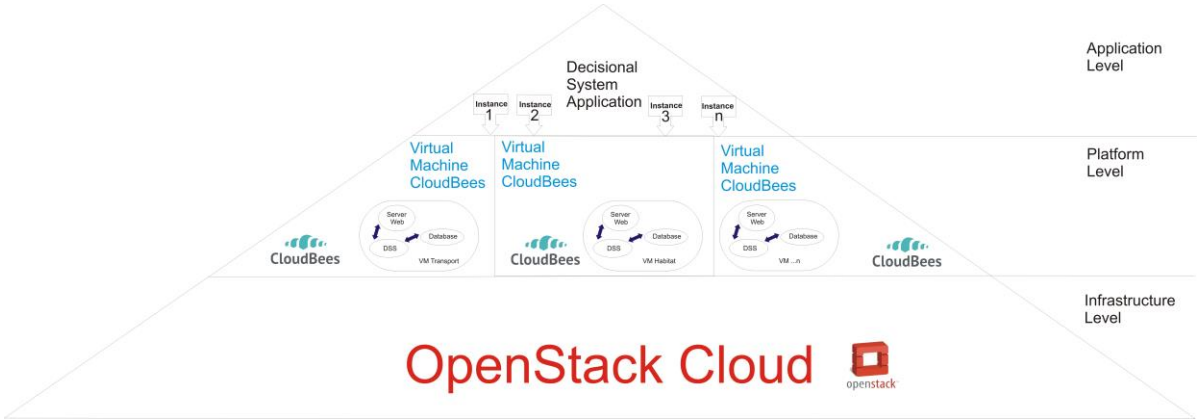


FIG. 6 – *Decisional system based Cloud.*

Our proposed Cloud architecture is composed of three levels:

Level 1: Infrastructure

Infrastructure level is is one of three main categories of cloud computing services, it includes the physical components that run applications and store data. So the infrastructure provides a layer of virtualized hardware that delivers the computing power and data centers required for applications to run, serving as a foundation for application and platform layer.

Several solutions of Cloud infrastructure are proposed, there exists proprietary solutions such as:

- Windows Blues (http://ww.windowsazure.com/),
- Amazon EC2 (http://aws.amazon.com/fr/ec2/),
- Dropbox (https://www.dropbox.com/)

They are mature clouds but not scalable since we do not have the opportunity to add additional modules. Moreover, these are commercial products that means that the data in most cases are not hosted in private, but at providers Cloud. So using these proprietary solutions there will be a security problem for people who have confidential data.

On the other hand, there exists different open source Cloud solutions, we quote: OpenNebula (http://archives.opennebula.org/documentation:rel4.4:intro) is an open-source management tool that offer a simple but feature-rich and flexible solution to build and manage enterprise clouds and virtualized data centers. Eucalyptus (https://www.eucalyptus.com/eucalyptus-cloud/iaas) allows easy installation of a cloud infrastructure. However, cloud computing aim is facilitate the development, but Eucalyptus does not allow this. This is why it is now abandoned. OpenStack (http://www.openstack.org/software/) is an open-source solution, flexible, scalable and mature. It creates and controls large group of virtual machines via a dashboard.

For our architecture Cloud we chose the OpenStack Cloud because it is a robust system allowing the creation of private Clouds and offers a development platform, so it allows developers to deploy a Platform Cloud in order to develop applications. OpenStack makes it possible to implement a virtual system of waiter and storage. Using its tools and components such as Nova, Swift, and Glance, we can create virtual machines or servers where each one of them corresponds to a domain of the city (transport, habitat, trade…). Figure 7 represents a virtual machine created by Openstack for transport domain.

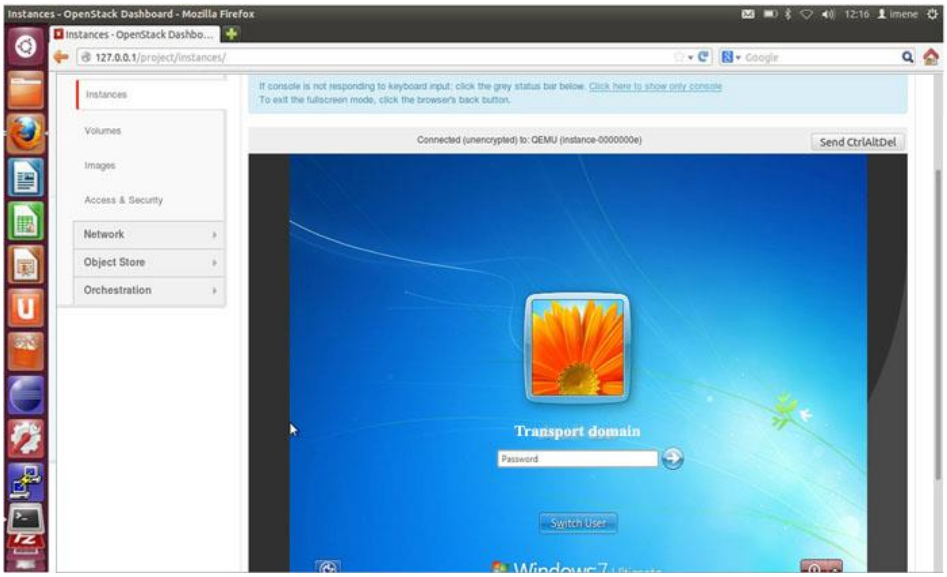


FIG. 7 – *Virtual machine by openstack for domain transport*.

Level 2: platform of deployment

It provides a platform and environment to allow developers to build applications and services over the internet. It allows them to create software applications using tools supplied by the provider.

There exist several solutions of platform Cloud that are suggested, we quote: Cloud Foundry (https://www.cloudfoundry.org/) is an open platform as a service; it allows offering a choice of Cloud, developing frameworks and application services. Cloud Foundry makes it faster and easier to build, test, deploy and scale applications. Cloudify (http://www.cloudify.cc/) is an open source solution for making his custom PaaS. This tool allows performing a set of tasks related to deploying and managing life cycle of an application on a cloud. Heroku (https://www.heroku.com/platform) is a cloud platform based on a managed container system, with integrated data services and a powerful ecosystem, for deploying and running modern apps. CloudBees (https://www.cloudbees.com/) provides an integrated and scalable platform based on standards for Java developers who want to develop and deploy web applications in a cloud computing environment.

For the second level of our architecture Cloud, we chose to use CloudBees as a platform of development because it can be deployed on OpenStack Cloud and integrate our decisional system which is implemented in Java language unlike other platform providers that are focused on programming languages like Ruby, PHP and Phyton. CloudBees allows us to create a custom platform, to deploy our application on a Cloud and manage it. On each virtual server created by OpenStack we deploy CloudBees, after that we use CloudBees deployed to create software virtual machine. So on the virtual machine of transport domain, we deploy CloudBees in which we deploy our decisional system. For each connection of a decision maker, CloudBees will allocate to him an instance of our decisional system.

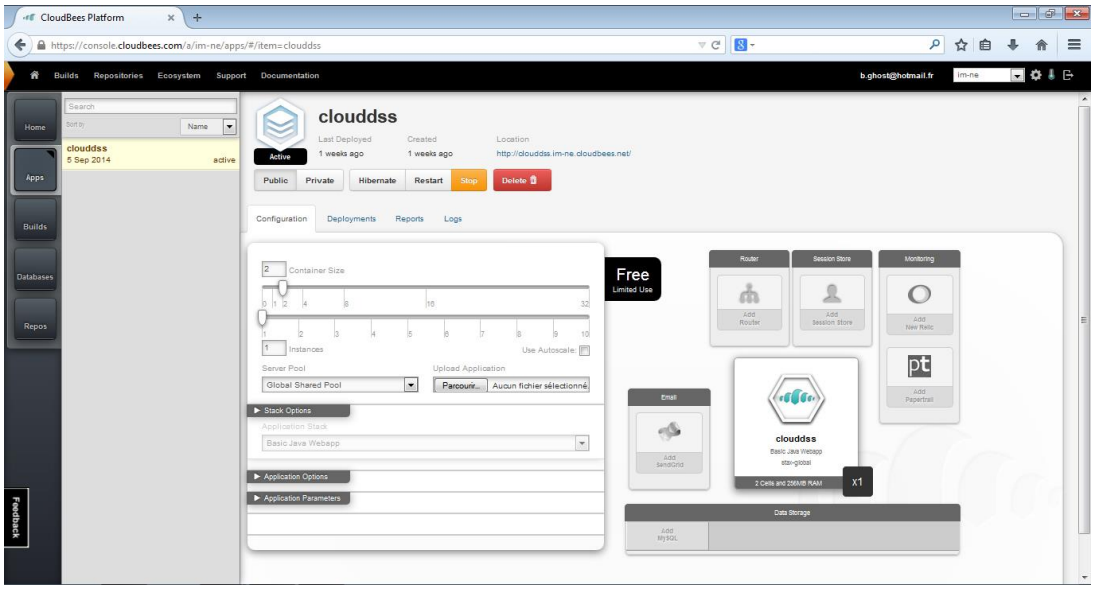


FIG. 8 – *Deployment of Decsional system on CloudBess.*

Level 3: application

It is the top most level of the Cloud architecture. It provides applications, programs, software and web tools to final user who can Access via any device. So , In order to evaluate and chose the best urban transport projects , decision maker can access the application (our decisional system) to participate in the decision-making process via computer, Smartphone, tablet or another device which is equipped with a web browser and an internet connection.

5 Conclusion

In the evaluation of urban projects process, it is important to have several alternatives, in which various factors are taken into account. In the past, the number of alternative was rather limited due to the difficulties in producing them. This is mainly due to the time-consuming procedures of creating scenarios as well as the evaluation that follows. Now, having integrating the decsional model on Cloud computing,the operation can be accomplished within a much shorter time frame . In this paper we proposed a decisional model based on a private Cloud. Our decisional model is based on multicriteria and negotiation method to help decision maker to chose the best urban transport project among several projects with contradictory alternatives.

References

Ahris. Y, Siti Zalina.A & Susilawati. S .(2009). Decision Support System for Urban Sustainability Planning in Malaysia. *Malaysian Journal of Environmental Management 10(1).*

Alshawi, M & Dawood, I, (2009). *Decision Support Systems (DSS) Model for the Housing Industry*. Paper presented at the Second International Conference on Developments in eSystems Engineering (DESE), Abu Dhabi.
Anonymous. Urban-planning. Retrieved January 13, 2015, 2015

Brans ,P. and Mareschal ,B.(2005). ROMETHEE methods," in Multiple Criteria Decision Analysis: State of the Art Surveys (J. Figueira, S. Greco,and M. Ehrgott, eds.), vol. 78 of International Series in Operations Research & Management Science , Springer.

Dur , Fatih, Yigitcanlar, Tan and Bunker, Jonathan M. (2009) A decision support system for sustainable urban development : the integrated land use and transportation indexing model. In: Proceedings for The Second Infrastructure Theme Postgraduate Conference 2009 - Rethinking Sustainable Development: Planning, Infrastrucutre Engineering, Design and Managing Urban Infrastructure. Queensland University of Technology, Brisbane, Queensland.

Fayech, B., Maouche, S., Hammadi, S., & Borne, P. (2002). Multi-agent decision-support system for an urban transportation network 2002 Proceedings of the 5th Biannual World Automation Congress (Vol. 14, pp. 27-32): IEEE.

Hasan, Mohamad K (2010). A Framework for Intelligent Decision Support System for Traffic Congestion Management System. Engineering, 2010, 2, 270-289
https://java.net/jira/secure/attachment/29265/CloudComputing.pdf

Jun, D., & Yikui, M. (2009). Intelligent Decision Support System for Road Network Planning. In Intelligent Information Technology Application Research Association Qi Luo, Technology Management Council (Ed.), ISECS International Colloquium on Computing, Communication, Control, and Management: CCCM 2009; Sanya, China, 8 - 9 August 2009; [including the 2009 IITA International Conference on Communication Systems, Networks and Applications (ICCSNA 2009) (Vol. 3): IEEE.

Longfei, W., & Hong, C. (2009). Coorperative Parking Negotiation and Guidance Based on Intelligent Agents International Conference on Computational Intelligence and Natural Computing (CINC '09). International Conference on (Vol. 2, pp. 76 - 79): IEEE.

Maktav, D., Jurgens, C., Siegmund, A., Sunar, F., Esbah, H., Kalkan, K.,Wolf, N. (2011). Multi-criteria Spatial Decision Support System for Valuation of Open Spaces for Urban Planning. In Aerospace and Electronic Systems Society Mustafa Ilarslan, Havacılık ve Uzay Teknolojileri Enstitüsu (Ed.), Proceedings of 5th International Conference on Recent Advances in Space Technologies (RAST 2011) (pp. 160- 163). Istanbul, Turkey: IEEE.

Mell, P, & Grance, T. (2011). The NIST definition of Cloud computing: Recommendations of the National Institute of Standards and Technology. Reports on Computer Systems Technology. http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

Moreira de Oliveira. T.H, Painho. M , Henriques. R.(2012). A spatial decision support system for the Portuguese public transportation sector. IWGS '12 Proceedings of the Third ACM SIGSPATIAL International Workshop on GeoStreaming. ACM

Muqing Liu. (2015). Urban Transport Project Prioritization Strategy in Developing Countries: A Scenario-Based Multi-Criteria Decision Analysis Perspective. Partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Graduate School of Arts and Sciences COLUMBIA UNIVERSITY.

Rhoton, J. (2011). Cloud computing explained (2nd Edition ed.). United Kingdom and United States: Recursive Press.

Schutte, I. C., & Brits, A. (2012). Prioritising transport infrastructure projects: towards a multi-criterion analysis. Southern African Business Review, 16(3), 97-117

Shi, Yongliang, & Li, Jingsheng. (2010). Improving the Decisional Context: New Integrated Decision Support System for Urban Traffic-related Environment Assessment and Control Proceedings of the 2010 International Conference on Mechanic Automation and Control Engineering (MACE) (pp. 1760 - 1763). Wuhan, China IEEE.

Sosinsky, B. (2011). Cloud computing bible. Indianapolis, IN: Wiley Publishing, Inc.
Sun Microsystems, Inc. (2009). Introduction to Cloud Computing architecture.

Yang, D & Mou, W. (1993). An Integrated Decision Support System in a Chinese chemical plant. Interfaces, 23(6), 93-100.

Yujing, Wang, Lin, Li, Hui, Zhu, & Zhihua, L.iu. (2011). Decision Model of Urban Infrastructure Projects Based on Comprehensive Evaluation *2nd IEEE International Conference on Emergency Management and Management Sciences (ICEMMS)* (pp. 895 - 898). Beijing, China: IEEE.

Résumé

La planification urbaine est la discipline de la planification sur le développement physique, social, économique et environnemental des régions métropolitaines, des municipalités et des quartiers. Elle consiste à développer l'utilisation des terres et les plans de construction ainsi que la construction locale et réglementations environnementales. Plusieurs problèmes dans le domaine de la planification de la ville exigent des solutions informatisées; l'un de ces problèmes est le problème du choix et l'évaluation des projets urbains. Dans cet article, nous avons proposé un système décisionnel basé sur le Cloud pour aider les différents décideurs de la ville à sélectionner le meilleur projet de transport urbain.
Mots clés. Système décisionnel, Cloud Computing, Virtualisation, projet de transport urbain.

Étude de la dynamique des phénomènes géographiques à travers une modélisation sémantique à base d'ontologie des objets spatio-temporels

Fethi Ghazouani*, Wassim Messaoudi*
Imed Riadh Farah*,**

*Laboratoire RIADI, ENSI, Université de Manouba, Tunisie
gfethi@yahoo.fr
wassim.messoudi@riadi.rnu.tn
**Département I.T.I, Telecom Bretagne, France
riadh.farah@ensi.rnu.tn

Résumé. Les images satellitaires se présentent comme un moyen efficace qui aide à l'étude des phénomènes spatio-temporels dans divers domaines de recherche, comme l'urbanisation, le suivi de l'environnement, l'étude écologique, etc. Ces données fournissent des informations multi-spectrales, multi-capteurs et multi-temporelles permettant une classification exacte de la couverture du sol. L'étude et la modélisation de ces données permettent une meilleure interprétation des phénomènes spatio-temporels et en conséquence de promouvoir une bonne gestion de l'occupation/l'utilisation des sols et d'améliorer les politiques de décisions sur divers processus de changement à savoir l'urbanisation et la déforestation. L'analyse de ces phénomènes spatio-temporels implique de connaître : (i) quels sont les éléments consécutifs qui les caractérisent c.à.d. identifier les différents types d'entités rencontrées, leurs propriétés géométriques et les attributs permettant de définir leurs sémantique, (ii) quelle est la répartition spatiale de ces éléments qui fait référence à la dimension spatiale des objets et (iii) à quel moment ces phénomènes surviennent qui fait naturellement référence à la dimension temporelle. En conséquence la modélisation de la dynamique des objets spatio-temporels doit tenir compte de ces trois composantes. C'est dans ce cadre que nous soulignons la difficulté de l'interprétation des phénomènes spatio-temporels et que nous proposons un modèle conceptuel basé sur les ontologies offrant la possibilité de comprendre la dynamique de ces phénomènes tout en tenant compte des dimensions spatiales, temporelles et sémantiques de ces objets. Ce cadre conceptuel pourra être utilisé comme un modèle décisionnel pour suivre les processus de changements qui peuvent se produire et pour analyser les risques.

Mots clés : objet spatio-temporel, dimension spatiale, dimension temporelle, dimension sémantique, objet dynamique, ontologies, BFO.

Étude de la dynamique phénomènes géographiques à travers une modélisation sémantique à base ontologie objets spatio-temporels

1 Introduction

L'analyse des phénomènes géographiques implique de connaître : (i) quels sont les éléments consécutifs qui les caractérisent "le Quoi", (ii) la répartition spatiale de ces éléments "le Où", et (iii) à quel moment ces phénomènes surviennent "le Quand" (Peuquet et Wentz (1994)). La composante "Quoi" consiste généralement à identifier les différents types d'entités rencontrées, leurs propriétés géométriques (i.e. leur forme) et les attributs permettant d'en calculer la sémantique. Ces entités peuvent être caractérisées par des frontières non clairement définies (entité dite fiat, par exemple une montagne) ou par des frontières connues (entité dite bona fide, par exemple un immeuble) (Smith (2001)). La composante "Où" (où le phénomène se produit) renvoi à des informations spatiales de localisation, données par exemple par la distance ou la position relative d'un objet par rapport à un autre (par exemple : inclusion, en dehors de). Cette composante est étroitement liée à la caractérisation des relations spatiales entre les entités : relations topologiques (Randell et al. (1992)), métriques (Pullar et Egenhofer (1988)) ou d'orientation (Frank (1996)). La troisième composante "Quand" fait naturellement référence à la dimension temporelle (Allen (1984)) dans laquelle s'inscrit le phénomène. L'étude peut être statique s'il s'agit d'un "arrêt sur image" limité à un instant ou une période en particulier, ou dynamique si elle prend en compte des changements se produisant sur un ensemble d'instants ou de périodes (Del Mondo et al. (2013)). A ces trois composantes, il faut ajouter la description des processus qui produisent ce (ces) phénomène(s) "le Comment" (Thériault et Claramunt (1999)). Cette dernière exprime la façon dont un phénomène se déroule, les différent processus et événements mis en jeu et leurs origines.

Dès lors, pour concevoir un cadre conceptuel d'information spatio-temporelle, il est nécessaire de considérer les trois dimensions décrivant les objets et les évènements : (1) la dimension spatiale qui décrit où le changement se produit, 2) la dimension temporelle qui décrit quand le changement se produit et finalement 3) la dimension sémantique décrit quoi se produit et/ou comment le changement se produit (Yuan (1999)). C'est dans ce cadre qui se situe notre problématique pour la modélisation de la dynamique des objets spatio-temporels pour le suivi de changement dans les images satellitaires.

Dans cet article, nous commençons par introduire brièvement quelques notions relatives aux objets spatio-temporels. Puis nous présentons les différentes approches proposées dans la littérature pour la modélisation de la dynamique des objets spatio-temporels. Nous verrons dans la suite en quoi l'utilisation des ontologies est importante pour la modélisation de dynamique des données spatio-temporelles. Nous présentons dans la suite une architecture multi-niveau basée sur les ontologies pour la modélisation de la dynamique des objets spatio-temporels dans les images satellitaires.

2 Notion de base sur les objets spatio-temporels

Une opinion partagée par les communautés scientifiques c'est qu'une entité géographique est un objet qui représente une abstraction d'un phénomène du monde réel avec une position de localisation sur la terre. Ainsi, une entité géographique est définie comme : l'ensemble composé par les dimensions spatiales, temporelles et sémantiques, dont différents types de relation peuvent être spécifiés (Pierkot et al. (2015)). Dans cette section nous présentons une brève description sur les différentes composantes définissant une entité géographique.

2.1 Dimension et relation spatiale

Généralement, la dimension spatiale d'une entité géographique est définie par une localisation et une forme géométrique. Cette dimension est étroitement liée à la caractérisation des relations spatiales entre les entités. Les relations spatiales peuvent être topologiques, de directions (d'orientations) ou de métriques (de distances). Ces relations spatiales peuvent être qualitatives (c'est-à-dire décrite avec des termes lexicaux) ou quantitatives (décrites avec des valeurs numériques). Hudelot et al. (2008) résume tous les types de relations spatiales dans un schéma d'ontologie commune (figure 1).

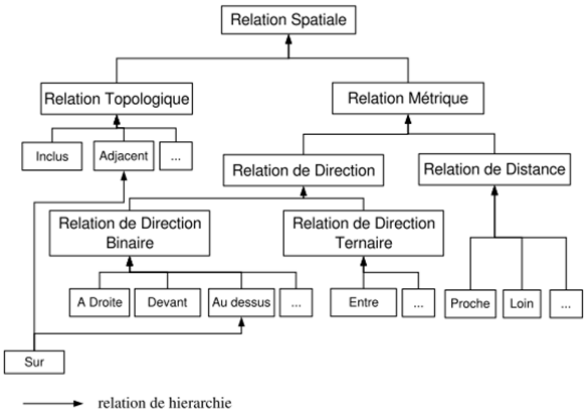


FIG. 1 – *Representation ontologique des relations spatiales Hudelot et al. (2008).*

2.2 Dimension et relation temporelle

Le temps est considéré comme des éléments discrets (à savoir une succession d'instants) dans lesquels des entités peuvent être observées. Quel que soit l'objet modélisé, le phénomène observé, il ne peut pas se produire une seconde fois exactement de la même manière (Hallot et Billen (2009)). La dimension temporelle est donc orientée. Par exemple, nous pouvons décrire l'évolution des entités le long des heurs, des jours, des semaines, des saisons, des années et ainsi de suite, dont chacun d'entre eux définit une granularité de temps (Del Mondo et al. (2013)). Les relations entre les entités dans le temps sont décrites par les relations temporelles. Comme les relations spatiales, ces relations temporelles peuvent être divisées en trois types : métrique, topologiques et structurelles. Allen (1984) a défini treize relations qualitatives possibles (figure 2) pour représenter les relations temporelles entre des intervalles de temps.

2.3 Dimension et relation sémantique

La dimension sémantique d'un objet a pour but de décrire les connaissances associées avec l'objet. En effet, elle permet de définir les caractéristiques de l'objet géographique à savoir son identité et ses propriétés. Cette dimension est définie par un ensemble de concepts qui

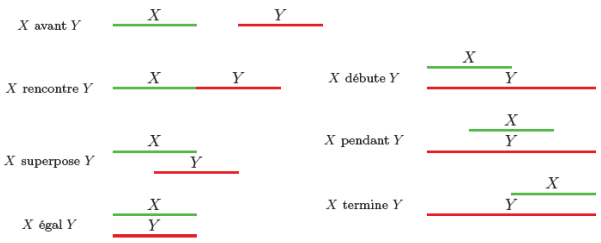


FIG. 2 – *Algèbre temporel d'Allen (Allen (1984))*

sont reliés au domaine étudié. Par exemple pour une image satellitaire, les concepts sont les propriétés physiques de l'image (bande spectrale, texture, etc.) ou la description du paysage (Pierkot et al. (2015)). Les relations sémantiques incluent toutes les autres associations qui peuvent exister entre les différents synonymes des concepts, tels que les relations « est-un (is-a) » et « partie-de ». Comme pour la dimension sémantique, la définition des relations sémantiques dépend du domaine et ces relations ne peuvent pas être explicitement spécifiques.

3 Travaux sur la modélisation des objets spatio-temporels

Nous présentons dans cette section quelques travaux pour la modélisation des objets spatio-temporels. En effet, Del Mondo et al. (2013) ont introduit une approche basée sur les graphes qui combine les concepts spatiaux, spatio-temporels et de filiation, pour représenter l'évolution des entités dans l'espace et le temps. Dans les travaux de Pierkot et al. (2015), les auteurs ont proposé un méta-modèle conceptuel de haut-niveau permettant de représenter les différentes dimensions (spatiale, temporelle et sémantique) de la connaissance spatio-temporelle. Le modèle est destiné à être utilisé comme un cadre conceptuel pour formaliser les connaissances spatio-temporelles dans le domaine de la télédétection.

L'ontologie est largement utilisée pour la modélisation des objets spatio-temporels. Les travaux existants proposent de combiner des présentations spatiales et temporelles pour modéliser les objets spatio-temporels et leurs relations (spatiale, temporelle et sémantique). Les représentations spatiales supportent plusieurs types de relations spatiales qualitatives (tels que nous l'avons présenté dans la figure 1). Les représentations temporelles incluent les approches ontologiques qui permettent d'ajouter la dimension temporelle à fin de représenter l'aspect dynamique. Parmi ces techniques nous citons : le graphe RDF spatio-temporel (Gutierrez et al. (2007)), la réification (Andronikos et al. (2009)), les relations n-aire (Hayes et Welty (2006)), versioning (Klein et Fensel (2001)) et l'approche 4D-fluent (Welty et Fikes (2006)). Batsakis et Petrakis (2010) ont présenté une approche pour la modélisation des aspects spatio-temporels des évènements et des objets qui évoluent dans le temps. Leur approche consiste à combiner la représentation 4D-fluent (pour la représentation temporelle) avec les relations spatiales directionnelles (pour la représentation spatiale). Dans un autre travail, Harbelot et al. (2013) ont proposé un modèle continuum spatio-temporel dans le but de représenter des entités géographiques dans le temps. Le modèle est une combinaison entre l'ontologie GeoSPARQL et l'ontologie 4D-Fluent. GeoSPARQL inclut une petite ontologie spatiale dans RDF/OWL per-

mettant une représentation des entités spatiales. L'ontologie 4D-Fluent permet d'associer des parties temporelles appelées timeslices à la représentation spatiale de l'objet.

Les objets et les phénomènes spatio-temporels géographiques peuvent être aussi modélisés avec ce qu'on appelle les ontologies fondamentales (appelées aussi ontologies de haut-niveau). Parmi ces ontologies nous citons, DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering) (Masolo et al. (2003)), BFO (Basic Formal Ontology) (Grenon et Smith (2004)), GFO (General Formal Ontology) (**?**) et d'autres. Ces ontologies offre une solution souhaitable pour l'analyse de ces types d'objets. En effet, dans ces ontologies, il y a une distinction fondamentale entre les entités statiques (les continuants ou les endurants) et les entités dynamiques (les occurrents ou les perdurants). Les continuants sont les objets qui persistent dans le temps. Ils incluent les objets physiques comme un arbre, un lac ou une rivière. Les occurrents sont les objets qui se produisent dans le temps tels que les évènements et les processus. Dans la littérature, l'approche BFO semble la plus répondue aux questions de la modélisation spatio-temporelle (Iwaniaka et al. (2013)), vue sa grande distinction des objets de la réalité. Un autre bénéfice de cette ontologie est qu'elle permet d'étendre la liste des processus et des évènements tout en ajoutant les régions temporelles et spatio-temporelles par rapport aux autres ontologies de haut-niveau.

Les travaux présentés précédemment souffrent de certaines limites. En effet, certaines approches ne mettent pas en considération toutes les dimensions associées à l'objet pour modéliser sa dynamique (Del Mondo et al. (2013)). Autre approches (Pierkot et al. (2015)) ne tiennent pas compte de la dynamique des objets dans l'étape de formalisation des connaissances. Dans les méthodes basées sur des combinaisons spatiales et temporelles, l'interrogation de l'information spatio-temporelle est compliquée par l'ajout d'un nouvel objet dans l'ontologie.

4 Architecture proposée pour la modélisation de la dynamique des objets spatio-temporels

Notre objectif consiste à modéliser les objets spatio-temporels à partir d'une série d'images satellitaires, tout en tenant compte de leurs dynamiques, dans le but d'étudier et de suivre les processus de changements qui peuvent se produire dans la couverture de sol. Nous proposons une architecture multi-niveau à base des ontologies (figure 3) pour la modélisation des objets spatio-temporelles dans une séries d'images satellitaires.

4.1 Les couches du modèle

Le cadre conceptuel de l'architecture proposée est constitué de trois couches : une couche fondement, une couche noyau et une couche de domaine (Ghazouani et al. (2015)). Nous présentons chaque couche dans les sections suivantes.

4.1.1 La couche fondement

L'ontologie BFO se présente efficace pour représenter les objets et les phénomènes spatio-temporels. Partant de cette hypothèse, notre focalisons consiste à mettre l'ontologie BFO dans un niveau supérieur dans l'architecture proposée à fin de l'utiliser comme un méta-modèle

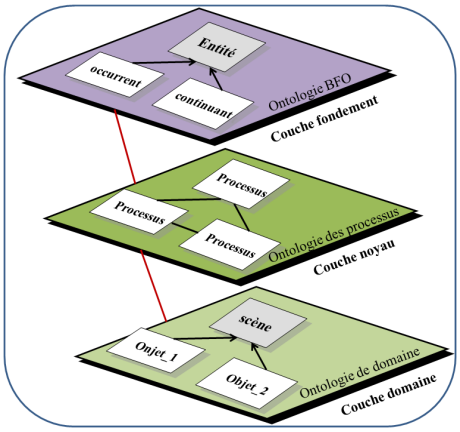


FIG. 3 – *Architecture générale de l'approche proposée.*

pour (i) raisonner sur les processus de changements et en conséquence la construction de l'ontologie noyau (couche noyau) et (ii) pour l'identification et la catégorisation des concepts, qui supportent ces processus, dans l'ontologie de domaine (la couche de domaine). L'ontologie BFO modélise les objets en deux ontologies : SNAP et SPAN. L'ontologie SNAP est conçue comme un modèle pour représenter la nature ontologique de continuants. Les continuant sont les entités qui durent pendant le temps c.à.d. les entités qui persistent identiquement même en subissant des changements de divers sortes (Grenon et Smith (2004)). L'ontologie SPAN comprend les régions spatio-temporellement étendues et les occurrents situés à ces régions. Les occurrent sont les entités qui se produisent dans le temps, ils incluent les processus, les évènements, les états et les changements (Grenon et Smith (2004)).

4.1.2 La couche noyau

Le rôle d'une ontologie spatio-temporelle est de capturer les processus non seulement dans une forme statistique, mais avant tout dans les relations "causes-et-effets" (Iwaniaka et al. (2013)). De ce fait, nous proposons de modéliser les processus géographiques et les relations qui les lient dans une ontologie que nous l'appelions "ontologie des processus". En effet, parmi les processus géographiques qui peuvent se produire sur le globe terrestre nous citons la déforestation, l'urbanisation, la dégradation des sols, etc. La déforestation est supposé être indiquée par la réduction de la couverture de la forêt. Cella est peut être suivi par une érosion continue des sols causée par le ruissellement des rivières. D'autre part, la déforestation est généralement suivie par un processus d'urbanisation ou par un processus de désertification. Le processus d'urbanisation est défini par la transformation d'une zone de forêt à une utilisation urbaine. La désertification est causée par un processus de dégradation continue qui entraine la détérioration temporaire ou permanente de la densité ou de la structure du couvert végétal. Dès lors, l'ontologie des processus est construite en suivant cet exemple de situation. Nous illustrons dans la figure 4 un exemple de modélisation ontologique des phénomènes géographiques et leurs relations.

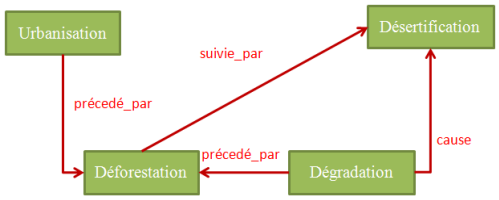


FIG. 4 – *Représentation ontologique des phénomènes géographiques.*

4.1.3 La couche de domaine

L'identification des objets et de leurs propriétés permet de connaitre qui est responsable de la dynamique des processus. Les continuants sont porteurs d'occurrents. Le changement de l'état donc d'un objet est une conséquence de déroulement des processus. C'est le rôle de l'ontologie de domaine d'identifier les objets ainsi que leurs propriétés. En effet, l'ontologie de domaine de télédétection permet une interprétation et une représentation sémantique des objets spatiaux à partir d'une scène d'image satellite. Des exemples de ces objets sont : zone de terrain, zone humide, rivière, zone urbaine, etc (Wassim et al. (2009)). Un exemple d'instance de cette ontologie est présenté dans la figure 5.

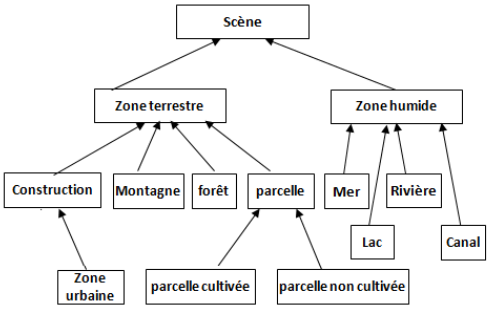


FIG. 5 – *Une instance de l'ontologie de domaine de télédétection*

Les images de télédétection fournissent des mesures et des observations, sur les objets acquis, qui peuvent être utilisées pour raisonner sur la dynamique des processus de changements. C'est en fonction de ces concepts que nous pouvons identifier les objets qui participent à un tel processus ou bien quel processus a changé l'état d'un objet. En fait, des indices comme les indices de végétation et les indices de variation d'eau sont souvent utilisés pour étudier et interpréter la dynamique des phénomènes spatio-temporels. Par exemple, l'indice NDFI (Normalized Difference Fraction Index, Indice Normalisé de Fraction de Différence) (Souza et al. (2005)) a été utilisé pour une détection avancée des dégâts sur la canopée de forêt causée par l'exploitation forestière sélective et par les incendies. Les valeurs élevées de NDFI indiquent la présence d'une forêt intacte, tandis qu'une forêt dégradée est obtenue pour une valeur réduite de cet indice. En conséquence, nous cherchons à enrichir l'ontologie de domaine de télédé-

tection par ces mesures et ces observations afin de l'adapter pour interpréter les phénomènes géographiques.

4.2 Les types de relations dans notre modèle

Dans le modèle proposé, des relations sont implémentées afin d'une part de représenter les relations entre les concepts (objets, processus, etc.) et d'autre part de relier les différents niveaux de l'approche proposée. Dans ce cas, deux types de relations peuvent être définis.

4.2.1 Relations intra-ontologiques

Une relation intra-ontologique est une relation qui peut exister entre les concepts d'une même ontologie. Par exemple les relations spatiales topologiques permettent de définir les relations spatiales entre les objets se trouvant dans la scène de l'image satellitaire. Si une rivière est située à côté d'une forêt, la relation entre ces deux concept sera définie par :

*Une forêt **est-à-côté** d'une rivière*

Des relations internes pouvant se trouver aussi entre les processus. Généralement ses relations entre les processus sont exprimées par des relations temporelles. Comme par exemple un processus peut causer, initier ou suivre un autre processus. Des exemples de telles relations sont les suivants :

*Une urbanisation **est-suivie** par une déforestation*

*Une dégradation **cause** une désertification*

Les relations sémantiques sont aussi utilisées pour lier les concepts d'une même ontologie. Les relations « est-un » (is-a) ou « partie-de » (part-of) sont maintenues pour exprimer les relations hiérarchiques entre les différents concepts. Dans les trois situations suivantes, le premier exemple présente la relation entre les deux concepts « zone humide » et « lac ». Nous avons aussi utilisé d'autres relations sémantiques pour lier les concepts à leurs propriétés. Exemples de telles relations sont : la relation « aPropriété », pour dire par exemple qu'une forêt a une propriété NDFI et la relation « aValeur » pour exprimer le fait que le concept NDFI a une qualité égale à une valeur donnée. Ces deux situations sont présentées par le deux derniers exemples.

*Un lac **est-une** une zone humide*

*Une forêt **aPropriété** NDFI*

*Le concept NDFI **aValeur** Val-1*

4.2.2 Relations inter-ontologiques (trans-ontologiques)

Les relations entre les concepts qui constituent deux ontologies distinctes sont appelées relations trans-ontologiques (Grenon et Smith (2004)). Si par exemple une forêt (concept de l'ontologie de domaine) est déboisée alors cet objet est considéré comme participant au processus de déforestation (concept dans l'ontologie de processus), dans ce cas une relation de

participation sera maintenue entre ces deux concepts et en conséquence les deux ontologies sont alignées. Cette relation est exprimée comme suit :

*Une déforestation **aParticipant** une forêt*

La relation sémantique « est-un (is-a) » est aussi considérée comme relation inter-ontologiques puisque nous l'avons utilisé pour connecter respectivement la couche noyau et la couche de domaine à la couche de fondement. Par exemple les processus géographiques seront reliés à la classe *Process* dans l'ontologie BFO. Cela permet de catégoriser ces concepts (les processus géographiques) en tant que processus. D'un autre côté, chaque processus géographique possède à un instant dans lequel il se produit et un intervalle de temps dans laquelle qui permet de définir sa début et sa terminaison. Deux concepts de base sont, dans l'ontologie BFO, sont utilisés pour catégoriser le temps : SpatioTemporalInstant et SpatioTemporalInterval. Dans ce cas, des relations temporelles à savoir « aTemps », « TempsDébut » ou « TempsFin » sont maintenues entre les processus et ces concepts à fin de présenter le déroulement de ces processus dans le temps. Ces exemples de situations sont présentés comme suit :

*La déforestation **est-un** un processus*

*La dégradation **a débuté** en été*

D'autres exemples de relations comme la création, la destruction, la détérioration, etc. peuvent être maintenus pour exprimer des relations, en particulier, entre les concepts de l'ontologie des processus (les processus) et ceux de domaine (les objets). Comme par exemples :

*Un processus de dégradation **détruit** une forêt*

*Un processus d'urbanisation **créé** une zone urbain*

Nous illustrons dans la figure 6 une instance de l'architecture proposée pour l'étude de la dynamique des phénomènes géographiques. Dans cette figure, les flèches en rouge représentent les relations intra-ontologiques et ceux en bleu représentent les relations trans-ontologiques.

4.3 Enrichissement axiomatique du modèle

L'enrichissement axiomatique permet de doter le modèle proposé par des mécanismes de raisonnement inférentiel. Il s'agit d'un ensemble de règles sémantiques qui sont essentiellement utilisées pour affirmer des nouveaux faits concernant les processus de changement. Nous présentons ci-après quelques exemples de règles ajoutées au modèle ainsi que leur prédicat :

R1 : Si une forêt f est située à côté d'une zone urbaine u, alors il y a une possibilité de dégradation.

$$urbaine(u) \wedge forêt(f) \wedge situéecoté(u, f) \rightarrow dégradation(d) \quad (1)$$

R2 : Si une parcelle x d'une forêt f est déboisée, alors la forêt f est un participant dans le processus de déforestation d

$$déboisée(parcelle(x) \wedge forêt(x, f)) \rightarrow déforestation(d) \wedge participant(x) \quad (2)$$

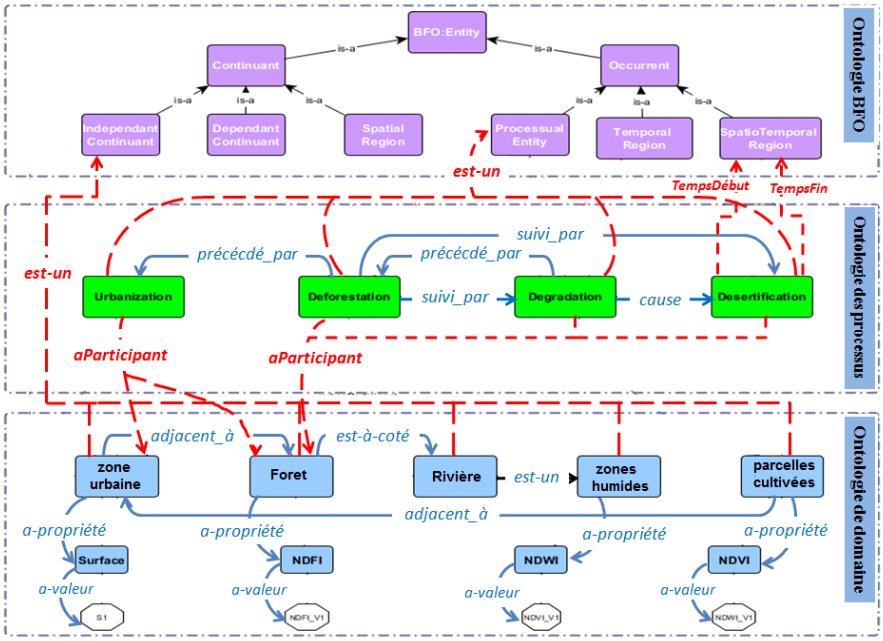


FIG. 6 – *Exemple d'une instance de l'architecture proposée*

R3 : Si à un temps t1 l'indice NDFI d'une forêt f pour une valeur V1 indique la production d'un processus de dégradation d, alors pour une valeur V2 supérieure à V1 à un temps t2 (t2>t1) le processus de désertification z a été produit.

$$NDFI(t1, V1) \wedge forêt(f) \wedge dégradation(d, f)$$
$$\rightarrow NDFI(t1, V2) \wedge désertification(z, f)(V2 > V1), (t2 > t1) \quad (3)$$

5 Conclusion

La modélisation de la dynamique des objets et des phénomènes spatio-temporels est une étape importante dans le processus d'aide à la décision. Elle permet aux décideurs de mieux analyser et interpréter le déroulement des ces processus géographiques et leur impact sur le globe terrestre. Cela va permettre par la suite le bon suivi et la planification efficace de ces phénomènes et en conséquence de promouvoir une bonne gestion de la surface de la terre et d'améliorer des décisions pour éviter certains risques. Ainsi, nous avons proposé dans cet article un modèle conceptuel multi-niveau basé sur les ontologies pour l'étude de la dynamique des phénomènes spatio-temporels à partir des images satellitaires. Le modèle proposé permet de modéliser les objets spatio-temporels tout en tenant compte des dimensions spatiales, temporelles et sémantiques associées à l'objet. En conséquence, des relations sont implémentées

entre les concepts de différents niveaux de l'architecture pour présenter la dynamique. Enfin, le modèle a été enrichi par un ensemble de règles pour raisonner sur les processus géographiques.

Comme un travail de futur, nous cherchons d'une première étape à un intégrer d'autre relations par l'importation d'autres ontologies tels que l'ontologie des relations et l'ontologie des relations temporelle. Dans une autre étape, l'enrichissement de notre modèle par d'autres règles sémantiques va permettre d'augmenter le mécanisme de raisonnement du modèle.

Références

Allen, J. F. (1984). Towards a general theory of action and time. *AI 23*(2), 123–154.

Andronikos, T., M. Stefanidakis, et I. Papadakis (2009). Adding temporal dimension to ontologies via owl reification. In *Panhellenic Conference on Informatics*, pp. 19–22.

Batsakis, S. et E. G. M. Petrakis (2010). SOWL : spatio-temporal representation, reasoning and querying over the semantic web. In *the 6th ICSS*, Graz, Austria, pp. 1–3.

Del Mondo, G., M. A. Rodríguez, C. Claramunt, L. Bravo, et R. Thibaud (2013). Modeling consistency of spatio-temporal graphs. *Data & Knowledge Engineering 84*, 59–80.

Frank, A. U. (1996). Qualitative spatial reasoning : Cardinal directions as an example. *International Journal of Geographical Information Science 10*(3), 269–290.

Ghazouani, F., W. Messaoudi, et I. R. Farah (2015). A multi-level ontological approach for change monitoring in remotely sensed imagery. In *the 7th International Joint Conference on KD, KE and KM*, pp. 435–440.

Grenon, P. et B. Smith (2004). Snap and span : Towards dynamic spatial ontology. *Spatial cognition and computation 4*(1), 69–104.

Gutierrez, C., C. Hurtado, et A. Vaisman (2007). Introducing time into rdf. *IEEE Transaction on Knowledge and Data Engineering 19*, 204–218.

Hallot, P. et R. Billen (2009). A pyramidal classification of st relationship models. In *3rd Workshop on Behaviour and Monitoring and Interpretation*.

Harbelot, B., H. Arenas, et C. Cruz (2013). A semantic model to query spatial- temporal data. In *the 6th International Workshop on IFGIS*, Petersburg, Russia.

Hayes, P. et C. Welty (2006). Defining n-ary relations on the semantic web.

Hudelot, C., J. Atif, et I. Bloch (2008). Fuzzy spatial relation ontology for image interpretation. *Fuzzy Sets and Systems 159*(15), 1929–1951.

Iwaniaka, A., J. Lukowicza, M. Strzeleckia, et I. Kaczmareka (2013). Ontology driven analysis of spatio-temporal phenomena, aimed at spatial planning and environmental forecasting. *Inter. Arch. of the Photogr., R. Sensing and Spat. Info. Sciences XL-7/W2*, 11–17.

Klein, M. et D. Fensel (2001). Ontology versioning for the semantic web. In *International Semantic Web Working Symposium (SWWS)*, California, USA, pp. 75–92.

Masolo, C., S. Borgo, A. Gangemi, N. Guarino, A. Oltramari, et L. Schneider (2003). The wonderweb library of foundational ontologies. wonderweb deliverable 18. Technical report.

Peuquet, D. et E. Wentz (1994). An approach for time-based analysis of spatiotemporal data. In *the 6th International Symposium of Spatial Data Handling, SDH'94*, pp. 489–504.

Pierkot, C., S. Andrés, J. F. Faure, et F. Seyler (2015). Formalizing spatiotemporal knowledge in remote sensing applications to improve image interpretation. *Journal of Spatial Information Science* (7), 77–98.

Pullar, D. V. et M. J. Egenhofer (1988). Towards the defaction and use of topological relations among spatial objects.

Randell, D. A., Z. Cui, et A. G. Cohn (1992). A spatial logic based on regions and connection. *KR 92*, 165–176.

Smith, B. (2001). Fiat objects. *Topoi 20*(2), 131–148.

Souza, C. M. J., D. A. Robertsb, et M. A. Cochrane (2005). Combining spectral and spatial information to map canopy damage from selective logging and forest fires. *Remote Sensing of Environment 98*, 329–343.

Thériault, M. et C. Claramunt (1999). La modélisation du temps et des processus dans les sig : Un moyen d'intégration pour la recherche interdisciplinaire. *Revue Internationale de Geomatique 9*, 67–103.

Wassim, M., F. I. Riadh, S. E. Karim, B. G. Henda, et S. Basel (2009). Spatio-temporal multimodality ontology for indexing and retrieving satellite images. In *COSI 2009 : colloque sur l'optimisation et les systèmes d'information*, Annaba, Algeria.

Welty, C. et R. Fikes (2006). A reusable ontology for uents in owl. In *Formal Ontology in Information Systems*, pp. 226–236.

Yuan, M. (1999). Use of a three-domain representation to enhance gis support for complex spatial-temporal queries. *Transactions in GIS 3*, 137–159.

Summary

Satellite images present an efficient way that help studying spatio-temporal phenomena in different research fields, such as urbanization, environment monitoring , ecological study, etc. These data provide multi-spectral, multi-sensors and multi-temporal information allowing an exact land-cover classification. The study and modeling of these data allow a better interpretation of spatio-temporal phenomena and consequently to promote effective management of land cover/use and to improve policy decisions on various change processes such as urbanization and deforestation. The analysis of spatio-temporal phenomena implies knowing: (i) what are the consecutive elements that characterize them, i.e. identify different types of encountered entities, their geometric properties and attributes which thus allow to define the semantics, (ii) what is the spatial distribution of these entities, that refers to the spatial dimension of objects and (iii) at what time these phenomena occur which naturally refers to the temporal dimension. Consequently, the dynamics modeling of spatio-temporal objects must take into account of these three components. In this context, we highlight the difficulty of interpreting spatial-temporal phenomena and we propose a conceptual model based on ontologies offering the ability to understand the dynamics of these phenomena taking account of the spatial dimensions, temporal and semantics of these objects. This framework can be used as a decision-making model to monitor change processes that may occur and to analyze the risks.
Keywords: spatio-temporal objet , spatial dimension, temporal dimension, semantic dimension, dynamic objet, ontologies, BFO.

Amélioration de l'Effectivité et du Temps de Réponse dans la Recherche d'Information

IMENE ZEMANI*,
LOUGMIRI ZEKRI**, MOHAMMED SENOUCI***

Université d'Oran1 Ahmed Benbella
Département D'informatique
BP 1524 El-M'naouer, Maraval, Oran 31000, Algérie

*imenezemani@yahoo.fr
**lougmiri@gmail.com
***msenouci@yahoo.fr

Résumé. De nos jours, la quantité d'information à travers le web est en accroissement de plus en plus grandissant. En effet, la recherche d'information (RI) s'étend sur différents axes. Par conséquent, les méthodes de la RI rencontrent des difficultés face à des documents de différents formats, taille et contenus. Les moteurs de recherche portent le défi de résoudre un nombre énorme des requêtes par minute, tout en respectant les critères du temps et de la qualité de réponse. Dans ce contexte, les algorithmes dits à arrêt au plutôt (early termination) sont une bonne solution pour répondre au premier critère mais reste plus ou moins faible vis-à-vis du deuxième critère. Une partie de ce papier est consacrée à la présentation des principaux algorithmes du domaine et met un accent sur l'algorithme WAND. Deux principales contributions de ce papier sont aussi présentées. La première consiste en la définition de certaines métriques fines lors du traitement des requêtes. Alors que la principale contribution est l'algorithme MWAND. Les expérimentations menées montrent que MWAND présente des gains meilleurs comparativement à WAND.
Mots clefs : Les algorithmes de réalisation anticipée, Top-K, WAND, MWAND.

1 Introduction

Les internautes aujourd'hui passent beaucoup de temps sur internet pour rechercher des informations ou des données, au même temps le volume d'information créé toutes les minutes devient de plus en plus important. Cette quantité d'information volumineuse rend la recherche d'information pertinente difficile à appréhender. Seuls les systèmes de recherche d'information (SRI) facilitent cette tâche de recherche en passant par un processus précis.

Ce processus se compose principalement d'un module d'indexation et un module d'évaluation. Le module d'indexation consiste à réaliser un index pour les documents et un procédé de médiation pour réécrire la requête afin qu'elle soit adaptable pour l'évaluation.

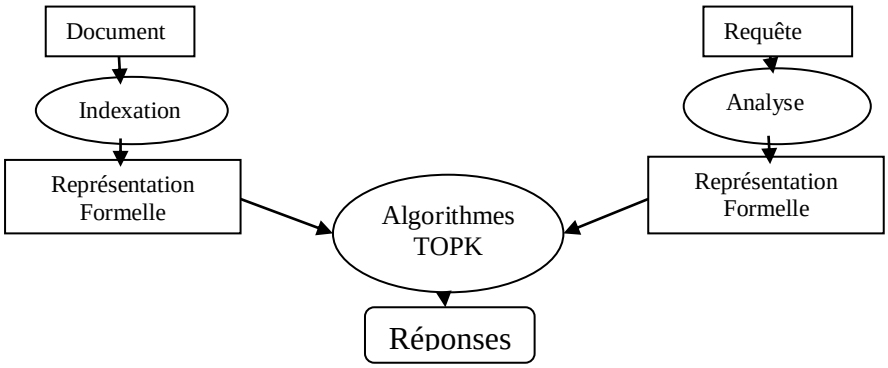


FIG. 1 – Processus de recherche d'information.

La figure 1 illustre globalement le processus de la recherche d'information derrière lequel se cachent les éléments de base :

- **La base des documents 'corpus de documents' :** est constituée d'un ensemble de documents notée ***Doc*** $_{id}$.
- **Requête :** c'est un ensemble de mots en langage naturel, qui permet à l'utilisateur d'exprimer son besoin.
- **Module de traitement de requêtes 'Analyse' :** transforme une requête d'un utilisateur en une requête simple en supprimant les mots vides.
- **L'index :** est un ensemble de structure de données. Il est toujours interrogé par les moteurs de recherche pour lui fournir un accès rapide aux informations. (nous discuterons de ce point dans les paragraphes suivants).

Les modèles de recherche d'information sont globalement de trois types. On cite alors le modèle booléen, où il s'agit de répondre aux requêtes conjonctives ou aux requêtes disjonctives. Le modèle vectoriel représente les documents et les requêtes sous forme de vecteurs. L'évaluation de requêtes dans ce modèle se base sur le calcul du TF-Idf, ou sur une variante de celui-ci. Le modèle probabiliste tire preuve de la théorie des probabilités et principalement sur la probabilité conditionnelle de Bayes. Il est à noter que l'évaluation est principalement approximative, où on admet que la réponse est avec une grande probabilité pertinente pour l'utilisateur.

Les algorithmes à arrêt anticipé sont souvent utilisés puisqu'il n'est pas judicieux de balayer les listes inversées vu la taille immense de ces listes. WAND est un algorithme anticipé et a pris beaucoup d'ampleur (Matthias et al, 2013). Il combine le modèle booléen avec le modèle probabiliste. Ayant remarqué que WAND est trop approximatif, ie pouvant présenter une faiblesse en termes de précision dans ses réponses. Dans ce papier, nous proposons une nouvelle version de l'algorithme WAND. L'objectif est de présenter une solution anticipée et précise. Pour ce faire, nous définissons des nouvelles métriques, plus fines. Nous montrerons que notre version est meilleure que WAND en termes des métriques déjà connues, comme la précision et les nouvelles métriques.

Le reste de ce papier se présente comme suit. La section 2 présente les travaux liés. Nous balayons le domaine de la recherche d'information. Nous présentons l'algorithme WAND en

détail. Nous survolons aussi quelques algorithmes du domaine. La section 3 présente MWAND, pour Modified WAND. La section 4 présente les expérimentations. Nous concluons ce papier dans la section 5 où nous parlons aussi des travaux futurs.

2 Travaux Liés

2.1 Stratégie d'évaluation

Turtle et Flood(Turtle et al,1995) ont classé les stratégies d'évaluation en deux classes. La stratégie TAAT (Term At A time) , cette classe permet de traiter les termes de la requête un par un. La stratégie DAAT (Document At A time) dans cette classe, chaque document est évalué complètement avant de passer au document suivant.

2.2 La Structure d'Index (Zobel et al, 2006)

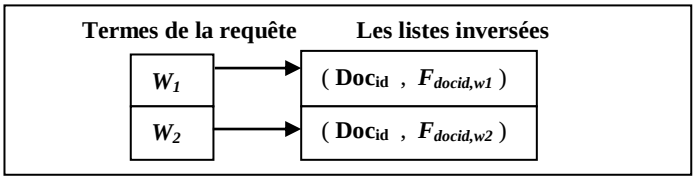


FIG. 2 – *Exemple d'un index inversé.*

L'indexation (appelée analyse pour une requête) est une étape très importante dans le processus de la RI. Elle représente les différentes collections des documents sous forme d'une représentation simple. Elle consiste à extraire pour chaque terme, de la collection, ses caractéristiques dans le document courant. Il existe plusieurs types d'index ; mais les structures les plus efficaces sont les index inversés et les index compressés :

- **L'index inversé** se compose de plusieurs listes inversées L_w**.** Chaque liste est attachée à un terme W_i et est représentée par au moins deux entrées :
  - **L'identifiant** Doc_{id}.cet identifiant est représenté par un nombre ordinal.
  - **La fréquence d'apparition** $F_{(docid,w)}$**.** Le nombre d'occurrences du terme W_i dans le document Doc_{id}**.**

En réalité d'autres paramètres peuvent être considérés ; tels que le lieu d'apparition du terme dans les documents (titre, résumé, corps,...), les positions, ainsi que le caractère d'écriture. La figure 2 illustre la manière d'utilisation des index inversés.

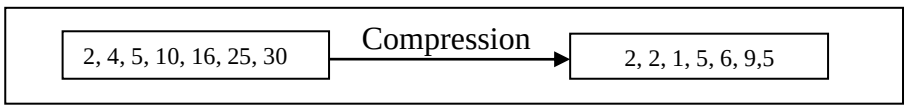


FIG. 3 – *Exemple de compression.*

- **L'index compressé** est une structure plus facile qui réduit le volume des données stockées dans les listes inversées. Son objectif est de compresser une séquence de Doc_{id}. La figure3 illustre un exemple de compression.

Une forme de compression est faite par le calcul des différences entre chaque Doc_{id} et Doc_{id-1} .NewPForDelta (New PFD). (Hao et al, 2009) est un exemple de méthode de compression.

2.3 Les algorithmes de réalisation anticipée

La technique de réalisation anticipée[1] accélère la recherche d'information et retourne l'ensemble des top-k sans passer par l'évaluation de tous les documents existants. Cette technique peut être réalisée dans les cas suivants :

- **L'arrêt tôt :** Dans ce cas, les documents qui ont une plus grande probabilité d'être parmi les ***TOP-K*** sont classés en premier. Par conséquent, on peut s'arrêter lorsqu'on aura les ***K*** meilleurs documents .On peut trouver dans cette catégorie les algorithmes FA (Fagin et al ,1999), TA et NRA (Fagin et al ,2001).
- **Skip dans les listes:**(Storhman et al ,2001) cette technique est appliquée au cas où les listes sont triées par le Doc_{id}**.** Son principe est de sauter plusieurs documents qui répondent à un certain critère. (WAND, BMW, LBMW) (Shan et al, 2012)

2.3.1 Le processus d'évaluation en deux niveaux WAND

WAND (pour Weak AND) est un système logique qui évalue les requêtes en deux niveaux. Le premier niveau est consacré à la recherche de l'ensemble des documents candidats, alors que le deuxième niveau vérifie et valide la pertinence du document. L'avantage de cet algorithme est dans l'évaluation préliminaire, où il est possible de sauter plusieurs documents sans les évaluer (Border ,2003). La figure 4 illustre les étapes de l'algorithme WAND.

Déroulement de l'algorithme WAND. L'algorithme WAND consiste à trier les termes de la requête selon l'ordre croissant des identificateurs du premier document de chaque liste ; ensuite, il cherche le terme pivot. Ce dernier est le premier terme pour lequel la somme de sa borne supérieure et les bornes supérieures des termes qui le précèdent est supérieure ou égale au seuil. Une fois que le document pivot est identifié, l'algorithme calcule le score approximatif (SA). Si ce score est inférieur au seuil alors il cherche un nouveau document pivot, sinon il passe à l'évaluation exacte. Celle-ci nécessite le calcul du score exact (SE). Si SE est supérieur ou égal au seuil alors ce document sera inséré dans la liste des TOPK [2] et supprimé des listes d'affectation.

[1] **" Early Termination" ou " l'arrêt au plus tôt " ou encore " réalisation anticipée":** Son objectif est de s'arrêter aussi rapidement que possible, lorsque certains critères sont vérifiés.
[2] **"La liste TOPK"** l'ensemble des **K** meilleurs documents retournés par l'algorithme.

1. Initialiser le seuil=0 ;
2. Calcul des Uppers Bounds des termes **;**
3. Répéter {
1ère Phase (Recherche des documents candidats)
a. Choix du document pivot ;
b. Calcul du score approximatif (SA) ;
c. Si : (SA < seuil) Alors retour à (a) ;
d. Sinon {
2ème Phase (Vérification et validation)
a. Calcul du score exact (SE)
b. Si : (SE < seuil) Alors retour à (a) ;
c. Si : (SE ≥ seuil) Alors {
a. Ajouter ce documents à la liste des TOP-K ;
b. Mise à jour du seuil.
}
}
Jusqu'à condition d'arrêt ;
}

FIG. 4 – *Déroulement du WAND.*

La figure 5 (Tonellotto, 2010) illustre la manière de sélection des documents :

- Le tri des listes d'affectation selon le premier Doc_{id} de chaque liste (11, 11, 22,23),
- la somme des bornes supérieures est supérieure ou égale au seuil (2+1+4 ≥ 6 ➔le terme t3 est le terme pivot ; Le document 22 est le document pivot).

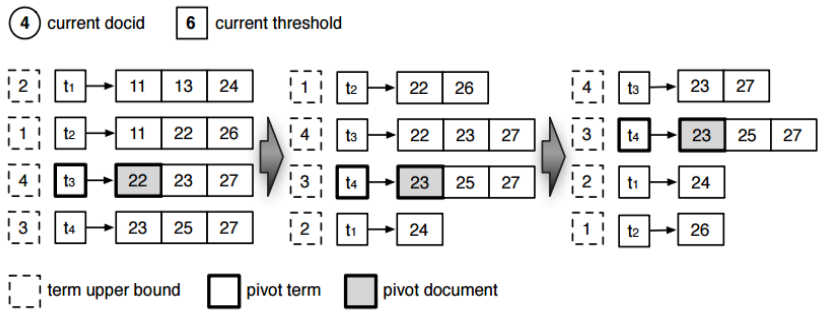


FIG. 5 – *Exemple de sélection du document pivot.*

Mise à jour du seuil. Le seuil est égal à zéro tant que la liste TOPK n'est pas pleine. Si la liste est pleine le seuil est égal au score minimal de la liste.

L'Insertion dans la liste TOPK. Pour insérer un nouveau document dans la liste TOPK, il faut que son score soit supérieur au score minimal de cette liste pour le remplacer. La figure 6 suivante illustre la manière d'insertion d'un document dans la liste TOPK.

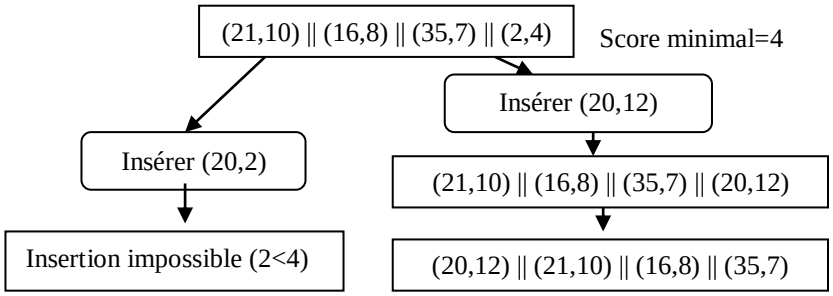


FIG. 6 – *Exemple d'insertion dans TOPK.*

Discussion. De par sa logique d'anticipation, WAND est rapide en termes de temps de réponse, toutefois, nous relevons certaines imperfections palpables en l'exécutant :

- l'algorithme souffre du problème d'effectivité puisqu'il retourne un ensemble de documents dans quelques fractions de seconde mais qui ne sont pas forcément des documents pertinents pour l'utilisateur,
- l'algorithme souffre de la perte de documents pertinents dans le passage entre les itérations. Par exemple, pour insérer un document dans la liste des TOPK, on supprime tous les documents avec un Doc_{id} inférieur à celui du document,
- Pour deux documents pertinents ayant le même score exact, WAND saute généralement celui ayant le plus petit identificateur. Pour avoir ce deuxième document pertinent, il faut élever la valeur de K, autrement, au lieu de demander un TOP-3, il faut demander un Top-7. Ce qui augmentera le temps de calcul,
- L'évaluation approximative inutile de plusieurs documents.

La logique d'anticipation proposée par WAND a été très attirante et a inspiré beaucoup de chercheurs. (Tonelloto et al ,2010) propose un algorithme à deux passes et intègre le calcul de la proximité entre les termes de la requête afin de retourner des documents plus pertinents. Les auteurs proposent de sauvegarder des couples de termes dans l'index. Bien que cette méthode soit performante en termes de pertinence elle reste non économique en termes d'espace de sauvegarde. (Ding et al, 2011) propose une structure d'index appelée **BMI** (pour **Bloc Max Index**) sur laquelle s'exécute WAND pour engendrer en fin l'algorithme **BMW**. Cette structure est un index compressé dont les listes L_w sont divisées en blocs contenant 64 ou 128 Doc_{id}, et chaque bloc peut être compressé ou décompressé séparément. Une table supplémentaire est ajoutée pour indiquer la contribution maximale de chaque bloc de l'index ainsi que les bornes de chaque bloc. Après cette modification dans la structure de l'index, la prochaine étape consiste à adapter WAND pour rechercher l'ensemble des TOP-K documents.

Les auteurs de (Dongdong Shan et al 2012) proposent un ensemble de méthodes qui s'exécutent sur BMI et qui combinent le score de WAND ou de MaxScore(Turtle et al,1995) avec le score de BM25(Robertson et al, 1998). Ces combinaisons ont généré les algorithmes Local Block-Max WAND (LBMW) et Local Block-Max MaxScore (LBMM). Les précédents systèmes sont tous centralisés, où une entité centrale exécute l'algorithme. Une telle architecture pose des problèmes de passages à l'échelle alors qu'une distribution de

WAND serait moins pénalisée. (Rojas et al, 2013) a pris l'initiative de réaliser une version à deux niveaux pour réduire la communication inter-processeurs et le coût de l'exécution.

2.4 Evaluation des performances

D'une façon générale, tout système de recherche d'information présente deux objectifs. Le premier est de retrouver tous les documents pertinents, et le deuxième est de rejeter tous les documents non pertinents. Ces deux objectifs sont évalués par les mesures de précision et de rappel définis ci-dessous. La figure 7 illustre sur un schéma ces définitions.

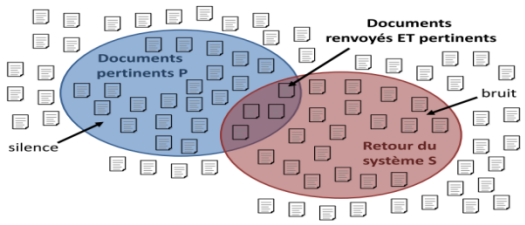


FIG. 7 – *Mesure d'évaluation d'un SRI.*

Les taux de précision et de rappel sont donnés par les formulations suivantes :

Rappel $= \frac{|P \cap S|}{|P|}$; Silence = 1-rappel ; Précision $= \frac{|P \cap S|}{|S|}$; *B*ruit = 1- précision

3 Propositions & Méthodologie

Pour faire face aux points de faiblesse de l'algorithme WAND, nous proposons comme première contribution un nouvel algorithme **MWAND (Modified WAND).** Celui-ci représente des résultats plus performants et plus effectifs. Outre les métriques bien connues, nous définissons de nouvelles métriques plus fines. Ces métriques serviront comme base de comparaison de ces algorithmes.

3.1 MWAND

MWAND se base sur l'algorithme WAND, mais porte plusieurs modifications pour améliorer le résultat. Il reçoit en entrée les listes d'affectation des termes de la requête. Dans la première phase, **MWAND** applique l'intersection entre les listes d'affectation pour extraire les documents pertinents. Cette étape permet ainsi d'éviter les tests inutiles qu'exécute WAND pour le remplissage des structures de données temporaires. La deuxième phase consiste en le remplissage de la liste des TOPK sans passer par l'évaluation approximative des documents. Nous lançons WAND sur l'ensemble des documents restants.

3.1.1 Le fonctionnement

Les fonctions principales dans le pseudo code **MWAND** sont les suivantes :

La fonction intersection. Prend en entrée les listes d'affectation des termes de la requête, et retourne une liste des documents pertinents, le nombre de documents retournés K' et les

nouvelles listes. Cette fonction consiste à pointer sur le premier document de la première liste et vérifie son existence dans les autres listes tout en avançant le pointeur de cette liste. Si un document figure dans toutes les listes d'affectation qui sont du nombre de termes de la requête, la fonction va le considérer comme un document pertinent et l'insère dans la liste des TOPK. La figure 8 illustre le fonctionnement de cette fonction.

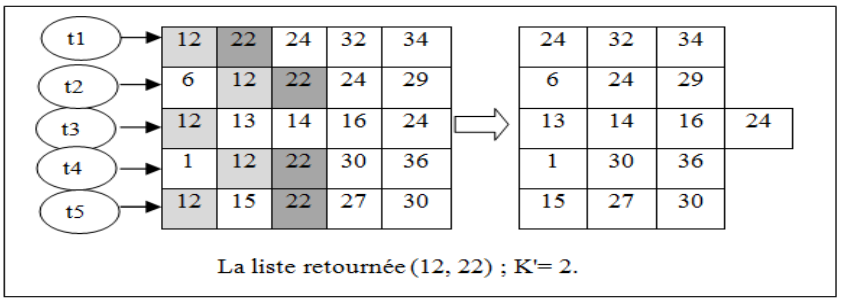


FIG. 8 – *Le déroulement de la fonction intersection.*

La fonction remplissage. Donne un avantage à l'algorithme MWAND par rapport au remplissage de liste TOPK et l'accélération du temps de réponse. Cette fonction prend en entrée les listes retournées par la fonction qui précède et la taille de la liste TOPK (K=K-K'). D'abord, les listes d'affectation seront triées selon $Curdoc_{id}$[3] en ordre croissant, pour retirer le $Curdoc_{id}$ maximal. Après une liste est remplie par tous les documents avec $Doc_{id} \leq Curdoc_{id}$ maximal, et triée en ordre croissant. La fonction permet d'insérer les K premiers éléments de cette liste dans la liste des TOPK. Puis on calcule le score exact de chaque document de cette liste. Cette fonction retourne, le seuil qui est représenté par le score minimal de la liste TOPK, les nouvelles listes d'affectation et la liste TopK. La figure 9 donne le procédé de remplissage de cette fonction.

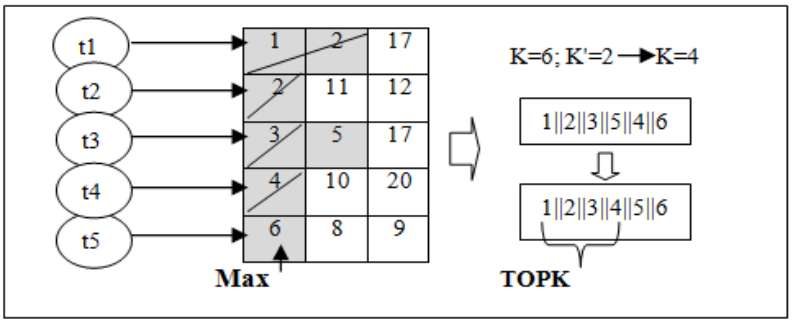


FIG.9 – *Le déroulement de la fonction remplissage.*

[3] "$Curdoc_{id}$": *le* doc_{id} *du premier élément de la liste.*

3.2 MWAND

Après l'exécution de l'algorithme WAND, nous avons constaté que l'algorithme souffre de la perte des documents pertinents. De là, nous avons eu l'idée d'ajouter des métriques fines:

- **La perte de suppression:** mesure la faiblesse de l'algorithme lorsqu'il supprime des documents pertinents sans les évaluer lors du passage entre les itérations.
- **La perte de l'arrêt tôt :** mesure la faiblesse de l'algorithme lorsqu'il arrête son exécution sans évaluer un ensemble de documents qui peut contenir des documents pertinents.
- **La fidélité :** montre la fortune de l'algorithme lorsqu'il classe les documents retournés par le même classement des documents pertinents.

Max. la valeur maximale de la liste TOPK
A. Nombre de documents qui ont le même classement ;
B. Nombre de documents du $\{P - (P \cap S)\}$ avec $Doc_{id} \leq$ Max
C. Nombre de documents du $\{P - (P \cap S)\}$ avec $Doc_{id} >$ Max

$$\textbf{Perte de Suppression} = \frac{B}{S} \;\; ; \textbf{ Perte l'Arrêt Tôt} = \frac{C}{S} \;\; ; \textbf{ Fidélité} = \frac{A}{S}$$

$$\textbf{Fidélité} \leq \textbf{Précision ;}$$

$$\textbf{Précision + Perte de Suppression+ Perte l'Arrêt Tôt=1}$$

Notre Objectif est d'augmenter la précision et la fidélité et diminuer le taux de perte.

4 Expérimentation

Pour mener nos expérimentations, nous avons mis au point les méthodes WAND et MWAND sur une machine (Processeur Intel(R) Core(TM) i7-4500 CPU@ 1,80 GHZ), dotée d'une capacité mémoire de 8 GB RAM sous Windows 8. L'application est développée en utilisant le langage de programmation JAVA sous l'environnement NetBeans. Pour évaluer la qualité des deux algorithmes, nous avons effectué nos expérimentations sur les documents du corpus Reuters-21578. Afin de tester les performances des différentes méthodes, nous avons lancé 550 000 requêtes test de longueurs 5 termes.

Nous avons lancé toutes ces requêtes sur le système. Pour chaque requête, nous avons enregistré les listes des documents leurs répondant. Un document d_i est plus pertinent qu'un document d_j pour une requête q s'il partage plus de terme avec elle. Si deux documents partagent le même nombre de termes avec q alors nous les classons selon la somme des fréquences de ces termes. Par la suite, nous avons lancé les deux algorithmes et nous avons collecté les résultats selon les métriques définies et sur le temps d'exécution en millisecondes.

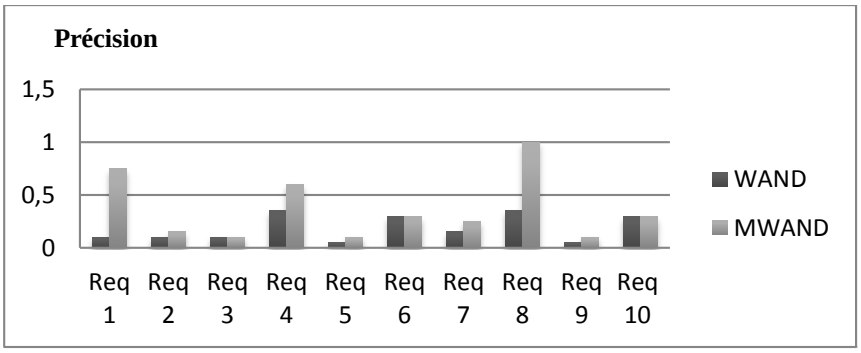


FIG.10 – *Comparaison entre WAND et MWAND en termes de précision*.

La figure 10 donne une comparaison entre WAND et MWAND en termes de précision. Cette figure montre d'une manière globale que MWAND présente plus de performance que WAND. Ce constat se justifie par la manière de sélection des documents et principalement dans la première phase où les documents partagés par les différentes listes sont récupérés.

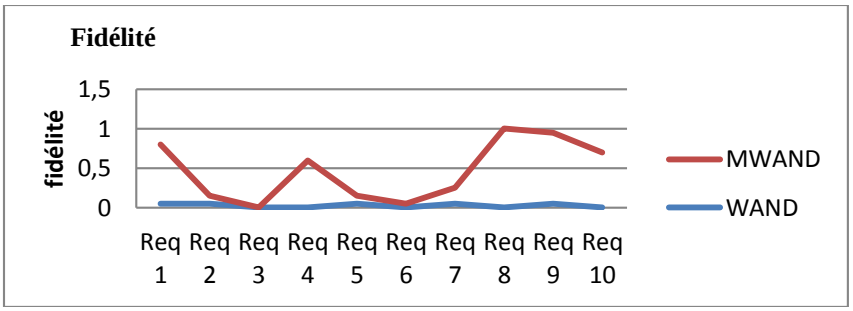


FIG.11 – *Comparaison entre WAND et MWAND en termes de fidélité.*

L'objectif de la deuxième expérimentation est de tester sa fidélité au système réel. L'objectif alors est de maximiser cette valeur. La figure 11 montre la comparaison entre ces deux algorithmes sur cette métrique. Nous remarquons d'une manière claire, MWAND a été plus fidèle que WAND vu que cette valeur a été maximale dans son cas. Nous revenons sur le fait que nous n'exécutons pas des sauts de listes d'une manière abusive. La fidélité est aussi maximisée suite à la considération des documents partagés, puisqu'ils sont sauvegardés assez tôt (dans la première phase).

En réalité, les métriques définies se complètent et si l'une s'élève l'autre s'abaisse. La fidélité et la précision sont liées. D'ailleurs, la fidélité est plus fine que la précision alors que la perte est une métrique qui justifie les autres métriques. Plus elle est minimale plus les deux autres s'élèvent. La figure 12 montre bien la comparaison entre MWAND et WAND en terme de perte et justifie les deux figures précédentes. Nous remarquons d'une manière globale que WAND a perdu plus d'information que MWAND. Cette perte se justifie par son caractère trop dynamique par rapport à MWAND. Les sauts de listes qu'il effectue sont très rapides, il parait qu'il privilégie plus le temps d'exécution et pertinence que d'être précis. En réalité, WAND considère qu'un document est pertinent s'il partage un terme avec la requête et que son TF-Idf est maximal au détriment du reste des termes. Par contre, MWAND est

plus attentif et tâche de récupérer un maximum de document partageant des termes avec la requête.

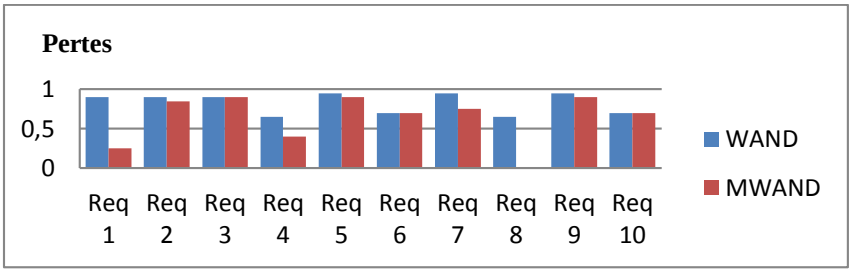


FIG.12 – *Comparaison entre WAND et MWAND en termes de perte.*

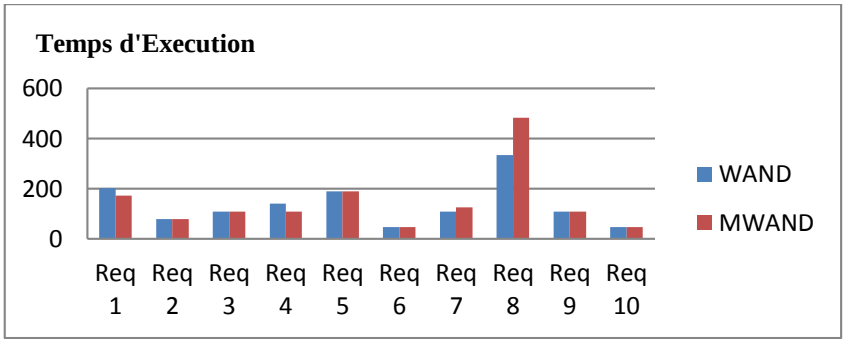


FIG.13 – *Comparaison entre WAND et MWAND en termes du temps d'exécution.*

La figure 13 présente la comparaison entre MWAND et WAND en termes du temps effectué pour localiser une réponse. D'une manière générale, le temps effectué par MWAND est plus grand que celui consommé par WAND. La justification est simple. Dans la théorie de l'anticipation, comme TA de (Fagin et al, 2001), l'algorithme s'arrête lorsque K objets sont découverts. WAND a hérité de cette théorie et s'arrête lorsqu'il remplit K document dans sa structure réservée pour ce but, au détriment de la précision, alors que MWAND est plus attentif.

5 Conclusion et Perspectives

La recherche d'information est un domaine passionnant. Il appelle des techniques intelligentes et pointues afin de répondre au mieux aux requêtes utilisateurs. L'objectif devient donc lors de la réalisation d'un système d'information est de maximiser la probabilité qu'un document satisfasse un utilisateur, car en réalité, il s'agit de la manipulation d'une information floue.

Dans ce papier, nous avons présenté un nouvel algorithme dont l'objectif est de répondre au mieux à la question posée. Avant de présenter notre approche, nous avons présenté des éléments importants liés à la terminologie utilisée dans ce domaine. Par la suite, nous avons

présenté l'algorithme WAND. Celui-ci a inspiré beaucoup de mondes suite à sa logique d'anticipation. Nous avons présenté un état de l'art des algorithmes tournant, adoptant et étendant cet algorithme. Pour ce qui est de notre, part, nous avons proposé MWAND suite à certaines imperfections du précédent algorithme et nous avons définies des métriques fines qui permettent de toucher à la qualité des réponses retournés. Les expérimentations ont montré que MWAND est compétitif.

Références

Broder, A.Z., Carmel, D., Herscovici, M., Soffer, A. and Zien, J. Efficient query evaluation using a two-level retrieval process (2003). CIKM'03 (New York, NY, USA, 426-434).

Ding.S, T. Suel. Faster Top-k Document Retrieval Using Block-Max Indexes (2011). SIGIR'11, July.

Fagin.R Combining fuzzy information from multiple systems (1999) . Journal of Computer and System Sciences 58, 83-99.

Fagin.R, A. Lotem, and M. Naor. Optimal aggregation algorithms for middleware (2001). In Symposium on Principles of Database Systems.

Hao.Y, Shuai.D, Torsten.S. Inverted index compression and query processing with optimized document ordering (2009). In Proceedings of the 18th International Conference on WWW.

Matthias.P , J. Shane Culpepper ,Moffat.A. ExploringThe Magic Of Wand (2013). ADCS '13, December 05 - 06 2013, Brisbane, QLD, Australia

Oscar, R. Gil-Costa, V. Mauricio. M. Distributing efficiently the Block-Max WAND algorithm (2013). International Conference on Computational Science, ICCS. Barcelona, Spain.

Robertson, S.E., Walker, S. and Hancock-Beaulieu, M. 1998. Okapi at TREC-7: Automatic Ad Hoc, Filtering, VLC and Interactive. TREC (1998), 199-210.

Shan.D, S.Ding, J.He, H.Yan, X.Li. Optimized Top-K Processing with Global Page Scores on Block-Max Indexes (2012). WSDM'12.

Strohman.A and W. Bruce Croft. Efficient document retrieval in main memory (2007). In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Tonellotto.N, Macdonald.C, Ounis.I Efficient Dynamic Pruning with Proximity Support. LSDSIR Workshop (July 2010). Geneva, Switzerland.

Turtle.H and J. Flood. Query evaluation: Strategies and optimizations (1995). Information Processing and Management, 31(6):831–850, 1995.

Zobel.J and A. Moffat. Inverted files for text search engines (2006) .ACM Computing Surveys, 38(2).

Summary

Today, the quantity of data in the web is very big. In effect the Retrieval Information extends over various axes. Consequently the methods of RI meet difficulties of various sizes

and contents. Search engines have to solve a huge number of requests per minute. To satisfy users it is necessary to respect two criterias, the time and the quality of results. In this context the algorithms of early termination (ET) are a good solution for the first criteria but not for the second. A part of this paper is dedicated to presentations of the mains algorithms of ET. Also Two contributions are presented in this paper .The first one consists of the definition of new metrics during the processing of requests .The second and main contribution is the algorithm MWAND. The experiments show that the MWAND is very better than the WAND.

Applications et enjeux des Big Data dans le contexte des défis mondiaux

Hamza Saouli*, Kazar Okba*
Dounya Kassimi*
** Laboratoire LINFI, Département d'informatique, Université de Biskra, 07000, Biskra, Algérie*
{hamza_saouli, okbakazar}@yahoo.fr
dounya_kassimi@yahoo.com

Résumé. Le bigdata est devenu un domaine de recherche et d'innovation très célèbre. L'attirance de ce domaine provient des techniques et approches qu'il offre pour le traitement de gigantesques volumes d'informations qui circulent sur Internet. Cet article présente les principales notions de base du bigdata ainsi que les domaines d'applications les plus fréquent dans notre vie quotidienne. Il explique également les défis et enjeux auxquels les chercheurs font fassent pour améliorer les techniques de traitement classique d'informations ainsi que les technologies et domaines étroitement liés au bigdata tel que le Cloud computing et l'Internet des Objets.
Mots clés : big data, enjeux et défis, domaines d'applications, sécurité, Cloud computing, Internet des objets.

1 Introduction

L'un des problèmes majeurs des traitements des flux de données sur Internet est le non structuration de ces derniers (notamment les données textuelles de nature complexe, les vidéos et les photos qui submerge le web), de ce fait il est devenu nécessaire de faire recours à des techniques de traitement de données sophistiquées qui répondent aux besoins du marché Internet.

Malgré que les techniques proposées par le bigdata (Tel que Hadoop et MapRduce) représentent la pierre angulaire de l'avenir du traitement d'information sur le net, ce domaine fait face à un ensemble de défis qu'on peut diviser en deux axes importants :

- la protection de la vie privée : qui représente l'aspect sécuritaire du traitement de données et qu'on peut également diviser en trois éléments : (1) Garantir l'intégrité des données à transmettre (2) Garantir l'acheminement continu (sans interruption) et sécurisé (sans interception par un programme malveillant) des données, et (3) Garantir l'accès authentifié pour tout type d'utilisateur.
- l'assurance de la qualité des données : selon une étude menée par le cabinet Forrester pour le compte de Xerox auprès de 330 entreprises, 55% de ces derniers

avouent ne pas être en mesure de garantir la qualité de données de leurs clients. Même les "datarati" souffrent de cette insuffisance pour 45 % d'entre eux[1].

Le reste du papier est organisé comme suit: Section 2: représente quelques notions de base. Section 3: détaille les principales technologies impliquées dans la mise en œuvre des systèmes bigdata. Section 4: explique le modèle bigdata utilisé pour le traitement et stockage des données. Sections 5: aborde les détails de l'aspect sécuritaire du bigdata. Section 6: met l'accent sur l'importance des bigdata pour les principaux domaines d'applications de notre vie. Section 7: aborde les défis que rencontrent les chercheurs sur les big data. Section 8: prédit l'avenir du bigdata par rapport aux technologies de pointe. Section 9: c'est la conclusion de ce travail.

2 Notions de base

Dans cette section nous présentons un ensemble de concepts qui permettent de bien clarifier la notion de bigdata.

2.1 Définition

Les données concernées par les bigdata peuvent être recueillies de différentes sources (media, site web, entreprises, organisations humanitaires, gouvernement …etc.) ce qui diversifie leurs natures (climatiques, environnementales, politiques, sportifs, … etc.). Certes, Cette diversité impose le développement de nouvelles règles, normes et technologies de gestions de données, mais aussi ouvre la porte à de nouvelles perspectives comme : la gestion sécuritaire de données à l'échelle planétaire, l'approfondissement de nos connaissances médicales sur le fonctionnement du cerveau humain, lier et analyser les évènements religieux et culturels à travers le monde pour éviter des conflits politiques, lier les phénomènes météorologiques à l'échelle planétaire afin de mieux préserver nos systèmes écologiques, économiser la consommation de l'énergie sur l'Internet par l'utilisation cohérente des techniques de traitement de données sur divers sites de stockage de bigdata, … etc[1].

2.2 Modèle 5V

Afin de qualifier un ensemble de données comme étant des bigdata, il faut qu'il réponde à trois principaux critères : Volume, Vitesse, et Variété, d'où vient la dénomination typologique *3V*. Mais il est courant, d'ajouter deux autres critères pour compléter la typologie *3V*, à savoir : Valeur et Véracité.

2.2.1 Volume

C'est à la fois le critère le plus clair et le plus relatif pour classer des données comme étant des bigdata. Puisque la question qui se pose est : quelle est le seuil au-dessus duquel on entrerait dans le monde du Big Data?[1].

2.2.2 Vitesse

Le critère de vitesse représente la capacité du système de traitement et stockage des bigdata à analyser et modifier les données en un temps record, voire même en temps réel. Mais comme pour le critère de volume, le problème de seuil se pose toujours[2].

2.2.3 Variété

La variété représente le critère le plus important et le défi le plus persistant des bigdata. Avec la diversité des sources de données et l'absence de standards pour l'unification de la représentation des données à l'échelle planétaire, des problèmes d'interopérabilité de données devient de plus en plus assidue, ce qui a même poussé le cabinet « New Vantage Partners » de proposé le replacement du terme Big Data par Mashup Data[2] Mohanty et al. (2015)

2.2.4 Valeur et Véracité

C'est deux critères ont pour objectifs de transformer le modèle 3V en 5V sous prétexte d'améliorer et compléter la façon de distinguer un système bigdata. La valeur, indique la capacité de chaque donnée à offrir un plus sémantique à la vue globale du bigdata dont elle fait partie. La véracité indique la capacité du système de traitement et stockage de données à prendre en compte la précision et l'exactitude de chaque donnée sémantiquement distincte. Mohanty et al. (2015). La figure 1 représente le modèle 5V ainsi que ses caractéristiques :

FIG. 1 - *Enjeux du modèle 5V.*

2.3 Fonctionnement des systèmes Bigdata

Un système de traitement et de stockage de bigdata doit assurer trois fonctionnalités principales ; chaque fonctionnalité a pour objectif d'enrichir, d'extraire et de transformer les data (données) en bigdata[2] Mohanty (2015):

2.3.1 La collecte de données

Avant tous il est nécessaire de surfer sur Internet pour recueillir toute les informations qui concernent le domaine des bigdata à traiter. Cette navigation permet de construire une base de données aussi large que possible, mais aussi une BD qui contient des informations issues de tout genre de sources : objet connecté, organisation privé ou publique, système ouvert (OpenData) … etc. L'hétérogénéité des sources de données représente un élément enrichissant des Bases bigdata malgré que cela puisse causer des problèmes d'interopérabilité comme on vient de l'expliquer dans l'élément précédant.

2.3.2 Agrégation

Le but de l'opération d'agrégation est d'homogénéiser les données collectées de différents sites afin de créer une base de bigdata aussi cohérente que possible, avec les données les plus pertinentes et qui offrent une valeur de plus à la vue globale de bigdata. L'Agrégation représente la fonctionnalité qui devrait résoudre le problème de Variétés de données afin de créer une base bigdata immédiatement opérationnelle[2].

2.3.3 Analyse

L'analyse consiste à utiliser un logiciel spécifique au secteur d'applications des bigdata précédemment agrégés et déjà interopérables. La figure 2 illustre le processus de création des bigdata :

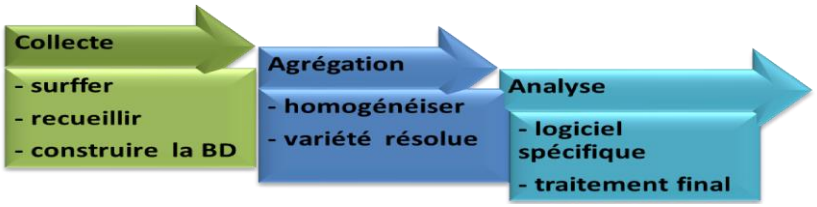


FIG. 2 - *Processus de génération de bigdata*

3 La révolution du bigdata

La célébrité et l'adoption rapide des bigdata, par les plus grandes sociétés de TI, provient principalement des avancé techniques et économiques (faible coût) qu'offre ce domaine. Cette adoption est véhiculée et concrétisée par un ensemble de technologies qui facilite le transport, le traitement et le stockage de données. On note essentiellement trois grandes technologies :

3.1 Cloud computing

Avant l'apparition du Cloud computing (qui représente une fédération de technologies destinées à perfectionner le niveau de qualité des services de stockage et traitement de l'information) l'information est stockée sur des entrepôts de données (datawarehouse) qu'il

faudrait liées physiquement pour pouvoir partager leurs données aux clients finaux Mohanty (2015).

Avec l'avènement du Cloud computing, une nouvelle couche de service appelé : Virtualisation-as-a-Service est apparue, cette couche permet la virtualisation du stockage et du traitement de données, ce qui rend l'accès aux entrepôts de données plus souple et sans besoin de rattachement physique. Outre, le Cloud computing offre la possibilité d'accéder aux données de façon instantanée et avec un taux de disponibilité record, ce qui le met à la tête des meilleurs supports technologiques des bigdata. Selon Stéphane Grumbach de l'INRIA d'ici 2020 un tiers des données sera migré vers le Cloud. Cette constatation est due essentiellement aux capacités de stockage et de traitement dont disposent les Datacenter hébergeant les systèmes d'exploitations Cloud et dont les bigdata ont besoin pour être traiter et stocker[2].

Cependant, l'acquisition d'infrastructure de stockage ou de traitement de bigdata nécessite l'investissement de budgets importants ce qui n'est pas offert aux petites et moyennes entreprises. C'est pour cela que ces entreprises font recours aux services Cloud qui leurs permet de louer et payer les unités de traitement et de stockage selon leurs besoins et par la suite économiser leurs dépenses Ramesh (2015).

3.2 Hadoop

A l'origine Hadoop a été créée par Doug Cutting en 2004 pour agrandir la taille de l'index de son moteur de recherche Open Source Nutch. Ensuite Hadoop s'est étendu pour se transformé en un environnent spécialisé dans le de traitement de grandes masses d'informations.

A l'encontre des autres technologies de traitement d'informations distribuées, Haddop se base sur une architecture Grid computing qui permet de distribuer les taches sur des clusters de serveurs pas forcément liés entre eux.

Malgré que les coûts d'adoption de Hadoop en tant que technologie de traitement des bigdata sont relativement élevés pour 45% des entreprises traitantes des bigdata (entre $100.000 et plus de $500.000) mais cela reste cinq fois moins coûteux que d'utiliser un data warehouse classique (sachant qu'un système basé Hadoop stocke cinq fois plus de données qu'un datawarehouse traditionnel) c'est pour ça que 98% des entreprises traitantes des bigdata adoptent Hadoop comme environnement de traitement de bigdata selon une étude effectuée par l'institut IDC6 Ramesh (2015).

3.3 NoSQL

Pour Not Only SQL,il représente la nouvelle génération des Systèmes de gestion de base de données (SGBD) conçue spécialement pour la gestion de grandes quantités d'informations, où il ne s'agit plus d'utiliser des tables pour le stockage d'informations, ni SQL pour l'interrogation de ces derniers. Apparu pour la première fois en 1998 et largement adopté vers la fin des années 2000, le NoSQL est devenu la pierre angulaire de la gestion des bigdata pour la plupart des entreprises sur Internet, à savoir : Facebook, Twitter, Google, Amazon.com, … etc Martha (2015). Les principales différences entre SQL et NoSQL qui ont permi à ce derniers d'être mieux adapté au bigdata sont :

Le non restriction des types de données : avec NoSQL on peut représenter des données aussi complexes qu'il y' on a, alors qu'avec SQL on est restreint au type de données connues (entier, réel, date … etc.)

La non structuration préalable de données : avec NoSQL on utilise des structures de données souples telle que les paires clé-valeurs ou les documment JSON ce qui nous permet de créer des métat-données au cours du traitement des bigdata et par conséquent avoir une grande souplesse pour la description des données. Alors qu'avec SQL et les bases de données relation on est obligé de déterminer les noms des champs dans chaque table au préalable ce qui rend la description des données plus rigides et moins maniable.

4 Traitement et stockage des Bigdata

La limite entre un système d'information destiné au bigdata et un SI classique se résume dans le simple fait que ce dernier représente des anomalies et des disfonctionnements quand le volume de données commence à dépasser un certain seuil. Le tableau suivant représente une étude comparative entre un SI de stockage classique et un SI bigdata quand il s'agit d'augmenter le volume de données Martha (2015):

SI Critères	Classique	Bigdata
Performance	Chute des performances	Hadoop : distribution intelligente des traitements
Evolutivité	évolutivité oscillatoire	Evolutivité assuré et ajout de nœud sans limite
Résilience	Allocation d'espace fixe et perte d'information	Résilience assurée et duplication d'informations

TAB. 1 - Principaux critères de différenciation entre SI classique et bigdata

Le tableau ci-dessus utilise trois critères principaux pour la différenciation entre un SI classique et bigdata :

4.1 Performances des unités de traitement :

Avec un système de gestion de bigdata tel qu'a Hadoop il est possible de distribuer les taches sur les nœuds du réseau de façon à alléger la charge de travail sur les nœuds surchargés. Cette distribution intelligente n'est pas possible avec un SI classique ce qui entraine des chutes de performance rapide, une fois qu'on dépasse le seuil bigdata, quel que soit le nombre d'unité de traitement ajouté au système.

4.2 Evolutivité des nœuds du réseau :

Avec un SI bigdata il est possible d'ajouter et d'assumer autant qu'on veut de nœuds dans le réseau, ce qui assure une évolutivité linéaire des ressources de traitement et de stockage

des bigdata. Cette linéarité est perdue quand le volume de données dépasse un certain seuil avec des SI classiques vue l'absence d'un environnement de gestion de cluster de nœuds tel qu'a hadoop.

5 Sécurité et Confidentialité

Dans cette section nous abordons le sujet le plus inquiétants des bigdata, à savoir la sécurité et la confidentialité des données. L'importance d'un tel sujet réside dans le fait qu'il s'agit de traiter de gigantesques masses de données ce qui augmente les risques d'inculquer des virus ou des programmes d'espionnages ce qui peut déstabiliser le système sécuritaire ou atteindre la vie privées des utilisateurs.

5.1 Sécurité

La sécurité des bigdata est un problème qui se pose notamment pour les infrastructures d'informations critiques IIC Sudarsan et al. (2015). Les IIC concernent des domaines telles que la défense, les transactions bancaires, gestion d'énergie, etc. avec les IIC le bigdata a comme problème essentiel de trouver les techniques nécessaires pour assurer la disponibilité de données. Puisque les autres problèmes de sécurité sont assurés par la fourniture d'une infrastructure menue de capacité matérielle et logicielle suffisante pour la protection des données. D'un autre côté, le fait d'isoler les infrastructures, physiquement d'Internet, n'est plus une solution efficace pour la protection de données, puisque ces derniers devraient être utilisés, en temps réel, par le large public. Un autre problème dont le bigdata devrait faire face pour les IIC, est la variété d'informations en amont et en aval d'Internet et qui devrait être traité à côté des informations stockées, de façon organisée, sur les serveurs et centre de données des IIC.

5.2 La confidentialité un gros souci de sécurité

La confidentialité est l'un des plus gros problèmes dont le bigdata devrait faire face. La confidentialité est le processus qui permet de créer des niveaux d'accès et de déterminer le type d'utilisateur qui peut accéder à de telles ou tell les données Jeong et Shin (2015).

Avec l'explosion en volume et en variété de données, la création d'algorithme de cryptage assez performant est devenue une tâche ardue. Outre, le bigdata est basé sur le principe de la valeur du partage et de l'enrichissement d'informations, ce qui s'oppose avec le principe de la restriction d'informations. Pour ces raisons-là, le plus grand défi des bigdata est la création de modèle de représentation de données assez performant pour pouvoir assurer la protection de la vie privée et la confidentialité des informations.

5.3 Technique de sécurité

Trois principales techniques de sécurisation de données se sont imposées pour la protection des systèmes bigdata[2] Sudarsan (2015):

- Big SQL d'IBM
- Impala de Cloudera
- Et Hive 0.13. d'Hortonworks.

Le tableau suivant, illustre la comparaison entre les trois techniques en termes de temps d'exécution[4] :

Techniques	Temp d'exécution	vitesse d'exécution
Big SQL	48 minutes et 28 secondes	3,6 fois plus vite qu'Impala et 5,4 fois plus vite que Hive
Impala	2 heures, 55 minutes et 36 secondes	2 fois plus vite que Hive
Hive	4 heures, 25 minutes et 49 secondes	Hive représente la technique la plus lente

TAB. 2 – Comparaison du temps d'exécution entre trois approches de sécurité bigdata

5.4 Disponibilité de données

La disponibilité est généralement assurée en utilisant un serveur de protection et sauvegarde de données, qui doit être isolé du système principal afin d'éviter de probable attaques Demchenko et al. (2014).

Avec des systèmes de traitement de big data, plusieurs défis s'imposent :

- La réduction des coûts de stockage de données sur les serveurs de sauvegarde.
- L'utilisation de plusieurs copies pour sauvegarder des données semble une solution peu pratique et très couteuse (en terme budgétaire et ressources), vue le volume gigantesque données à traiter et stocker.
- L'apparition de l'Internet des objets (Internet of Things en Anglais) a considérablement augmenté le nombre d'appareils connectés sur la toile, ces derniers ont besoins d'accès continu et en temps réel aux données stockées sur les serveurs bigdata, ce qui met en péril l'isolation des serveurs de sauvegarde et nous poussent à la reconsidérer.

6 Domaines d'applications du bigdata

Dans cette section nous présentons quelques principaux domaines d'applications du bigdata Boinepell (2015) Krishna (2015):

6.1 Agriculture

D'ici 2050 on prévoie le dépassement de 9 milliards d'êtres humains sur le globe, ce qui rend l'agriculture un domaine prioritaire pour gérer les besoins alimentaires de la population

mondiale. Le big data représente un atout considérable pour l'organisation de l'agriculture à travers le monde, notamment pour la gestion de l'irrigation.

6.2 Assurance

La possibilité de récolter de gigantesques masses d'informations qui concernent la vie des individus permet de concevoir un modèle de vie pour chacun d'eux : hygiène de vie, conduite de voiture, amende, consommation électrique, relation professionnelle …. Etc. Ces modèles de vie permettent aux agences d'assurances d'améliorer leurs offres, d'optimiser leurs méthodes, et même de mener des enquêtes plus précises.

6.3 Marketing

Avec le marketing on est amené à gérer de gigantesques masses d'informations qui proviennent de divers sites et réseaux sociaux que des clients potentiels peuvent visiter. Mais ce qui révolutionne vraiment le marketing de nous jours c'est l'omniprésence de capteurs publics sur les centres commerciaux, métros, aéroports et universités, et qui sont destinés à capté le comportement des consommateurs, ce qu'il achète, ce à quoi ils s'intéressent, et les produits qu'il ne trouve pas aux marchés, ce qui permet d'analyser et étudie leurs besoins en temps réel afin de produire des solutions et méthodes de marketing plus efficaces.

6.4 Gestion de catastrophes naturelles

L'une des applications les plus intéressantes des Bigdata, est la possibilité d'analyser des données météorologiques en temps réel, ce traitement permet de suivre et de visualiser le déplacement des ouragans et de prédire les endroits géographiques où ces derniers vont frapper.

FIG. 3- *Domaines d'applications du big data.*

7 Enjeux et recherche

Le Bigdata fait face à un ensemble d'enjeux qui doivent être pris en charge par la communauté académique et industrielle. Ces challenges tournent autour de trois axes :

7.1 Enjeux de gestion

Ce type d'enjeux concerne la gestion du stockage et du traitement de données. Les enjeux consistent à :

- Analyser et comprendre le volume gigantesque de données pour pouvoir créer les métadonnées qui décrivent de façon précise et correcte ces derniers Demchenko et al. (2014).
- Analyser et comprendre les liens et relations entre les données pour optimiser leurs schémas de stockage.
- Gérer la duplication des données pour réduire le coût.

7.2 Enjeux sécuritaires

Les enjeux sécuritaires concernent la réduction des risques qui environnent l'infrastructure de stockage du bigdata. Les enjeux consistent à Sudarsan et al. (2015):

- Trouver les mécanismes qui assurent un excellent niveau de disponibilité
- Gérer la duplication des données pour assurer un excellent niveau d'intégrité.
- S'assurer de l'authentification des sources de données.
- Réduire les besoins en ressources pour les algorithmes de sécurisation de données.

8 L'Emergence du futur

L'émergence de la fédération des technologies Cloud computing a permis d'élever, de façon massive, le niveau de qualité des services de stockage et de traitement de données. Le Cloud permet d'offrir des softwares de traitement en ligne (Software-as-a-Service), des plateformes de création de logiciel en ligne (Platform-as-a-Service), et des infrastructures de stockage gérées avec de puissants outils de virtualisation (Infrastructre-as-a-Service et Virtulisation-as-a-Service).

D'un autre coté l'émergence de l'Internet des objets (Internet of Things) a permis la liaison de centaines de milliards (environ 700 milliard d'objets) d'objets sur le Net. Ces objets génèrent de massives quantités de données et accroissent le trafic sur la toile d'Internet Sudarsan et al. (2015).

8.1 Big data et Cloud computing

l'adoption du Cloud computing non pas seulement comme infrastructure de stockage et de traitement de données, mais aussi comme un moyen d'interface entre consommateurs et services, permet de réduire les coût de stockage et assurer un excellent niveau de disponibilité mais aussi offre la possibilité aux détenteurs de données de choisir le type de déploiement des ressources Cloud pour assurer le niveau de sécurité et de confidentialité qu'il désirent. C'est ainsi qu'un consommateur ou fournisseur de données peut choisir : un déploiement privé si les données sont de nature secrètes (ex. données qui concernent la défense nationale), un déploiement public s'il désire un service accessible par tout type d'internaute, ou un déploiement hybride s'il désir restreindre l'accès à un type particulier d'internautes (ex. données au sein d'une entreprise de marketing). Le Cloud computing représente le seul para-

digme, et offre le seul type d'infrastructure, qui peut faire face aux défis des bigdata (Section 7). Le Cloud computing est la future Infrastructure des bigdata.

8.2 Bigdata et Internet des objets :

L'apparition de l'internet des objets commence déjà à bouleverser l'architecture de l'internet traditionnel. Avec l'apparition de nouveaux protocoles qui s'assure de l'interconnexion des objets et la gestion des données engendrées par ces objets. L'accumulation des données crées par des objets menus de capteurs comme : les moteurs à diésel, machines, électroménagers, détecteurs d'incendies, etc., lance le défis de la gestion de gigantesques volumes de données, lesquelles d'origine, ne représentent que de petits morceaux de données regroupées et accumulées à partir de centaines de milliards d'objets interconnectés. La nature des données engendrées par ces objets nécessite de nouvelles techniques de recueillement, d'organisation et du traitement et auxquelles le domaine du bigdata doit répondre. L'Internet des objets représente le futur défis des big data.

9 Conclusion

L'émergence du big data comme un nouveau domaine qui s'occupe du traitement des grandes masses d'informations a permis l'ouverture de nouvelles portes au chercheurs ainsi qu'aux entreprises professionnelles pour créer et tester leurs approches et produits à l'échelle planétaire.

L'article présent les principaux enjeux sécuritaires qui diversifient les systèmes d'informations classiques du big data notamment en ce qui concerne la confidentialité et système de sauvegarde. Nous avons également présenté l'impact de l'avènement du bigdata sur quelques domaines qui touche la vie quotidienne de chacun de nous, à savoir : l'agriculture, le marketing, propagation des épidémies, etc.

Enfin nous pensons que les futures approches et solutions des problèmes rencontrés avec le bigdata se trouvent essentiellement dans la fédération des technologies Cloud computing et que les principaux défis, à moyen et à cours termes, se trouvent dans le domaine de l'Internet des objets.

REFERENCES

Boinepell, H. (2015). Applications of Big Data. *Serie Studies in Big Data,* 11:161-179.

Demchenko, Y., C. Nao, C. D. Laat, P. Membre, and D. Gordijenko. (2014). Big Security for Big Data: Addressing Security Challenges for the Big Data Infrastructure. *Series Lecture notes in computer science,* 8425: 76-94.

Jeong, Y.-S., and S.-S. Shin (2015). An Efficient Authentication Scheme to Protect User Privacy in Seamless Big Data Services. *Journal of Wireless Personal Communications,* 86: 7-19.

Krishna, P. R. (2015). Big Data Search and Mining. *Serie Studies in Big Data,* 11:93-120.

Martha, V. (2015). Big Data Processing Algorithms. *Serie Studies in Big Data,* 11:61-91.

Mohanty, H. (2015). Big Data: An Introduction. *Serie Studies in Big Data*, 11:1-28.

Mohanty, H., P. Bhuyan, and D. Chenthati (2015) . *Studies in Big Data 11: Big Data A Primer* . New York, Springer Verlag.

Ramesh, B. (2015). Big Data Architecture . *Serie Studies in Big Data*, 11:29-59.

Sudarsan, S. D., R. P. Jetley et S. Ramaswamy (2015). Security and Privacy of Big Data. *Serie Studies in Big Data*, 11:121-136.

Zal, A. (2015). Opportunities and Security Challenges of Big Data. *Series Palgrave Macmillan's Studies in Cybercrime and Cybersecurity*, 181-197.

Zhang, X., and S. Xiang. (2015). Data Quality, Analytics, and Privacy in Big Data. *Serie Studies in Big Data*, 9: 393-418.

[1] http://franceblog.emc.com/securite-intelligente-approche-big-data-securite-41 ,Consulté le 27/07/2015.

[2] http://www.bigdataparis.com/2014/article-informatica-2.php ,consulté le 28/07/2015.

[3] http://www.ad-exchange.fr/ . Consulté le 14/01/2016

[4]http://www.zdnet.fr/partenaires/ibm/big-data-et-le-choix-sql-sur-hadoop-comparatif-des-performances-39816782.htm , Consulté le : 04/09/2015

Summary

Bigdata has become a famous area of research and innovation. The attraction to this area comes from the techniques and approaches it offers for treating gigantic information capacity, circulating on the Internet. This article presents the main basics of Big Data and the areas of the most common application in our daily live. It also explains the issues that the researchers need to improve in order to enhance the conventional information processing techniques, technologies and fields closely related to Big Data, like Cloud computing and Internet of Things.

Data mining pour la construction de communautés d'utilisateurs

Nour El Houda Boulkrinat[1,2], Habiba Drias[2], Hakima Mellah[1]
Hassina Khellouf[1], Aida Bouchabou

[1] CERIST, CEntre de Recherche sur l'Information Scientifique et Technique,
05, Rue de Frères Aissou Ben Aknoun, Alger, Algérie
[2] USTHB, Université des sciences et de la technologie Houari-Boumédiène
BP 32 El Alia 16111 Bab Ezzouar, Alger, Algérie
nboulkrinat@cerist.dz, h_drias@hotmail.fr, hmellah@cerist.dz,
hassina_khellouf28@hotmail.fr

Résumé. Les communautés d'utilisateurs sont tenues ensemble par un intérêt commun dans un champ de savoir et sont conduites par un désir et un besoin de partager des idées, des expériences, des modèles, des outils et des meilleures pratiques. Dans ce travail nous présentons notre solution pour la construction de communautés d'utilisateurs en tenant compte des centres d'intérêts et en se basant sur la technologie du data mining. Les centres d'intérêts seront introduits à partir d'une ontologie dédie au domaine de l'informatique. Quant au data mining, il sera utilisé pour la segmentation des communautés via l'utilisation de l'algorithme K-means. Cette construction de communautés aidera le système de recherche d'information à l'acquisition implicite des nouveaux profils par la sélection ou l'adaptation d'un profil prédéfini, et à la reformulation des requêtes de l'utilisateur en fonction de nouvelles connaissances acquissent à partir de la communauté où il appartient.

1 Introduction

La recherche d'information sociale est une nouvelle voie dans laquelle le processus de recherche incorpore le contexte social des documents et le contexte social des utilisateurs, en effet, une manière plus naturelle de communiquer. Cette notion est déployée particulièrement dans le World Wide Web (Kirsch, 2005). Ainsi, le Web2.0 a permis de former différents types de réseaux sociaux à grande échelle, qui sont maintenant reconnus comme un moyen important pour la diffusion de l'information.

L'explosion des réseaux sociaux a permis l'émergence d'une nouvelle branche de la Recherche d'Information : la RI sociale qui a permet de personnaliser la recherche collaborative : actuellement, il y a un manque d'outils qui permettent une combinaison de la recherche collaborative et personnalisée, c'est à dire la production des résultats de la recherche adaptée à un utilisateur particulier, en utilisant l'information recueillie auprès de sa communauté. La plupart des efforts de la recherche collaborative sont basés sur l'utilisation des contributions d'une communauté d'utilisateurs, les traitements de tous les membres de la communauté sont égaux.

Tamine et al. (2006) définit une communauté comme un groupe de personnes qui interagissent entre elles, partagent et utilisent des informations en relation avec leurs centres d'intérêts, caractéristiques démographiques ou activités professionnelles, communes. Ces centres d'intérêt traduisent les domaines des connaissances et d'expertise ciblés régulièrement par la communauté durant ses sessions de recherche, ils sont définis par des mots clés. Ils sont généralement représentés par le modèle vectoriel où chaque centre est représenté par une liste de termes représentatifs.

Plusieurs techniques et méthodes ont était proposées pour la détection ou la construction de communautés, l'une des techniques la plus estimée à l'heure actuelle est celle du Data Mining, le Data Mining, est un ensemble de techniques d'exploration de données permettant d'extraire d'une collection de données des connaissances sous forme de modèles de description afin de décrire le comportement actuel et / ou de prédire le comportement futur des données.

Dans la recherche d'information personnalisée, l'acquisition du profil utilisateur peut être explicite et/ou implicite. L'approche explicite consiste à obtenir les informations directement de l'utilisateur, et implicite largement motivée par les travaux actuels dans le domaine, implique l'exploitation des données du comportement de l'utilisateur pour inférer son profil. Le principal avantage de cette approche est qu'elle ne nécessite aucune implication directe de l'utilisateur, ni de temps passé à émettre des jugements, ni un effort d'attention particulier lors de sa recherche. En effet, toute interaction de l'utilisateur avec le système est considérée comme une estimation de son jugement d'intérêts. L'acquisition du profil peut se faire à partir d'une base qui contient un ensemble de profils prédéfini.

L'objectif de ce travail est la construction de communautés pour la recherche d'information en utilisant la technique de K-means. Les communautés d'utilisateurs aideront le système de recherche d'information à satisfaire les besoins des utilisateurs en termes d'information, comme la reformulation et la réécriture des requêtes de l'utilisateur en fonction de nouvelles connaissances acquissent à partir de la communauté où il appartient. La notion de communauté participera aussi, à l'acquisition implicite des nouveaux profils par la sélection ou l'adaptation d'un profil prédéfini (existant dans une communauté).

Après cette introduction, le reste de ce papier est organisé comme suit : dans la deuxième section nous présentons les travaux existants, la section 3 sera consacrée à la présentation de notre approche ; dans laquelle nous présentons l'algorithme proposé pour la construction de communautés. Nous terminerons dans la section 4, par une conclusion dans laquelle nous énumérons les travaux futurs.

2 Travaux existants

La détection de communautés a fait l'objet de nombreux travaux de recherche, qui constituent des états de l'art complets: (Fortunato 2009) (Yang et al., 2010) (Papadopoulos et al., 2011) (khatoon et Banu, 2015). La plupart de ces études classent les articles et les travaux de recherche en fonction du type de l'algorithme. On peut distinguer d'une part les techniques d'apprentissage non supervisé, qui exploitent les attributs décrivant les objets, comme la classification hiérarchique ou les k-means, et d'autre part celles qui considèrent les relations entre les différents objets comme c'est généralement le cas dans le partitionnement de graphes (Fortunato 2009).

L'objectif de la détection de communautés dans les graphes, ou encore dans les réseaux sociaux, est de créer une partition des sommets, en tenant compte des relations qui existent entre les sommets dans le graphe, de telle sorte que les communautés soient composées de sommets fortement connectés et qu'elles soient peu reliées entre elles (Combe, 2013).

L'objectif du Clustering (classification non supervisée) est de fractionner l'ensemble hétérogène d'objets de la base de données en un certain nombre de sous-ensembles plus homogènes, appelés clusters. Pour cela les clusters doivent contenir des objets qui partagent un haut degré de similarité (maximisation de la similarité intra-cluster) et que ces clusters doivent être suffisamment larges (minimisation de la similarité inter-cluster). La similarité des objets est en général mesurée en termes de distance géométrique entre les objets.

Le clustering par partitionnement à pour but de regrouper un ensemble d'objets en cluster (classes ou groupes) tels que chaque objet appartient à au moins un cluster. Les divers algorithmes appartenant à cette famille d'approches, différent par la fonction de qualité utilisée pour évaluer la pertinence d'un partitionnement, ainsi que par la distance choisie pour effectuer les regroupements (clustering) (Aït el Hadj, 2013).

L'approche décrite dans cet article est réalisée par l'utilisation d'un algorithme classique de classification K-means. Cet algorithme se basera sur une mesure de similarité proposée qui prendra en considération les centres d'intérêts des utilisateurs.

3 Approche proposée

Nous proposons un système à base de profils utilisateurs (schématisé par la Figure 1). A cet effet, la première étape dans notre système sera l'acquisition des données sur l'utilisateur et la deuxième sera la construction de communautés via l'exploitation des centres d'intérêts. La construction de la base de profils est faite par l'acquisition des données sur l'utilisateur. Cependant, deux techniques d'acquisition existent: explicite et implicite. Dans ce travail, la méthode explicite a été adoptée. Cette technique consiste à obtenir les informations sur l'utilisateur à partir du remplissage d'un formulaire pour collecter ses données dans la base de profils. Ces informations sont des données personnelles et démographiques, les préférences et les centres d'intérêt représentent par une ontologie.

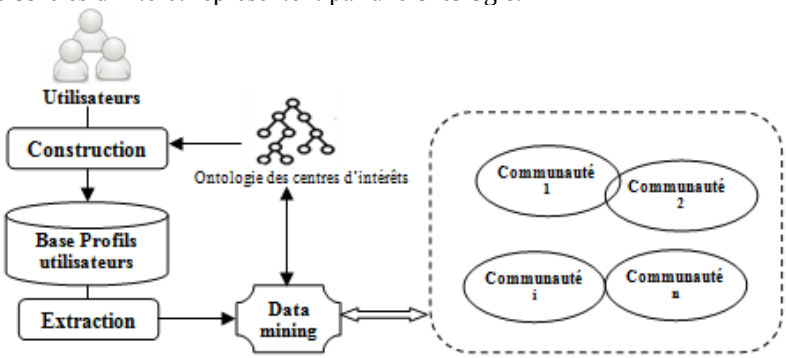


FIG. 1 – *Architecture du système proposé*

Le centre d'intérêt d'un utilisateur exprime le domaine d'expertise de l'utilisateur ou son périmètre d'exploration, les centres d'intérêts sont représentés, dans cet article, par une onto-

logie proposée de domaine de l'informatique. La figure suivante (Figure 2) donne un aperçu sur cette ontologie.

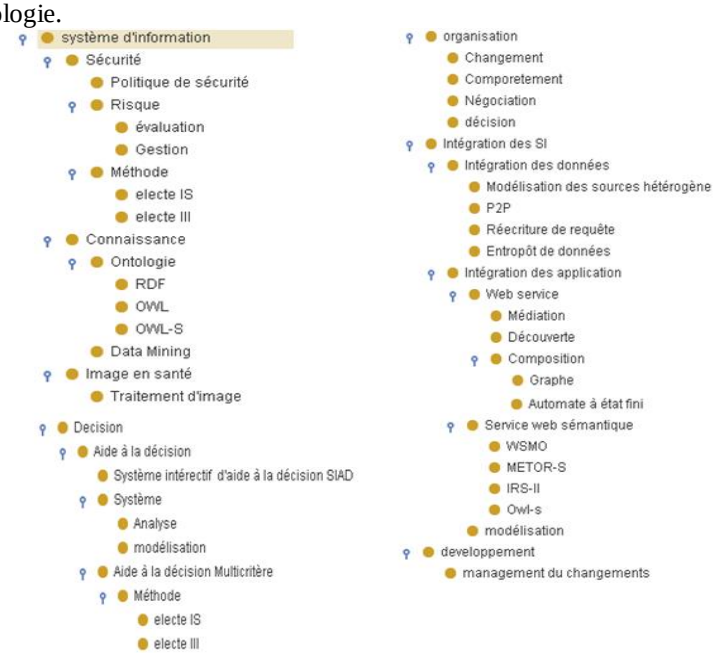


FIG. 2 – *Aperçu de l'ontologie des centres d'intérêts*

L'utilisateur peut sélectionner ses centres d'intérêt à partir de l'ontologie lors de son inscription ou les introduire manuellement si son domaine d'intérêt n'existe pas dans l'ontologie.

Dans ce qui suit, nous allons présenter les étapes de construction des communautés en prenant en considération les centres d'intérêts des utilisateurs :

3.1 Préparation des données

Nous avons une base de profils (Figure 1) qui contient un ensemble de tables tel que *: utilisateur, document, préférences, requête, mot_cle, centre_d'interet*. A partir de cette base de profils, nous allons procéder à la préparation et exploitation des données, ensuite nous choisirons les données pertinentes pour les étapes de la construction de communautés.

3.1.1 Description des données

Les données (la table « Centre_profil» de la relation plusieurs à plusieurs entre la table profil et la table centre_intérêt) que nous allons exploiter dans notre cas sont :

- **Utilisateur** (représenté par la table ***profil***) qui est représenté par un identificateur unique et ses variables décrivent son comportement (Id_user, Pseudo, mot_de_pasee, Nom, Prenom, email, profession, domaine_act, Visibilité)
- **Centres d'intérêts** (représenté par la table ***centre_intérêt***) qui sont représentés par des concepts dans une ontologie relative au domaine de l'informatique (Uri_concept).

3.1.2 Création de la base analytique

Afin de pouvoir appliquer les techniques analytiques issues du data mining, nous avons besoin d'une base sous la forme d'une matrice. A cet effet, nous proposons une base analytique sous la forme d'une matrice (lignes - colonnes) où chaque utilisateur représente une ligne, et les variables : centres d'intérêt seront représentés par les colonnes.

Soit une matrice de données ***mat[n][m]*** présenté par le Tableau 1, où les valeurs de **m** variables (**m** centres d'intérêts) sont stockées pour un ensemble de **n** objets (**n** profils). La case C_{ij} représente la valeur de la variable C_j pour l'objet P_i tel que :

$$\mathbf{C_{ij}} = \begin{cases} \mathbf{1}, & \text{si profil } \mathbf{P_i} \text{ posséde le centre d'intérêt } \mathbf{C_j} \\ \mathbf{0}, & \text{sinon} \end{cases}$$

Chaque profil P_i est représenté comme un vecteur à ***m*** centre d'intérêt $(C_{i1}, \ldots, C_{im})$.

P/C	$\mathbf{C_1 \ldots \ldots \ldots C_j \ldots \ldots \ldots C_m}$
$\mathbf{P_1}$	$C_{11} \ldots \ldots \ldots C_{1j} \ldots \ldots \ldots C_{1m}$
.	... ...
$\mathbf{P_i}$	$C_{i1} \ldots \ldots \ldots C_{ij} \ldots \ldots \ldots C_{im}$
.	... ...
$\mathbf{P_n}$	$C_{n1} \ldots \ldots \ldots C_{nj} \ldots \ldots \ldots C_{nm}$

TAB. 1 – *Matrice de données de **n** profils et **m** centres d'intérêts*

3.2 Segmentation des profils

Le processus du clustering (segmentation) comporte trois étapes majeures : la préparation des données, le choix de l'algorithme, et l'application et la validation de l'algorithme de clustering.

- **Nombre de clusters désirés :** l'un des problèmes major lors du clustering est le choix du nombre de clusters qui doit être décidé avant le lancement du traitement. A cet effet, nous proposons que le nombre de clusters soit égal au nombre de concepts pères de l'ontologie proposée (Figure 2), dans ce cas **Nbr-clusters=6**.
- **Nombre des données disponibles :** nous avons une matrice de **21** profils et **90** centres d'intérêts ***mat[21][90]*** où les variables sont *qualitatives* de type *ordonné*.

3.2.1 Choix de l'algorithme

La quantité d'objets à traiter est un premier facteur de décision pour le choix de l'algorithme, dans notre cas les données sont des profils d'utilisateurs c-à-d des données de très grande taille, et pour cela il est possible d'avoir recours à des méthodes plus complexes nécessitant un temps de traitement plus important comme les méthodes de partitionnement **(K-means \ K-moyennes)**. Proposé en 1967 par McQueen (Celeux, 1989), l'algorithme des centres mobiles (K-means) est l'algorithme de clustering le plus connu et le plus utilisé, tout en étant très efficace et simple.

a. Le pseudo-code de l'algorithme de K-means

Nous adaptons l'algorithme de K-means et nous considérons dans notre cas:
Cluster = Communauté ; Données = {profil, centres d'intérêts}.

Algorithme 1 : Pseudo-code de l'algorithme K-means

Entrée : ***K :*** le nombre de communautés (clusters) désirés,
Dist *:* une mesure de similarité sur l'ensemble des profils à traiter ***X***.
Sortie : Une partition $\mathbf{C_{final}}$= *{C_1 … C_k}*, avec chaque C_i représente l'une des communautés.
Etape 0 :

1. Initialisation par tirage aléatoire des ***X*** profils, de ***K*** centres $\mathbf{x^*_{1,0} \ldots\ldots x^*_{k,0}}$.
2. Constitution d'une partition initiale C_0= *{C_1 … C_k}* par allocation de chaque profil $x_i \in X$ au centre le plus proche :

$$\mathbf{C_1 = \{x_i \in X |\ Dist\ (x_i, x^*_{1,0}) = \max_{h=1,\ldots k} Dist(x_1, x^*_{h,0})\}} \quad /^{**}Formule\ (1)^{**}/$$

3. Calcul des **centroïdes** des ***K*** clusters obtenues $\mathbf{x^*_{1,1} \ldots\ldots\ldots\ldots x^*_{k,1}}$

Etape T :

4. Constitution d'une nouvelle partition c_t= {c_1 … c_k} par allocation de chaque objet $x_i \in X$ au centre le plus proche :

$$\mathbf{C_1 = \{x_i \in X |\ Dist\ (x_i, x^*_{1,t}) = \max_{h=1,\ldots k} Dist(x_1, x^*_{h,t})\}}$$

5. Calcul des **centroïdes** des ***K*** communautés obtenues $\mathbf{x^*_{1,t+1} \ldots\ldots\ldots\ldots x^*_{k,t+1}}$
6. Répéter les étapes **(4)** et (**5**) tant que des changements s'opèrent d'un schéma $\mathbf{C_t}$ à un schéma $\mathbf{C_{t+1}}$ ou jusqu'à un nombre τ d'itérations.
7. Retourner la partition final $\mathbf{C_{final}}$.

b. *Distance*

Afin de fonctionner l'algorithme avec des données catégoriques (qualitatives), nous proposons la distance de similarité suivante :
Etant donné deux profils de type de données quantitatives, la similarité entre eux est définie comme la somme des centres d'intérêts communs correspondant entre eux, comme illustré par la formule (1) ci-dessous :

$$\boldsymbol{Dist\ (x, y) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \delta(x_i, y_j)}{\mu}} \qquad \mathbf{(1)}$$

Tel que:

- ***x*** et ***y*** sont deux profils représentés comme des vecteurs à **n** et **m** centres d'intérêt respectivement.
- $\mathbf{x_i}$ et $\mathbf{y_j}$ sont les centres d'intérêts de deux profils ***x*** et ***y***.
- $\boldsymbol{\delta(x_i, y_j) = \begin{cases} 0, & x_i \neq y_j \\ 1, & x_i = y_j \end{cases}}$
- $\boldsymbol{\mu = n + m}$ est la somme des centres d'intérêts de deux profils.
- **Dist** : plus la mesure est grande, plus les profils sont similaires.

c. Centroïdes

Nous déterminons les nouveaux centres de gravité (**Centroïdes)** pour chaque cluster à l'aide de la formule ci-dessous :

$$\max_{i \in [1,N]} (\boldsymbol{Dist\ (P_i, G_j)}) \qquad \mathbf{(2)}$$

Tel que : $\boldsymbol{G_j}$ **:** graine initiale et $\boldsymbol{P_i}$ **:** profils de même graine.

d. Condition d'arrêt

- La partition n'est pas modifier lors d'une itération ; ou
- Les nouveaux centroïdes ne changent pas.

3.2.2 Application et validation de l'algorithme de clustering

Le système est développé en utilisant le langage de programmation java, et la base de données sous oracle 10g. Nous avons adapté Protégé-2000 pour le développement de l'ontologie qui est implémentée en langage OWL.

a. Déroulement de l'algorithme

Nous fixons à priori le nombre de clusters attendus, donc on fixe **k** à **6.**

- Nous choisissons d'abord au hasard les centres de gravité et nous construisons les clusters initiaux autour de ces centres, les **six** premières graines prennent donc les **six** premiers profils (**P_1, P_2, P_3, P_4, P_5, P_6**) ;
- Nous calculons la distance en utilisant la formule (1), et nous comparons entre chaque profil avec chaque grain **:**
-

$$Dist\ (G, Y) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\delta(\mathrm{Gi}, y_j)}{\mu}$$

Où : $\delta(G_i, y_j)$ est calculé en comparons le concept père de centre d'intérêt du gaine G_i avec le concept père de centre d'intérêt du profil y_j. Pour cela, nous utilisons la procédure de comparaison entre les centres d'intérêt, présenté comme suit :

a.1 Comparaison des concepts : fait appel à la procédure **GetConceptsPeres**, cette procédure vérifiée l'emplacement du centre d'intérêt par rapport aux concepts de l'ontologie et retourner son concept père.

Algorithme 2 : Récupération des concepts pères à partir de l'ontologie

```
Liste_concepts  L← {C1,…..CN} ; /** liste des concepts de profil utilisateur **/
Liste_concepts_Super  Lcs ;        /** Liste de tous les Supers concepts d'un concept donné **/
Liste_concepts_Pere  Lcp ;         /** Liste de résultats **/
Chemin_ontologie CHO = (C:\\……\\Protege\\nom_ontologie) ;  /**Chemin d'accès à
                                                              L'ontologie protégé**/
Debut
Procédure GetConceptsPeres (liste L)  /**permet d'obtenir les concepts pères
                                        de la liste L à  partir de l'ontologie **/
Pour  chaque Ci de  L
 FAIRE
   Lcs = GetSuperClasse(Ci)   /** permet d'obtenir tous les super concepts Ci  de  l'ontologie**/
   Pour  j=1 à largeur (Lcs)
      Faire
         Si (Lcs [j] = '' THING '') Alors Lcp= Lcs [j+1];
                                    Sinon Afficher ('Concept non trouvé dans l'ontologie') ;
       Fait
   Fait
Fin
```

a.2 Comparaison concept par concept: comparer les concepts pères entre chaque pair des profils.

Algorithme 3 : Comparaison concept par concept entre chaque pair des profils

```
Liste Lcp1 ← {Père1,….. Pèren}  /** liste des concepts pères de profil utilisateur 1 **/
Liste Lcp2← {Père1,….. Père m}  /** liste des concepts pères de profil utilisateur 2 **/
Vecteur Vij                     /** vecteur de résultat de comparaison {0 ou 1} **/
Procédure comparaison (Liste Lcp1, Liste Lcp2)
Début
  Pour i=1 à n faire
    Pour j=1 à m faire
      Si (père i = père j) Alors Vij = 1 ;
                           Sinon Vij = 0 ;
    Fait ;
  Fait ;
FIN
```

Exemple : pour expliquer l'algorithme k-means et compléter le tableau suivant (Tableau 2) nous avons:

***Profil 1*= P_1** (RDF, OWL, Classification, Raisonnement basé sur mémoire, Weka) ;
Où **P_1** est la graine **G_1**.

Profil 2= P_2 (Politique de Sécurité, Evaluation des Risques, Gestion des Risques).

En appliquant la comparaison des concepts :

Comparaison (GetConceptsPeres (RDF), GetConceptsPeres (Politique de Sécurité)) → 0
Comparaison (GetConceptsPeres (RDF), GetConceptsPeres (Evaluation des Risques)) → 0
⋮

En appliquant les calcules pour tous les profils nous obtiendrons le tableau suivant :

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}	P_{13}	P_{14}	P_{15}	P_{16}	P_{17}	P_{18}	P_{19}	P_{20}	P_{21}
G_1	2,5	0	0,5	0	0	0	0	1,6	0	0	0	0	1,8	0	0	0	0	2,2	0	1,3	0
G_2	0	1,5	0	0	0	0	0	0	0,8	0	0	1,2	0	0	0	0	0	0	0	0	0,8
G_3	0,5	0	1,2	0	0	0	0	1,1	0	0	0	0	0,4	0	1,3	0	0	0,5	0	0,3	0,8
G_4	0	0	0	1,5	0	0	0,9	0	0	0	1,5	0	0	0	0	0	1	0	1	0	0
G_5	0	0	0	0	1,7	0,4	0,3	0	0,8	1,5	0	0	0	0,4	0	1,7	1,1	0	0,2	1,1	0
G_6	0	0	0	0	0,4	2	1,4	0	0	0	0	0	0	2	0	0	0	0	1,2	0	0
Max	2,5	1,5	1,2	1,5	1,7	2	1,4	1,6	0,8	1,5	1,5	1,2	1,8	2	1,3	1,7	1,1	2,2	1,2	1,3	0,8
Aff	G_1	G_2	G_3	G_4	G_5	G_6	G_6	G_1	G_2;G_5	G_5	G_4	G_2	G_1	G_6	G_3	G_5	G_5	G_1	G_6	G_1	G_2;G_3

TAB. 2 – *Calcul des distances entre chaque graine et chaque profil*

- **Max**: c'est la valeur maximale parmi les distances disponibles.
- **Aff**: c'est le numéro du graine où les profils seront affectés.

D'après cette première affectation, nous avons les associations suivantes :

- **Graine1(P_1)** : < P_1, P_8, P_{13}, P_{18}, P_{20}>
- **Graine2(P_2)** : < P_2 P_9, P_{12}, P_{21} >
- **Graine3(P_3)** : < P_3, P_{15}, P_{21} >
- **Graine4(P_4)** : < P_4, P_{11} >
- **Graine5(P_5)** : < P_5, P_9, P_{10}, P_{16}, P_{17} >
- **Graine6(P_6)** : < P_6, P_7, P_{14}, P_{19} >

a.3 Pondération de la distance

D'après le Tableau 2, nous trouvons des distance plus proches entre eux, pour cela nous avons effectué une pondération des distances ayant une différence **D** ; tel que **D** entre [**0 ; 0,5**].

Nous obtenions les nouvelles associations :

- **Graine1(P_1)** : < P_1, P_8, P_{13}, P_{18}, P_{20}>
- **Graine2(P_2)** : < P_2, P_9, P_{12}, P_{21}>.
- **Graine3(P_3)** : < P_3, **P_8**, P_{15}, P_{21}>
- **Graine4(P_4)** : < P_4, **P_7**, P_{11}, **P_{17}**, **P_{19}**>.
- **Graine5(P_5)** : < P_5, P_9, P_{10}, P_{16}, P_{17}, **P_{20}** >
- **Graine6(P_6)** : < P_6, P_7, P_{14}, P_{19} >.

Pour obtenir les nouveaux centroïdes, nous prenons le $\mathbf{\underset{1..N}{Max}\ Dist(Pi, Gj)}$ de chaque cluster.

- Max_Graine1 = **P_1** - Max_Graine2 = **P_2**
- Max_Graine3 = **P_{15}** - Max_Graine4 = **P_4**
- Max_Graine5 = **P_{16}** - Max_Graine6 = **P_6**

Nous réitérons le processus précédent à ces nouvelles valeurs, ce qui nous permet d'obtenir le tableau suivant :

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}	P_{13}	P_{14}	P_{15}	P_{16}	P_{17}	P_{18}	P_{19}	P_{20}	P_{21}
G_1	2,5	0	0,5	0	0	0	0	1,6	0	0	0	0	1,8	0	0	0	0	2,2	0	1,3	0
G_2	0	1,5	0	0	0	0	0	0	0,8	0	0	1,2	0	0	0	0	0	0	0	0	0,8
G_{15}	0	0	1,3	0	0	0	0	0,9	0	0	0	0	0	0	1,5	0	0	0	0	0	0,8
G_4	0	0	0	1,5	0	0	0,9	0	0	0	1,5	0	0	0	0	0	1	0	1	0	0
G_{16}	0	0	0	0	1,7	0	0	0	1	1,7	0	0	0	0	0	2	1,2	0	0	1,2	0
G_6	0	0	0	0	0,4	2	1,4	0	0	0	0	0	0	2	0	0	0	0	1,2	0	0
Max	**2,5**	**1,5**	**1,3**	**1,5**	**1,7**	**2**	**1,4**	**1,6**	**1**	**1,7**	**1,5**	**1,2**	**1,8**	**2**	**1,5**	**2**	**1,2**	**2,2**	**1,2**	**1,3**	**0,8**
Aff	G_1	G_2	G_{15}	G_4	G_{16}	G_6	G_6	G_1	G_{16}	G_{16}	G_4	G_2	G_1	G_6	G_{15}	G_{16}	G_{16}	G_1	G_6	G_1	G_2 ; G_{15}

TAB. 3 – *Calcul des distances entre chaque profil et les nouvelles graines*

Après cette deuxième affectation, nous obtienderons les associations suivantes :

- **Graine1(P_1)** : < P_1, P_8, P_{13}, P_{18}, P_{20}>
- **Graine2(P_2)** : < P_2, P_9, P_{12}, P_{21}>
- **Graine3(P_{15})** : < P_3, P_{15}, P_{21}>
- **Graine4(P_4)** : < P_4, P_7, P_{11}, P_{17}, P_{19}>
- **Graine5(P_{16})** : < P_5, P_9, P_{10}, P_{16}, P_{17}, P_{20} >
- **Graine6(P_6)** : < P_6, P_7, P_{14}, P_{19} >

La figure suivante représente les profils dans leurs communautés :

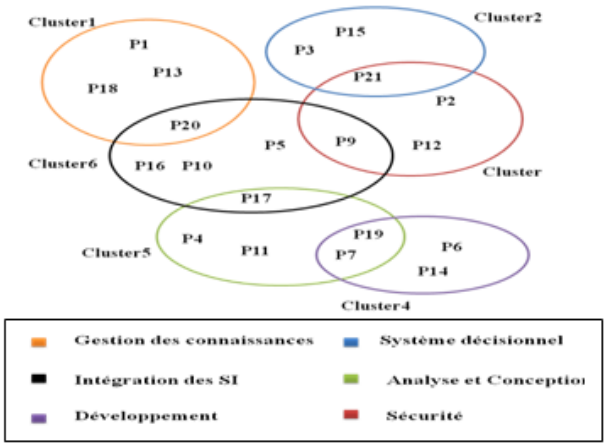


FIG.3 – *Les profils en communautés*

b. Evaluation des résultats

L'évaluation du résultat permet d'estimer la qualité des clusters, c'est-à-dire sa capacité à déterminer correctement les valeurs qu'il est censé avoir appris sur des cas nouveaux. Un critère général pour évaluer les résultats du clustering consiste à comparer la partition calculée avec une partition "correcte". Dans notre travail, nous avons utilisé la **Mesure par silhouette de la qualité des clusters** (Rousseeuw, 1987). Afin de mesurer la qualité de la classification, un indice de cohésion et de séparation entre les classes appelé « le coefficient de la silhouette moyenne (SC) » est calculé de la manière suivante :

$$SC_k = \frac{1}{N}\sum_{i=1}^{N} S(i)$$

Le but de Silhouette est de vérifier si chaque profil a été bien classé. Pour cela, et pour chaque profil **i** de la partition, on calcule la valeur suivante :

$$-1 \leq S(i) = \frac{B_i - A_i}{\max(A_i, B_i)} \leq 1$$

- ***k*** : le nombre de communautés (k = 6) ;
- ***N*** : le nombre total des profils d'une communauté ;
- A_i : représente la distance moyenne qui sépare le profil i des autres profils du cluster à laquelle il appartient ;
- B_i : représente la distance moyenne qui sépare le profil i des profils appartenant au cluster le plus proche.

Cluster le plus proche = **max dist (x,g_k)** : Distance maximale entre le profil (x) et le centre de gravité de tous les autres clusters (g_k).

- Quand **S(i)** est proche de 1, le profil est bien classé : la distance qui le sépare de la communauté la plus proche est très inférieure à celle qui le sépare de sa communauté.
- Par contre, si **S(i)** est égal de -1, cela veut dire que le profil est mal classé.

- Mais si **S(i)** est proche de 0 alors il pourrait également être classé dans la communauté la plus proche (Rousseeuw, 1987). En appliquant les calculs pour toutes les communautés, nous obtiendrons les résultats (nous donnant les résultats de 2 communautés) suivants :

	Communauté$_1$ "Gestion des connaissances" Centre_gravité =P_1				
	P_1	P_8	P_{13}	P_{18}	P_{20}
S (i)	1	-0,3	1	1	-0,1
$\boldsymbol{SC_1} = \frac{1}{5}\sum_{i=1}^{5} \boldsymbol{S(i)}$ = 0,52					

	Communauté$_2$ " Sécurité " Centre_gravité =P_2			
	P_2	P_9	P_{12}	P_{21}
S (i)	1	0,1	1	0,2
$\boldsymbol{SC_2} = \frac{1}{4}\sum_{i=1}^{4} \boldsymbol{S(i)}$ = 0,58				

TAB. 4 – *Résultats de la qualité de Communauté$_1$ et Communauté$_2$ par le coefficient de la silhouette*

Nous remarquons d'après les résultats obtenus du Tableau 4 que :
S (P$_1$) = ***S (P$_{13}$)*** = ***S (P$_{18}$)*** =**1,** signifie que les profils ***P$_1$***, ***P$_{13}$***, ***P$_{18}$*** sont bien classés (la distance qui les sépare de la communauté la plus proche est égal à **0**) c'est-à-dire ***P$_1$***, ***P$_{13}$***, ***P$_{18}$*** n'apparient qu'à la communauté$_1$ " **Gestion des connaissances** ".
Et ***S (P$_8$)*** = -0,3, signifie que profil **P$_8$** est classé dans la communauté la plus proche (Communauté$_3$ " **Système décisionnel** ").

Rousseeuw (1987) propose l'interprétation suivante du coefficient SC : $SC = \max_k SC_k$

- SC ≥ 0.7 : forte structure des clusters ;
- 0.5 ≤ SC ≤ 0.7 : structure raisonnable ;
- 0.25 ≤ SC ≤ 0.5 : structure faible ;
- SC ≤ 0.25 : pas de structure.

	Communauté$_1$	**Communauté$_2$**	**Communauté$_3$**	**Communauté$_4$**	**Communauté$_5$**	**Communauté$_6$**
SC_k	**0,52**	**0,58**	**0,17**	**0,52**	**0,25**	**0,4**

TAB. 5 – *Coefficient de la silhouette moyenne de six communautés*

Donc suivant le résultat de **SC = 0,58**, les communautés sont de structure raisonnablement.

4 Conclusion

Cet article présente notre solution pour la construction de communautés d'utilisateurs, l'avantage de cette construction est qu'elle permettra une recherche d'information selon les centres intérêts de la communauté où l'utilisateur appartient. L'acquisition implicite des nouveaux profils se basera sur le modèle de la communauté (exemple de profils), et la reformulation ou l'enrichissement des requêtes se fera par les centres d'intérêts de la communauté où les utilisateurs appartiennent.

Nous avons, à cet effet, procédé à une segmentation des profils en valeur des centres d'intérêts où nous avons appliqué l'algorithme K-means. Pour valider notre démarche, nous avons appliqué la mesure par silhouette de la qualité des clusters. Les résultats étaient satisfaisants.

Nous envisageons dans les travaux futurs, d'élaborer des règles d'affectations à l'aide des techniques de classification pour permettre la classification d'un nouveau profil utilisateur dans une des communautés (exemple : K-plus proche voisins), et de tester notre solution avec d'autres algorithmes, pour effectuer une étude comparative.

Références

Aït el Hadj, F.S (2013). *Approche de détection de communautés chevauchantes dans des réseaux bipartis*. Thèse de doctorat. Université mouloud mammeri de tizi-ouzou.

Celeux, G., Diday, E., Govaert, G., Lechevallier, Y., Ralam-Bondrainy, H. (1989). *Classification Automatique des Données*. Bordas, Paris.

Combe, D. (2013). *Détection de communautés dans les réseaux d'information utilisant liens et attributs*. Thèse de doctorat Université Jean Monnet - Saint-Etienne.

Fortunato, S. (2009). *Community Detection in Graphs*. Physics Reports 486 (3-5) (June 3): 103.

Khatoon, M., Banu, W. A (2015). A Survey on Community Detection Methods in Social Networks. I.J. Education and Management Engineering, 2015, 1, 8-18. Doi: 10.5815/ijeme.2015.01.02

Kirsch, S. M. (2005). *Social Information Retrieval*. Thèse de doctorat, Université de Rheinische Friedrich-Wilhelms.

Papadopoulos, S., Kompatsiaris, Y., Vakali, A., and Spyridonos, P. (2011). *Community Detection in Social Media*. Data Mining and Knowledge Discovery (June): 1–40.

Rousseeuw, P.J. (1987). *Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis*. Journal of Computational and Applied Mathematics, 20.

Tamine-Lechani, L., Boughanem, M., Chrisment, C. (2006). *Accès personnalisé à l'information: Vers un modèle basé sur les diagrammes d'influence*. Information - Interaction - Intelligence (I3), vol. 6, no 1, p. 69–90, Cépaduès Editions.

Yang, B., Dayou, L., Jiming, L., and Borko, F. (2010). *Discovering Communities from Social Networks: Methodologies and Applications*. Ed. Borko Furht. Boston, MA: Springer US. doi:10.1007/978-1-4419-7142-5.

Summary

User communities are held together by a common interest in a field of knowledge and are driven by a desire and a need to share ideas, experiences, models, tools and best practices.

In this work, we present our solution for the detection of user communities taking into account the interests and based on data mining technology. The interests will be introduced from an ontology dedicated to the field of computing. The data mining will be used for segmentation of communities through the use of K-means algorithm.

This construction of communities will help the information retrieval system in the implied acquisition of new profiles by selecting or adapting a predefined profile, and the reformulation of user queries based on the new knowledge acquire from the community where it belongs.

Elaboration d'un modèle artificielle de filtrage de SPAM basé sur les fonctions rénales humaines

Mohamed Amine BOUDIA*, Reda Mohamed HAMOU**, Abdelmalek AMINE***

Laboratoire de Gestion des Connaissances et des Données Complexes (GeCoDe Lab)
Department d'informatique, , Université Dr. Tahar Moulay Saida, Algeria

mamiamounti@yahoo.fr * hamoureda@yahoo.fr ** abd_amine1@yahoo.fr ***

Résumé. Dans le monde connecté d'aujourd'hui ; La communication électronique a pris un grand élan, l'une des méthodes les plus utilisées, est le courrier électronique pour son efficacité et rentabili-té. Les e-mails indésirables (messages indésirables) ou SPAMS sont largement répandues, constituent une part importante dans la boîte de réception et peuvent même être considérer comme une attaque (contenir des virus ou trojan). La détection et le filtrage de SPAM est un enjeu majeur pour la communauté Internet, car il réduit le gaspillage des ressources (ressources spatiales et temporelles) : les e-mails de SPAM doivent être filtré et séparé de celles non-SPAM (HAM).
Dans cet article, nous proposons une méta-heuristique basée sur le Système rénal pour la détection et le filtrage des SPAM. Nous nous sommes inspirés du modèle naturel du système rénal surtout de sa fonctionnalité de purification de sang et le filtrage des toxines ainsi que la régularisation de la pression artérielle. Les messages sont représentés par un sac mots en N-Gram qui est indépendant des langues (parce qu'un email peut être reçu dans n'importe quelle langue).

Mots-clés : Détection, Filtrage, SPAM, Système Rénal.

1 Introduction

La démocratisation a fait d'internet un outil très puissant pour le partage de l'information; elle fournit email, chat, et audio / vidéo conférence pour la communication notamment l'email qui est largement utilisé pour les communications officielles et non officielles, car il est gratuit pour les utilisateurs et il fournit également un transfert de fichiers.

Selon le rapport le plus récent (2014) du Radicati Group (groupe fournissant des recherches quantitatives et qualitatives détaillées sur le courrier électronique, la sécurité, la messagerie instantanée, les réseaux sociaux, l'archivage des données, la conformité réglementaire, les technologies sans fil, les technologies du web, les communications unifiées) il y avait : 2.9 billions de comptes emails actifs dans le monde; 2.4 milliards de personnes qui utilisent les mails régulièrement et elles seront 3 % de plus par an de 2013 à 2017 pour dépasser les 2,7 milliards de personnes et 67 000 milliards est le nombre de mails envoyés en 2013 soit 182,9 milliards chaque jour dans le monde en moyenne. Ce nombre passera à 206,6 milliards en 2017.

Le courrier non sollicité (SPAM), peut atteindre selon les mêmes rapports de Radicati Group plus que 89,1 % soit 262 millions de SPAMS par jour. Ce qui rend le spamming un phénomène mondial. Selon le CNIL (la Commission Nationale de l'Informatique et des Libertés): «Le" spamming "est d'envoyer du courrier électronique massif et parfois répété, non sollicités, à des personnes avec lesquelles l'expéditeur n'a eu aucun contact et dont il a capté l'adresse électronique de façon irrégulière. ». Bien que la décision "SPAM / non-SPAM" est le plus souvent facile à prendre pour un être humain. Le nombre d'email en circulation que nous venons de citer, empêche d'aborder le tri manuel des e-mails.

La littérature donne deux grandes approches pour le filtrage et la détection de SPAM: l'approche basée sur l'apprentissage machine et l'approche non basée sur l'apprentissage machine. La première approche se base sur la sélection des caractéristiques qui est une étape importante dans les systèmes de classification. Elle vise la réduction du nombre de caractéristiques tout en essayant de préserver ou d'améliorer la performance du classifieur utilisé. Tandis que la deuxième approche se repose sur de nombreuses techniques et algorithmes existants : l'analyse de contenu, les listes noires, listes blanches et l'authentification de boîte aux lettres ainsi sur des heuristiques et métaheuristiques.

Dans le corps humain, un processus important à la survie se produit automatiquement, qui est la purification du sang par le système rénal. L'homme peut mourir si le taux des toxines et les matières indésirable se trouvant dans le sang dépassent un certain seuil; le système rénal filtre et purifie le sang d'une manière automatique et d'une façon très délicate et précise ; un autre rôle du système rénal est la régularisation de la pression artérielle.

Nous proposons une méthode inspirée du système rénal pour la détection et le filtrage du SPAM avec une hybridation des deux approches (basé et non basé sur l'apprentissage) ainsi que l'utilisation de plusieurs technique dans le même système de filtrage de SPAM notamment : l'analyse de contenu, les listes noires, listes blanches. Une autre partie de notre approche contrôle le flux des emails qui représente un des rôles du système rénal afin de minimiser le risque d'attaque DDoS (déni de service).

Notre approche consiste à combiner différentes propriétés positives de ces techniques de filtrage à différents niveaux en les déployant dans une approche hybride. Notre étude vise à répondre aux questions suivantes : Le meilleur filtre qu'on peut trouver dans la nature est celui du rein humain c'est une de nos motivation qui nous ont poussé à s'inspirer de ce filtre naturelle. C nous pouvons mimer ce phénomène nous pensons que nous attendrons des résultats probants dans le domaine des filtrages de spam par exemple.

2 Représentation des données

Les algorithmes d'apprentissage ne peuvent pas traiter directement les données non structurées: image, vidéo, et bien évidemment les textes ; ce qui nous oblige à passer par une étape d'indexation. L'indexation est tout simplement une représentation de texte en numérique, mais ce passage du texte au numérique est très délicat et très important au même temps : une mauvaise représentation entraîne sûrement de mauvais résultats.

Dans cette étape on représente chaque texte par un vecteur, dont les composantes sont des termes de tous les textes, et à qui on associe un poids. De cette façon on aura un vecteur qui représente le texte et qui est exploitable par les algorithmes d'apprentissage en même temps.

La caractéristique principale de la représentation vectorielle est que chaque langue est associée à une dimension particulière dans l'espace vectoriel. Deux textes utilisant les mêmes segments textuels sont donc projetées sur des vecteurs identiques (Hamou, 2010) .

Plusieurs approches pour la représentation des textes existent dans la littérature, parmi lesquels la représentation en sac des mots qui est la plus simple et la plus utilisée, la représentation sac de phrases, les racines lexicales et bien sûr la représentation n-grammes qui est une représentation indépendante du langage naturel (Shannon, 1948).

2.1 Un nettoyage basique

Aucun nettoyage ne doit être appliqué sur le texte, car les caractères spéciaux (#, \, [,].....) et les émoticônes peuvent être utile pour détecter un SPAM , la seul opération qui va être appliqué sur le texte est de remplacer les espaces les saut de ligne et les tabulation par « _ » puis de remplacer les suite succesive «_ » par une seul pour éviter les token (après tokinisation) à taille égale à zéro.

2.2 Choix de terme

Dans notre étude on utilisera une des deux représentations :

- représentation en sac de mots (N-Gramme caractère à N= 2 , 3 ou 4) :
- représentation en sac de mots (N-Gramme mot à N= 1) :

De nombreux travaux ont montré l'efficacité des n-grammes comme méthode de représentation des textes (Dunning,1994), (Damashek, 1995), (Miller et al., 1999); (Teytaud and Jalam, 2001);. Cette technique présente des points forts, par rapport à d'autres techniques :

- Les n-grammes capturent automatiquement les racines des mots les plus fréquents sans passer par l'étape de recherche des racines lexicales.
- L'indépendance de langue comme le montre
  - ✓ Les n-grammes de caractères sont tolérants aux fautes d'orthographes et au bruit pouvant être causés lors de l'utilisation de lecteurs optiques
  - ✓ Ses deux représentations sont introduites dans le cadre du modèle vectoriel.

2.3 Pondération

Une fois que la matrice document –terme (document = texte (email ou message), terme = token) est prête, On calcule la pondération du matrice document-terme en utilisant un des codages connus (tf-idf, ou tfc) .

Le poids d'un terme t_k dans le texte i (message (email)) est calculé ainsi :

2.3.1 TF-IDF

$$TfIdf(t_k, p_i) = Nb * \log(A / B)$$

Nb : le nombre d'occurrences du terme tk dans le texte i (message (email)) ;
A : le nombre total de documents(message (email)) dans le corpus
B : le nombre des messages (emails) dans les quels le terme tk apparaît au moins une fois.

2.3.2 TFC

$$\mathrm{tfc}(t_k , p_i) = \frac{\mathrm{tf} - \mathrm{idf}\,(t_k , p_i)}{\sqrt{\sum_{i=1}^{|p|} \mathrm{tf} - \mathrm{idf}\,(t_k , p_i)^2}}$$

Afin d'analyser automatiquement un ensemble de données, il est nécessaire d'avoir un opérateur pour évaluer avec précision les similitudes ou les différences qui existent dans les données. Sur cette base, nous utiliserons les distances et les mesures de similarité. On dit que deux documents sont proches l'un de l'autre si la distance entre eux est faible. On dit que deux documents se ressemblent s'ils sont similaires. Si la distance entre le document D1 et le document D2 est faible alors leur similarité est grande.

3 Notre Approche

Notre travail consiste à modéliser une méthode bio-inspiré qui est le Système rénal aux problèmes informatique en occurrence la détection et filtrage des SPAM. Avant d'expliquer et de détailler notre approche on doit d'abord décrire le modèle naturel du fonctionnement du système rénal et mettre la lumière sur les aspects qui nous ont orienté à choisir cette méta-heuristique. Puis on dresse un tableau de modélisation (modèle naturel vers le modèle artificiel) enfin on expliquera le modèle artificiel qui est la pulpe de notre approche.

3.1 Modèle Naturel

3.1.1 Pourquoi le système rénal pour filtrage de SPAM?

Le rôle du système rénal est : nettoyer, purifier le sang et régler la pression artérielle. Notre choix de modéliser le système rénal pour le filtrage de SPAM à été suite a cette superposition

	Le système rénal	Le filtrage de SPAM
Entrée INPUT	Sang	Les emails
Résultat OUTPUT	Deux sorties : • Sang purifié (retour à la circulation sanguine) • Urine : vers la vessie puis l'extérieur du corps En plus la régularisation de la pression artérielle	Deux classes : • HAM • SPAM
Type de processus	Automatique et continu (temporellement)	Automatique et continu
Tolérance pannes	Un sujet humain peut vivre avec un seul rein	Doit tolérer aux pannes et aux bugs

TABLEAU 1 : SUPERPOSITION ENTRE LE SYSTEME RENAL ET LE FILTRAGE DE SPAM

Cette superposition donne une idée préalable sur la faisabilité de cette modélisation, le fonctionnement du système rénal est automatique (indépendant du cerveau) et très précis : l'erreur peut être fatale ou gravement pathologique pour le sujet humain.

3.1.2 La physiologie Système Rénal

Le système rénal est constitué de : Deux reins : notons qu'un sujet humain peut vivre avec un seul rein, Deux Uretères, Une vessie et Un urètre .

Nous allons focaliser sur les reins car c'est l'unité fonctionnelle du Système rénal : le rein est situé dans la loge rétro péritonéale. Le rôle principal du rein est d'assurer l'élimination des produits toxiques et substance exogènes du sang. Et il joue un rôle dans la régularisation de la pression artérielle. Le rein est formé de 2 zones distinctes : la médullaire profonde centrale qui est formée par 8 cônes (4 à 18) appelés les pyramides de Malpighi dont l'extrémité interne dénommée papille, fait saillie dans les calices et dont la base externe jouxte le cortex périphérique; et le cortex périphérique.

Les pyramides de Malpighi contiennent des milliers de néphrons sachant que le néphron constitue l'unité fonctionnelle du rein. Le néphron est l'unité fonctionnelle du rein. Chaque rein comporte plus d'un million de néphrons situés dans le tissu interstitiel et cheminent les vaisseaux et les nerfs. Chaque néphron comporte deux parties (Figure 1) :

- Corpuscule Rénal : qui est constitué de : Glomérule et Capsule de Bowman (Figure 2) : un filtre entre le sang et l'urine (seule les composants à éliminé peuvent passer pour produire l'urine primitive) ressemble à une passoire permettant de retenir une partie des éléments qu'il contient de grande taille et de laisser passer les autres (les plus petits).
- Tubule Rénal constitué de : Tubule Contourné Proximal, Anse de Henlé, Tubule Contourné Distal et Tubule Collecte

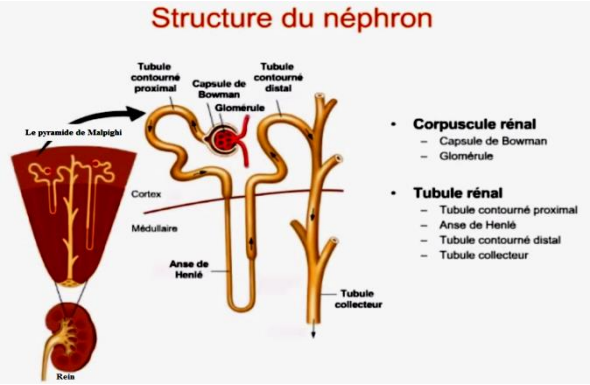


FIGURE 1 : STRUCTURE DU NEPHRON

FIGURE 2 : CAPSULE DE BOWMAN

Figure 3 : différents types de néphrons

Il existe deux types de néphrons:

- Le néphron cortex : caractérisé par son anse de Henlé court
- Le néphron juxta médullaires : caractérisé par son anse de Henlé long (ce type de néphron sert à alimenter le rein)

3.1.3 Fonctionnement du Système Rénal

Chaque néphron est associé à deux lits capillaires : le glomérule responsable de la filtration glomérulaire afin de produire l'urine primitive, et le lit capillaire péritubulaire où les processus de réabsorption et de sécrétion ce produisent, une fois ces deux processus achevé on aura l'urine définitive.

Le fonctionnement de Système rénal se divise sur deux étapes :

Etape 1 : La filtration glomérulaire

Le sang arrive par l'artériole afférente et entre dans la chambre glomérulaire (glomérule + capsule de bowman) pour subir la filtration glomérulaire. La filtration glomérulaire est un processus mécanique passif non sélectif: ne consommant pas de l'énergie, la pression artérielle dans le capillaire glomérulaire représente l'élément dynamique de la filtration. Il n'est pas sélectif car toute molécule qui a une taille inférieure au trou de capsule de Bowman sera filtrée. La filtration glomérulaire s'arrête lorsque la pression artérielle chute au dessous du 60 mm Hg. Une fois l'étape achevée, l'urine primitive suit le trajet de Tubule Rénal où sa composition sera modifiée lors de la deuxième étape ; le sang filtré primitivement rejoint l'artériole efférente, cette dernière (l'artériole efférente) va contourner le Tubule Rénal formant ainsi le capillaire péritubulaire.

Etape 2 : transfert tubulaire

La composition de l'urine primitive produite par la filtration glomérulaire sera modifiée dans le tubule rénal par deux processus qui se passe en parallèle:

- ✓ Réabsorption : qui consiste en un transfert de certains constituants de l'urine primitive vers le capillaire péritubulaire (l'eau, les sels minéraux, glucose …)
- ✓ La sécrétion : les substances toxique ou exogènes qui on échappé à la filtration glomérulaire sont ajouté à l'urine tubulaire

Au niveau du tubule collecteur où se fera la régulation définitive du volume et de l'acidité des urines par le stimulus ADH (pour faire la deuxième réabsorption de l'eau et acidifié l'urine) ou une inhibition ADH (afin de diluer l'urine et régularisation de l'équilibre des liquide dans le sang).

A la fin de processus on aura : de l'urine définitive qui va être conduite par les uretères vers la vessie puis vers l'extérieur du corps

Le Sang purifié qui va rejoindre la circulation sanguine générale (le bout final du capillaires péritubulaire rejoint veine intralobulaire) . Notant que la pression artérielle sera stabilisée à la normale en amont.

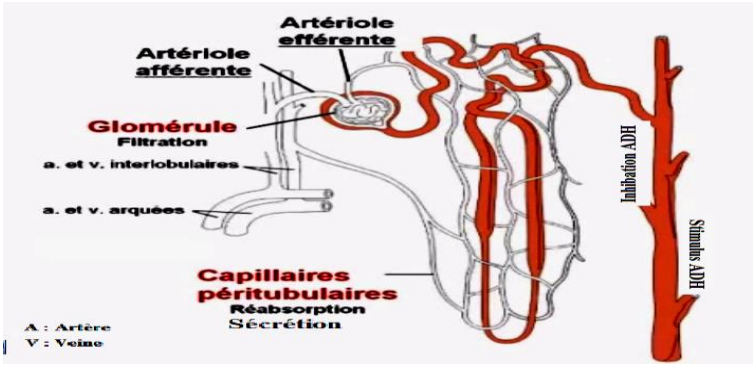


FIGURE 4 : FONCTIONNEMENT DE NEPHRON

3.2 Tableau de modélisation : le passage du modèle naturel au modèle artificielle

Modèle Naturelle	Modèle Artificielle
Consommation rénale = 10 -15 % de O^2	Base d'apprentissage = 10-15 % des SpamBase
Néphron (plusieurs néphrons)	Agent filtre (système de filtrage)
Capsule de Bowman (diamètre de filtre)	Capsule de Bowman artificielle (Seuil de filtrage)
Sang venant dans Artériole Afférente = 20% du sang cardiaque / nombre de néphron Chaque minute	Message entrant à traiter = 20% du nombre de messages total /par le nombre d'agent Chaque itération
La filtration glomérulaire	La filtration glomérulaire artificielle(Naive Bayes)
Transfert tubulaire ✓ Réabsorption ✓ Sécrétion	Optimisation: K-Means/(- : moins de, + : plus de) ✓ (- Faux Négative) (+ de Vrai Positive) ✓ (- Faux Positive) (+ de Vrai Négative)
Résultat final : ✓ Sang purifié: rejoint la circulation sanguine ✓ Urine définitive : extérieur du corps	Résultat final : ✓ SPAM définitive (après l'optimisation) ✓ HAM définitive (après l'optimisation)
Régularisation de pression artérielle, maintenir le fonctionnement de filtration glomérulaire.	Anti Attaque DDoS , déclencher et stopper le processus de filtrage
✓ Néphron cortex ✓ Néphron juxtamédullaires	✓ Pas de mise à jour de la base d'apprentissage ✓ Mettre à jour de la base d'apprentissage
Stimulus ADH = Urine concentrée	Boite noire et la bloque liste (plus de SPAM)
Inhibition ADH = Urine diluée	Boite Blanche (Moins de SPAM)

TABLEAU 2 : PASSAGE DU MODELE NATURELLE AU MODELE ARTIFICIELLE

Dans le tableau ci-dessus nous expliquons d'une façon générale la correspondance entre le modèle naturel du système rénal et celui du modèle artificiel qui est le filtrage des SPAM.

3.3 Modèle Artificiel

Dans notre approche dédiée à la détection et filtrage des SPAM nous proposons de modéliser le système rénal d'où le modèle artificiel passe par trois états :

3.3.1 Etat initial

Pour résoudre le problème du filtrage et de détection de SPAM nous utiliserons une base de test et d'une base d'apprentissage. La base d'apprentissage représente la consommation du rein de l'O2 qui est égale de 10 à 15 % du sang qui lui est amené (Tableau 2) . D'autre part, la base de test est l'ensemble des messages (SPAM et HAM) à traiter.

Initialement les reins doivent être fonctionnels et alimentés pour exercer leur rôle, dans le modèle artificiel nous commençons à remplir la base d'apprentissage à 15 % du nombre de

messages de SpamBase avec critère que la proportion des SPAM dans la base d'apprentissage sera plus important que celle des HAM [30].

Nous proposons de lancer l'application en deux threads sur le serveur, chaque thread représentera un rein pour but : d'accélérer le processus de filtrage et de donner une tolérance aux pannes : même si un thread s'arrête accidentellement ou intentionnellement l'autre thread assure le service minimum durant le temps de maintenance de la panne.

Les néphrons sont identiques : les mêmes paramètres sont généralisés pour l'ensemble de néphrons. Puisque les paramètres des néphrons sont identiques : Si un email considéré comme un SPAM par le premier néphron alors il va aussi être considéré comme un SPAM par les autre néphrons et vice versa. Chaque message (email) passe par un seul néphron aléatoirement dans chaque itération.

Dans notre modèle artificielle ; 15% de nombre de néphrons artificielle (agent filtrant) de chaque rein sont utilisés pour la mise à jour de la base d'apprentissage par les messages après avoir subit le processus de filtrage et d'optimisation, en équivalence des néphrons juxtamédullaires qui alimente le rein.

3.3.2 Etat d'activité

L'arrivée de sang (message) par l'artériole afférente augmente la pression artérielle au sein du glomérule et déclenche le processus de filtration de sang (filtrage de SPAM) ; ce processus est divisé en deux phase :

Etape 1 : Filtration glomérule artificielle:

Avant de commencer cette étape on doit affecter à chaque message un score de la telle sorte que les HAM auront un score élevé et les SPAM auront un score faible, afin de copier le modèle naturelle, ou les HAM seront les grand molécule qui ne seront pas filtré par la capsule de Bowmen (la passoire) et les SPAM seront les petite molécule qui passerai par la capsule de Bowmen (la passoire). Il existe plusieurs techniques et méthode de scoring, nous avons choisis d'utiliser le classifieur bayésien. Le score de chaque message sera égal à sa probabilité Naves Bayes d'être un HAM.

Question 1:Pourquoi avoir choisi le classifieur bayésien?

Réponse 1: « Notre choix est justifié par le fait que les performances du classifieur bayésien sont fortement et négativement corrélées avec le nombre de classes»

Le seul changement que nous avons introduit au classifieur bayésien est l'étape l'affectation de classe. Où nous allons définir un diamètre des trous de la capsule de Bowmen (passoire) ; C'est le seuil de filtrage. Les messages ayant un score plus grand ou égale au seuil restent dans le sang et rejoignent l'artériole efférante donc considérée comme HAM primitive tandis que les messages ayant un score plus petit au seuil de filtrage seront filtrés (passe de l'autre coté du capsule de Bowman ou passoire) et rejoindront le tubule rénal donc ils seront considérer comme SPAM primitive. Le filtrage glomérulaire naturelle produit jusqu'à 180 Litres d'urine primitive par jour, cela doit être repris dans le modèle artificielle par le choix d'un seuil de filtrage très élevé afin d'avoir un filtrage de SPAM très sévère.

A la fin de cette première étape on aura :

- ✓ L'ensemble de messages SPAM primitifs qui représente l'urine primitive qui rejoint le tubule rénal.
- ✓ L'ensemble de messages HAM primitifs qui représente le sang filtré primitivement et qui rejoint l'artériole efférente.

Etape 2 : Transfert Tubulaire (Optimisation) ;

Le transfert tubulaire se fait par deux processus : réabsorption et sécrétion au niveau du tubule rénal; ces deux processus se feront d'une manière automatique. Au niveau du tubule collecteur deux opération se passent en occurrence Stimulus ADH afin de récupérer l'eau et Inhibition ADH pour rejeter l'eau, ces deux opérations se font par une hormone commandée par le cerveau donc semi-automatique.

Dans notre travail, et en parallèle au modèle naturelle, nous proposons l'utilisation d'algorithme de classification de façon à ce que son état initial sera le résultat de la première étape, puisque le transfert tubulaire prends pour entrée les résultats de la filtration glomérulaire. Cet algorithme doit représenter les deux processus réabsorption et sécrétion ; nous proposons l'algorithme K-Means avec k=2, initialement les centroïde seront calculé par la classification issu de la première étape (filtration glomérulaire) comme suit : le centroïde des HAM sera calculé par l'ensemble des messages HAM primitives ; de même le centroïde des SPAM sera calculé par l'ensemble des messages SPAM primitives.

Question2: Pourquoi l'algorithme K-Means pour l'étape de transfert tubulaire?

Réponse2 : « K-Means est l'algorithme qui reflète le transfert tubulaire, le transfert tubulaire= sécrétion + réabsorption (en même temps)

- ✓ réabsorption : transfert à partir de l'urine primitive vers le sang primitive (SPAM primitive vers HAM primitive)
- ✓ sécrétion : transfert à partir du sang primitif vers l'urine primitive (HAM primitive vers SPAM primitive)

La philosophie de l'algorithme K-Means est que à chaque itération : les documents changent de classe pour rejoindre la classe dont ils sont proches de son centroïde(classe), dans notre cas pour K=2 (SPAM et HAM) : dans chaque itération des messages classés comme HAM change de classe vers le SPAM (sécrétion) et d'autre messages change de classe allant du SPAM vers HAM (réabsorption) »

Dans le modèle naturel ; à la fin des deux processus de transfert tubulaire (réabsorption + sécrétion) deux opérations suivent : le Stimulus ADH et /ou l'inhibition ADH .

En modèle artificiel, on va être représenté le Stimulus ADH par la technique de liste blanche (qui sera générée par l'user ou le fournisseur de service) et inhibition ADH sera représenté par la technique de liste noire (qui sera générée par l'user ou le fournisseur de service).

Question 3: $(\text{Stimulus ADH})_{\text{modèle naturelle}}=(\text{Liste Blanche})_{\text{modèle artificielle}}$ & $(\text{inhibition ADH})_{\text{modèle naturelle}} = (\text{Liste Noir})_{\text{modèle artificielle}}$, pourquoi ?

Réponse 3: « Le stimulus ADH vise à acidifier l'urine en lui retirant l'eau par des enzyme (semi-automatique) autrement dit : retirer que l'eau (et pas les autres composant comme glucose…ect) du l'urine vers le sang, en parallèle Les liste Blanche retire que quelque message considérer comme SPAM mais qu'ils sont réellement des HAM (Faux Négatif), de la classe SPAM vers la classe HAM. Le inhibition ADH vise à diluer l'urine en lui ajoutant l'eau par des enzyme (semi-automatique) autrement dit : ajouter que l'eau (et pas les autres composant comme glucose etc…) à l'urine à partir du sang (enlever l'eau du sang et l'ajouter vers l'urine), en parallèle les listes Noire ne retire que quelques messages considérée comme HAM mais qui sont réellement des SPAM (Faux Positif), de la classe HAM vers la classe SPAM.»

Etape 4: Mise à jour de la base d'apprentissage

Dans le modèle artificielle 15% des néphrons sont juxtamédullaires, les messages qui sont traités par ces derniers serviront à la mise à jour de la base d'apprentissage. La base

d'apprentissage ne dois pas grandir pour éviter le sur-apprentissage donc on doit faire un écrasement du message le plus répété pour créer la diversité dans la base d'apprentissage afin que le maximum de cas sera englobé par cette dernière (base d'apprentissage).

Nous supposons que HAM1 et HAM2 (respectivement SPAM 1 et SPAM2) sont les plus similaire parmi l'ensemble des messages HAM (respectivement SPAM), le message à écraser (sortant) est celui le plus proche au centroïde HAM (respectivement SPAM) et on l'appellera le Message HAM (respectivement SPAM) sortant, Le message HAM sortant sera écrasé et remplacé par une copie du HAM qui est dans le lit capillaire péritubulaire respectivement Le message SPAM sortant sera écrasé et remplacé par une copie du SPAM qui est dans l'anse de Henlé.

3.3.3 L'état final

Après un nombre fini et suffisant d'itérations ; L'ensemble de message SPAM représente l'urine définitive qui sera conduite vers l'extérieur de l'homme de façon à ne pas revenir à la circulation sanguine. L'ensemble de messages HAM représente le sang purifié qui va rejoindre la veine intralobulaire.

La base d'apprentissage sera mise à jour et les recommandations de l'user ou fournisseur de services et seront prises on considération (Liste noire et liste blanche) en faisant partie de la base d'apprentissage (assez réduite (15% de Spam Base) donc il aura un poids dans la prise de décision par le classifieur bayésien (calcul de probabilité)).

La régularisation de la pression artérielle sert dans le modèle artificiel à limiter le nombre de message à traiter par itération et cela pourra servir comme solution à la paralysie du système que peuvent provoquer les Spammeurs par une attaque DDoS (déni de service). A la fin du flux de message la pression glomérulaire sera égale à zéro ce qui remettra le filtrage de SPAM en repos. Si une vrac de message arrive par l'artériole afférente alors la pression glomérulaire remonte et sera différente de zéro ce qui re-déclenchera le système de filtrage de SPAM.

En récapitulation à la fin du notre processus que nous proposerons, les résultats seront :

- ✓ Ensemble de messages SPAM qui représente l'urine définitive.
- ✓ Ensemble de messages HAM qui représente le sang purifié.
- ✓ Une base d'apprentissage mise à jour pour la prochaine utilisation (rein alimenté).

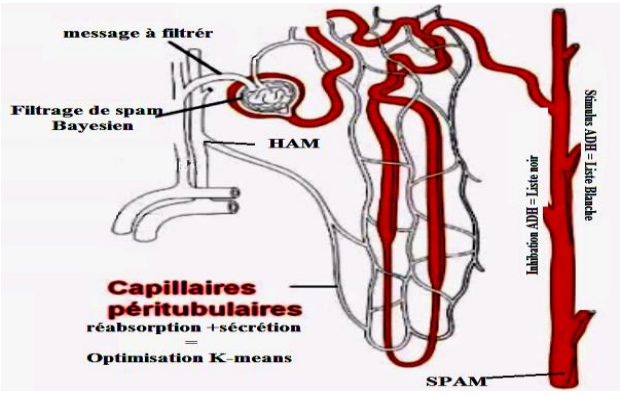


FIGURE 5 : MODELE ARTIFICIELLE VS MODELE NATURELLE

4 Conclusion et perspective

Dans cet article, nous avons proposé une nouvelle méthode bio-inspirée en l'occurrence le système rénal, nous avons élaboré un modèle artificiel pour la détection de SPAM basée sur les fonctions rénales humaines et nous avons prouvé que ce modèle basé sur le système rénal est capable de détecter et de filtrer le SPAM.

La nouvelle approche que nous proposons est robuste, tout d'abord parce que nous utilisons la notion de classifieur bayésien qui se comporte bien avec les classification à nombre de classe réduit (dans notre cas nombre de classe est égale à 2) la seul différence est que l'affectation de classe se fait par un seuil de filtrage et non pas par la probabilité la plus majoritaire ; le deuxième point fort de notre approche est que l'état initial de l'algorithme d'optimisation par K-Means n'est pas aléatoire mais se base sur les résultat de la première étape. Le troisième point fort est l'hybridation des approches de détection et filtrage de SPAM (approche fondée sur l'apprentissage et approche non-fondée sur l'apprentissage) par l'utilisation de deux algorithmes de classification pour l'approche fondée sur l'apprentissage en l'occurrence le classifieur bayésien et K-means et deux techniques de la deuxième approche en occurrence Liste Noire et Liste Blanche.

Un autre point fort de notre approche est que la base d'apprentissage n'est pas statique, en effet la base d'apprentissage se met à jour à chaque itération afin d'assurer une grande représentativité des cas possibles et d'éliminer les répétitions et les cas similaires. Cette mis à jour vise à prendre en charge les messages que nous avons récupérer/supprimé par les technique liste blanche/liste noire. Donc après un nombre d'itérations suffisantes nous pensons que la base d'apprentissage les prendra en considération.

Notre approche assure la tolérance aux pannes, en effet en lançant deux threads, même si un des deux s'arrête intentionnellement ou accidentellement l'autre thread assura le service minimum en attendant la maintenance de l'autre thread. Ainsi que notre approche à une résistance à l'attaque DDoS par la régularisation de pression de traitement, même on surchargeant le serveur par un nombre exponentiel de message notre approche ne traite qu'une part à la fois.

Un autre avantage de notre approche consiste à une optimisation spatiale et temporelle ; l'optimisation spatial se montre par la mise en pause quand la pression glomérulaire est égale à zéro (nombre message à traiter est égal à zéro) et dés qu'un message se présente au traitement le thread reprends l'activité ; tandis que l'optimisation temporelle se fait d'abord par le partage de tâche entre les deux threads, et par le grand nombre de néphrons qui travaille en parallèle.

Dans l'avenir, nous prévoyons d'expérimenter notre modèle artificielle pour le filtrage des SPAM basé sur les fonctions rénales humaines et comparer sa performance avec d'autre algorithmes et techniques de filtrage de SPAM. Nous envisageons après les expérimentations d'apporter quelque amélioration selon les résultats en appuyant sur le point fort et nous essayons de rectifier au maximum les points faibles. Nous prévoyons aussi de crée un modèle de fonctionnement en parallélisme propre à notre approche.

Référence

1. Mueller, Scott H. 2009. spam.abuse.net. Fight Spam on the Internet! [Online] spam.abuse.net, 18 April 2009. [Cited: 21 April 2009.] http://spam.abuse.net/.
2. Van Staden, F., & Venter, H. S. (2009). The State of the Art of Spam and Anti-Spam Strategies and a Possible Solution using Digital Forensics. In ISSA (pp. 437-454).
3. Faynberg, I., Lu, H. L., Perlman, R., & Zeltsan, Z. (2010). U.S. Patent No. 7,752,440. Washington, DC: U.S. Patent and Trademark Office.
4. Obied, A., & Alhajj, R. (2009). Collection and Analysis of Web-based Exploits and Malware. Journal of Applied Intelligence, 30(2), 112-120.
5. Ruan, G., & Tan, Y. (2010). A three-layer back-propagation neural network for spam detection using artificial immune concentration. Soft Computing, 14(2), 139-150.
6. Hamou, R. M., Amine, A., & Boudia, A. (2013). A New Meta-Heuristic Based on Social Bees for Detection and Filtering of Spam. International Journal of Applied Metaheuristic Computing (IJAMC), 4(3), 15-33.
7. Guzella, T. S., & Caminhas, W. M. (2009). A review of machine learning approaches to spam filtering. Expert Systems with Applications, 36(7), 10206-10222.
8. Lakshmi, R. D., & Radha, N. (2010, September). Spam classification using supervised learning techniques. In Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India (p. 66). ACM.
9. Sharma, V., & Lewis, S. (2013). U.S. Patent No. 8,489,689. Washington, DC: U.S. Patent and Trademark Office.
10. Dash, M., & Liu, H. (1997). Feature selection for classification. Intelligent data analysis, 1(3), 131-156.
11. Hamou, R. M., Amine, A., & Lokbani, A. C. (2013). Study of Sensitive Parameters of PSO Application to Clustering of Texts. International Journal of Applied Evolutionary Computation (IJAEC), 4(2), 41-55
12. Saporta, G. (2011). Probabilités, analyse des données et statistique. Editions Technip.
13. Lokbani, A. C., Lehireche, A., & Hamou, R. M. (2013). Experimentation of Data Mining Technique for System's Security: A Comparative Study. In Advances in Swarm Intelligence (pp. 248-257). Springer Berlin Heidelberg.

Abstract. In today's world of globalization and borderless technology, the explosion in communication has revolutionized the field of electronic communication. The e-mail is therefore one of the most used methods for its efficiency and profitability. In The undesirables emails (SPAM) are widely spread as they play an important part in the inbox. So that such emails must be filtered and separated from those non-SPAMS (HAMS). Several recent studies have provided the importance of filtering of SPAM is a major interest.

In the present paper, we propose and experiment a new and original meta-heuristic based on the renal system for detection and filtering spam. The natural model of the renal system is taken as an inspiration for its purification of blood, the filtering of toxins as well as the regularization of the blood pressure, we propose to use two models to apply a Bayesien classification on textual data: Bernoulli or Multinomial model.

Keywords : Detection, filtering, SPAM, kidney system.

A new approach based on the power saves of social bees for automatic summaries of texts by extraction

Mohamed Amine BOUDIA*, Reda Mohamed HAMOU**,
Abdelmalek AMINE***, Mohamed Elhadi Rahmani****

Laboratoire de Gestion des Connaissances et des Données Complexes (GeCoDe Lab)
Department d'informatique, Université Dr. Tahar Moulay Saida, Algeria

mamiamounti@yahoo.fr * hamoureda@yahoo.fr **,
abd_amine1@yahoo.fr ***, r_m_elhadi@yahoo.fr*****

Abstract. In this paper, we propose a new approach for automatic text summarization by extraction based on Saving Energy Function where the first step constitute to use two techniques of extraction : scoring of phrases, and similarity that aims to eliminate redundant phrases without losing the theme of the text. While the second step aims to optimize the results of the previous layer by the metaheuristic based on Bee Algorithm. the objective function of the optimization is to maximize the sum of similarity between phrases of the candidate summary in order to keep the theme of the text, minimize the sum of scores in order to increase the summarization rate, this optimization also will give a candidate's summary where the order of the phrases changes compared to the original text.T he third and final layer aims to choose the best summary from the candidate summaries generated by bee optimization, we opted for the technique of voting with a simple majority.

Keywords : Automatic Summary Extraction, Data Mining, Bee Algorithm, optimization, Scoring, similarity, Saving energy, text analysis, text mining.

1 Introduction

Every day, the mass of electronic textual information is increasing dramatically making it more and more difficult access to relevant information without the use of special tools. Additionally, access to the content of the texts by rapid and effective ways has become a necessity.

A summary of a text is an effective way to represent the contents of the texts and allow quick access to their semantic content. The purpose of a summarization is to produce an abridged text covering most of the content from the source text. « We cannot imagine our daily life, one day without summaries ». Newspaper head-lines, the first paragraph of a newspaper article, newsletters, weather, tables of results of sports competitions and library are all summarized. Even in the research, the authors of scientific articles must accompany their scientific articles by a summary written by them-selves. Automatic summaries can be used to reduce the search time to find the relevant documents or to reduce the treatment of long texts by identifying the key information.

To make an automatic summary, the current literature presents three approaches: summarisation by extraction, summarisation by understanding or summarisation by classification.

Our current work uses automatic summarization by extraction as it is a simple method to implement that gives good results; only in the previous work, produce the automatic summary by extraction approach consists to use only one technique at a time (Scoring of phrase, Similarity between phrase or prototype) and respects the order of the phrases in the original document. Thus, our work seeks to answer the following questions:

- What is the contribution of the use of two methods of summarization at the same time on the quality of summary?
- Can the bio-inspired method based on Energy Savings of Bees brings more for the automatic summary and increase the quality of the summary?

2 State of the Art

Automatic summarization appeared earlier as a field of research in computer science from the axis of NLP (automatic language processing), HP Luhn proposed in 1958 a first approach to the development of automatic abstracts from extracting phrases. In the early 1960s, HP Edmundson and other participants in the project TRW (Thompson Ramo Wooldridge Inc.) Proposed a new system of automatic summarization where it combined several criteria to assess the relevance of phrases to extract.

From the 1980s, theories have emerged to describe the various treatments involved in the human cognitive system in the activities of reading and text understanding, in particular the model of Kintsch and Van Dijk explained in more construction of a summary.

These theories had then greatly inspired the architecture of the automatic summary of the time. The influence of psychological theories constituted a new step in the automatic summarization compared to the previous techniques, henceforth we "understand" the text, using the knowledge from deeper cognitive structures like scripts, scheme, frames... one of these early works inspired by research in psychology was that of G. DeJong with the FRUMP system. Other important works continued to appear at that time such as SUSY, TOPIC, SCISSOR and PAULINE.

Systems for automatic summarization by an understanding of this era were strongly influenced by all the works that are done on reading comprehension and knowledge representation in cognitive psychology and artificial intelligence. Some works, such as SUSY, where even very ambitious, they considered many treatments (syntactic, semantic, etc..) Which, even today, have never been fully made, for example, the implementation of the propositional representation of text.

In some cases, the same system is evaluated multiple times with different combinations of evaluation criteria, changes in coefficients of importance, changes in coefficients for learning, etc. These are not always theoretical or cognitive considerations that guide choices in the right combination of criteria pertinences, the values of coefficients of importance, etc.

This phenomenon is even more pronounced in techniques with learning. An important event during the 90s was the TIPSTER program, led mainly by DARPA (Defense Advanced Research Projects Agency), which began in 1991 by studying the different information retrieval techniques (algorithms, software architecture, etc..). In its third and final phase, from 1996 to 1998, the TIPSTER program was interested in automatic summarization and

has an "official" evaluation of proposed at the time in the field solutions. In 1998, the first major evaluation SUMMAC (TIPSTER Text Summarization Evaluation) has been done on different systems by defining two frames of evaluation.

Finally, for the years 2003 and 2004, there seems to be a slowdown in the field of automatic summarization as well in academia, where jobs are scarcer than in firms that do not innovate in this area (see the quiet development of their product). This fact is also obviously observable in the reduction of seminars or conferences on this domain. Since 2001, following the work of Inderjeet Mani, no reference on automatic summarization has been created in English. In 2004, a number of the French journal TAL appeared in the automatic summarization works, but only gather some articles on some very specific works, and does not offer a sufficient overview of the French area. In general, this phenomenon of slowing automatic summary is parallel accompanied by new approaches to information retrieval, textual search and navigation intra and inter textual.

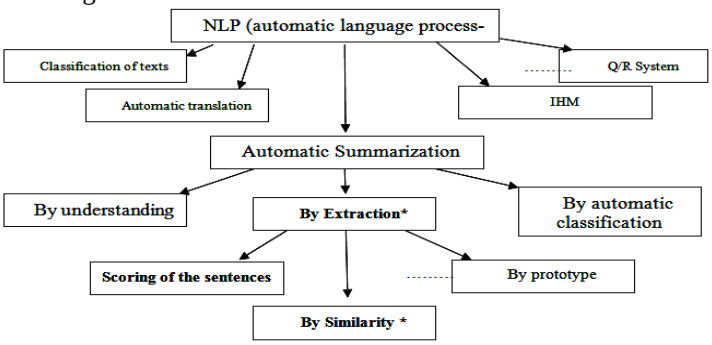


FIGURE 1. THE TECHNIQUES OF AUTOMATIC SUMMARIZATION.

3 Data representation

The machine learning algorithms cannot process directly the unstructured data: image, video, and of course the texts written in natural language. Thus, we are obliged to pass by an indexing step. Indexing step is simply a representation of the text as a vector where each entry corresponds to a different word and the number at that entry corresponds to how many times that word was present in the document (or some function of it); this is very delicate and very important at the same time: a poor or bad representation will lead certainly to bad results.

3.1 A basic cleaning

Stop words will not be removed, because the method of automatic summarization by extraction aims to extract the most informative phrases without modifying them: if we remove Stop words without information on their morph syntactic impact on the phrases, we risk having an inconsistent summary in a morphological part.

Then cleaning is to remove emoticons, to replace spaces with "_" and remove special characters (#, \, [,] .....).

3.2 Choice of term

For automatic summarization by extraction, we will need two representations:

- Bag of words representation.
- Bag of sentences representation.

Both representations are introduced in the vector model.

Finally the "word-phrases" occurrence matrix will be generated after the two previous performances, the size of this matrix is equal to (the number of words in the text)*(the number of words in the text), p_{ik} weight is the number occurrence of the word i in the phrase j;

3.3 Weighting

Once the "Word-Phrase" matrix is ready, we calculate the weighting of "Word-Phrase" matrix (we name it M) by using one of the encodings (tf-idf, or tfc) with a small modification to adapt it with the concept of a mono-document summarization.

The weight of a term t_k in the phrase pi is calculated as:

3.3.1 TF-IDF.

$$TfIdf(t_k, i) = Nb * \log(A / B) \quad (1)$$

Nb : The number of occurrences of the term tk in the text i;

A : The total number of phrase in the text;

B : The number of phrase in which the tk term appears at least once.

3.3.2 TFC.

$$tfc(t_k, p_i) = \frac{tf-idf(t_k, p_i)}{\sqrt{\sum_{i=1}^{|p|} tf-idf(t_k, p_i)^2}} \quad (2)$$

After calculating the frequency of each word, a weight is assigned to each phrase. The generated summary is then generated by displaying the highest score of the source document phrases.

4 Our proposed approach

The textual unit that is most similar to the other unit's potential for carried the theme of the text. which means to produce a good summary that has the same theme of the original text we have to maximize the sum of similarity between-sentences and at the same time reduce the sentences of sum score to increase the rate of reduction.

4.1 First Step: Primitive summary

To create a summary by extraction, it is necessary to identify textual units (phrases, clauses, phrases, paragraphs) considered salient (relevant), then we select the textual units that carry the main ideas of the document content with some order to finally build a resume.

4.1.1 Sub-step 1 : Scoring.

After a vectorization of text like we have seen in previous section (Pretreatment and Weighting), a weight or score is assigned to each phrase. The generated summary is then generated by displaying the highest score of the source document phrases.

This score of a phrase is equal to the sum of the words in this phrase:

$$\text{SCORE}\,(p_i) = \sum_{k=0}^{nbr\_word} \text{Mik} \qquad (3)$$

$\boldsymbol{M_{ik}}$ **:** weighting of "Word-Phrase" matrix **M** using tf-idf or tfc

"Suggested process claims on the principle that high-frequency words in a document are important words".

The final step is to select the N first phrases that have the highest weight and which are considered the most relevant. The process of extracting the first N phrases intended to build the summary is defined either by a threshold, in this case, the score of the phrase must be greater than or equal to the threshold in order that this phrase will be extracted the second method is to fix a number N of phrase to be extracted, all phases will be ranked in descending order according of their score, and we take only the first N phrases.

4.1.2. Sub-step 2 :Elimination of rehearsals and theme detection: using SIMILARITY method summarization by extraction.

The result of the previous sub-step is a set of phrases which is a high score. Just we have a possibility that two or more phrases have a high score but they are similar, so we proceed to the elimination of phrases that resembling. The similarity between the phrases that have been selected at the end of the previous step with known metrics (cosine Euclidean .....).

The similarity is also used to detect the phrase that has more relation with the theme of the text. According to the domain experts, it is the phrase which is most similar to the other phrases door themed text.

4.2 Second Step : Optimization using Saving energy model of Bees

4.2.1 Natural Model.

According to the lifestyle of the species, there are several types of bees. As a definition, the term "bee" is one of the common names of the European honeybee (Apis mellifera). It can also be used for any other domesticated bee by humans. The term "social bee" means a species of bee that lives in colonies, Among all categories of bees exist (Faustino, Silva-Matos, Mateus, & Zucchi, 2002). A very highly organized colony always composed of workers, drones and one queen. The workers are exclusively female bees, the most numerous of the colony (about 30,000 to 70,000 per hive). From the 22nd day of their life and until his death, she will go from flower to flower to collect nectar, pollen and propolis. Consequently, it becomes forager and brings food to the hive.

In the life cycle of bee workers, the largest labs time is during the period of collecting the nectar, pollen and propolis in the flower field. The worker bee uses a model to save energy that is spent to keep high efficiency. Roughly, is done as follows: Choose the flowers that initially are of low or no pheromone to start collecting, Minimize the weight that she carries

to every flower and look for flower with less pollen and Select the next flower in its movement in order to minimize the distances that are traveled.

A social bees communicate with the others in the taking pollen task, movement or attack; communication can be done by two different methods; by dance move or by the concentration of hormonal substances. Each dance or each concentration of substances has a specific meaning to others bees.

4.2.2 The Artificial model.

As it is known in physics, the energy is computed in term of power by time (equation 4). Also the power is equal to the applied force multiplied by the rate of distance by time (equation 5).

$$\text{Energy} = \text{power} * \text{time} \quad (4)$$

$$\text{Power} = \sum \text{force} * \frac{\sum Distance}{Time} \quad (5)$$

From this two previous equations we can conclude that the final law of energy is equal to force multiplied by the distance (equation 6)

$$\text{Energy} = \sum \text{force} * \sum \text{distance} \quad (6)$$

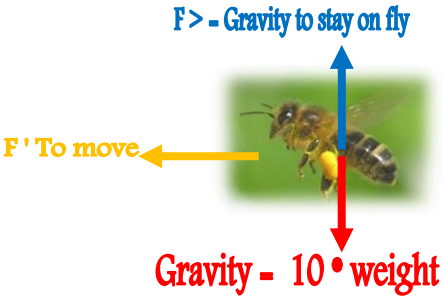


FIGURE 2. FORCE APPLY ON A BEE IN COLLECTION OF POLLEN TASK

Our inspiration is based on the force that a bee applies to stay stable in a specific level. As it known that the force applied on the bee is the gravity force which is generally related to the weight of the bee and the pollen which he collect. For that the bee must apply at least the same force on the opposite way in order to preserve its current level.

F' will be not take we will not focus on F ' because it is not our purpose in this paper.

Weight of pollen represents in our artificial model a sum score of sentence that the bee holds. From that we conclude that the force of an artificial bee in equal to the score of a phrase. Meanwhile, there is another known law that links the distance to the similarity of two objects. Simply the similarity is equal to the rate of 1 by the given distance (equation 7)

$$\sum \text{Similarity} = \frac{1}{\sum Distance} \quad (7)$$

However, for good functioning of such bee, it must conserve the maximum of its energy. To do that, a natural bee tries to get the nectar from the most closed flowers to each other. In other words, it tries to minimize distance between sources. Consequently, by looking to the equations 3 and 4, and the inspiration we made in the previous paragraph, we conclude a new fitness function for automatic summarization approach using bees' algorithm. This function

is based on minimizing loss of energy of a bee using the similarity between phrases and its scores. Our fitness function is demonstrated in equation 8 below.

$$\text{energy} = R * \frac{\sum score_{phrase}}{\sum \text{similarity}} \quad (8)$$

Where R is a constant that represents minimum carrying capacity that a bee could transfer. The objective of using this constant is to keep the fitness function in a positive position and avoid a fitness that is equal to 0.

↘ Energy ⇨ ↘ $\sum score_{phrase}$ and ↗ $\sum$ similarity

The minimization of energy is equivalent to minimizing the sum of the scores of phrases in order to minimize the summarization rate, and maximize the similarity and to preserve the theme of the candidate summary, while respecting the dice constrained utility and semantics represented by the interval :[lower threshold of summarization rate, upper threshold of summarization rate].

- **The utility constraint**: produce a summary automatically with the summarization rate higher than "upper threshold of summarization rate" is not helpful.
- **The semantic constraint**: produce a summary Automatically with summarization rate smaller than "lower threshold of summarization rate" Leads to losing a lot of semantics.

Our Artificial model is described as follow:

Environment : a flower grid (N * N). N is the square root of the number of phrases after step 1 (pre-summary and the elimination of similar phrases) where each flower is representing a phrase, the pickets have different sizes that representing the score of the corresponding phrase. Initially, there is not pheromons in enviromment; The number of bees is equal to or less than the number of phrase, initially each bee is placed on a flower (phrase) randomly.

Pollen collection and movement : According to the fitness function (energy), taken Pollen and movement will be made. We associated to the worker bee i in iteration j a Pij weight initialized to zero and it equal to the sum of the weights of k SCORE phrases whose worker bee i have visited during the iteration j.

Communication : Each bee leaves a hormonal trace on each visited flower so that other bees will not take this part of the way. Each bess keeps in mind, the best visited paths, after a number of bees iterations, every bee returns the best paths.

Path : It is a series of flower visited in chronological order, and is a summarization. Recall that each flower is a phrase (See the artificial environment).

End of the optimisation of the workers bees : when the number of iterations performed reached the maximum number of iterations, each bee returns all paths (each path, is a candidate summary) and was associated with each path (candidate summary) a set of evaluations indices and launching a voting algorithm compared these evaluation indices to choose the best candidate summary as final result in the next layer.

Candidates summaries generated by the previous layer will be evaluated by several evaluations metric, and then we will classify by pairs. R1 and R2 are two candidate summaries rate by N metric evaluation, the number of associated point to R1 represents the number of evaluation indicating that the quality of R1 is greater than or equal to R2(respectively to R2). The summary with most points will win the duel and will face another until there are no more challenger. Summary will be declared the winner as the best summary.

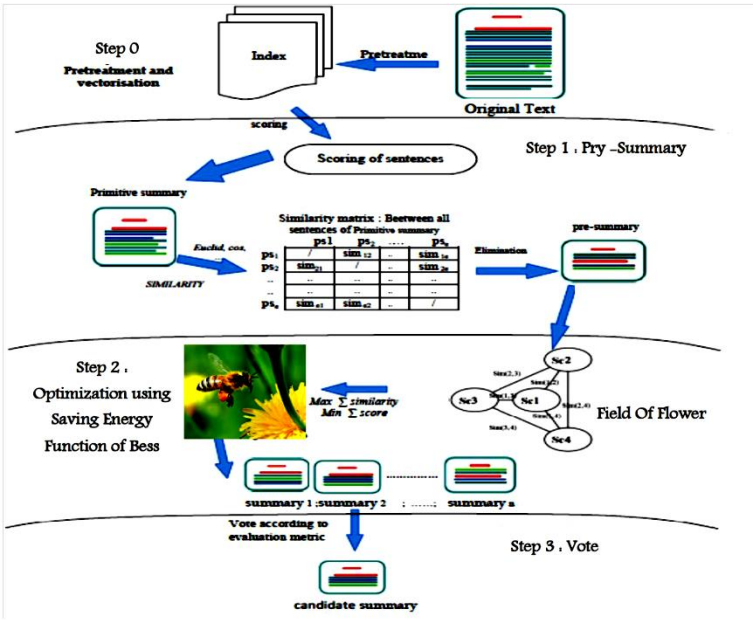


FIGURE 3. PROCESS OF THE PROPOSED APPROACH.

5 Experimentation

This work consists of four modules. The first carries the vectorization of the text and applies of scoring technique, the second module applies the similarities between phrases to eliminate repetition, the third module is to optimize the results with the socials bees and launches it's on the canvas. Finally, the fourth module choose the best abstract obtained by the previous module.

5.1 The Data Used

It was used as the text corpus "Hurricane" in French, which contains a title and 20 phrases and 313 words. As Reference summary we use three references summaries produced successively by Summarizer CORTEX, Essential Summarizer, and a summary produced by a human expert.

5.2 Results

After experiments we have grouped the results in the table below.

In Yellow: the local optimal candidate summary before optimization, just for illustration
In Green: the candidate summary before optimization used for optimization with Bees,
We used two parameter combination:

Threshold Scoring Similarity		0,70				Threshold Scoring Similarity		0.725			
Threshold Similarity	Evaluation metric	Cortex	REG	Human	Reduce rate	Threshold Similarity	Evaluation metric	Cortex	REG	Human	Reduce rate
0.65	ROUGE-SU(2)	0.7025	0.7222	0.5711	31,28%	0.65	ROUGE-SU(2)	0.6710	0.6940	0.5318	34,78%
	F-Measure	0.5433	0.4999	0.3933			F-Measure	0.6363	0.5441	0.5309	
0.70	ROUGE-SU(2)	0.6939	0.7131	0.5238	32,87%	0.70	ROUGE-SU(2)	0.6684	0.6932	0.5810	35,07%
	F-Measure	0.5800	0.5218	0.4530			F-Measure	0.6831	0.6215	0.5789	
0.75	ROUGE-SU(2)	0.6874	0.6975	0.6055	33,51%	0.75	ROUGE-SU(2)	0.6666	0.6899	0.5809	35,21%
	F-Measure	0.6074	0.5381	0.5247			F-Measure	0.6251	0.5733	0.5583	
Threshold Scoring Similarity		0.75				Threshold Scoring Similarity		0.775			
Threshold Similarity	Evaluation metric	Cortex	REG	Human	Reduce rate	Threshold Similarity	Evaluation metric	Cortex	REG	Human	Reduce rate
0.65	ROUGE-SU(2)	0.6597	0.6741	0.5518	37,22%	0.65	ROUGE-SU(2)	0.5519	0.6043	0.4933	53,33%
	F-Measure	0.6148	0.5726	0.5495			F-Measure	0.6643	0.6149	0.6666	
0.70	ROUGE-SU(2)	0.6503	0.6685	0.5426	38,42%	0.70	ROUGE-SU(2)	0.5257	0.5895	0.4825	55,43%
	F-Measure	0.6159	0.5699	0.5401			F-Measure	0.7263	0.6650	0.6767	
0.75	ROUGE-SU(2)	0.6426	0.6678	0.5417	39,45%	0.75	ROUGE-SU(2)	0.4411	0.4759	0.4747	58,13%
	F-Measure	0.6075	0.5689	0.5372			F-Measure	0.7709	0.6866	0.7245	

TABLE 1. STEP 1 : BEFORE OPTIMISATION WITH BEE ALGORITHM(PRE-SUMMARIZATION)

Combine 1 :

Threshold higher discount rate 75%=0,75.
Threshold lower discount rate 50%=0,65.
Number of iterations = 500.
Number of bees = 10.

Combine 2 :

Threshold higher discount rate 50%=0,80.
Threshold lower discount rate 30%= 0,50.
Number of iterations = 500.
Number of Bees = 10.

	Evaluation metric	REG	Cortex	Human	Nbr Word	Nbr Phrase	Reduced Rate	Time Execution
Befor Optimization	ROUGE-SU(2)	0.6684	0.6932	0.5810	217	15	35,07%	718 ms
	F-Measure	0.6831	0.6215	0.5789				
Optimization Bee Algorithm (Combine 1)	ROUGE-SU(2)	0.7859	0.7924	0.6338	207	13	35,31 %	3146 ms
	F-Measure	0.6831	0.6821	0.6289				
Optimization Bee Algorithm (Combine 2)	ROUGE-SU(2)	0.7211	0.7153	0.5949	195	12	39,06 %	3119 ms
	F-Measure	0.6912	0.6339	0.5814				

TABLE 2. OPTIMIZATION OF THE OPTIMAL SUMMARY CANDIDATE FROM STEP 1

We conducted a series of experiments to find and fix the most optimal parameters of algorithm bee. According to the experimental set of results when we set the target parameter values, it has turn out that: increasing number of iterations and the increase in bees influences the execution time, the candidate summary quality is not reached by the change of these two parameters

6 Interpretation and Discussion

We experimented document "Hurricane" using the coding TFC for the first layer (scoring) and several similarity distances (second stage) to try to detect the sensitive parameters to have the best results, we validated the result by the metric ROUGE by comparing our candidates summaries with the references summaries produced by REG, COTREX and a human expert.

ROUGE is a metric intrinsic semi-automatic evaluation based on the number of co-occurrence between a summary candidate and one or more reference summaries divided by the size of the latter. Its weakness is that it is based on the summary references and neglects the original text. The values given in ROUGE for a summary, a negligible reduction rate is high. This high value is due to the increased number of term co-occurrence between the candidate summary and abstract references.

The F-measure is one of the most robust metric used for the evaluation of classification; the F-measure is a combination of Recall and precession. For adaptation that is added to the F-measure of the strength that made, we will proceed with an extrinsic evaluation at the beginning, and continues with an intrinsic valuation: thus, a hybrid evaluation. For automatic summary reduced rate of reduction, F-Measure offers better feedback than ROUGE because it takes into account the absence of the term. Unlike ROUGE evaluation summary, high reduction rate candidate can be distorted because the false negative maximum value is what will give good evaluations for generally poor summary (high reduction rate leads to an increase in entropy of information)The accuracy indicates the purity of the candidate summary while the reminder interprets the candidate summary resemblance to the summary reference.

In table 1 : first sub-table (top left corner) indicates incoherence between the F-Measure evaluation metric and ROUGE. This incoherence is resulting from a false evaluation of ROUGE because the ROUGE metric is weak against summary with negligible reduction rate: in fact a summary has low reduction rate will have a big number of occurrences between him and a set of reference summaries (more larger than a summary has greatly reduced rates).

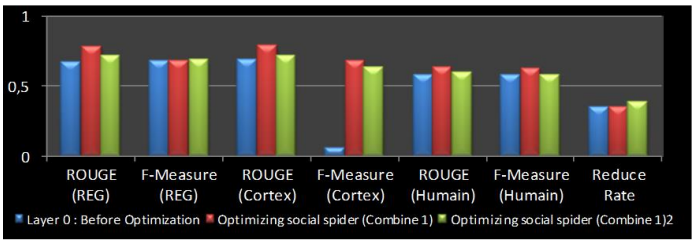


FIGURE 4. OPTIMIZATION OF THE OPTIMAL SUMMARY CANDIDATE FROM STEP 1

Always in Table 1: the second sub-table (top right corner) and the third sub-table (bottom left) shows complete coherence between the three evaluation indexes: reduction rate, F-Measure and ROUGE. While the 4th sub table shows another inconsistency between the evaluation ROUGE and F-Measure; because F-Measure is weakness against summary with a very high reduction ratio: the number of true negative is the maximal value, then F-Measure is big but that does not really reflect the quality of summary candidate who has lost much of semantic because it was much reduced.

The graphs below show explicitly that the first combine of parameter optimization with bees return results better compared to the first combine. This is explained by the given interval of "utility and semantics" represented by two thresholds: upper and lower discount rate is reduced, which allows well-directed the bees. From the graphs, it is suggested that the interval of "utility and semantics" must be as wide as possible in order to have a better optimization of social Bees and generate automatic summarization.

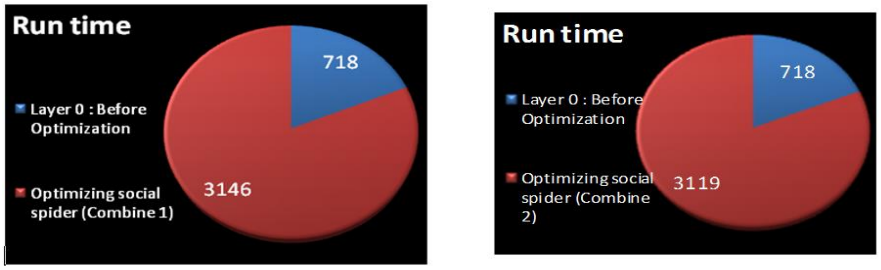


FIGURE 5. EXECUTION TIME OPTIMIZATION 1 SUMMARY CANDIDATE SCORE THRESHOLD = 0.65, THRESHOLD OF SIMILARITY = 0.70

We note that the execution time optimization combined with the first is greater than the second combines this means that the search field first combines is greater than the second combination.

In the following tables, we will compare the results of this approach : "Saving Energy Function of Workers Bees for automatic summarization by extraction" with previous results: A new multi-layered approach for automatic text summaries Mono-Document based on social spiders.

	The Worst case		*The Best case*	
	Bee algorithm (Our Approach)	Social Spider for Summarization	Bee algorithm (Our Approach)	Social Spider for Summarization
F-Mesure (Cortex)	**0.7153**	0.66	**0.7924**	0.78
ROUGE(Cortex)	**0.6339**	0.62	**0.6821**	0.66
F-Mesure (REG)	**0.7211**	0.68	**0.7859**	0.75
ROUGE(REG)	**0.6912**	0.64	**0.6831**	0.65
F-Mesure (Human)	**0.5949**	0.47	**0.6338**	0.62
ROUGE(Human)	**0.5814**	0.55	**0.6289**	0.60
Reduce Rate	**39,06%**	40,25%	**35,31%**	39,29%

TABLE 3. RESULT OF EVALUATION OF THE TWO APPROACHES (OUR APPROCHE (BEE ALGORITHM), SOCIAL SPIDER FOR SUMMURAZATION)

Based on some evaluation criteria, Table 3 is presented for comparison between the two algorithms studied. In terms of results and based on several evaluation measure, we find that our model based on Saving Energy Function of Workers Bees for automatic summarization gives the best results even in the worst case.

7 Conclusion and perspective

In this article, we presented new ideas. Firstly, we have used two techniques of extraction summary after another to improve the rate of reduction without loss of semantics.

The second idea is the use of a biomimetic approach that has the representation of strength graph; in occurrence a Saving Energy Function of Bees for Automatic Summarization by Extraction and we use the communication module.

Given the results obtained, our approach based on a biomimetic approach (Bee Algorithm) can help solve one of the problems of textual data exploration and visualization will.

In the light of these results, facts and opinions, we will try to improve this approach using the WordNet thesaurus, and use a summary based on feelings using the SentiWordNet. For nature still has not revealed all the secrets, we will also try to explore other biomimetic methods.

Reference

1. Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of research and development, 2(2), 159-165.
2. Edmundson, H. P. (1963). Automatic Abstracting, TRW Computer Division, Thompson Ram Wooldridge. Inc., Canoga Park, CA.
3. Van Dijk, T. A. (1985). Handbook of discourse analysis. In Discourse and dialogue.
4. Fum, D., Guida, G., & Tasso, C. (1982, July). Forward and backward reasoning in automatic abstracting. In Proceedings of the 9th conference on Computational linguistics-Volume 1 (pp. 83-88). Academia Praha.
5. Salton, G., Singhal, A., Mitra, M., & Buckley, C. (1997). Automatic text structuring and summarization. Information Processing& Management, 33(2), 193-207.
6. Mitra, M., Buckley, C., Singhal, A., &Cardie, C. (1997, June). An Analysis of Statistical and Syntactic Phrases. In RIAO (Vol. 97, pp. 200-214).
7. Kim, S. N., Medelyan, O., Kan, M. Y., & Baldwin, T. (2010, July). Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles. In Proceedings of the 5th International Workshop on Semantic Evaluation (pp. 21-26). Association for Computational Linguistics.
8. Boudin, F., & Morin, E. (2013, June). Keyphrase Extraction for N-best ranking in multiphrase compression. In North American Chapter of the Association for Computational Linguistics (NAACL).
9. Hamou, R. M., Amine, A., &Rahmani, M. (2012). A new biomimetic approach based on social spiders for clustering of text. In Software Engineering Research, Management and Applications 2012 (pp. 17-30). Springer Berlin Heidelberg.
10. Hamou, R. M., Amine, A., & Boudia, A. (2013). A New Meta-Heuristic Based on Social Bees for Detection and Filtering of Spam. International Journal of Applied Metaheuristic Computing (IJAMC), 4(3), 15-33.
11. Boudia, M. A., Hamou, R. M., Amine, A., Rahmani, M. E., & Rahmani, A. (2015). A New Multi-layered Approach for Au-tomatic Text Summaries Mono-Document Based on Social Spiders. In Computer Science and Its Applications (pp. 193-204). Springer International Publishing.

Unsupervised Classification of Unstructured Data using Filters Combination by Workers Social Bees with 3D Visualisation

Hadj Ahmed Bouarara[*], Reda Mohamed Hamou[**], Abdelmalek Amine[***]
GeCode Laboratory, Department of Computer Science
Tahar Moulay University of Saida Algeria

Abstract— With the advent of the World Wide Web, the amount of textual document exchanged and stored in electronic form does not ceases to increase, where the automatic text classification (clustering) has become a crucial problem that is the basis of several domains such as information retrieval, and knowledge extraction ... etc. In this paper we present a new bio-mimetic approach called filters combination by workers bees FC-WB for text clustering composed of four steps: texts vectoring, using a variation of segmentation methods (n-gram characters and bag of words) and mixed weighting(TF*IDF) for coding the component of each text; classification step based on the cooperation between artificial workers bees (researcher, security and cleaner) where each agent is responsible for filter in order to ensure that the similar text will be in the same hive(cluster). The queen agent represents the centroid of the hive; tests phase using the data set Reuters 21785 and a panoply of validation tools (f-measure, entropy, cluster number and response time). We compare the performance of our approach with the performances of other approaches existed in the literature (2D Cellular Automata, Artificial Immune System (AIS) and Social Spiders (SS)); The navigation step provides a 3D visualisation of the results obtained (hive and apiary).

Keywords: Clustering, Worker Bees, Validation tools, 3D navigation, N-gram.

I. Introduction and background

With the arrival of the internet and the evolution of communication means such as social networks, hence as the expansion of technical and information processing system equipment in many areas (marketing, biology... etc.) have given a birth to a new concepts like big data that reflects the large amount of information available online and offline, where the digital society was enriched every day with a new substance which makes managing it hard, for this reason we need to develop tools for searching, classify, store, update and analyse available data, we need tools to help us to find within a reasonable time the desired information, performing certain chores in our place and facilitate our study. Search engines, automatic indexing and management system databases are examples among many techniques developed.

In order to handle the rising quantity of unstructured documents and acquire the most knowledge possible, it becomes necessary to organize and classified them.

The manual processing of data is very costly in time and personnel. For instance, we are facing to a set of texts of different fields (medicine, computer, biological, linguistic). We ask a human to classify them without the help of a domain expert. This procedure requires a great deal of time

that is why we are expecting to develop automated methods for the automatic classification of text (clustering) that presents the context of our study.

To solve this problem, various systems have emerged based on classical techniques that have proven their limit and were confronted with multiple difficulties

- the selection of parameters(similarity measure, text representation method and threshold)
- The initialisation of the cluster number
- Response time

With the number of researchers that exist and source of inspiration available around us and especially biology and life-living that fits into the world of bio-inspired, a set of techniques that have invaded the domain of text processing which prove their performances facing to different problems, in this paper we have introduced a new inspired model called ***filter combination by worker bees FC-WB*** to develop a solution to the problem of text clustering that represents a topicality challenge in the scientific middle.

II. Stat of the art

Automatic classification has attracted considerable attention from research and industry. Several documents have been published on the subject that follow globally the same process [1]: I) text representation, ii) construction of distance matrix, iii) modelling, iiii) evaluation of results.

A. Classical clustering techniques (CCTs)

The CCTs can be classified into two large families:

1) Hierarchical classification family (HC)

The HC is a recursive family that we cannot alter the source program in running. It establishes a hierarchy of clusters in a tree format based on the movement of granularity. We can distinguish two types of approaches: agglomerative (bottom) and down (divisive) [2].

2) Non-hierarchical classification family (NHC)

The NHC has a goal is to minimize the distance between objects of the same cluster, leading to decompose the set of all individuals in K disjoint sub-sets, basing on fixing the class number in advance. The most popular algorithms that follow this family are Self-Organizing Map (SOM) [3] and k-means [4].

B. Bio-inspired text clustering (BITC) techniques

The clustering problem can be reformulated as an optimization problem. There are several techniques inspired by the aliveness of living and behaviour of insects applied for the clustering problem as indicate Fig. 1.

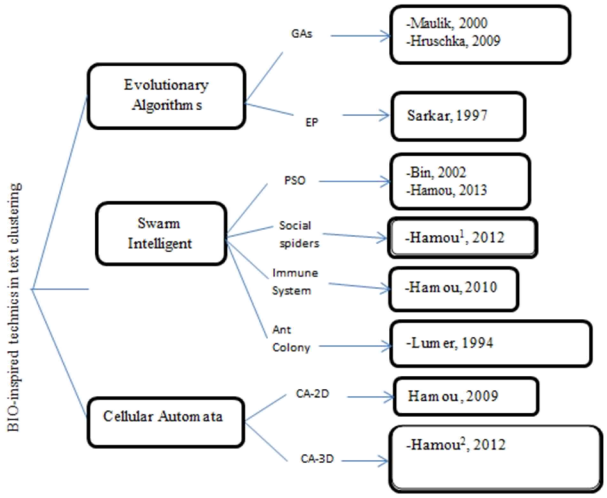


Figure .1. BIO-inspired technics of text clustering (BTTC)

1) Evolutionary Algorithms Texts Clustering (EA-TC)

The Evolutionary Algorithms (EA) dates from the 1970s based on the Darwinism theory using the biological evolution and the reproduction operators (selection, crossover and mutation). Various techniques have been established following the functioning of this family, such as Genetic Algorithms (GAs), Genetic Programming (GP) and Evolutionary Strategy (ES). The work proposed in [5] by using the GAs in order to determine the initial cluster number that represent a crucial problem for the partitioning techniques. In [6] the idea exhibited based on the representation of a chromosome with 2 lines and N columns where N is the number of documents to be classified. Otherwise, Sarkar and his friend in [7] have introduced a technique based on Genetic Programming (GP) allows for grouping data in a small clusters with the elimination of the solutions locally optimal and a functioning independent to the initialisation of the class centres.

2) Swarm Intelligent Texts Clustering (SI-TC)

Touching on the SI techniques such as Particle Swarm Optimization (PSO) algorithm that is the most used by researchers. In [12] the authors were studied the sensitive parameters that affect the quality of outcomes. In [9] the proposed approach was grounded on the behaviour of bee method based on a single cosine distance measure and bag of words as text representation inspired from the principle of quality flowers and the flowers that have more nectar must be visited several times this method based on a wriggling measure.

We noted in the literature that Hamou and all had published some papers using swarm inspired techniques for text clustering like in [13] by exhibiting the immune system using the antigens and antibodies, therefore as the mutation operator, selection and cloning steps for the problem of text

clustering. View the behaviour of social spiders based on their movement, weaving, operation of displacement and return to the rear in [11] the authors formulate this principle to resolve the problem of data unsupervised classification.

The first work introducing the principle of Ant Colony Optimization (ACO) was described in [16]. The work contained in [14] treats the problem of k-means by adapting the ACO but in [15] a combination of ACO and a meta-heuristic method taboo search was detailed.

3) *Cellular Automata Texts Clustering (CA-TC)*

The CA is seen as a good example where the environment, times and physical variables take discrete values. The environment is represented by a matrix of cells that can takes one, two or three dimension depends on the problem. In [9] Hamou and his friends had proved that 2D Cellular Automata (2D-CA) can yield acceptable results, but this method offered a lot of limitations, whether the amelioration presented in [10] using the 3D Cellular Automata (3D-CA) as a solution to these limitations, which ensure a safe theatrical performance of data and a best partitioning.

III. Our idea

View the biological behaviour of the bees we had the idea of transforming the collective spirit of the worker bees to the machine in order to guarantee the full functioning of the hive and fill the need of the queen bee for solve the problem of texts clustering. The documents (honey) the most similar must be in the same cluster (hive) and the most dissimilar in different clusters (hives).

A. Bees' lifestyle

Bees live in the form of a society organized that work for the right functioning of the hive. The bees are classified according to the specific roles:

1) Queen bee (QB):

The QB is the important factor of the hive represent the mother of bees' workers that belong to the same hive. It is the only fertile female individual in the colony that has the capacity of launching a new hive.

2) Workers bees (WBs)

The WBs are the most numerous that ensures the right functioning of the hive by changing role gradually as they grow older.

a. Cleaner bee (WRB):

The WCB represents the bee of the first three days. It permits the cleaning of the hive.

b. Security bee (WSB):

The WSB has the mission of regulating the temperature inside the hive. It represents the guardian of the hive and communicates through its antennas with the bees of the same colony. Those who are not part of the colony are rejected.

c. *Researcher bee (WFB):*

The WRB roams the countryside around the hive to collect nectar, pollen and water for the production of honey indispensable to the survival of the hive.

B. *filter combination by worker bees (FC-WB)* texts clustering:

Our system is constituted from 4 steps as shows in FIG .2. 5 which will be detailed in the next paragraphs

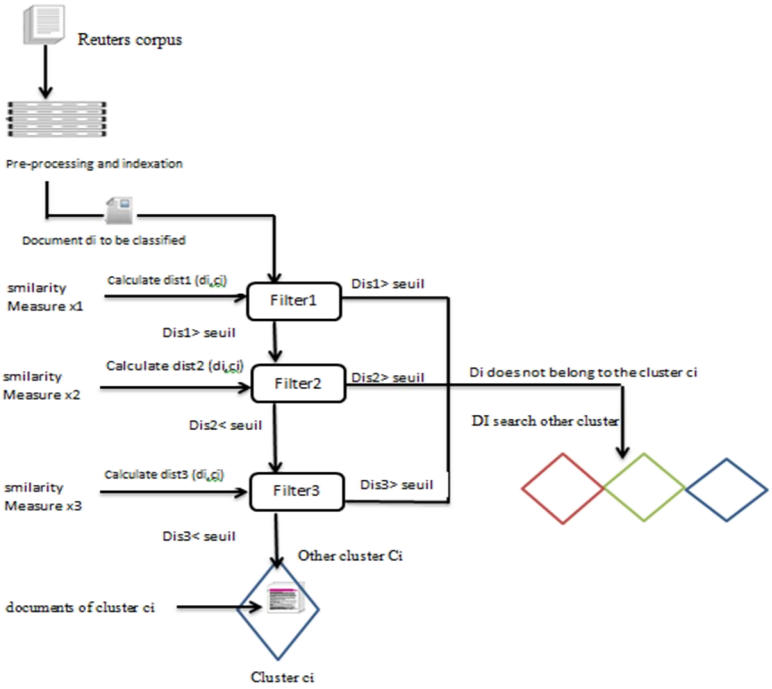


Figure .2. Steps of attributing document due to a cluster

1) *Texts Preparation and indexation:*

This step is done according to the following process:

a. Text cleaning used to eliminate all non-alphabetic characters and stop words.
b. Text representation allows the transit of a text into a list of smaller units called terms such as
   i) Bag of words consists of breaking up the text into a list of words and each word is established by a set of characters tied together.
   ii) N-gram characters represents a window of n characters that moves iteratively step by step in the text and each n-gram captured will be stored in a list.
c. Indexation: applied in order to transforms each text into a vectors using the weighting TF-IDF for calculating the importance of each term in the vector. The corpus will be transformed into a matrix documents * terms.

d. Distance matrix: We calculate a distance matrix of N*N with N: number of documents to be classified. The distances measures used in our work are: Chebychev(cheb), Euclidian(euc), cosine(cos), Manhattan(Man).

2) BEES' Artificial life

We use a set of artificial agents to ensure the right functioning of the automatic text classification. The principle is to run to the corpus (flowers) and bring the documents with two objective the most similar documents are inward the same cluster (hive) and the dissimilar documents are in different clusters.

a. *Queen bee agent QBA* The QBA is the centroid of each cluster (hive).

b. *Worker bees agents (WBAs)*

The QBA generate a WBAs that ensure the smooth functioning of text clustering. We can identify three types of worker bee according to the role covered by each one of them.

i) Researcher bee agent (RBA)

The RBA is placed in the first dam (filter 1) characterized by the vector of the centroid and a distance measure x1 with a role that are: select a document D_i from the corpus, calculate and normalise the distance between the document D_i and the centroid C_i distance1 (D_i, C_i).

ii) Security bee agent (SBA)

The SBA Designed for the second dam characterized by the vector of the centroid and a distance measure x2 with a role is to calculate and normalize the distance between the document D_i (that had passed the preceding filter) and the centroid distance2 (D_i, C_i).

iii) Cleaner bee agent (CBA)

The CBA is designed for the third dam(filter) characterized by a measure of distance x3. It has a role is to calculate and normalize the distance between the document D_i that had passed the two previous obstacles and the centroidC_i, distance3 (D_i, C_i).

The distance measure for every sort of bee must be dissimilar to the others. We must normalise each distance calculated by each agent bee in order to make all distances between 0 and 1.

c. Clustering strategy

The document must pass by three filters to reach the hive (clusters) in each filtering a type of bee agent intervenes following the next steps:

- Initially: a document D_1 is randomly selected among the documents of the corpus and it was placed in a cluster c1 (hive) as the centroid and a threshold must be fixed in advance (between 0 and 1).
- Filter 1 controlled by the collector agent.
- Filter 2 controlled by the surveillance agent.
- Filter 3 controlled by cleaner agent.

- Adaptation function (AD): The AD is used in order to select the cluster where the document D_i will be added: AD=$\frac{distance1(di,ci)+distance2(di,ci)+distance3(di,ci)}{3}$

The document DI belongs to the cluster with the smallest average distance less than the threshold specified in forward motion

- Creation of new cluster: If a document D_i cannot access to any cluster (the average distance of Di for each cluster are superior then the threshold)
- Update: for each new document added to the cluster centroid of this cluster will be updating

Algorithm

Threshold: Integer
x1, x2, x3: distance measures;
AD: average distance;
D_i: document i from corpus;
C_i: clusters i
Begin
Choose 3 distance measures and establish a distance matrix for each one of them
Find the maximum distance $dist_{max}$ and the minimum distance $dist_{min}$ from the3 distance matrix
Choose randomly $D_i \in$ corpus as the centroid of first cluster C_1
For each document di from corpus do
For each cluster C_i do
Compute distance1 (D_i, C_i), distance2 (D_i, C_i), distance3 (D_i, C_i)
Normalization of all the distances (1,2 and 3) distN=$\frac{distance_X - dist_{min}}{dist_{max} - dist_{min}}$
Compute AD(C_i) AD(D_i, C_i)=$\frac{dist1(D_i,C_i)+distance2(di,ci)+distance3(di,ci)}{3}$
endfor
The document di belongs to the cluster which has the lowest AD
if there is no average distance AD(D_i, C_i) <threshold then create a new cluster CI with di as centroid.
endfor
for each new document added to the cluster
the centroid of this cluster will be updating
end

3) *Experimentation:*

In front we demonstrate and talk over our solutions we aim to detail the data set and the metrics of evaluation used for our experiment

a. Reuters 21578

It is a benchmark contains 22 files where the first 21 files contain 1,000 documents and last one contains 578 documents with a sum of 21578 documents. To test our algorithm we decided to use just the 20 first files and take 50 documents from each file with a total of 1000 documents.

b. Validation Tools:

Bearing on the evaluation of our system we use the f-measure and entropy based on the traditional metrics (recall and precision).

IV. Results

In order to present the results of our tests, we conduct a comparison in 4 steps following the next experimental protocol:

The first confrontation put in competition the results got by our approach with an assortment of parameters:

- Text representations (2, 3, 4 and 5 gram characters and bag of words)
- Different distances measures combination ((EUC, MAN, COS) (CHEB, MNH, COS) (EUC, CHEB, COS) (EUC, CHEB, COS))
- We had tested several thresholds, but in this paper we choose three thresholds (0.6, 0.7, and 0.8) that give the best results.

We specify in each test two parameters and varied the third, calculating the f-measure, entropy, the cluster number and time execution for the purpose to identify the sensitive parameters. The solutions obtained are grouped in the accompanying tables:

TABLE 1: CLASSIFICATION OF 1000 DOCUMENTS: THRESHOLD = 0.6

	4gram				*5gram*				*Bag of words*			
	#class	*F-measure*	*Entropy*	*TE(ms)*	*#class*	*F-measure*	*Entropy*	*TE(ms)*	*#class*	*F-measure*	*Entropy*	*TE(ms)*
(EUC, MAN, COS)	***76***	***0.37***	***0.19***	***6981***	***61***	***0.23***	***0.10***	***4789***	***71***	***0.31***	***0.16***	***6945***
(CHEB, MNH, COS)	***61***	***0.42***	***0.17***	***7005***	***63***	***0.56***	***0.08***	***5598***	***101***	***0.32***	***0.09***	***7013***
(EUC,MAN,CHEB)	***67***	***0.27***	***0.16***	***6697***	***59***	***0.36***	***0.12***	***4945***	***88***	***0.21***	***0.19***	***6358***
(EUC,CHEB, ,COS)	***57***	***0.49***	***0.07***	***6834***	***66***	***0.69***	***0.07***	***5439***	***94***	***0.29***	***0.18***	***6758***

The analysis of the table 1, mention that the best performance (blue boxes) is rendered using the combination of (EUC, CHEB, COS) and 5–gram characters). The lowest performance is delivered (red boxes) using the combination of (EUC, MAN, CHEB and bag of words)

TABLE 2: CLASSIFICATION OF 1000 DOCUMENTS: THRESHOLD = 0.7

	4gram				*5gram*				*Bag of words*			
	#class	*F-measure*	*Entropy*	*TE(ms)*	*#class*	*F-measure*	*Entropy*	*TE(ms)*	*#class*	*F-measure*	*Entropy*	*TE(ms)*
(EUC, MAN, COS)	***55***	***0.28***	***0.21***	***5130***	***38***	***0.46***	***0.091***	***4598***	***64***	***0.23***	***0.17***	***5414***
(CHEB, MNH, COS)	***34***	***0.35***	***0.17***	***4276***	***29***	***0.23***	***0.17***	***4439***	***84***	***0.39***	***0.10***	***6826***
(EUC,MAN,CHEB)	***53***	***0.40***	***0.10***	***5371***	***31***	***0.36***	***0.15***	***5371***	***63***	***0.18***	***0.23***	***5213***
(EUC,CHEB, ,COS)	***29***	***0.55***	***0.08***	***4897***	***26***	***0.48***	***0.06***	***3945***	***73***	***0.29***	***0.19***	***5109***

The table 2 shows that with a threshold of 0.7 the best performance is rendered using the combination ((EUC, TCHEB, COS) and 4-gram characters). The badly performance is provided by (EUC, MAN, TCHEB) and bag of words.

Table 3: classification of 1000 documents: threshold = 0.8

	4gram				*5gram*				*Bag of words*			
	#class	*F-measure*	*Entropy*	*TE(ms)*	*#class*	*F-measure*	*Entropy*	*TE(ms)*	*#class*	*F-measure*	*Entropy*	*TE(ms)*
(EUC, MAN, COS)	**51**	**0.48**	**0.05**	**4623**	**17**	**0.49**	**0.10**	**3508**	**47**	**0.36**	**0.08**	**5024**
(CHEB, MNH, COS)	**19**	**0.45**	**0.08**	**3461**	**20**	**0.30**	**0.14**	**3451**	**52**	**0.26**	**0.19**	**4474**
(EUC,MAN,CHEB)	**38**	**0.29**	**0.12**	**3914**	**19**	**0.41**	**0.09**	**3267**	**56**	**0.39**	**0.15**	**5358**
(EUC,CHEB, ,COS)	**23**	**0.54**	**0.09**	**3458**	**12**	**0.63**	**0.10**	**3129**	**61**	**0.27**	**0.12**	**4758**

The table 3 exhibits the results obtained with a threshold= 0.8 where the best performance was given by the combination 5-gram + (EUC, TCHEB, COS) blue boxes.

c. Discussion:

The variation in the choice of the parameters (threshold, texts representation, the combination of distances) was an interesting experience. After different tests the question that arises: what is the configuration that allows to have the best performances? To answer this issue and make a decision that'll be used by other researchers, we must study the influence of each parameter.

- The influence of text representation

As regards the results received in the previous tables. We can ensure that our approach yields better performance using the n-gram characters as a text representation with N = 5 compared to the suitcase of bag of words method that makes a weak performance because the word is ambiguous. We did not use the linguistic processing to remove this ambiguity (we had a low entropy because we did not used a dimension reduction).

- In terms of similarity measure

The best results are given by the combination that contains (Cosine and chebychev). The performances of our approach using two combinations with the same two distance measures among three, are very close because they ensure that both filters have the same requirement.

The threshold plays a real significant part in the stability of results. The most appropriate threshold is 0.6 and whenever we increase the threshold, the number of clusters decrease, because when we increase the possibility that a document di spends the three filters. Automatically, we will have fewer opportunities that a new cluster will be produced.

- In term of execution time and the number of classes

After the various experiences we observe clearly that the execution time is directly related to the number of clusters. The execution time augments when the number of clusters is large for a single reason is that for each new cluster created requires more tests (at least 3 tests).

Generally by analysing the previous results that represents a recapitulation we note that the best classification result were obtained by the parameters (5grm+ (EUC, TCHEB, COS) validated by the high f-measure and the lowest entropy.

d. Comparison

In the 2nd comparison, we compare the best results provided by our approach, with a reference results of three bio-inspired methods of text clustering, namely social spiders (SS), artificial immune system (AIS) and cellular automata 2D (CA 2D) obtained by Hamou and al in [11], [13] and in [8] tested on the Reuters 21578 dataset. This comparison is presented in the following table in order to give our results some credibility.

Table 4: comparison with (Cellular automata 2D, Artificial Immune System and Social Spiders)

	ET(MS)	#class	F-measure	Entropy
Our approach	5439	66	0.69	0.07
Social Spider (SS)	3025	32	0.672	0.19
Artificial Immune System(AIS)	2075	2	0.49	0.23
Cellular Automata 2D	386	157	0.50	0.21

In terms of execution time the table 4 shows that our approach was very slow compared to others algorithms because his operation is based on three measures of distance and several filtering and we did not use a dimension reduction unlike the others algorithms based on a single distance measure with a simple operation otherwise in term of f-measure our approach has better performance compared to (AIS, SS and CA 2D).

V. 3D navigation

The 3D visualization is a topical subject shared by several engineering and computing science. This part of our system is the most important that allows the visualization of the results in 3D format based on the 3D java navigation which provides images with impressive realism characterized by a set of functionalities such as zoom (front and rear) and rotation in order to view the different angles and positions, it is Constituted of two phases: apiary visualization that ensures to see all the clusters obtained by an overall vision as presented in fig .5. Other hand the hive visualization which represents a detail view of each cluster as shown in Fig 6.

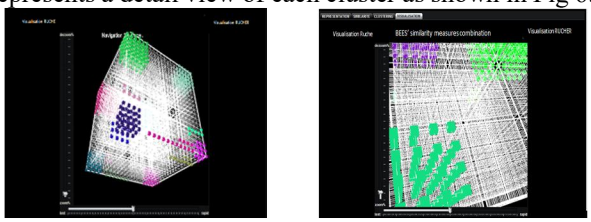


Figure .5. 3D visualization of the clustering

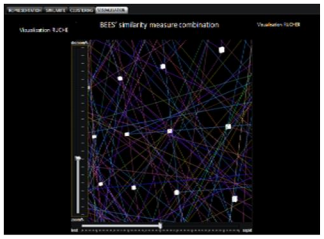


Figure .6. visualisation of the hive

VI. CONCLUSION:

Our work represents a novel approach called filters combination by workers bees (FC-WB) based on coupling a combination of distance measures and the life of worker bees, with the

protection of each cluster by 3 dams to ensure that only the most similar documents will be together. Where each agent bee of the cluster CI (hive) has a specific role is to rule out all the documents that are dissimilar to the documents that belongs to the cluster. The solutions obtained are positive and confirms the idea of testing this new approach validated by a variation of validation tools (recall, precision, f-measure, entropy).

AS perspective, we will use other parameters (representation of texts and other weighting as TFC), the application of our approach for the segmentation and clustering images and trying to develop a team of clustering where each player represents a bio-inspired technique such as (Genetic algorithms (Gas), social spiders (SS), cellular automata 2D and 3D.....etc.) which treats the problem of the diversity of data.

References:

[1] Buhmann, J. (2003). Data clustering and learning. In Arbib, M., editor, The Handbook of Brain Theory and Neural Networks, pages 308–312. Cambridge, MA: The MIT Press

[2] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical statistics and probability, volume 1, pages 281–297, Berkeley. University of California Press

[3] Nasraoui, O., Gonzalez, F., Cardona, C., & Dasgupta, D. (2002). Artificial immune systems and data mining : Bridging the gap with scalability and improve learning. National Science Foundation Workshop on Next Generation Data Mining (NSFNGDM), 110-121.

[4] Selim, S. Z., & Ismail, M. A. (1984). K-means type algorithms: A generalized convergence theorem and characterization of local optimality. IEEE Transactions on Pattern Analysis and Machine Intelligence archive, 6(1), 81-87.

[5] Srikanth, R., George, R., Warsi, N., Prabhu, D., Petry, F. E., & Buckles, B. P. (1995). A variable-length genetic algorithm for clustering and classification. P attern Recognition Letters, 789–800.

[6] Maulik, U., & Bandyopadhyay, S. (2000). Genetic algorithm-based clustering technique. Pattern Recognition, 1455–1465.

[7] Sarkar, M., Yegnanarayana, B., Khemani, D.(1997). A clustering algorithm using an evolutionary programming based approach. Pattern Recognition Letters 18: 975-986.

[8] Hamou, R. M., Lehireche, A., Lokbani, A. C., & Rahmani, M. (2010, October). Text clustering by 2D cellular automata based on the N-grams. In Cryptography and Network Security, Data Mining and Knowledge Discovery, E-Commerce & Its Applications and Embedded Systems (CDEE), 2010 First ACIS International Symposium on (pp. 271-277). IEEE.

[9] Nihal, AbdelHamid, M., Abdel Halim, M. B., & Waleed Fakhr, M. (2013). BEES ALGORITHM-BASED DOCUMENT CLUSTERING. ICIT 2013 The 6 th International Conference on Information Technology, 253-259.

[10] Hamou, R. M., Amine, A., Lokbani, A. C., & Simonet, M. (2012). Visualization and clustering by 3D cellular automata: Application to unstructured data. ArXiv preprint arXiv: 1211.5766.

[11] Hamou, R. M., Amine, A., & Rahmani, M. (2012). A new biomimetic approach based on social spiders for clustering of text. In Software Engineering Research, Management and Applications 2012 (pp. 17-30). Springer Berlin Heidelberg.

[12] Hamou, R. M., Amine, A., & Lokbani, A. C. (2013). Study of Sensitive Parameters of PSO Application to Clustering of Texts. International Journal of Applied Evolutionary Computation (IJAEC), 4(2), 41-55.

[13] Hamou, R. M., Lehireche, A., Lokbani, A. C., & Rahmani, M. (2010). Text Clustering Based on the N-Grams by Bio Inspired Method (Immune Systems). International Refereed Research Journal Researchers Worls, 1(1).

[14] Gnanapriya, S., & Ranjani, P. S. (2013, May). INITIALIZATION K-MEANS USING ANT COLONY OPTIMIZATION. International Journal of Engeneering and Science & Technologie, 2(2).

[15] Vaijayanthi, P., Natarajan, A. M., & Murugadoss, R. (2012, March). Ants for Document Clustering. IJCSI International Journal of Computer Science Issues, 9(2).

Indexing-based link discovery in Linked Data

Khayra BENCHERIF *, Mimoun MALKI **, and Soumia BERRAHAL *

* EEDIS Laboratory, Djilali Liabes University of Sidi Bel-Abbes, Algeria
** Heigh School of Computer of Sidi Bel-Abbes (ESI,Sidi Bel-Abbes), Algeria

Abstract. Linked Data, which is considered as a variant of the semantic web technologies, is a publishing paradigm for making data and not just human-readable documents fully accessible and inter-linkable anywhere on the internet. This allows establishing a global data space based on open standards - the web of data. In this context, different kinds of semantic links can be established between data. A number of Linked Data sources put links owl: sameAs pointing to other sources and other do not. In order to facilitate the establishment of these links, link discovery frameworks have been developed. However, the main problem of these frameworks is the runtime complexity which is very high due to the large number of instances in the source and target data source. In this paper, we present a lossless approach to discover typed links between data sets in Linked Open Data Cloud in very efficient time. Our approach improves the first algorithm of LIMES by indexing task to reduce the number of comparisons and hence the runtime complexity. Moreover, we use the WordNet in order to improve the indexing result and therefore the effectiveness of our system. We evaluate our approach using real data sets and show that it has a smaller number of comparisons in the matching process. In addition, we compare the runtime of our approach with that of LIMES and SILK frameworks and show that our approach outperforms them when mapping large data sets.
Keywords: Linked Data, link discovery, indexing, matching.

1 Introduction

Linked Data is a paradigm that links items in multiple data sources to construct the web of data as a single data space. Over the latest years, the number of data sources in Linked Open Data Cloud project [1](LOD Cloud) is rapidly increasing and therefore it is necessary to establish typed links between items in these data sources to facilitate the combination of information from different sources. These links are generated by calculating the similarity between entities from different data sources. Heath and Bizer (2011)

Currently, many researches help to discover typed links between URIs that represent the same real word object in different data sources (Glaser et al. (2009),Raimond et al. (2008),Cervantes (2013),Scharffe et al. (2009),Volz et al. (2009),Isele et al. (2011),Ngomo and Auer (2011),Ngomo (2011),Axel-Cyrille and Ngomo (2012),Nikolov et al. (2012), Dreéler and Ngomo.

1. http://lod-cloud.net/ 01.05.2015

(2014)). Time-efficient link discovery frameworks attempt to reduce the runtime complexity in the matching task that can be measured by the number of comparisons necessary to complete this task. In this paper, we present a lossless approach that allows optimizing the runtime of link discovery in LOD Cloud. Our approach improves the first algorithm of LIMES (Link Discovery Framework for metric spaces) by (1) using the WordNet to find all synonyms of each source instance and then index them in the target knowledge base using structured inverted indices to rapidly obtain possible candidate results from entities that have been indexed in the LOD Cloud, and (2) removing each exemplar from the target dataset to delete duplicate exemplars and therefore reduce the number of comparisons in the matching process.

The rest of this paper is organized as follows: First, we briefly provide background on Linked Data, indexing task and the matching process in section 2. Then, we review the state of the art in link discovery in section 3. Section 4 gives an overview of our approach. In section 5, we present our experimentation. Finally, we conclude our approach and we describe our future work.

2 Background

In this section, we provide a brief introduction to Linked Data, indexing task and the definition of the matching process as well as the metric space.

2.1 Linked Data

Tim Berners-Lee Berners-Lee (2006) proposed a set of practices and introduced four rules to share and connect structured data on the web Heath and Bizer (2011):

1. Use URIs to names the things,
2. Use HTTP URIs with the goal that individuals can look up those names,
3. Provide useful information by using RDF and SPARQL When someone looks up a URI,
4. Include links to other URIs, so that they can discover more things.

The first and the second rules allows identifying things with URIs and dereferencing them over HTTP protocol; the third rule describe the content of an object with a single data model RDF; and the fourth rule allows creating links between objects (using owl:sameAs connections to represent identity links).

2.2 Indexing task

A semantic search system is an information retrieval system that performs the matching of queries and potential results at a conceptual level, it provides search over billions of documents stored on millions of computers Pound et al. (2010). In the context of semantic web, it is necessary for querying Linked Data at different levels of granularity to retrieve information from various sources for the mismatch between the user request and the response of an information search system. The user enters a keyword query and waits for a ranked list of web pages. Most of search engines use inverted indexes to find potential results of a given query, these indexes do not contain synonyms and cannot differentiate between homonyms.

Linked Open Data project makes possible to index and explore many structured data on the web to navigate through its large data sets . For example, Semantic Search workshop evaluated indexing task by extracting a set of queries from a commercial search engine in 2010 [2] and 2011 [3]. In Blanco et al. (2011), the authors takes advantage of indexing and ranking methods over rdf data sets used at the Billion Triple Challenge [4]. In Tonon et al. (2012), the authors implement a mixed architecture that combines unstructured inverted indices with a structured graph database to improve instance matching task. FLBSM Gupta et al. (2014) is an approach that used Fuzzy Logic to develop a new similarity measure based on TF/IDF measure.

2.3 Matching

In the LOD Cloud, many data sets describing instances have been created and published on the web. Instance Matching is defined as the task of identifying two instances following different schemas (or ontologies) but referring to the same real-world object.
Definition 1: "Given two sets S (source) and T (target) of instances, a metric $m : S \times T \rightarrow [0, \infty[$ and a threshold $\theta \in [0, \infty[$, the goal of instance matching task is to compute the set $M = \{(s, t)|m(s, t) \leq \theta\}$ " Euzenat and Shvaiko (2013).

Examples of metrics on strings include the Levenshtein, qGram and jaccard distance (for more details on these metrics see Doan et al. (2012)). The runtime complexity of matching can be measured by the number of comparisons needed to complete this task (generally, it needs $O(|S||T|)$ comparisons.
Definition 2: "Suppose a metric space $M = (D, d)$ defined for a domain of objects (or the objects' keys or indexed features) D and a total (distance) function d. In this metric space, the properties of the function $d : D \times D \rightarrow \mathbb{R}$, some runtime called the metric space postulates are typically characterized as : " Zezula et al. (2006)

1. $\forall x,\ y \in D,\ d(x,\ y) \geq 0$ (non-negativity),
2. $\forall x,\ y \in D,\ d(x,\ y) = d(y,\ x)$ (symmetry),
3. $\forall x, y \in D,\ x = y \Leftrightarrow d(x,\ y) = 0$ (identity),
4. $\forall x,\ y,\ z \in D,\ d(x,\ z) \leq d(x,\ y) + d(y,\ z)$ (triangle inequality).

3 Related work

In the context of Linked Data, linking is concerned with establishing typed links between entities in data sets . Over the last years, many approaches have been developed to discover typed links between the different knowledge bases; they can be subdivided into two fundamental categories: domain- particular and general Ngomo and Auer (2011).

3.1 Domain-particular

RKB Knowledge base (RKBCRS) Glaser et al. (2009) allows discovering links between entities from the domain of academics. RKBExplorer extracts RDF data from heterogeneous

2. http://km.aifb.kit.edu/ws/semsearch10/ 01.05.2015
3. https://km.aifb.kit.edu/ws/semsearch11/ 01.05.2015
4. http://challenge.semanticweb.org/ 01.05.2015

	RKB	GNAT	RDF-AI	SILK	LIMES
WordNet	No	No	Yes	No	No
String similarity	Yes	Yes	Yes	Yes	Yes
Runtime complexity	$O(\|S\|\|T\|)$	$O(\|S'\|\|T\|)$, S' is SPARQL results.	$O(\|S\|\|T\|)$	$O((\|S\|+\|T\|)\|T\|)$	$O((\|E\|+\|S\|)\|T\|)$

TAB. 1 – *Comparing approaches of link discovery.*

data sources, and it populates its knowledge bases with instances from the AKT ontology [5]. GNAT Raimond et al. (2008) is a tool that was developed for the music domain. It implements the online graph matching algorithm (OGMA) to discover equivalent resources. In Cervantes (2013),the authors present FLORA project to generate a financial dataset. FLORA uses both LIMES and SILK frameworks to discover different pertinent links in the LOD Cloud.

3.2 General

RDF-AI Scharffe et al. (2009) uses a sequence alignment algorithm to match strings and the WordNet for computing a semantic similarity between words. Thus, the mapping by using this tool can be very runtime-consuming. Another link discovery framework is SILK Volz et al. (2009). It implements many approaches to minimize the runtime necessary for mapping instances from knowledge bases. In addition to implementing rough index pre-matching to achieve a quasi-linear runtime complexity, SILK also implements a lossless blocking algorithm called MultiBlock Isele et al. (2011) to reduce its runtime. It utilizes a multidimensional index during which similar objects are located close to one another. LIMES framework Ngomo and Auer (2011) uses the mathematical characteristics (the triangle inequality) of a metric space to reduce the number of comparisons in the matching process. Then, LIMES integrates and extend PPJoin+ ,HYPPO Ngomo (2011), and HR^3 Axel-Cyrille and Ngomo (2012) algorithms that depend on space tiling in spaces with measures that can be divided into independent measures across the dimensions of the problem at hand. KnoFuss Nikolov et al. (2012) implements blocking methods to attain adequate runtime. This method uses a genetic algorithm to compute the similarity to expand the quality of the resulting links. In Dreéler and Ngomo. (2014), the authors improve Jaro-Winkler measures that are used to compare person names by giving equations that allow disposing a large number of computations.

3.3 Discussion

As shown in Table 1, the state-of-the-art approaches have the problem of time complexity that is vey high; due to the large number of comparisons between the source and target knowledge base. In most approaches, each source instance must been compared $|T|$ times, and in the LIMES framework, it must be compared $\sqrt{|T|}$ times even if it has not a candidates matches. In order to resolve this problem and to reduce the number of comparisons of each source instance, we present an approach that improves LIMES by indexing task using the WordNet.

5. http://www.aktors.org/publications/ontology/ 01.10.2014

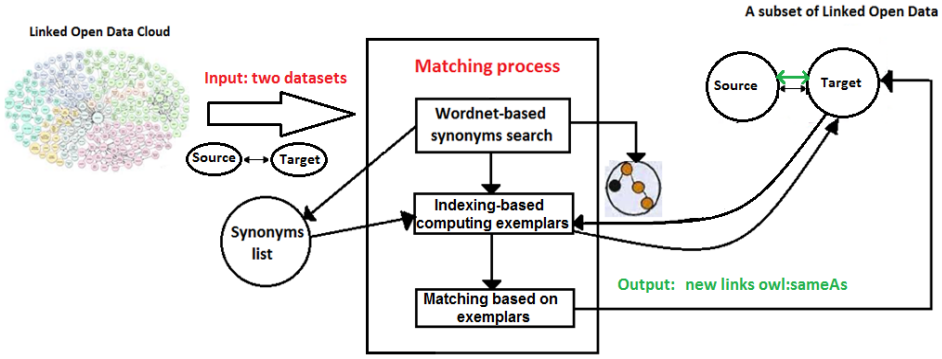


FIG. 1 – *Our system architecture*

4 Proposed framework

In order to discover owl: sameAs links between URIs in an efficient time, we present an approach that improves the first algorithm of LIMES by WordNet based indexing task.
Figure 1 depicts the architecture of our system; it takes as input two data sets S (source) and T (target) from LOD Cloud, and generates owl: sameAs links as output.

4.1 LIMES algorithms

In order to reduce the number of comparisons, LIMES calculates the approximates of the similarity between instances by using the mathematical characteristics of metric spaces (the triangle inequality: $m(x;y) - m(y;z) > \theta \Rightarrow m(x;z) > \theta$). Then, it uses these estimates to sort out the instances pairs that cannot suffice the matching condition. LIMES presents two core algorithms:

Algorithm 1 of LIMES: Computing exemplars

Input: number of exemplars n, target dataset T
Output: Set E of exemplars and their matching to the instances in T
1. Pick random point $e_1 \in T$;
2. Set $E = E \cup \{e_1\}$;
3. Compute the distance from e1 to all $t \in T$;
while $|E| < n$ do
    4. Get a random point e' such that $e' \in argmax_t \sum_{t \in T'} \sum_{e \in E} m(t,e)$
    5. $E = E \cup \{e'\}$;
    6. Compute the distance from $e'$to all $t \in T$;
end
7. Map each point in $t \in T$to one of the exemplars $e \in E$ such that $m(t,e)$ is minimal;
8. Return E;

Algorithm 1: Given the source S, the target T and the threshold θ, the first algorithm of LIMES computes a set of exemplars E (we note that the exemplars are as dissimilar as possible) for T and matches each point $t \in T$ to the exemplar closest to it. The complexity of this algorithm is $O(|E||T|)$.

Algorithm 2 of LIMES: The matching based on exemplars

Input: Set of exemplars E, point $s \in S$, threshold θ
Output: matching M for s
1. $M = \emptyset$;
for $e \in |E|$ do
    if $m(s, e) \leq \theta$ then
        2. $M = M \cup \{e\}$
    end
    for $i = 1...|L_e|$ do
        if $(m(s, e) - m(e, \lambda_i^e)) \leq \theta$ then
            if $m(s, \lambda_i^e) \leq \theta$then
                3. $M = M \cup \{\lambda_i^e\}$
            end
        else
            break;
        end
    end
end
4. return M;

Algorithm 2: For each $s \in S$ and each $e \in E$, the distance $m(s; e)$ is computed. The list L_e contains the instances related with an exemplar $e \in E$ in step 7 of the first algorithm, and $\lambda_1^e, ..., \lambda_m^e$ are the elements of L_e. The complexity of this algorithm is $O((|E|+|S|)|T|)$. Thus, we detect two main drawbacks of LIMES: (1) Each instance in the source dataset is compared n (the number of exemplars in LIMES was $\sqrt{|T|}$) times even if it has not a candidates matches. (2) When the first algorithm of LIMES adds an exemplar to the list of exemplars, it does not remove this exemplar from the target dataset and therefore, there may be a duplication of exemplars.

4.2 Example

In this section, we describe an example that shows all steps of our approach. We take as input Drugbank (source dataset) and DBpedia (target dataset) and generate owl:sameAs between them as output. First, we search all synonyms of each instance in Drugbank by using the WordNet database [6]. For example, the synonyms of <http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/people> in the WordNet are: **people, citizenry, multitude, masses, hoi polloi and the great unwashed**. Then, we index these synonyms in DBpedia to give all candidates matches: <http://dbpedia.org/resource/The_People>, <http://dbpedia.org/resource/The_people>, <http://dbpedia.org/ resource/ People%21>, <http://dbpedia.org/resource/People>, <http://dbpedia.org/resource/These_People>, <http://dbpedia.org

6. https://WordNet.princeton.edu/WordNet/download/ 01.05.2015

	Set of exemplars	Set of the rest of instances in T'	Result
people	The_People, People_For, People_Are_People, For_the_people	The_people, People%21, People, These_People, For_the_People, For_The_People, By_the_People, Will_of_the_People, Of_The_People, By_The_People, By_the_people, By_the_People%2C_ for_the_People, People_are_People, People_are_people, People_to_People, People%2C_People	The_People, People
citizenry	Citizenry , Citizens_%E2%80%93_ Party _of_the_Citizenry	Citizens%E2%80%93 Party_of_the_Citizenry, Citizens-Party_of_the_Citizenry, Citizens_-_ Party_of_the_Citizenry	Citizenry
multitude	Multitude , Multitude:_War_and_Democracy_in_the_Age_of _Empire	Feeding_the_multitude, Feeding_of_the_multitude, A_Multitude_of_ Casualties, In_Your_Multitude, The_Assembled_ Multitude, Assembled_Multitude, Big_Daddy_Multitude	Multitude
masses	{ }	{ }	{ }
hoi polloi	Hoi_polloi	The_hoi_polloi	Hoi_polloi, The_hoi_polloi
the great unwashed	The_Great_Unwashed	Great_Unwashed, Unwashed_biodiesel	The_Great_ Unwashed, Great_Unwashed

TAB. 2 – *Indexing based computing exemplars and the matching based on exemplars.*

/resource/For_the_People>, <http://dbpedia.org/resource/For_The_People>, <http://dbpedia.org/resource/By_the _People>, <http://dbpedia.org/resource/People_For>, <http://dbpedia.org/resource/Will_of_the_People>, <http://dbpedia.org/resource/For_the _people>, <http://dbpedia.org/resource/Of_The _People>, <http://dbpedia.org/resource/By_The _People> , <http://dbpedia.org/resource/By_the_people> , <http://dbpedia.org/resource/By_the_People%2C_for_the _People> , <http://dbpedia.org/resource/People_Are_People>, <http://dbpedia.org/resource/People_are_People>, <http://dbpedia.org/resource/People_are_people>, <http://dbpedia.org/resource/People_to_People>, <http://dbpedia.org/resource/People%2C_People>, <http://dbpedia.org/resource/Citizenry>, <http://dbpedia.org/resource/Citizens%E2%80%93Party_of_the_Citizenry>, <http://dbpedia.org/resource/Citizens_%E2%80%93_Party_of_the_Citizenry>, <http://dbpedia.org/resource/Citizens-Party_of_the_Citizenry>. Then, we compute the set of exemplars and we remove them from the list of candidates matches to give the final result of the matching (see table 2). We use several similarity metrics [7] in our algorithm of matching (levenshtein, qGram...etc). Consequently, the similar instances of <http://www4.wiwiss.fu-berlin.de/drugbank/resource /drugbank/people> are: <http://dbpedia.org/resource/The_People>, <http://dbpedia.org/resource/ People>, <http://dbpedia.org/resource/Citizenry> , <http://dbpedia.org/resource/ Multitude>, <http://dbpedia.org/resource/Hoi_polloi>, <http://dbpedia.org/resource/The_hoi_polloi>, <http://dbpedia.org/resource/ The_Great_ Unwashed>, <http://dbpedia.org/resource/Great_Unwashed>.

7. http://simmetrics.sf.net 01.05.2015

Algorithm 3 Indexing based computing exemplars

Input: Point $s \in S$, number of exemplars n, the WordNet, target dataset T
Output: Set E of exemplars and their matching to the instances in T
1. Search all synonyms of s in the WordNet and put them in $v = v \cup \{s\}$.
2. Index the elements of v in T and put the results in T'.
3. Select random point $e_1 \in T'$;
4. Set $E = E \cup \{e_1\}$;
5. Remove e1 from T';
6. Compute the distance from e1 to all $t \in T'$;
while $|E| < n$ do
    7. Get a random point e' such that $e' \in argmax_t \sum_{t \in T'} \sum_{e \in E} m(t, e)$
    8. $E = E \cup \{e'\}$;
    9. Remove e' from T'
    10. Compute the distance from e' to all $t \in T'$;
end
11. Map each point in $t \in T'$ to one of the exemplars $e \in E$ such that $m(t, e)$ is minimal;
12. Return E;

4.3 WordNet-based synonyms search

The WordNet is a lexical database that clusters English words in sets of equivalents words called synsets. It gives a short definition, examples, homonyms and hyponyms of these synonyms. Wordnet can be seen as a mix of thesaurus and dictionary; it establishes semantic distance between two concept by providing six measures of similarity and three measures of relatedness Pedersen et al. (2004). In this step, we search all synonyms of each instance of the source dataset by using the WordNet database for improving the result of the indexing task and finding all candidates matches.

4.4 Indexing based computing exemplars

In order to reduce the time complexity of the matching process, we improve the first algorithm of LIMES by: (1) **WordNet based indexing:** receives as input a SPARQL endpoints or LOD dumps of the source and the target data sets and retrieves all corresponding triples as output. We use structured inverted indices Demartini et al. (2013) by using Lucene [8] search engine to obtain a ranked list of candidate matches from the target data set (see algorithm 3: step 1 and 2). Since the indexes do not contain synonyms, we use the WordNet to improve the result of indexing task. (2) **Removing each exemplar from the target dataset:** when the first algorithm of LIMES adds an exemplar to the list of exemplar, it does not remove this exemplar from the target dataset and therefore there may be a duplication of exemplars; this can increase the number of comparisons. In order to resolve this problem, we add an instruction that remove each exemplar from the target dataset to reduce the number of comparisons as shown in algorithm 3: step 5 and 9).
Algorithm 3 takes as input a point $s \in S$ and a target dataset T. First, we search all synonyms of each instance $s \in S$ and we put them in a list $v = v \cup \{s\}$. Then we index the elements of

8. http://lucene.apache.org/ 01.05.2015

Source instance $s \in S$	$\|T\|$	candidates matches	Number of comparisons in our approach	Number of comparisons in LIMES	Difference (LIMES, our approach
<http://dbpedia.org/resource/People>	4346	20	4	65	61
<http://dbpedia.org/resource/Albert_Einstein>	4772	15	3	69	66
<http://dbpedia.org/resource/Organisation>	12701	5	2	112	110
<http://dbpedia.org/resource/America>	5000	7	2	70	68
<http://dbpedia.org/resource/Cameroun>	1000	0	0	32	32

TAB. 3 – *comparisons between our approach and LIMES depending on the number of comparisons of each approach.*

v in the target dataset to retrieve all candidate matches from the target dataset and we put them in a list T' (step 2). In steps 3, 4, 5, we initialize E by selecting a random point e_1 in the metric space (T', m) ($E = \{e_1\}$), then we remove e_1 from T'. Then, we compute the similarity from the exemplar e_1 to every other point $t \in T'$(step 4). As the size of E is less than n, we pick a point $e' \in T$ such that the sum of the distances from e' to the exemplars $e \in E$ is maximal. Then, we add this point to E and we remove it from T' (step 6 and 7). In step 8, we compute the distance from e' to all other points in T. Finally, we map each point in T' to the exemplar to which it is most similar. Therefore, the complexity of our algorithm is $O(|E|(|T'| - |E|))$.

4.5 Matching based on exemplars

In this step, we reuse the second algorithm of LIMES. The list L_e is the set of exemplars found in step 9 of the second algorithm of our approach, for more details see Ngomo and Auer (2011). The time complexity of our approach is $O((|E| + |S|)(|T'| - |E||))$ where $T' \subseteq T$.

5 Experimentation

In this section, we experimentally evaluate the performance of our system on real data sets (DBpedia [9], LinkedCT [10], Drugbank [11], Bio2RDF [12]) by comparing the number of comparisons and the runtime in our approach, LIMES version 0.6 and SILK version 2.6. We run all experiments on a 64 bits system with a 2.5GHz Intel Core i5 machine with 8 GB RAM.
As shown in Table 3, LIMES framework uses all instances of the target knowledge base to calculate the number of exemplars ($\sqrt{|T|}$); while our framework uses a portion of the target knowledge base that is calculated by indexing task to reduces the search space of candidate matches. For example, if the size of our knowledge base is 1000 then the number of exemplars is $\sqrt{|1000|}$=32 and therefore each source instance will be compared 32 times using LIMES framework. However, each source instance is compared 4 times by using our system for the

9. http://dbpedia.org/sparql 10.01.2015
10. http://data.linkedct.org/sparql 10.01.2015
11. http://www4.wiwiss.fu-berlin.de/drugbank/sparql 10.01.2015
12. http://mesh.bio2rdf.org/sparql 10.01.2015

Source S	Target T	$\|S\|$	$\|T\|$	Our approach	LIMES			SILK
					th=0.80	th=0.90	th=0.95	
DBpedia	LinkedCT	9509	9509	2.43	15	10	6	25
Drugbank	DBpedia	5291	5291	1.28	12	8	5	18
Bio2RDF	DBpedia	50031	74458	3.62	78	52	31	130
Drugbank	LinkedCT	3544	3544	0.99	6	5	2	11

TAB. 4 – *The runtime of our approach, LIMES and SILK. All times are given in seconds.*

previous example (as the result of indexing task is ranked, we take the top 20 instances and therefore number of examplars is $\sqrt{|20|}$=4). The total runtime of our approach is calculated like that: the time to retrieve the instances from the source knowledge base, the time to index the sources instances in the target knowledge base by using WordNet, the time to compare the instances of the source with their of the target knowledge bases, the time to put up the results. Table 4 illustrates the different sizes of the sources and targets knowledge bases as well as the runtime of LIMES, SILK and our approach. The results demonstrate that our system outper-

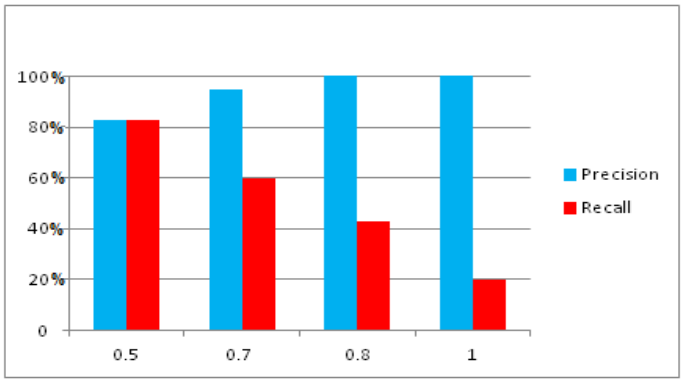


FIG. 2 – *Precision and recall as compared to the threshold* $\theta \in [0, 1]$. *The x-axis shows the threshold* θ*, the y-axis the values of precision and recall.*

forms LIMES and SILK when mapping large data sets.
In order to evaluate the effectiveness of our approach for instance matching, we select a subset from DBpedia, drugbunk, LinkedCT, and we validate the result of the matching process by an expert. In order to compute the precision and the recall, we compare the set of matching and non matching data for each source instance provided by our system and by the expert. The precision is defined as: $P = \frac{tp}{(tp+fp)}$, and the recall is defined as: $R = \frac{tp}{(tp+fn)}$. Where:
True positive (tp): In the case where the expert and our approach select the matches.
False positive (fp): In the case where our system selects a match while the expert does not.
False negative (fn): In the case where the expert selects a match while our approach does not.
In our evaluation, we will focus on the precision since it is the most useful metric in practice. Figure 2 shows how precision and recall could vary according to the threshold th. So, the precision is maximum if the threshold nearing to the value 1.0, while the recall is low (because

while we select a high threshold, the number of the matches retrieved by the system is low). This means that our system returns substantially more results in the matching process due to the use of both WordNet and string similarity measure to compute the similarity.

6 Conclusion and future work

In this paper, we have presented an approach to discover owl: sameAs links between data sources in the LOD Cloud in very efficient time. Our approach used the WordNet to improve indexing task and the algorithms of LIMES. We evaluated our approach with real data sets and we showed that it outperforms state-of-the-art approaches with respect to the number of comparisons. In our future works, we use this framework to facilitate integrating ontologies in the LOD Cloud.

References

Axel-Cyrille and N. Ngomo (2012). Link discovery with guaranteed reduction ratio in affine spaces with minkowski measures. In *International SemanticWeb Conference*, pp. 378–393.

Berners-Lee, T. (2006). Linked data - design issues. `http://www.w3.org/DesignIssues/LinkedData.html`.

Blanco, R., P. Mika, and S. Vigna (2011). Effective and efficient entity search in rdf data. In *Proceedings of the 10th International Conference on The Semantic Web - Volume Part I*, ISWC'11, Berlin, Heidelberg, pp. 83–97. Springer-Verlag.

Cervantes, J. L. S. (2013). Discovering and linking financial data on the web.

Demartini, G., D. E. Difallah, and P. Cudré-Mauroux (2013). Large-scale linked data integration using probabilistic reasoning and crowdsourcing. *The VLDB Journal 22*(5), 665–687.

Doan, A., A. Y. Halevy, and Z. G. Ives (2012). *Principles of Data Integration.*

Dreéler, K. and A.-C. N. Ngomo. (2014). On the efficient execution of bounded jaro-winkler distances.

Euzenat, J. and P. Shvaiko (2013). *Ontology Matching, Second Edition*. Springer.

Glaser, H., I. C. Millard, W.-K. Sung, S. Lee, P. Kim, and B.-J. You (2009). Research on linked data and co-reference resolution. Technical report, University of Southampton.

Gupta, Y., A. Saini, A. K. Saxena, and A. Sharan (2014). Fuzzy logic based similarity measure for information retrieval system performance improvement. In *Distributed Computing and Internet Technology - 10th International Conference, ICDCIT 2014.*, pp. 224–232.

Heath, T. and C. Bizer (2011). *Linked Data: Evolving the Web into a Global Data Space*.

Isele, R., A. Jentzsch, and C. Bizer (2011). Efficient multidimensional blocking for link discovery without losing recall.

Ngomo, A.-C. N. (2011). A time-efficient hybrid approach to link discovery.

Ngomo, A.-C. N. and S. Auer (2011). Limes: A time-efficient approach for large-scale link discovery on the web of data. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, IJCAI'11, pp. 2312–2317.

Nikolov, A., M. d'Aquin, and E. Motta (2012). Unsupervised learning of link discovery configuration. In *Proceedings of the 9th International Conference on The Semantic Web: Research and Applications*, ESWC'12, Berlin, Heidelberg, pp. 119–133. Springer-Verlag.

Pedersen, T., S. Patwardhan, and J. Michelizzi (2004). Wordnet::similarity: Measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, HLT-NAACL–Demonstrations '04, Stroudsburg, PA, USA, pp. 38–41.

Pound, J., P. Mika, and H. Zaragoza (2010). Ad-hoc object retrieval in the web of data. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, New York, NY, USA, pp. 771–780. ACM.

Raimond, Y., C. Sutton, and M. Sandler (2008). Automatic interlinking of music datasets on the semantic web. In *Proceedings of the 1st Workshop about Linked Data on the Web*. en

Scharffe, F., Y. Liu, and C. Zhou (2009). Rdf-ai: an architecture for rdf datasets matching, fusion and interlink. In *Proc. IJCAI 2009 workshop on Identity, reference, and knowledge representation (IR-KR), Pasadena (CA US)*.

Tonon, A., G. Demartini, and P. Cudré-Mauroux (2012). Combining inverted indices and structured search for ad-hoc object retrieval. In *Proceedings of the 35th International ACM SIGIR Conference*, New York, NY, USA, pp. 125–134.

Volz, J., C. Bizer, M. Gaedke, and G. Kobilarov (2009). Discovering and maintaining links on the web of data. In *The Semantic Web - ISWC 2009*, Volume 5823, pp. 650–665.

Zezula, P., G. Amato, V. Dohnal, and M. Batko (2006). *Similarity Search: The Metric Space Approach*, Volume 32 of *Advances in Database Systems*. Springer.

Résumé

Linked Data est un paradigme de publication pour rendre les données et pas simplement des documents accessibles et inter-cliquables sur l'internet. Cela permet la création d'un espace de données global basé sur des normes ouvertes- le web des données. Dans ce contexte, différents types de liens sémantiques peuvent être établies entre les données. Un certain nombre de sources de données liées mettent des liens owl: sameAs pointant vers d'autres sources et d'autres ne les font pas. Afin de faciliter la mise en place de ces liens, les frameworks de découverte des liens ont été développés. Néanmoins, le problème principal de ces frameworks est le temps d'exécution qui est très élevé en raison du grand nombre d'instances dans les ensembles des données source et cible. Dans cet article, nous présentons une approche qui permet de découvrir des liens typés entre les ensembles de données de Linked Open Data Cloud en temps très efficace. Notre approche améliore le premier algorithme de LIMES par la tâche d'indexation afin de réduire le nombre de comparaisons et par conséquent le temps d'exécution. De plus, nous utilisons le WordNet pour améliorer le résultat d'indexation et par conséquent l'efficacité de notre système. Nous évaluons notre travail sur des ensembles de données réelles et nous montrons que notre approche a un plus petit nombre de comparaisons dans le processus de matching. En outre, nous comparons le temps d'exécution de notre approche avec celle de LIMES et SILK et nous montrons que notre approche est la plus rapide.
Keywords: Linked Data, la découverte des liens, l'indexation, le matching.

Extraction de connaissances à partir des séries temporelles d'images satellites pour l'interprétation des changements aléatoires

Ali Ben Abbes[*,***] Imed Riadh Farah[*,**]

[*]Laboratoire RIADI, Campus universtaire de la Mannouba 2010, Tunisie
ali.benabbes@isima.fr
http://www.riadi.rnu.tn
[**]Departement ITI, 29238 Brest Cedex, France
riadh.farah@ensi.rnu.tn
http://departements.telecom-bretagne.eu/iti/
[***]Laboratoire LIMOS, Campus universtaire cézeaux,Aubiere 63000, France

Résumé. Les séries temporelles d'images satellites (STIS) représentent une source d'informations pertinente pour interpréter l'occupation du sol et comprendre les changements des zones géographiques. Ces changements peuvent être périodiques, progressifs ou aléatoires. Toutefois, découvrir des connaissances impose de répondre à plusieurs défis qui sont liés aux caractéristiques des STIS et à leurs contraintes. Dans cet article, nous nous intéressons à l'interprétation des changements aléatoires. Notre contribution consiste à proposer un modèle pour l'extraction des connaissances valides et fiables à partir des STIS par une méthode de fouille de données. L'apprentissage est assuré par des algorithmes génétiques (AG) dans le but d'extraire des règles d'association floues pour expliquer les changements aléatoires en intégrant d'autres sources comme la température et la précipitation. Notre modèle est validé sur une série d'images MODIS NDVI.

1 Introduction

Au cours de ces dernières années, les nouvelles technologies de la télédétection fournissent aux chercheurs un nombre considérable d'images satellites. Ces STIS ont été considérées dans de nombreuses applications telles que la géologie, l'environnement, l'écologie terrestre, la croissance urbaine et la surveillance des forêts (Ahmed et Ahmed, 2012; Bonansea et al., 2015; Brooks et al., 2015). De ce foisonnement, sont nées des nouvelles applications tentant d'extraire non seulement une information valide et fiable, mais plus précisément des connaissances significatives et réutilisables permettant d'appuyer la prise de décision (Petitjean et al., 2012; Flamary et al., 2015). Les STIS jouent un rôle important dans le suivi des phénomènes dynamiques et leurs évolution au cours de temps. Les STIS ont été utilisés avec succès dans la détection du changement de l'occupation du sol (Lillesand et al., 2014; Rotem-Mindali et al., 2015; Thakkar et al., 2015). Les changements qui apparaissent sur les zones géographiques peuvent être de nature très différente et provenir de phénomènes brutaux ou lents (Martínez et

Gilabert, 2009; Hutchinson et al., 2014; Jamali et al., 2015; Verbesselt et al., 2010). Les changements périodiques qui présentent des variations relativement régulières, rythmées à moyen terme, des changements progressifs tels que la variabilité interannuelle du climat et des changements brusque causés par des troubles tels que la déforestation, les inondations et les incendies. Dans la littérature plusieurs travaux ont été proposés pour l'identification des différents types des changements (Martínez et Gilabert, 2009; Hutchinson et al., 2014; Verbesselt et al., 2010; Jamali et al., 2015). Dans ce papier, nous nous sommes intéressés à l'analyse des changements aléatoires par une méthode de fouille de données. Le but de ce travail est d'extraire de connaissances qui expliquent les liens entre les changements aléatoires et les facteurs climatiques. L'approche proposée consiste à détecter les variations aléatoires après une décomposition additive des STIS ensuite un AG est appliqué pour extraire des règles d'association floues qui vont permettre de décrire les variations aléatoires détectées en expliquant leurs causes et leurs influences. Le reste de cet article est organisé comme suit. La motivation de l'utilisation des AG est donnée dans la deuxième section. L'approche d'extraction de connaissances à partir des STIS est expliquée dans la troisième section. Les résultats expérimentaux sont indiqués dans la quatrième section. Nous finirons avec une conclusion et quelques travaux futurs.

2 les AGs pour l'extraction de règles d'associations floues

Parmi les types de techniques de représentation de connaissances, les règles de production sont couramment employées en raison de leurs nombreux avantages. Les règles sont relativement simples à construire, elles permettent un prototypage rapide, et des tests peuvent commencer par quelques règles et sont un moyen naturel de résumer la connaissance humaine. Dans cette recherche, les règles de production sont utilisées pour la représentation des connaissances. Des travaux récents ont montré que les AG peuvent être utilisés avec succès, pour la découverte de règles d'association floues. A titre d'exmple, (Wakabi-Waiswa et al., 2011)ont proposé une nouvelle méthode d'extraction des règles d'association en utilisant les AGs avec cinq mesures de la qualité de règle. (Mankad et al., 2011) ont présenté un système conçu pour identifier les différentes compétences des élèves dans le domaine de l'éducation représentée comme règle utilisant une approche génétique floue. Un système génétique à base de règles floues pondérée a été proposé dans (Dutta et Burdwan, 2009) et appliqué dans le processus de classement des données de l'iris. Le choix d'algorithme génétique est justifié dans plusieurs recherches dans la littérature qui ont montré que :

- Plusieurs mesures de la qualité de règle peuvent être utilisées avec les AGs telle que la compréhensibilité, la confiance, J - mesure, la surprise et le gain.
- Les algorithmes existants de règles d'extraction sont des calculs très coûteux parce que trouver tous les motifs fréquents dans une base de données consiste à rechercher tous les jeux d'éléments possibles.

Sur cette base, nous adoptons un AG conçu spécifiquement pour l'interprétation des changements aléatoires dans les STIS.

3 Modélisation à base de connaissance pour l'interprétation des changements aléatoires

L'essentiel de l'interprétation, dans notre cas, est d'expliquer les causes des changements aléatoires affectant une zone géographique (végétation ou zone urbaine). L'objectif principal est d'extraire des connaissances sur les liens entre les variations aléatoires et un ensemble de facteurs extérieurs (facteurs climatiques par exemple). L'organigramme de l'approche proposée est illustré dans la figure 1.

L'image que nous transmet le satellite est bruitée. Pour cela, il faut effectuer de prétraitements pour la rendre exploitable. Parmi ces prétraitements, nous citons : les corrections géométriques, atmosphériques, radiométriques.

La réalisation de notre objectif passe par deux phases : La première phase consiste à détecter les variations aléatoires et la deuxième sert à extraire l'ensemble des règles d'associations floues.

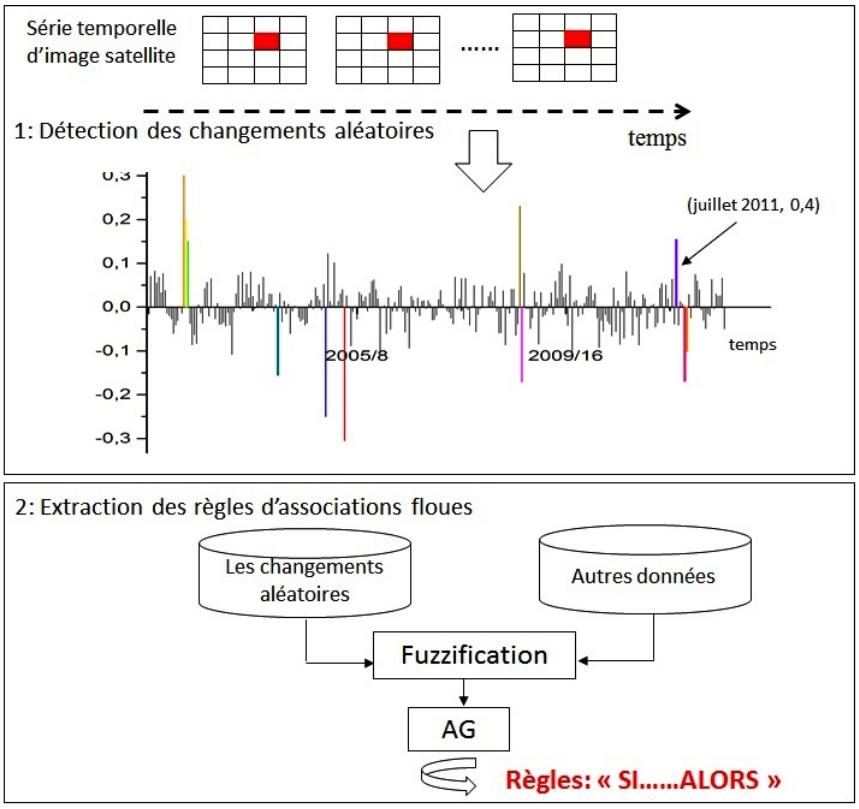


FIG. 1 – *Modèle d'extraction des règles floues*

3.1 Détection des variations aléatoires

Le processus d'extraction de connaissances passe généralement par deux étapes : collecte et prétraitement des données. Dans notre contexte, ces étapes sont nécessaires pour détecter les variations aléatoires. Dans nos travaux nous avons proposé une méthodologie pour la détection des variations aléatoires dans les STIS. (Ben Abbes et al., 2014, 2015). Cette méthodologie intègre une décomposition additive de la série S_t en trois composantes (Hutchinson et al., 2014). Elle permet d'identifier la variation de chaque composante et l'interaction entre elles.

$$S_t = Ct_t + Cs_t + Ca_t \tag{1}$$

Avec :
- Ct_t : La composante tendancielle
- Cs_t : la composante saisonnière
- Ca_t : La composante aléatoire

La composante aléatoire est définie comme étant un bruit (Martínez et Gilabert, 2009; Hutchinson et al., 2014; Verbesselt et al., 2010; Jamali et al., 2015). Par conséquent, elle est ignorée. Par contre, elle peut englober des données précieuses ce qui a été déjà prouvé dans d'autres domaines (Grieser et al., 2002). Ces valeurs inattendues peuvent être des erreurs d'observation, des événements improbables qui peuvent se produire par hasard ou des événements improbables spéciaux qui n'ont pas eu lieu par hasard. Ces derniers sont appelés des variations aléatoires dans cet article. Le critère de détection est basé sur l'hypothèse que les résidus considérés sans événements aléatoires, suivent une loi normale. Tout d'abord la valeur trouvée à la plus grande distance de la moyenne réelle est sélectionnée comme variation aléatoire potentielle. Par la suite on calcule la moyenne et la variance des résidus restants. La série réduite est soumise à l'épreuve de Jarque-Bera pour examiner si elle diffère sensiblement de bruit gaussien. Ces étapes sont répétées jusqu'à la vérification de l'hypothèse nulle. Pour chaque variation aléatoire détectée, nous sélectionnons leurs caractéristiques (par exemple amplitude et de temps).

3.2 Extraction des règles d'associations floues

Les règles d'association extraites, décrivent les variations aléatoires détectées dans notre série. Chaque variation aléatoire dispose d'un ensemble de causes pouvant influencer l'activité d'une zone géographique.

L'algorithme proposé fonctionne sur deux phases. Tout d'abord, les attributs à valeurs réelles sont fuzzifiées. Ensuite, un AG est utilisé pour extraire des règles de production floues. Une règle floue est spécifié comme ci-dessous :

$$(x_1\, est A_1 k) \wedge (x_2\, est A_2 k) \wedge \ldots \wedge (x_n\, est A_n k) \rightarrow D_m \tag{2}$$

Avec x_1, x_2 et ...x_n sont des facteurs extérieurs.

La motivation d'utiliser la logique floue a été exprimée par (Zadeh, 1973) de la manière suivante : Dans une situation donnée, la capacité de l'esprit humain à raisonner en termes flous est d'un grand avantage même dans une énorme quantité d'informations. En fait, Un système de logique floue traite l'imprécision des variables d'entrée et de sortie en définissant

des ensembles flous qui peuvent être exprimées dans les variables linguistiques (par exemple , les petites, moyennes et grandes).

La fuzzification consiste à transformer les valeurs nettes dans les grades de membres pour les termes linguistiques des ensembles flous. La fonction d'appartenance est utilisée pour associer une note à chaque terme linguistique.

L'AG est basé sur les phases suivantes :
1. Initialisation : une population initiale de N chromosomes est tirée aléatoirement
2. Évaluation : chaque chromosome est décodé puis évalué
3. Sélection : création d'une nouvelle population de N chromosomes par l'utilisation d'une méthode de sélection appropriée.
4. Reproduction : possibilité de croisement et mutation au sein de la nouvelle population
5. Retour à la phase d'évaluation jusqu'à l'arrêt de l'algorithme

La Fonction d'évaluation, ou de fitness des règles d'association floues se fait en adoptant les mesures suivantes : le support, la confiance, l'intérêt, le lift et le JMesure. Ensuite, nous fusionnons ces mesures dans une seule fonction. Soit A est l'antécédent et C la Conséquence.

– Support : exprime la probabilité d'occurrence d'une règle dans l'ensemble des données, comme illustré dans la formule 1 :

$$Sup(A \to C) = \frac{freq(A \to C)}{N} \tag{3}$$

– Confiance : exprime le pourcentage des transactions contenant A et produisant C.

$$Conf(A \to C) = \frac{Sup(A \to C)}{Sup(A)} \tag{4}$$

– Intérêt (Interestingness) : proposé par freitas et al. Pour calculer l'intérêt et la pertinence d'une règle.

$$intret = \frac{RAI + CAI}{2} \tag{5}$$

Avec :
Le degré de pertinence d'antécédent de la règle (RAI) est :

$$RAI = 1 - \left[\frac{\frac{\sum_{i=1}^{n} InfoGain(A_i)}{n}}{log_2(|G_k|)} \right] \tag{6}$$

Avec $InfoGain(A_i)$ est le gain d'information de chaque attribut. Le degré de pertinence de la conséquence de la règle (CAI) est :

$$CAI = (1 - Pr(G_{kl}))^{1/\beta} \tag{7}$$

Avec :
– G_{kl} est la probabilité à priori de la valeur de l'attribut de but.
– β est un paramètre spécifié par l'utilisateur.
– $1/\beta$ est un paramètre pour réduire l'influence de la conséquence de la règle.

– *Jmesure* : exprime la dissemblance entre la distribution priori$f(y)$, i.e. $f(Y = y)$ et $f(Y \neq y)$, et la distribution à posteriori $f(Y|\overrightarrow{X})$. est calculé comme suit :

$$J_M = f(x)(f(y|x)ln(\frac{f(y|x)}{f(y)}) + (1?f(y|x)).\frac{1-f(y|x)}{1-f(y)}) \tag{8}$$

– *Lift* : peut être considéré comme le rapport de l'appui observé à celle attendue si A et C étaient statistiquement indépendants :

$$Lift(X \rightarrow Y) = \frac{\sigma(X \subset Y)}{\sigma(X) * \sigma(Y)} \tag{9}$$

Enfin, la valeur de fitness (qualité) est calculée comme suit :

$$F(x) = \frac{\omega_s * S + \omega_c * C + \omega_l * L + \omega_i * I + \omega_j * J_m}{\omega_s + \omega_c + \omega_i + \omega_l + \omega_j} \tag{10}$$

avec ω_s, ω_c, ω_l, ω_i et ω_j sont les poids de chaque mesure.

4 Expérimentation

Dans les deux sections précédentes, nous avons présenté une approche générique pour l'interprétation des changements aléatoires dans les STIS. Pour mettre au point notre contribution, nous avons réalisé un prototype qui permet de matérialiser toutes les phases de l'approche proposée en exploitant une série d'images satellitales. Nous nous intéressons au cas de la végétation. La base d'image utilisée correspond à la région « Ain-Drahem, Tunisie » provenant du capteur MODIS (spectroradiomètre imageur à résolution modérée) pour la période allant du 18/02/2000 jusqu'au 17/11/2012. En recueillant 23 images par an à une résolution spatiale de 250 mètres. Ces images sont valables gratuitement sur le site : http ://glovis.usgs.gov/. Ces images présentent l'évolution de l'indice de végétation par différence normalisé (NDVI) comme illustré dans la figure 2. NDVI est un indice spécifié pour la description de la végétation. Il est calculé à partir des canaux rouges (R) et proches infra rouge (PIR). NDVI = (PIR-R)/PIR+R). Et dans le but de chercher les causes de changements aléatoires, nous avons collectés des données climatiques de la même zone. Ces données sont récupérées de l'institut national météorologique tunisien. La série des données NDVI subi une décomposition additive en trois composantes comme illustré dans la figure 2.

Les variations aléatoires, ainsi que les données climatiques vont subir une fuzzification qui consiste à transformer les valeurs nettes dans les grades de membres pour les termes linguistiques des ensembles flous comme illustré dans le tableau 1. La fonction d'appartenance est utilisée pour associer une note à chaque terme linguistique. Cette étape est faite par un expert

TAB. 1 – *Liste des attributs et leurs valeurs*

	Nom de l'attribut	Valeurs de l'attribut
1	Température	Très faible, faible, moyenne et forte
2	Précipitation	Faible, moyenne, forte et très forte
3	Variation aléatoire	Faible, moyenne et forte

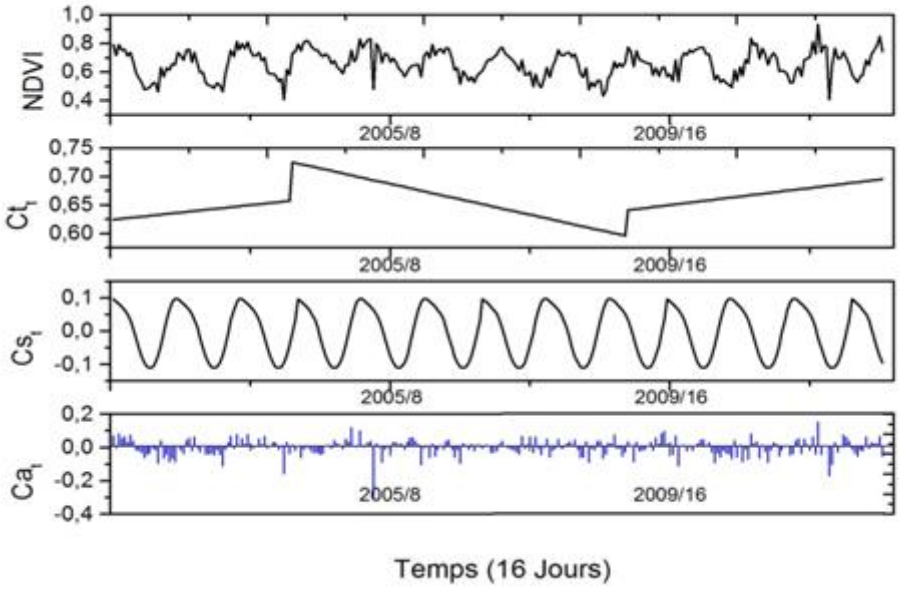


FIG. 2 – *Série originale et résultats de décomposition : Ct_t : La composante tendancielle , Cs_t : la composante saisonnière et Ca_t : La composante aléatoire*

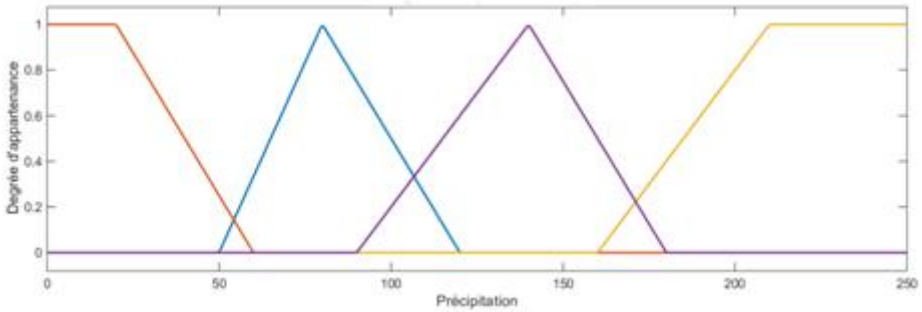


FIG. 3 – *Fuzzification de la précipitation*

Cette étape est faite par un expert

Les règles d'association extraites décrivent les variations aléatoires détectées dans notre série comme illustré dans la figure 4. Chaque variation aléatoire dispose d'un ensemble de causes et peut influencer l'activité de la végétation de la phrénologie.

Un exemple d'exécution de l'AG qui génère un sous-ensemble de règles donné ci-dessous :

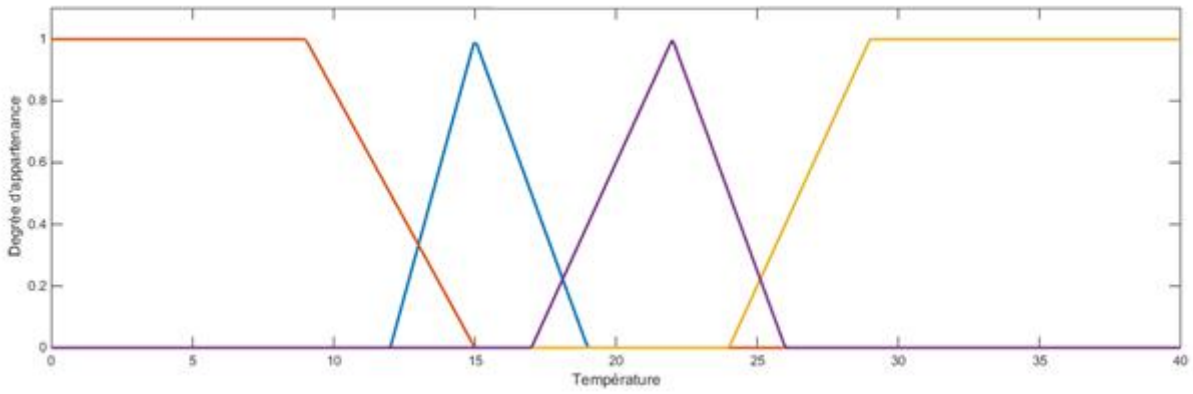


FIG. 4 – *Fuzzification de la température*

- Règle 1 : Si la température est moyenne, la variation aléatoire est faible.
- Règle 2 : Si la température est moyenne et la précipitation est forte, la variation aléatoire est faible.
- Règle 3 : Si la précipitation est moyenne, la variation aléatoire est moyenne.
- Règle 4 : Si la température est moyenne et la précipitation est moyenne, la variation aléatoire est moyenne.

Le tableau 2 présente quelques exemples de règles obtenus pour chacune des trois valeurs de l'attribut but variation aléatoire. Nous indiquons pour chaque règle la valeur de cinq mesures de qualités utilisées.

TAB. 2 – *interprétation des règles*

Règle	**JMesure**	**Support**	**Confiance**	**Lift**	**Intèrêt**
But : Variation aléatoire=Forte					
Si (Précipitation=moyenne) Alors (Variation aléatoire=forte)	0.024	0.375	0.166	0.410	0.067
But : Variation aléatoire=Moyenne					
Si (Température=forte) Alors (Variation aléatoire=Moyenne)	0.032	0.312	0.400	1.828	0.087
Si ((Précipitation=moyenne) et (Température=moyenne)) Alors (Variation aléatoire=Moyenne)	0.017	0.281	.0333	1.523	0.072
But : Variation aléatoire=Faible					
Si (Température=moyenne) Alors (Variation aléatoire=Faible)	0.015	0.125	0.500	1.777	0.140
Si ((Précipitation=forte) et (Température=moyenne)) Alors (Variation aléatoire=Faible)	0.015	0.120	0.550	1.512	0.200

Nous donnons dans le tableau 2 les deux meilleures règles : (par rapport à la mesure Jm)

obtenus pour chacune des trois. Nous indiquons pour chaque règle la valeur de mesure Jm, la valeur du support, du lift, intérêt et la confiance. Si l'on considère le critère de la mesure Jm, nous constatons que :

- Pour le but variation aléatoire=moyenne : R3 et R4, nous obtenons la meilleur règle ainsi que d'autres règles de bonne qualité.
- Pour le but variation aléatoire=faible : R1 et R2 nous obtenons la meilleur règle qui ne présentent que des règles avec des mesures Jm quasi nulles (Jm=0.01), de faible support (=0.12) alors que d'une confiance moyenne (=0.5).
- Pour le but variation aléatoire=forte une seule règle.

Ainsi pour le critère de la mesure Jm, l'algorithme proposé apparait performant sur ce type de données. Ce tableau indique également la valeur d'autres mesures pour ces règles et montre qu'une règle très bonne par rapport à un critère, peut l'être beaucoup moins par rapport à un autre. Ce tableau montre aussi l'importance fondamentale du choix du critère en fonction des règles recherchés.

5 Conclusion

Dans cet article, nous avons proposé une approche pour l'extraction des connaissances à partir des STIS. Nous avons axé notre travail sur l'interprétation des changements aléatoires. L'originalité principale de notre étude réside dans l'utilisation des algorithmes génétiques pour l'extraction des règles d'associations floues pertinentes qui décrivent les liens cachés entre les facteurs extérieurs et leurs influences sur la génération des variations aléatoires. Ces règles, aident les experts à analyser et prévoir ces changements. La validation du modèle a fournit des résultats satisfaisantes. Plusieurs axes de recherches peuvent être envisagés de notre travail sur le plan méthodologique et applicatif. Sur le plan méthodologique nous envisageons d'intégrer un module intelligent pour la prédiction des changements aléatoires en se basant sur notre base de connaissances. Sur le plan applicatif, nous comptons faire d'autres tests pour améliorer la performance de notre méthodologie et l'appliquer dans d'autres domaines d'applications.

Références

Ahmed, B. et R. Ahmed (2012). Modeling urban land cover growth dynamics using multi-temporal satellite images : A case study of dhaka, bangladesh. *ISPRS International Journal of Geo-Information 1*(1), 3–31.

Ben Abbes, A., H. Essid, I. R. Farah, et V. Barra (2014). An adaptive multiplicative decomposition of non stationary multi-temporal satellite images : Application to urban changes detection. In *IPAS 2014 : First International Image Processing, Applications and Systems Conference*, pp. 1 – 7.

Ben Abbes, A., H. Essid, I. R. Farah, et V. Barra (2015). Rare events detection in NDVI time-series using jarque-bera test. In *IGARSS 2015 : IEEE International Geoscience and Remote Sensing Symposium*.

Bonansea, M., M. C. Rodriguez, L. Pinotti, et S. Ferrero (2015). Using multi-temporal landsat imagery and linear mixed models for assessing water quality parameters in río tercero reservoir (argentina). *Remote Sensing of Environment 158*, 28–41.

Brooks, C., A. Grimm, R. Shuchman, M. Sayers, et N. Jessee (2015). A satellite-based multitemporal assessment of the extent of nuisance cladophora and related submerged aquatic vegetation for the laurentian great lakes. *Remote Sensing of Environment 157*, 58–71.

Dutta, D. et G. N. Burdwan (2009). Discovering prediction rules for predicting quality of students using multi objective genetic algorithm from student database. In *Proceedings of International Conference on recent trends in Computing and Communications (FACT 2009), KCG College of Technology, Chennai, India*, pp. 19–24.

Flamary, R., M. Fauvel, M. Dalla Mura, et S. Valero (2015). Analysis of multitemporal classification techniques for forecasting image time series. *Geoscience and Remote Sensing Letters, IEEE 12*(5), 953–957.

Grieser, J., S. Trömel, et C.-D. Schönwiese (2002). Statistical time series decomposition into significant components and application to european temperature. *Theoretical and applied climatology 71*(3-4), 171–183.

Hutchinson, J., A. Jacquin, S. Hutchinson, et J. Verbesselt (2014). Monitoring vegetation change and dynamics on us army training lands using satellite image time series analysis. *Journal of Environmental Management*.

Jamali, S., P. Jönsson, L. Eklundh, J. Ardö, et J. Seaquist (2015). Detecting changes in vegetation trends using time series segmentation. *Remote Sensing of Environment 156*, 182–195.

Lillesand, T., R. W. Kiefer, et J. Chipman (2014). *Remote sensing and image interpretation*. John Wiley & Sons.

Mankad, K., P. S. Sajja, et R. Akerkar (2011). Evolving rules using genetic fuzzy approach : an educational case study. *International Journal on Soft Computing 2*(1), 35–46.

Martínez, B. et M. A. Gilabert (2009). Vegetation dynamics from ndvi time series analysis using the wavelet transform. *Remote Sensing of Environment 113*(9), 1823–1842.

Petitjean, F., C. Kurtz, N. Passat, et P. Gançarski (2012). Spatio-temporal reasoning for the classification of satellite image time series. *Pattern Recognition Letters 33*(13), 1805–1815.

Rotem-Mindali, O., Y. Michael, D. Helman, et I. M. Lensky (2015). The role of local land-use on the urban heat island effect of tel aviv as assessed from satellite remote sensing. *Applied Geography 56*, 145–153.

Thakkar, A., V. Desai, A. Patel, et M. Potdar (2015). Land use/land cover classification using remote sensing data and derived indices in a heterogeneous landscape of a khan-kali watershed, gujarat. *Asian Journal of Geoinformatics 14*(4).

Verbesselt, J., R. Hyndman, A. Zeileis, et D. Culvenor (2010). Phenological change detection while accounting for abrupt and gradual trends in satellite image time series. *Remote Sensing of Environment 114*(12), 2970–2980.

Wakabi-Waiswa, P. P., V. Baryamureeba, et K. Sarukesi (2011). Optimized association rule mining with genetic algorithms. In *Natural Computation (ICNC), 2011 Seventh International Conference on*, Volume 2, pp. 1116–1120. IEEE.

Zadeh, L. A. (1973). Outline of a new approach to the analysis of complex systems and decision processes. *Systems, Man and Cybernetics, IEEE Transactions on* (1), 28–44.

Summary

Satellite image time series (SITS) represent a source of valuable information for studying land-use. In fact, land-use interpretation from high temporal resolution remotely sensed data is important to promote better decisions for sustainable management land cover. Change in land use can be divided into three classes: seasonal, trend or random. This latter is the subject of this paper. The purpose is to present a datamining method for extracting valid and reliable knowledge to explain random change. However, discovering knowledge required to address several challenges related to the characteristics of SITS and their constraints. Indeed, for learning is based on genetic algorithm (GA). We validate this work on a series of MODIS images.

Méthode Déterministe pour la Fragmentation Horizontale des Entrepôts de Données

Mohamed Barr*, Kamel Boukhalfa** Karima Hocine**

*Ecole Nationale Supérieure d'Informatique
m_barr@esi.dz
**Université des Sciences et de la Technologie Houari Boumediene
{kboukhalfa, kbouibede}@usthb.dz

Résumé. L'appartenance du problème de sélection des structures d'optimisation des entrepôts de données relationnels à la classe des problèmes NP-Complets, rend leur résolution à l'aide des méthodes exhaustives, une tâche difficile. Ces structures peuvent être utilisées dans un contexte centralisé, distribué ou parallèle. Dans la littérature, il existe peu de travaux consacrés à la prise en charge des problèmes de sélection de ces structures ou techniques qui se basent sur des méthodes exactes. Dans ce présent travail, nous avons proposé une nouvelle méthode déterministe pour obtenir des schémas de fragmentation horizontale qui optimise une charge des requêtes. La méthode a été validée en utilisant le Benchmark APB1.

Mots Clés. Entrepôts de Données, Conception Physique, Fragmentation Horizontale, Optimisation.

1 Introduction

Dans (Gardarin, 2003), l'auteur propose cinq étapes fondamentales pour la conception d'une base de données : (1) perception du monde réel et capture des besoins, (2) élaboration du schéma conceptuel, (3) conception du schéma logique, (4) affinement du schéma logique, et (5) élaboration du schéma physique. Les bonnes performances d'une application sont étroitement liées à la maitrise la dernière étape en fonction de la prise en considération des transactions pour identifier les patterns d'accès fréquents. Puis, il découle le bon choix des structures physiques telles que le partitionnement, l'indexation, la matérialisation des vues, etc.

Le partitionnement (ou fragmentation) dans les bases de données relationnels a été toujours considéré comme un sujet de recherche d'actualité compte tenu, d'une part, de sa nature non redondante qui ne nécessite ni espace de stockage supplémentaire ni mise à jour, et d'autre part, de ses apports en terme d'amélioration de la disponibilité, la facilité de gestion, et la performance dans les entrepôts de donnée (Bellatreche, 2000; Boukhalfa, 2009; Mahboubi, 2008; Özsu & Valduriez, 1991; Aouiche, Boussaid, & Bentayeb, 2005).

Dans notre présent travail, nous nous intéressons à la prise en charge du problème de sélection d'un schéma de Fragmentation Horizontale (FH) dans les entrepôts de données (ED) relationnels en utilisant une méthode déterministe qui explore les correspondances exactes des prédicats de sélections contenus dans la charge des requêtes, au niveau des transactions

de la table des faits. Théoriquement, sur la base de la méthode des prédicats (Dimovski, Velinov, & Sahpaski, 2010), pour *n* prédicats de sélection, nous pouvons générer 2^n fragments différents pour un schéma qui introduit la totalité des *n* prédicats. Ce nombre représente un chiffre extrêmement important si *n* est élevé ce qui complique le problème de distribution de l'entrepôt fragmenté sur les nœuds du réseau. Dans (Derrar, Nacer, & Boussaid, 2013), les auteurs rappellent la complexité du problème d'allocation qui s'élève à $O(k^{2n})$ pour un réseau de k nœuds avec n prédicats simples de sélection. Dans le présent travail, nous proposons une méthode déterministe basée sur les relations entre prédicats aux niveaux des transactions de la table des faits pour fragmenter l'ED. Nous nous sommes intéressés dans un premier temps au problème de FH dans un contexte centralisé et une charge des requêtes stable.

La section 2 présente un état de l'art. Nous détaillons notre approche dans la section 3 et nous présentons l'étude expérimentale dans la section 4. La section 5 conclut le papier et présente quelques perspectives.

2 Etat de l'art : Travaux sur la FH

Avant de présenter les travaux effectués sur la FH, nous abordons la complexité du processus de FH. Dans (Boukhalfa, 2009), le nombre de schémas de FH possibles basé sur n prédicats est égal au nombre de Bell ($B(n) \approx n^n$ lorsque n est très grand). Ce nombre de schémas extrêmement important montre la complexité exponentielle du problème de sélection de la FH. Nous rappelons ici que le nombre B(n) ci-dessus exprimé concerne uniquement un seul attribut, et calculé uniquement pour une FH primaire de la table des faits. Si k attributs sont impliqués dans l'exploration des schémas de FH, le nombre devient beaucoup plus élevé.

Travaux de Bellatreche et al : La méthodologie de FH proposée par l'auteur s'applique aux ED en étoile. Les auteurs proposent de fragmenter les tables de dimension ensuite utiliser leurs schémas de FH pour fragmenter la table des faits. L'algorithme utilisé commence par l'extraction des prédicats simples de sélection, puis il calcule leurs sélectivités. Il répartit ensuite les prédicats sur les tables de dimensions correspondantes (les tables de dimension n'ayant pas de prédicats de sélection sont écartées du processus de FH, les autres sont fragmentées). La dernière étape consiste à la FH dérivée de la table des faits en se basant sur la FH primaire des tables de dimension.

Le problème rencontré dans cette méthode réside dans la génération d'un nombre important de fragments. Pour remédier à ce problème, l'auteur a proposé un algorithme glouton qui part d'un choix aléatoire d'une table de dimension, et procède à la FH tout en respectant un nombre de fragments N à ne pas dépasser à condition que le schéma généré réduit le coût d'exécution des requêtes, dans le cas contraire le schéma sera rejeté.

Travaux de Boukhalfa et al. : Dans ces travaux, le nombre de fragments est contrôlé par un seuil maximum fixé par l'administrateur. La démarche se base sur un modèle de coût qui estime le nombre d'E/S nécessaire pour exécuter chaque requête sur l'ED. L'approche exploite trois types de connaissances: les informations sur l'ED (tailles des différentes tables, nombre de pages nécessaires pour stocker chaque table etc.), les informations système (taille du cache, la taille de la page système, etc.), les informations relatives à la charge des requêtes (fréquences, facteurs de sélectivité, etc.).

La démarche adoptée par les auteurs passe par quatre étapes : (1) Une préparation de la FH qui s'occupe de l'identification des prédicats à partir de la charge des requêtes, puis les attributs auxquels appartiennent les prédicats sont découpés en sous-domaines. Cette étape aboutit à un ensemble de prédicats minimal et complet ; (2) chaque sous-domaine est assigné à un numéro qui représente le code d'un gène dans une solution ou « individu » au sens de l'algorithme génétique (initialement une table de dimension fragmentée suivant un attribut de n sous-domaines, contiendra n fragments) ; (3) génération d'une solution initiale à base de l'algorithme d'affinités entre les différents sous-domaines et (4) Chercher la meilleure solution en utilisant une méta-heuristique. Deux heuristiques à trajectoires ont été utilisées (Hill Climbing et Recuit Simulé) et une heuristique à population (Algorithme Génétique).

Travaux de Barr et al. : Le travail réalisé dans (Barr & Bellatreche, 2010), a traité le problème de sélection de schéma de FH en se basant sur la méta-heuristique de colonie de fourmis où le problème a été assimilé à un problème de sac à dos tel que les prédicats représentent les objets à mettre dans le sac. Un fragment F_i quelconque a comme poids son coût de chargement. A chaque prédicat correspond un profit qui est égal au coût de chargement de toute la table des faits (-) moins celui du fragment F_i. Des expérimentations ont été réalisées sur le Benchmark APB1 sous le SGBD Oracle. Le calibrage de l'algorithme (le dépôt, l'évaporation, l'importance de phéromone, la visibilité) a abouti aux choix des prédicats qui ont alimenté le modèle d'Özsu. Le schéma généré fournit un certain nombre de fragments volumineux qui ont été éclatés d'une manière intuitive qui a générée une réduction du coût global de la charge des requêtes d'une manière significative (Barr, 2011; Barr & Boukhalfa, 2013).

Travail de Bouchakri et al : Dans (Boukhalfa, 2009), une similarité entre la FH et les Index binaires a été mis en évidence. Les auteurs, ont identifié un ensemble de problèmes sur les travaux qui ont combiné les deux techniques à la fois : (1) tous les attributs sont introduits dans le processus de FH dans le cadre du mode combiné. Ce qui augmente le nombre de schémas de fragmentation ; (2) réalisation séparée de sélection des deux techniques ; (3) la sélection des IJB après la FH peut offrir des index inutiles et (4) pas de distinction significative des attributs préférables pour les IJB par rapport à ceux choisis pour la FH.

Pour remédier au problème de chevauchement de partage des attributs entre la sélection des IJB et la FH, les auteurs dans (Bouchakri, Bellatreche, & Boukhalfa, 2010) ont tenu compte de trois facteurs essentiels pour favoriser un ensemble d'attributs pour définir l'une des deux techniques : la cardinalité de l'attribut, la fréquence de son utilisation dans la charge des requêtes et le volume de données chargé relativement à cet attribut.

La méthode utilisée dans ce travail consiste à répartir l'ensemble des attributs sur deux classes Class_IJB et Class_FH en se basant sur l'algorithme de classification « k-means ».

Travail de Gacem et al. : (Gacem & Boukhalfa, 2013), proposent une méthode de FH dans le contexte d'une charge de requête volumineuse. Le principe général de la méthode part de l'hypothèse qui suppose que l'importance de la taille d'une charge de requêtes introduisant un nombre important de prédicats simples, mène à une explosion de schémas de FH possibles. En effet, le travail réalisé vise l'aboutissement à deux objectifs : la réduction de la taille de la charge et la sélection d'un schéma de FH optimisant en même temps la charge réduite et celle d'origine. Les auteurs ont définis trois relations entre les différentes requêtes de la charge : (1) l'intersection qui représente le cas de l'accès aux données communes des tables (faits ou dimensions) pour des requêtes quelconques ; (2) l'exclusion qui traduit une intersection vide des requêtes, et (3) l'inclusion entre requêtes qui met en évidence la conte-

nance d'un ensemble de données retourné par une requête par rapport à une autre requête. Les auteurs utilisent l'algorithme K-means pour classer les requêtes. Pour chaque classe de requête obtenue, une requête représentative est élue. Ces requêtes sont utilisées pour fragmenter l'ED.

Travail de Derrar et al. : Dans (Derrar et al., 2013), les auteurs ont proposé une approche basée sur les statistiques d'accès aux données de l'ED pour une FH dynamique de l'ED. Ils signalent que les algorithmes utilisés dans la littérature traitent le problème de FH pour une charge de requêtes statiques, et tout changement dans la charge implique une réadaptation de l'algorithme (Bellatreche & Boukhalfa, 2005).

L'approche est constituée de deux éléments principaux qui sont le critère d'évaluation et la structure de donnée utilisée pour observer et sauvegarder l'accès aux données. Le critère d'évaluation permet d'estimer l'exécution ou pas du processus de ré-fragmentation. Il peut être le coût d'accès aux données, le coût de transfert de données entre fragments, ou le coût de transfert s'il s'agit d'un contexte distribué. La structure la plus appropriée pour les auteurs est l'histogramme.

Les travaux proposés dans la littérature pour la FH d'un ED utilisent souvent des métaheuristiques qui se basent souvent sur des modèles théoriques qui ne prennent pas en considération le contenu réel de l'ED. Par conséquent, plusieurs fragments générés sont vides car aucune donnée dans l'entrepôt ne vérifie la clause de prédicats utilisée lors de sa création. Cela augmente le coût de maintenance d'un nombre de fragments important sans une réelle raison d'être de ces derniers. Pour remédier à ces problèmes, nous proposons, dans ce travail, une approche déterministe qui se base sur les correspondances de prédicats existant réellement dans l'ED.

3 Présentation générale de notre méthode

Notre approche comporte deux phases essentielles, la première concerne l'affectation des prédicats aux niveaux des lignes ou enregistrements de la table des faits en fonction des clés primaires des tables de dimension (clés étrangères de la table des faits). Cette phase a en entrée l'ED non fragmenté et la charge des requêtes contenant les prédicats simples de sélection qui représentent les objets de base pour la méthode de FH que nous avons proposé. En sortie de la première phase, nous avons l'ensemble des groupes similaires qui contiennent les lignes avec une forte similarité au sens où ces lignes vérifient les mêmes prédicats simples de sélection. Le rangement du même groupe dans un fragment à part représente le fond de notre méthode de FH.

3.1 Méthode de fragmentation

Le principe de notre méthode de FH se base sur l'identification des collections similaires en fonction des valeurs des prédicats contenus dans la charge des requêtes. Ces collections contiennent des enregistrements de la table des faits regroupés en fonction des valeurs des attributs des tables de dimensions jointes à la tables des faits via les clés étrangères (clés primaires des tables de dimensions).

Pour formaliser notre méthode d'une manière plus précise, nous pouvons l'assimiler à un problème d'affectation défini comme suit :

– Tout prédicat de sélection P, relié à un attribut A d'une table de dimension D, contenu dans la charge des requêtes et dont le résultat de son interrogation par une requête Q n'est pas vide, alors il existe une ligne L (ou un enregistrement) de la table des faits F où ce prédicat est affecté (via la clé étrangère, clé primaire respectives de D et de F) à cette ligne
– Un prédicat P de valeur V, ne peut pas appartenir à la fois à deux fragments différentes, mais le même prédicat P peut figurer dans n lignes différentes d'une manière différée, où Le nombre total de lignes n (où il appartient) divisé par la taille |F| de F, représente la sélectivité de P
– La somme des lignes de l'ensemble des prédicats couvrant tout le domaine d'un attribut A, est égale à la taille |F| de F.

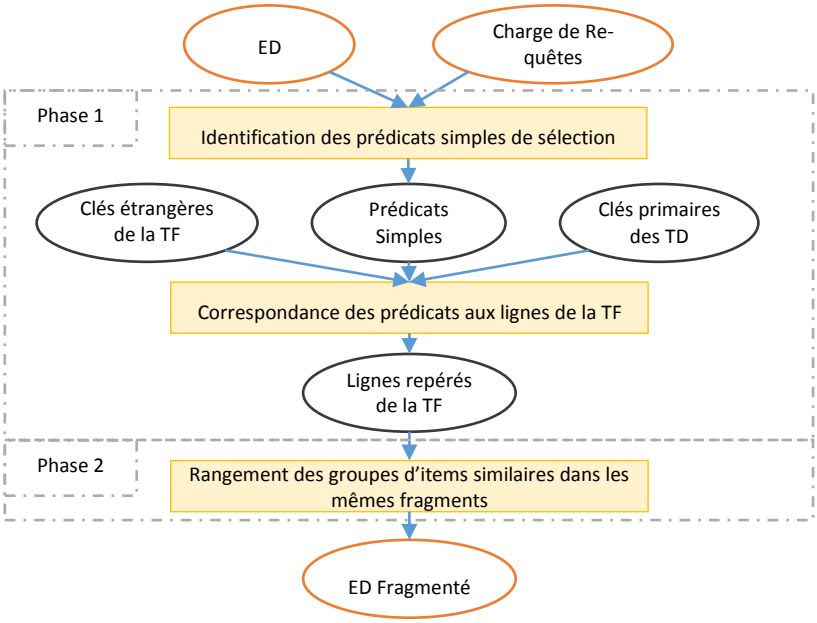


FIG 1. *Schéma global de la solution proposée*

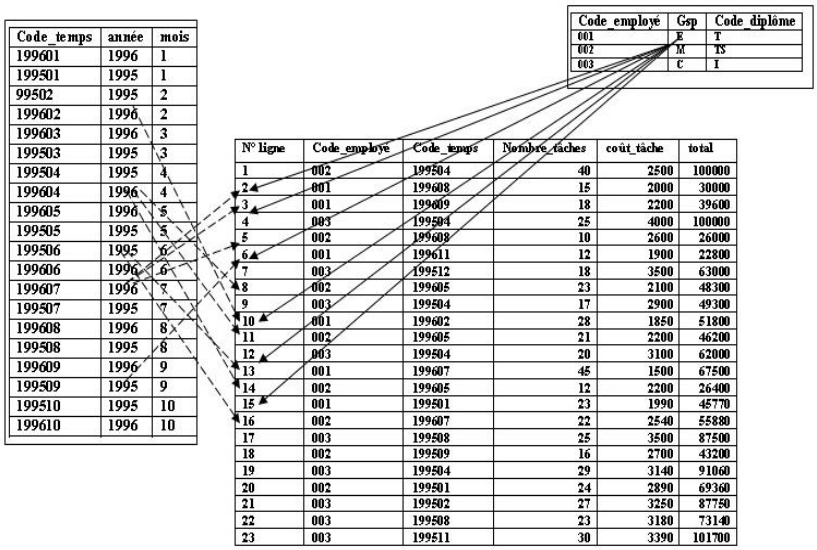
Le principe de la méthode se résume tout simplement par le rangement de toutes les collections similaires dans le même fragment du schéma. Nous entendons par collections similaires la correspondance d'un même ensemble de prédicats ou de conjonctions de prédicats à un nombre de lignes de la table des faits.

Exemple. Soit le modèle de données composé d'une table de faits « Réalisations » et deux tables de dimensions « Employé » et « Temps », définies comme suit (les instances sont dans FIG 2) :
Employé (Code_employé, GSP,Code_diplôme)
Temps (Code_temps, année, mois)
Réalisations (Code_employé, Code_temps, Nombre_tâches, coût_tâche, total)

Étape d'Affectation des prédicats aux lignes de la table des faits : Nous remarquons dans l'exemple de la FIG 2, comment la valeur « E » (qui signifie « Exécution » pour la catégorie Groupe Socio-Professionnelle des employés) de l'attribut GSP de la table de dimension « Employé » est affectée aux lignes Li de la table des faits « Réalisations». En effet dans cet exemple les valeurs :

- « E » (Exécution) va appartenir à l'ensemble des lignes E = {2,3,6,10,13,15}
- « M » (Maitrise), sera affectée à l'ensemble des lignes M= {1,5, 8, 11, 14, 16, 18, 20}
- « C » (Cadre), va correspondre à l'ensemble de lignes C={4, 7, 9 , 12 , 17 , 19, 21 , 22 , 23}

Pour le deuxième attribut de la table Employé, nous remarquons que les valeurs « T » (Techniciens), « TS » (Techniciens Supérieurs), et « I » (Ingénieurs », seront affectées respectivement aux lignes d'affectation de « E », « M », et « C », car dans cet exemple chaque Groupe Socio – Professionnelle correspond à une seule valeur du code de diplôme.



Code_temps	année	mois
199601	1996	1
199501	1995	1
99502	1995	2
199602	1996	2
199603	1996	3
199503	1995	3
199504	1995	4
199604	1996	4
199605	1996	5
199505	1995	5
199506	1995	6
199606	1996	6
199607	1996	7
199507	1995	7
199608	1996	8
199508	1995	8
199609	1996	9
199509	1995	9
199510	1995	10
199610	1996	10

Code_employé	Gsp	Code_diplôme
001	E	T
002	M	TS
003	C	I

N° ligne	Code_employé	Code_temps	Nombre_tâches	coût_tâche	total
1	002	199504	40	2500	100000
2	001	199608	15	2000	30000
3	001	199609	18	2200	39600
4	003	199504	25	4000	100000
5	002	199608	10	2600	26000
6	001	199611	12	1900	22800
7	003	199512	18	3500	63000
8	002	199605	23	2100	48300
9	003	199504	17	2900	49300
10	001	199602	28	1850	51800
11	002	199605	21	2200	46200
12	003	199504	20	3100	62000
13	001	199607	45	1500	67500
14	002	199605	12	2200	26400
15	001	199501	23	1990	45770
16	002	199607	22	2540	55880
17	003	199508	25	3500	87500
18	002	199509	16	2700	43200
19	003	199504	29	3140	91060
20	002	199501	24	2890	69360
21	003	199502	27	3250	87750
22	003	199508	23	3180	73140
23	003	199511	30	3390	101700

FIG 2. *Exemple d'un ED d'une table de faits et deux tables de dimension*

Étape de Rangement des lignes de la table des faits dans des partitions : Au sens du prédicat Employé.Gsp='E' ; les lignes de l'ensemble E sont similaires et seront rangées dans un même fragment. Idem, pour les prédicats, Employé.Gsp= « M » et Employé.Gsp= « C », les lignes des ensembles M et C vont appartenir aux mêmes fragments.

Éléments statistiques de l'exemple : Nous avons la cardinalité de la table « Réalisations » (Card(R)=Card(E) + Card(M) + Card(C)=23)
R= E∪M∪C (complétude)
E∩M∩C=∅ (Disjonction)
Soit une requête Q_1 exprimée comme suit :

Select code_employé, sum(total) from Réalisations R, Employé E
Where R.code_employé=E.Code_employé and E.Gsp='M' group by code_employé

Si nous imaginons une information fictive reliée aux différents ensemble E, M et C, appelée respectivement IFE, IFM et IFC.
Supposons que les nouvelles relations (fragments de la table Réalisations) RE, RM et RC liées à E, M et C seront identifiées par IFE, IFM et IFC.
La requête Q_1 peut être écrite de la façon suivante :
Select code_employé, sum (total) from RM where Id_RM=IRM group by code_employé.
Si maintenant, nous introduisons les prédicats correspondants aux valeurs de l'attribut année de la table Temps.
AN1= ensemble des lignes de l'année 1995 ={1, 4, 7, 9, 12, 15, 17, 18, 19, 20,21,22,23}
AN2= ensemble des lignes de l'année 1996 ={2, 3, 5, 6, 8, 10, 11, 13, 14, 16}
Il est clair que la cardinalité Card(Réalisations)= Card(AN1) + Card(AN2) =23 (Complétude) et AN1 ∩ AN2 = ∅ (Disjonction)
Par rapport à un processus de FH qui implique uniquement les deux prédicats *Temps.Année=1995* et *Temps.Année=1996*, la table des faits « réalisations » peut être fragmentée en deux fragments RAN1 et RAN2.
Pour l'implication de deux attributs, le rangement des prédicats sera effectué tout en assurant les opérations ensemblistes d'intersection suivantes. Nous notons par L les lignes d'affectation dans la table des faits, d'un ensemble de prédicats énuméré.
*L(AN1) ∩ L (E) = {**15**}, L(AN1) ∩ L(M) = {**1,18,20**}, L(AN1) ∩ L(C)={**4, 7, 9, 12 , 17 , 19, 21 , 22 , 23**}, L(AN2) ∩L (E) ={**2, 3, 6, 10, 13**}, L(AN2) ∩ L(M) = {**5, 8, 11, 14, 16** }*
*L(AN2) ∩ L(C) = **∅***
Au vu d'un processus de FH introduisant les prédicats des deux attributs Employé.Gsp et Temps.Année, nous avons obtenus six (06) fragments dont un fragment vide.

3.2 Algorithme proposé

Les algorithmes utilisés dans la littérature pour élaborer un schéma de FH sont basés soit sur les prédicats, soit sur l'affinité des prédicats, ou bien sur un modèle de coût. La majorité des travaux basés sur ces algorithmes ne tient pas compte des relations qui existent au niveau transactions. ALG1 représente l'algorithme que nous proposons.

```
INPUT : ED de données non fragmenté, Ensemble de prédicats Pi de la charge
Pour chaque prédicat Pi d'une table de dimension Dk
i=1
Faire
    Lire la clé primaire de Dk
Pour j allant de 1 à |F|
Faire
        Identifier la ligne Lj correspondante à Pi
        Affecter Pi à Lj
Fin Faire
    i=i+1
Fin Faire
 Ranger les combinaisons similaires de prédicats dans les mêmes fragments
```

ALG1. *Pseudo-algorithme proposé pour la FH*

Par rapport à l'approche basée sur les prédicats, notre méthode ne passe pas forcément par les étapes de : (1) génération d'un ensemble complet et minimal de prédicats, (2) génération de l'ensemble des minterms et (3) simplification des minterms.

Pour la première étape, notre algorithme affecte les prédicats d'une manière réelle aux différentes lignes de la table des faits d'une part, et se base sur le respect des intégrités du modèle relationnel utilisé qui garantit l'absence des minterms contradictoires qui n'ont pas besoin d'être éliminés, d'une autre part. Pour ce qui est de la deuxième et la troisième phase, notre algorithme, génère l'ensemble des minterms effectif.

Concernant la deuxième approche qui regroupe les prédicats subissant les mêmes fréquences d'exécution des requêtes qui a été critiquée dans (Derrar et al., 2013). En effet, la matrice d'affinité utilisée ne relie que les attributs ou les prédicats deux à deux. Notre démarche suppose que le fait de rassembler les prédicats qui vont ensemble dans les enregistrements reflète un nouveau sens d'affinité tel que si une requête quelconque Q invoque l'ED suivant un prédicat P_1 à une fréquence f, alors tout prédicat P_2 liés avec P_1, aura la même fréquence d'exécution dans le cadre du même fragment. Ce type d'affinité aux niveaux enregistrements n'a pas la même logique d'affinité utilisée dans la méthode basée sur l'affinité, mais ça peut intéresser le décideur pour avoir périodiquement les prédicats qui vont ensemble dans un but métier.

En comparant notre méthode à celle basée sur un modèle de coût, nous notons que le fait de ranger les mêmes conjonctions de prédicats dans les mêmes fragments veut dire que la somme des tailles (en nombre de lignes) de ces fragments donne exactement le nombre d'occurrences d'un prédicat quelconque appartenant à cette collection, et par conséquent l'invocation de toutes les lignes où figure le prédicat suffit pour remplacer la partie de restriction de la requête interrogeant la table de faits pour ce prédicat.

Pour la partie jointure, notre méthode détermine l'affectation d'un prédicat à une ligne en fonction de la clé primaire de la table de dimension pour l'attribut auquel il correspond. D'où la non nécessité de joindre la table des faits fragmentée à la table de dimension à laquelle est associé le prédicat, après FH, sauf pour la récupération du libellé relatif à la valeur du prédicat.

4 Expérimentations

Nous avons implémenté notre méthode sur un ED relationnel tiré du Benchmark APB1. TAB 1 présente les caractéristiques de l'ED étudié.

Table	Type	Clé primaire	Clé étrangère	Taille
Acvars	Faits			24786000
Prodlevel	Dimension	Code_level	Product_level	9000
Custlevel	Dimension	Customer_level	Store_level	900
Chanlevel	Dimension	Chanal_level	Base_level	9
Timelevel	Dimension	Tid	Time_level	24

TAB 1. *Caractéristiques de l'ED généré à partir du Benchmark APB1*

Nous avons utilisé une charge de 80 requêtes couvrant trente-cinq prédicats de sélection simples appartenant à l'ensemble des neuf attributs autres que les clés primaires des tables de dimension.

Nous avons généré 74 schémas de FH différents, et nous avons évalué la performance de chaque solution. Les résultats obtenus sont dressés dans TAB 2. Nous remarquons que le taux de réduction en nombre d'entrées/sorties est proportionnel au nombre de fragments. Plus le schéma couvre un nombre important de prédicats, meilleur est le résultat. Le meilleur schéma obtenu en termes de nombre d'entrées/sorties minimal est celui qui correspond au couple de nombre de fragments et nombre d'entrées/sorties respectifs (4896, 258382).

N° Solution	# de Prédicats	# de Fragments	# E/S	% de Réduction	N° Solution	# de Prédicats	# de Fragments	# E/S	% de Réduction
1	35	4896	258382	89	38	23	168	708304	69
2	34	4480	408731	82	39	22	156	735356	68
3	33	4080	437411	81	40	13	144	1827383	21
4	34	4080	342724	85	41	21	144	762409	67
5	33	3740	470851	80	42	23	136	772481	66
6	32	3672	466090	80	43	20	132	789462	66
7	31	3264	501597	78	44	12	120	1854435	19
8	33	3264	370152	84	45	19	120	902511	61
9	32	2992	498279	78	46	11	96	1881488	18
10	30	2856	567375	75	47	18	96	928428	60
11	29	2448	608854	74	48	10	72	1908541	17
12	32	2448	406973	82	49	17	72	953790	59
13	28	2040	653693	72	50	9	48	1935594	16
14	27	1632	681917	70	51	16	48	1001328	56
15	31	1632	460706	80	52	8	24	1962646	15
16	24	1224	872410	62	53	14	24	1046979	54
17	25	1224	856957	63	54	15	24	1026690	55
18	26	1224	797330	65	55	7	18	1996616	13
19	23	816	894079	61	56	6	12	2024956	12
20	30	816	486255	79	57	11	12	1438592	37
21	29	748	614382	73	58	12	12	1539373	33
22	29	612	514768	78	59	13	12	1519008	34
23	21	408	958405	58	60	10	11	1454045	37
24	22	408	914444	60	61	9	10	1519512	34
25	28	408	571431	75	62	8	9	1726387	25
26	19	288	1671829	27	63	7	8	1754612	24
27	20	288	1180685	49	64	6	7	1838259	20
28	24	272	744147	68	65	5	6	1894671	18
29	18	264	1697191	26	66	5	6	2053289	11
30	17	240	1722553	25	67	4	5	1971427	14
31	16	216	1747915	24	68	4	5	2171737	6
32	26	204	627146	73	69	3	4	2028914	12
33	27	204	599755	74	70	3	4	2222833	3
34	15	192	1773277	23	71	2	3	2143249	7
35	25	192	654198	72	72	2	3	2248384	2
36	24	180	681251	70	73	1	2	2213443	4
37	14	168	1800330	22	74	1	2	2273933	1

TAB. 2. *Résultats obtenus*

4.1 Nombre théorique de fragments

Comme nous l'vons mentionné, parmi les avantages de la méthode utilisée la non nécessité d'énumérer tous les minterms possibles pour procéder ensuite à l'élimination de ceux qui sont contradictoires. TAB 3. présente le nombre important des minterms théorique comparé à celui pratique trouvé par notre algorithme. Les résultats montrent que les nombres de fragments trouvés par notre algorithme sont beaucoup plus petits à ceux théoriquement possibles. Par exemple, pour 35 fragments, théoriquement 2^{35} = 34 359 738 368 fragments sont possibles or que réellement, 4896 seulement sont envisageables.

4.2 Passage à l'échelle

Les résultats obtenus dans le tableau 2, représentent des solutions intégrant des schémas de FH par rapport à une charge de requêtes contenant 35 prédicats simples de sélection. Nous essayons dans ce qui suit d'introduire le maximum de prédicats et voyons comment le nombre de fragments augmente.

# de prédicats	1125	520	1121	1109	999	984	905	605	35
# théorique de fragments	-	3,432 10^{156}	-	-	5,358 10^{300}	1,635 10^{296}	2,705 10^{272}	1,328 10^{182}	3,436 10^{10}
# réel de fragments	4499679	2248284	4499679	529374	433	433	433	433	4896

TAB 3. *Comparaison entre nombres théoriques et obtenus par la méthode proposée*[1]

Parmi les problèmes rencontrés suite à un processus de FH dans les ED relationnels, nous citons celui de l'explosion des nombres de fragments. Dans le modèle d' Özsu, le nombre s'élève à 2^n ; Dans (Boukhalfa, Bellatreche, & Pascal, 2008), le nombre est égal au produits des fragments des tables de dimension. Ce qui nous donne pour l'ensemble des prédicats du Benchmark APB1 un chiffre égal à 74 519 000 000 000 fragments possibles.

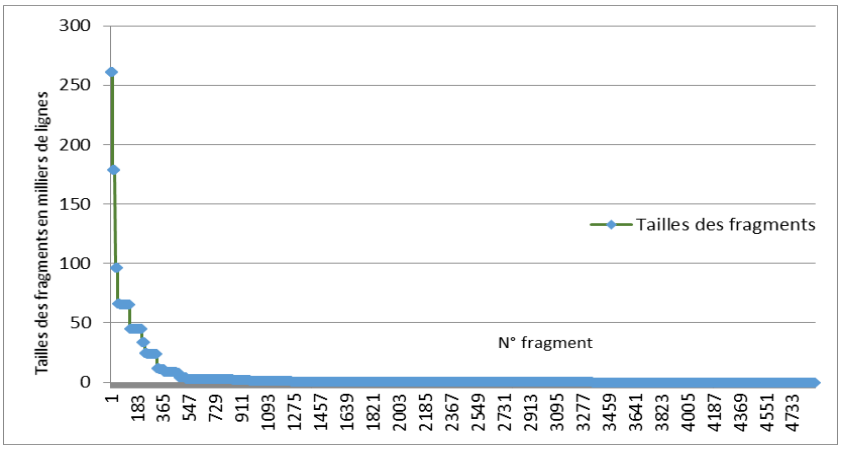


FIG 3. *Taille des fragments*

[1] Les valeurs marquées « - » signifient que la valeur n'a pas pu être calculée (Excel, Oracle)

4.3 Finesse des fragments en tailles

La philosophie principale de la technique de FH est basée sur le morcèlement des tables de l'ED afin de charger en mémoire à chaque fois seulement les lignes pertinentes interrogées par les requêtes. Pour montrer cette idée, nous visualisons les tailles des fragments obtenues suite au processus de FH que nous avons effectué de l'ED du benchmark APB1. Le fragment le plus volumineux contient 261084 lignes et le plus fin contient uniquement une seule ligne (voir FIG 3).

5 Conclusion et perspectives

En conclusion, nous pouvons rappeler que la mise en pratique de la méthode proposée nous a montré qu'il y a un écart extraordinaire entre le nombre de fragments théorique cité dans le modèle d' Özsu et celui déterminé pratiquement. Cela signifie que notre méthode peut représenter une parfaite amélioration de celle d' Özsu. Notre approche a permet de considérer un nouveau sens d'affinité entre les prédicats au niveau transactions qui donne une indication expressive métier pour les managers, et dépasse l'affinité qui rassemble les prédicats seulement en fonction des fréquences des requêtes, chose qui a été critiqué par des chercheurs dans le domaine. Par ailleurs, la mise en exergue de tous les prédicats lors du passage à l'échelle, nous a montré le nombre important de fragments générés qui risque de dégrader la performance à cause de multiples opérations d'union.

Nous envisageons comme perspectives la généralisation de la solution dans le cadre distribué ainsi qu'une association avec d'autres structures comme les index.

Bibliographie

Aouiche, K., Boussaid, O., & Bentayeb, F. (2005). Automatic Selection of Bitmap Join Indexes in Data Warehouses. In *7th International Conference on Data Warehousing and Knowledge Discovery (DAWAK 05)* (p. 64‑73).

Barr, M. (2011). A new Approach based on Activity Sort Ants for resolving of the Horizontal Fragmentation in Relational Data warehouses. In *2011 International Conference on Computer Science and Logistics Engineering (ICCSLE)*. Zhangzhou, China.

Barr, M., & Bellatreche, L. (2010). A new approach based on ants for solving the problem of horizontal fragmentation in relational data warehouses (p. 411‑415). IEEE. http://doi.org/10.1109/ICMWI.2010.5648104

Barr, M., & Boukhalfa, K. (2013). Ants Based Self - Monitoring Approach for The data warehouses Horizontal Partitioning Evolution. *Journal of Convergence Information Technology*, *8*(4), 748‑755. http://doi.org/10.4156/jcit.vol8.issue4.86

Bellatreche, L. (2000). *Utilisation des vues matérialisées, des index et de la Fragmentation dans la Conception Logique et Physique d'un entrepôts de données* (PhD. Thesis). Blaise Pascal University, France.

Bellatreche, L., & Boukhalfa, K. (2005). An Evolutionary Approach to Schema Partitioning Selection in a Data Warehouse Environment. *Proceeding of the International Conference on Data Warehousing and Knowledge Discovery (DAWAK'2005)*, 115‑125.

Bouchakri, R., Bellatreche, L., & Boukhalfa, K. (2010). Une Approche par K-means de Sélection Multiple de Structures d'Optimisation dans les Entrepôts de Données. In *6ème*

Journée Francophone sur les Entrepôts de données et l'Analyse en ligne (EDA'10), Revue des Nouvelles Technologies.
Boukhalfa, K. (2009). *De la Conception Physique aux Outils d'Administration et de Tuning des Entrepôts de Données* (PhD. Thesis). Poitiers University, France.
Derrar, H., Nacer, M. A., & Boussaid, O. (2013). Exploiting data access for dynamic fragmentation in data warehouse. *International Journal of Intelligent Information and Database Systems, 7*(1), 34. http://doi.org/10.1504/IJIIDS.2013.051736
Dimovski, A., Velinov, G., & Sahpaski, D. (2010). Horizontal Partitioning by Predicate Abstraction and Its Application to Data Warehouse Design. In B. Catania, M. Ivanović, & B. Thalheim (éd.), *Advances in Databases and Information Systems* (Vol. 6295, p. 164‑175). Berlin, Heidelberg: Springer Berlin Heidelberg. Consulté à l'adresse http://link.springer.com/10.1007/978-3-642-15576-5_14
Gacem, A., & Boukhalfa, K. (2013). Nouvelle Approche Scalable Dédiée au Charges Volumineuses pour la Fragmentation des Entrepôts de Données. In *Advances in Decisional Systems ASD*. Marrakech, Maroc.
Gardarin, G. (2003). *Bases de données*. Paris: Eyrolles.
Mahboubi, H. (2008, décembre). *Optimisation de la performance des entrepôts de données XML par fragmentation et répartition* (Theses). Université Lumière - Lyon II. Consulté à l'adresse https://tel.archives-ouvertes.fr/tel-00350301
Özsu, M. T., & Valduriez, P. (1991). *Principles of Distributed Database Systems*. Prentice Hall.

Summary

The NP-Completeness of the optimization structures selection problem in relational data warehouses makes their resolution using exhaustive methods a difficult task. These structures can be used in a centralized, distributed or parallel context. In the literature, there is little work to handle selection problems of these structures based on exact methods. In the present work, we proposed a new deterministic method to obtain horizontal fragmentation schema that optimizes a workload of queries. The proposed method was validated using the Benchmark APB1.

New data selection approach for faster SVMs.

Chaibi Sonia*,
*Department of Computer Science .
Networks and Systems Laboratory - LRS.
Badji Mokhtar University-UBMA- Annaba, Algeria.
Chaibi_dz@yahoo.fr

Abstract. Support Vector Machines have been extensively applied to deal with various data classification problems. However, given a large scale real world problem SVM's Training process requires large memory and long CPU time which rends SVM impractical solution. To circumvent this problem we propose a preprocessing procedure based on the k-means Algorithm. Its role is to allow selecting samples that are more likely to be support vectors which are distributed near the decision boundary and participate in the constitution of the margin thereabout. In this paper we propose to speed up the training process of Support vector machines (SVMs) by using small number of samples extracted from the original training data set. We aim choice samples that are the most close to the separation hyperplane to construct the reduced training set. Clustering algorithms have been widely used to reduce the training data of SVMs. K-Means clustering is the most frequently used one, it has shown a considerable success to reduce SVM's complexity. In this paper a new combination SVM/K-means is proposed to speed up SVM training. Empirical studies on the Banana dataset show our algorithm to be effective in reducing the size of training set without significantly compromising testing accuracy. The proposed approach reduce 80% of the time or more in our experiments

Keywords—Support vector machines; K-means Clustering; Speeding up SVM's Training, Linear Discriminant Analysis, Feature Dimension Reduction.

1 Introduction

Since their introduction in the late seventies (Vapnik, 1995), Support Vector Machines (SVMs) marked the beginning of a new area in the learning from examples paradigm. SVMs have attracted recent attention from the pattern recognition community due to a number of theoretical and computational merits derived from the Statistical Learning Theory Vapnik (Cortes and Vapnik, 1995; Vapnik, 1995) developed by Vladimir Vapnik at AT&T.

Although SVMs have shown potential and promising performance in classification (Burges, 1998; Smola and Schölkopf, 2004; Cristianini and Shawe-Taylor, 2010), they have been limited by speed particularly when we have to train a large data set. A huge quadratic optimization problem needs to be solved in order to find the separation hyperplan. Prior work

in speeding up SVMs training can be categorized into two approaches: algorithmic approaches and data-processing approaches.

Algorithmic approaches focus on modifying SVM classifier so that it could deal with large data sets within an acceptable time. They try to accelerate the training process by giving up the exact SVM restoring to an approximate SVM. One set of such algorithms is the noteworthy sequential minimal optimization (SMO) algorithm (Platt, 1998).This one uses only two training items in each small optimization problem. Other algorithms of this categories are the Joachims' SVMlight (Joachims, 1999), which introduced shrinking and kernel caching, and the working set selection and heuristics used by LIBSVM (Chang, 2011). Despite this research, SVM training time is still significant for large training sets.

Another way is to reduce the large data set training to a small size which can be treated by a simple SVM. Using Clustering is an effective way to reduce a large data set. Some results (Nghi and Mai, 2011; Schohn and Cohen, 2000; Tong and Koller, 2002) show that clustering technique can help to decrease complexity of SVM training. The K-means clustering was used frequently and it has shown a considerable success to reduce SVM's data training (Nguyen and Phung, 2010; Prabhu, 2011).

However, the use of clustering to reduce the data training set to the centers set, eliminate the essential data used to calculate the separation hyperplane. Also the separation hyperplane will be deformed if we replace support vectors by their cluster's centers. While (Li and al, 2010) use centers generated by K-Means clustering as training set and focus on the determination of input parameters without considering the effectiveness of the reduction approach we propose an algorithm based on the use of Kmeans clustering within an iterative procedure defined to allow reducing data far from the boundary decision and keeping data close to it.

The remainder of the paper is organized as follows; section II gives a brief introduction to the theoretical background with reference to classification principles of SVM .After the introduction of K-means clustering in section III, our proposition is described in section VI. Section V is devoted to experimental results. We conclude our study in section V.

2 Support Vector Machines

The dataset of supervised learning consists of *M* pairs of input/output (labels).

$$(x_i, y_i), i = 1,2, \dots, M \qquad (1)$$

Suppose the training samples and testing samples are with indexes i_t for training set; t = 1; 2;…T and i_s for testing set; s = 1; 2;…S, respectively. A supervised learning algorithm seeks a mapping from inputs to outputs based on the training set and predicts any outputs based on the mapping. If the output is a categorical or a nominal value, then this will become a classification problem. Classification is a common task in machine learning. Consider, for example, a simple two classes classification task with linearly separable data:

$$p_r(i) \in \{-1, +1\} \qquad (2)$$

SVM attempts to find the separating hyperplane

$$w.x + b = 0. \qquad (3)$$

with the largest margin (Vapnik, 98) satisfying the following constraints:

$$\begin{cases} w.x_{it} \geq 1 \; for \; l_{it} = 1 \\ w.x_{it} \leq -1 \; for \; l_{it} = -1 \end{cases} \qquad (4)$$

For linear separable case, in which *w* is the normal vector to the hyperplane. The constraints (4) can be combined into:

$$l_{it}(w.x_{it} + b) \geq 1 \qquad (5)$$

The requirements for the separating hyperplane with the largest margin can formulate the problem into the following optimization model:

$$\begin{cases} \text{minimize } \|w\|^2 \\ \text{subject to} \\ l_{it}\,(w, x_{it}) + b \geq 1 \quad \text{in which } t = 1,2 \ldots, T. \end{cases} \qquad (7)$$

The dual form of (6) by introducing Lagrange multipliers is

$$\begin{cases} maximize \; \sum_{i_t} \alpha_{it} - \frac{1}{2} \sum_{i_t, j_t} \alpha_{i_t} \alpha_{j_t} l_{i_t} l_{j_t} x_{i_t} \cdot x_{j_t} \\ \text{subject to} \\ \sum_{i_t} \alpha_{it} \, l_{i_t} = 0. \quad \alpha_{it} \geq 0, \\ 0 \leq \alpha_{it} \leq c, \; t = 1,2 \ldots, T. \end{cases} \qquad (8)$$

Where

$$w = \sum_{i_t} \alpha_{it} \, l_{i_t} x_{i_t}. \qquad (9)$$

The samples x_{i_t} satisfying $\alpha_{i_t} > 0$ are called the support vectors. By substituting (9) into (3), the solution for the separating hyperplane is:

$$f(x) = \sum_{i_t} \alpha_{it} \, l_{i_t} x_{i_t} \cdot x + b \qquad (10)$$

where $w = \sum_{i_t} \alpha_{it} \, l_{i_t} x_{i_t}$. (9)

The samples x_{i_t} satisfying $\alpha_{i_t} > 0$ are called the support vectors.

By substituting (9) into (3), the solution for the separating hyperplane is:

$$f(x) = \sum_{i_t} \alpha_{it}\ l_{i_t} x_{i_t} \cdot x + b \qquad (10)$$

The brilliance of equation (10) is that it just relies on the inner product between training points and testing points. It allows SVM to be easily generalized to nonlinear SVM. If *f(x)* is not a linear function about the data, the nonlinear SVM can be obtained by introducing a kernel function:

$$k(x_{i_t}, x) = \varphi(\alpha_{it}) \cdot \varphi(x). \qquad (11)$$

Fig. 1. Linear separating hyperplane for the separable case (Burges, 98).The support vector are circled.

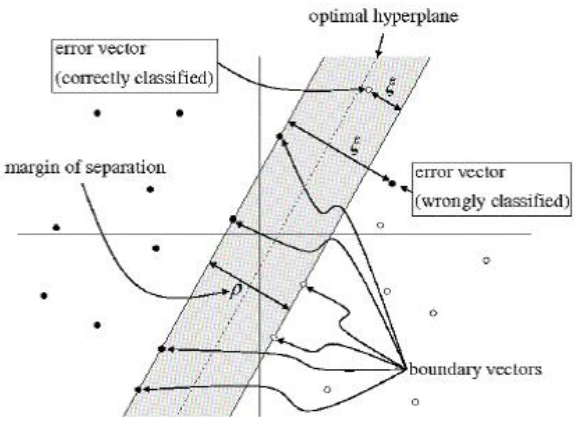


FIG. 1 – *The separation Hyperplane(Burges, 1998).*

We thus implicitly map the original data into a higher dimensional feature space *F*, where is the mapping from the original space to the feature space. In *F*, $\varphi($ are linearly separable. The separating hyperplane in feature space F is easily generalized into the following form:

$$f(x) = \sum_{i_t} \alpha_{it}\ l_{i_t} \varphi(x_{i_t}) \cdot \varphi(x) + b = \sum_{i_t} \alpha_{it}\ l_{i_t} k(x_{i_t}, x) + \quad (12)$$

By introducing the kernel function, the mapping φ does not need to be explicitly known which reduces much of the computational complexity. For more details refer to (Burges, 1998; Smola and al, 2004).A function is a valid kernel if there exists a mapping φ satisfying (11). Mercer's condition (Burges, 98) gives us the condition about what kind of functions are valid kernels. Actually, some common used kernels are as follows:

The polynomial kernels:

$$k(x_i, x_j) = (x_i . x_j + 1)^d \quad (13)$$

The Radial Basis kernels (RBF):

$$k(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|\right)^2 \quad (14)$$

and Gaussian RBF kernel :

$$k(x_i, x_j) = \exp\left(\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (15)$$

3 K-means Clustering Algorithm

K-means is a commonly used partitioning based clustering technique that tries to find a user specified number of clusters (k), which are represented by their centroids, by minimizing the square error function (Prabhu, 2011).. The K-means algorithm is one of the partitioning based, nonhierarchical clustering methods. Given a set of numeric objects X and an integer number k, the K-means algorithm searches for a partition of X into k clusters that minimizes the within groups sum of squared errors.

The steps of the Kmeans algorithm are written below:

Input: $X = \{d1, d2, \ldots\ldots, dn\}$ // set of n data items.

Output: A set of *k* clusters

Step 1: Initialization: choose randomly *K* input vectors (data points) to initialize the clusters.

Step 2: Nearest-neighbor search: for each input vector, find the cluster center that is closest, and assign that input vector to the corresponding cluster.

Step 3: Mean update: update the cluster centers in each cluster using the mean (centroid) of the input vectors assigned to that cluster

Step 4: Stopping rule: repeat steps 2 and 3 until no more change in the value of the means.

4 The proposed K-means Selection.

Motivated by results realized by Burges (Burges, 1998) in which it has been proved that the decision boundary dependent on only a portion of the training samples that lie close to the decision boundary called support vectors. We can be sure that removing non-support vectors does not affect SVM training results.

Our proposition lead to reduce the training samples set to a small one composed from samples which are likely to be the support vectors .Our Algorithm is based on an iterative

procedure permits progressively to lead to this samples set by removing, in each iteration , instances which are far from the separation hyperplane and keeping instances close to this one. The training data set will be updated each iteration. We begin with the all training samples and after the last iteration we obtained the optimal training samples set to be used to run the sequential minimal optimization (SMO) algorithm.

The procedure consists of two essential steps: clustering and de-clustering step. Clustering step apply K-means on the training data set , once clusters are defined ,we divided them into two groups of clusters: Mixed-clusters and Homogeneous-clusters. The Homogeneous-Clusters are clusters composed from data of same labels and Mixed-Clusters are composed from data of different labels. De-clustering step consist of getting back samples of Mixed-clusters.

The new training dataset is so composed of: centers of Homogeneous-clusters and the data of Mixed-Clusters then in the next iteration the K-means clustering will be applied again on this new training data set using a new value of the compression rate.

At each iteration, the new training data set becomes more important than the last one because at each iteration we get closer to the separation zone. To avoid removing important data we decrease the size of clusters to be generated in the next iteration .In other words; we increase the number of centers and the value of the compression rate.

To ensure the validity of our approach, three parameters need to be determined experimentally and relatively to the data set size. The first one is the initial compression rate value (ICR) ,the final value of the compression rate(FCR) and the incrementing value of the compression rate(CRI) .The definition of these parameters influence on the precision of the approach and on the time of execution. The following procedure describes our solution.

Input : the training data set T , The initial value of initial compression rate value (ICR) ,the final value of the rate(FCR), the value of the step of rate increasing(CRI) .

Output :The reduced training data set to be used by SMO.

Step1: Define the training data set T and the initial compression ratio (ICR) which is around 30 % of the whole input set;

CR = ICR.

Step2: Clustering: Apply *k* -mean algorithm on T to produce clusters and their centers with the compression rate equal CR.

Step3: Classify the obtained clusters into two sets:

- Cluster with only same labeled data; denoted by and the set of their centers is denoted .
- Cluster with both positives and negatives labeled data or mix-labeled data, denoted by the set of their centers is denoted .

Step4: Declustering :get back samples belong to Mixed labeled Clusters

Step5 : Update the training data set T : T = { U M};

Step6: Increase the compression ratio CR= CR +CRI,

Step7: if CR < CRF then stop, Otherwise go to 2.

Step8: apply the SMO on the last training data set obtained T.

It should be pointed out that though our approach utilizes K-means clustering technology, there are some remarkable differences between our method and the other ones in(Lyhyaoui and al, 1999),(Li and al, 2010),(Wang and al 2005)]clustering-based sample selection algorithms that are summarized as follows; we do not compress all the training set but only samples likely to be not close to the separation hyper plane. Our compression rate is not fixed but it changes from an iterations to the next, to guarantee getting back essential data respecting to the new size of the training data.

5 Experimental Results :

In this section we apply our method on two datasets .The Riply's training dataset, which contains 250 points, and compare the results achieved with those achieved using a standard SVM and the method of (Wang and al 2005).

All code for this experiment was written in Matlab, and all simulations were performed on a Intel Pentium Dual-Core processor with 4GB of RAM running Ubuntu 6.06 .The comparison of the novel SMO with the original one is demonstrated in Table 1. Parameter setting for standard SMO are C =30 in Equation (9) and $\sigma = 1$ in Equation (15).

Firstly, to prove the importance of recovering Mixed regions we have execute one iteration of our algorithm on the Riply's data and we have used the Mixed regions and centers of homogeneous regions to be the training set of SMO. Results for different value of the number of centers are summarized in the flowing table:

		K-means clustering				
	SVM	**10**	**20**	**30**	**40**	**50**
CPU Time(s)	0.4138	0.1800	0.1743	**0.1519**	0.1883	0.1909
Error	0.0940	0.1150	0.1000	**0.0930**	0.1050	0.0970

TAB1- *Results for one iteration with different number of clusters.*

Figure2 illustrate the decision boundary of the two SVM's classifiers ,and the boundary decision of the results obtained from one iteration of the proposed algorithm for the number of clusters equal 10.

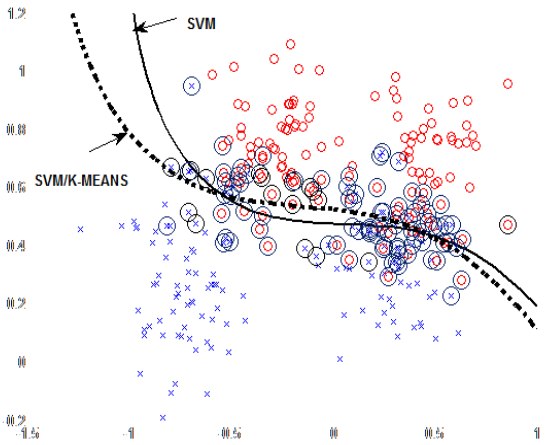


FIG. 2 – *Decision boundaries built from one iteration of K-means selection.*

Iteration	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Compression Rate(CR)	0.1	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65
Number of centers (K)	25	20	29	38	39	44	45	58	62	73	78	91
Training Data Size (DS)	**250**	**152**	**143**	**140**	**131**	**131**	**129**	**127**	**124**	**124**	**118**	**112**

TABLE2. RESULTS OF OUR APPROACHES ON REPLY DATA SET FOR THE VALUES OF PARAMETERS **CRI = 0 .1 ,CRF =0.7** AND **ICR = 0.05** ARE:

After the last iteration we applied SMO on the obtained training set which is about 112 samples:

	SVM	**SVM/KEMANS**
TRN size	250	112
NSV	116	45
CPU Time (s)	0.4138	0.1588
Error	0.0930	0.0940

TABLE3. RESULTS OF THE PROPOSED APPROACH ON RIPLY.

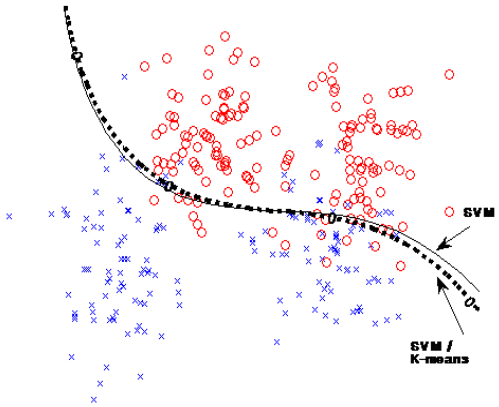


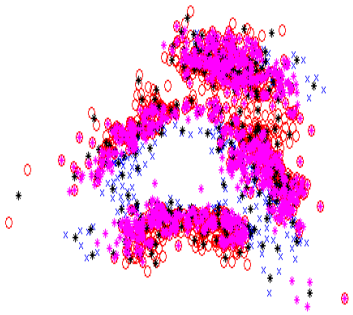
FIG. 3–*Decision boundaries built from the proposed K-means selection on Riply dataset.*

The second dataset used is the Banana dataset which is composed from 4770 training samples and 530 testing samples. Steps of the proposed data selection approach are summarized in Table4.

Iteration	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Compression Rate(CR)	0.01	0.11	0.21	0.31	0.41	0.51	0.61
Number of centers (K)	48	408	378	427	449	441	421
Training Data Size (DS)	4770	3708	1797	1377	1095	864	789

TABLE4. RESULTS OF ON BANANA FOR THE VALUES OF PARAMETERS **CRI = 0 .01 ,CRF =0.7** AND **ICR = 0.** 1 .

The three next figures illustrate the details of the fourth iteration of the proposed approach applied on Banana dataset.



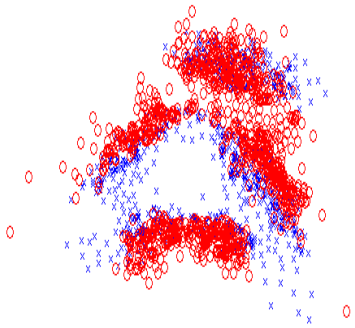


FIG 3-The reduced data after the third iteration on Banana.

FIG.4 - The Mixed region of the fourth iteration, black points are the centers of clusters, bold points are the mixed region

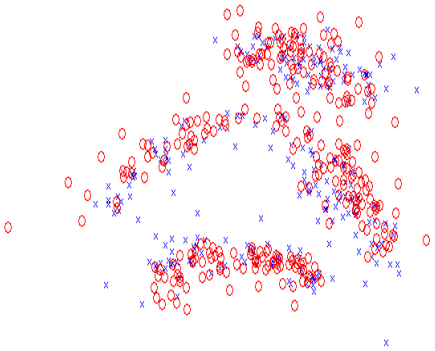


FIG5-The Reduced data after The fourth iterations.

After the last iteration we applied SMO on the obtained training composed of 789 samples which are about 17% of the whole training dataset:

	SVM	**SVM/KEMANS**
TRN size	4770	789
NSV	**1107**	**742**

CPU Time (s)	359.2292	30.5782
Error	**0.1019**	**0.1113**

TABL. 5 - RESULTS OF THE PROPOSED APPROACH ON BANANA.

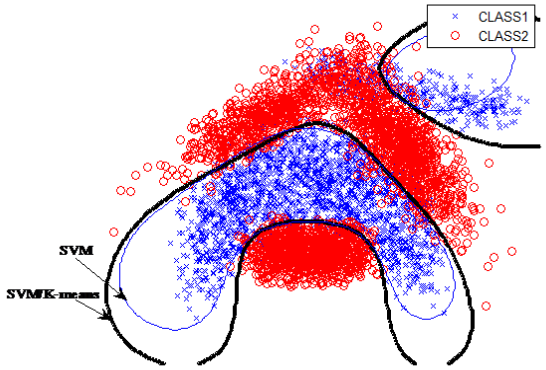


FIG6- The Reduced data after four iterations.

6 CONCLUSION

In this paper, we have developed a new data selection method which applies K-means clustering in order to speed up training time of Support Vector Machines (SVMs). The number of training data is reduced by extracting and using small number of representatives of the original training data set. Applied on Banana dataset, our approach shows that it is possible to reduce training set even up to 80%, without affecting the correctly classified rate.

The time taken for SVM training is reduced with the reduced set training samples. This results shows that the proposed technique is a valid way to reduce the size of the SVM training set with a minimum effects on precision.

This work shows that the proposed method can be applied successfully on large dataset and results can be important. This is what we envisage doing in futures works.

References

Abdul Nazeer, K.A., Sebastian, M. P.: "Improving The Accuracy and Efficiency of The K-means Clustering Algorithm"; World Congress On Engineering, WCE2009, July 1-3, London, U.K 2009.

Burges, C.J.C.: "Simplified Support Vector Decision Rules" Int. Conf. on Machine Learning, 2(1), 7(1996). 71-77.

Burges, C.J.C.: "A Tutorial on Support Vector Machines for Pattern Recognition"; Data Mining and Knowledge Discovery, 2(2), 1998, 121–167.

Cervantes, J., Li, X., Yu, W.: "Support Vector Machines Classification Based On Fuzzy Clustering for Large Data Sets ";J. Advances in Artificial Intelligence, Lecture Notes in Computer Science, 4293, 2006, 572-582

Chang, C.C, Lin, C.J.: "LIBSVM : A Library For Support Vector Machines"; ACM Transactions on Intelligent Systems and Technology, 2,3 ,April 2011. 2:27:1-27:27.

Chen, J., Liu, C. L. : "Instance Selection for Speeding up Multi-class SVMs With Neighborhoods"; Asian. Conf. on Pattern Recognition (ACPR), Beijing, China, 2011.264-268.

Cortes, C., Vapnik, V.:"Support Vector Networks"; J. Machine Learning, 20, 3 (1995), 273-297.

Cossato, H .P. E., Botto, Durdanovic, L., I., Vapnik, V.: "Parallel Support Vector Machines :The Cascade SVM"; Advances in Neural Information Processing Systems 17,2005 ,521-528.

Cristianini, N., Shawe-Taylor, J.:"An Introduction to Support Vector Machines and Other Kernel-based Learning Methods"; Cambridge University Press 2010, ISBN 978-0-521-78019-3, 1-189

Joachims, T.: " Making Large Scale Support Vector Machine Learning Practical"; Advances in Kernel Methods: Support Vector Learning"; MIT Press (1999), 169-184.

Li, X., Cervantes J. W., Yu, W.: "A Novel SVM Classification Method For Large Data Sets "; Proc. The IEEE International Conference on Granular Computing, IEEE-ICGC 2010, 297-302.

Lyhyaoui, A., Martinez, M., Mora, I., Vazquez, M., Sancho, J., and Figueiras-Vaidal, A. R. : "Sample Selection Via Clustering To Construct Support Vector Like Classifiers"; IEEE Transactions on Neural Networks,(1999), 10(6), 1474-1481.

Nghi, D.H., Mai, L.C: "Data Selection For Support Vector Machines "; Int. Conf. on Information and Electronics Engineering. 6 (2011). 28-32.

Nguyen, G. H., Phung, S.L.:" Efficient SVM Training with Reduced Weighted Samples"; World Congress on Computational Intelligence ,WCCI 2010, USA, 1764-1768.

Panda, N., Chang, E.Y., Wu, G.: " Concept Boundary Detection For Speeding Up SVMs"; 23rd Int. Conf. on Machine learning(ICML '06) 2006, 681-688.

Platt, J. : "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines"; Microsoft, Redmond, Washington, Tech. Rep. 98-14, April 1998.

Prabhu, P., Anbazhagan, N.: "Improving The Performance of K-means Clustering for High Dimensional Data Set"; Int, J. of Computer Science and Engineering(IJCSE),3 ,6 June 2011, 2317-2322;

Schohn, G., Cohen, D.: " Less is More : Active Learning with Support Vector Machines ";Proc. Int. Conf, in Machine Learning ICML '00, 2000, 839-846.

Shin, H. J., Cho, S.Z.: "Neighborhood Property Based Pattern Selection for Support Vector Machines"; J.Neural Computation ,19, 3, March 2007 , 816-55.

Smola, A.J., Schölkopf, B.: "Tutorial On Support Vector Regression"; J. Statistics and Computing, 14, 3(2004), 199–222.

Tong , S., Koller, D.: Support Vector Machines Active Learning with Application to Text Classification"; J. Machine Learning, 2 (2002) , 45-66.

Vapnik, V.N.,. "The Nature of Statistical Learning Theory". New York, USA: Springer-Verlag, 1995.

Wang , J., Wu, X., Zhang, C., : "Support Vector Machines based on K-means Clustering for Real Time Business Intelligent Systems"; J. Business Intelligence and Data Mining ,1, 1(2005) .54-64.

Xia, X.L., Lyu, M. R., Lok, T.M., Huang, G.B.: "Methods of Decreasing The Number Of Support Vectors Via K -Mean Clustering"; J. Advances in Intelligent Computing, P, LNCS 3644, 717 – 726.

Résumé :

Les Machines à Vecteurs Supports ont été largement appliqués classification des données. Cependant, le processus d'apprentissage des SVM ,une foi appliqué au problèmes de taille importante de données, nécessite un espace mémoire et un temps CPU importants ce qui rends les SVM impraticable. Pour contourner ce problème, nous proposons une procédure de pré-traitement utilisant le K-moyenne, son rôle est de permettre la sélection d'échantillons qui sont les plus susceptibles d'être des vecteurs supports et qui sont distribués près de la frontière de décision. Dans cet article, nous proposons d'accélérer le processus de d'apprentissage des machines à vecteurs de support (SVM) en utilisant un petit nombre d'échantillons extraits de l'ensemble d'apprentissage initiale. Les algorithmes de classification ont été largement utilisés pour réduire les données d'apprentissage des SVMs. Le regroupement K-moyenne est le plus fréquemment utilisé, il a montré un succès considérable pour réduire la complexité des SVMs. Dans cet article, une nouvelle combinaison SVM / K-moyenne est proposé pour accélérer l'apprentissage des SVM. Des études expérimentales utilisant la base Banana montrent que notre approche proposée réduit 80% du temps d'exécution sans affecter trop la précision des tests.

Vers une Solution de Protection des Données Entreposées dans le Cloud basée sur les Systèmes Multi Agents

Sara Rhazlane*, Nouria Harbi**, Nadia Kabachi**, Hassan Badir*

*LabTIC ENSA Tanger
BP 1818 Tanger Principal Tanger, Maroc
sara.rhazlane@gmail.com hbadir@uae.ac.ma
**Laboratoire ERIC
5, avenue Pierre Mendès France 69676 Bron Cedex, France
nouria.harbi@univ-lyon2.fr nadia.kabachi@univ-lyon1.fr

Résumé. Un des services fournis en Cloud est le Database-as-Service, qui implique l'hébergement par des fournisseurs Cloud, des données des entreprises en prenant en charge les exigences de stockage et de performance. Cependant, ces services prometteurs, soulèvent la question de la protection des données en termes de confidentialité, disponibilité et intégrité. Il s'agit dans ce travail d'exploiter les caractéristiques des systèmes multi agents pour proposer une solution sécurisée optimale de stockage et d'exploitation des données entreposées dans le cloud. Seules les données sensibles seront concernées par un chiffrement rendant les données non intègres avant le stockage. Des agents multi profils, adaptatifs et clonables seront utilisés dans notre architecture pour optimiser l'exploitation des données chiffrées dans le cloud.

Mots-clés. Cloud Computing, Sécurité, Protection des données, Chiffrement/Déchiffrement, Systèmes Multi Agents.

1 Introduction

Depuis l'apparition de la virtualisation, de l'externalisation, le développement des réseaux à haut débit, le principe du paiement à l'usage et la démocratisation de l'informatique, un nouveau paradigme a fait son apparition, le Cloud Computing.

Le Database-as-Service, l'un des services fournis par le Cloud, soulève à l'instar des autres services offerts, la question de la protection des données et la conformité aux réglementations. En effet, une grande partie de la sécurité des services en ligne est assurée par des moteurs d'application mais les vulnérabilités de ces derniers, couplées aux failles issues du développement peuvent entrainer des attaques. De plus les systèmes de gestion de bases de données, particulièrement privilégiés des pirates, contiennent des données privatives stockées dans des serveurs distants et contrôlés par des fournisseurs de service pas forcement fiables. Leur sécurisation est donc indispensable et nécessite une vigilance extrême.

Une des solutions, pour répondre à ces préoccupations et bénéficier des avantages et du potentiel du Cloud Computing tout en ayant une visibilité et un contrôle sur la confidentialité des données, est le chiffrement de ces dernières.

Nous présentons dans ce papier une proposition qui peut être considérée comme une première étape pour la mise en œuvre d'une solution de protection des données hébergées

dans le Cloud via une architecture équilibrée, adaptative et de haute disponibilité qui permettra de prendre en charge des requêtes sur les données chiffrées à l'aide des systèmes multi-agents.

Les objectifs de ce travail sont accompagnés de plusieurs considérations. D'une part, il existe une possibilité d'effectuer un chiffrage sur une partie des données en utilisant différentes clés de chiffrement, gérées par le client loin du serveur Cloud, aussi bien que le déchiffrement qui se fait au niveau client. D'autre part, l'utilisation des agents adaptatifs est justifiée par le fait qu'ils permettent en étant autonomes et communicants, d'être déployés sur plusieurs machines dans une architecture distribuée.

Notre travail sera structuré de la manière suivante : nous allons commencer par un état de l'art des travaux de recherche réalisés dans le domaine. Cette partie explicative sera suivie de la présentation de notre contribution accompagnée d'un test de faisabilité, avant de finir avec des conclusions et perspectives.

2 Etat de l'art

La question de sécurité des données a été étudiée dans Singh (2015), qui présente une analyse des diverses questions de sécurité classées en sept catégories. Les violations, la confidentialité et les pertes de données sont les principaux problèmes de sécurité et une des préoccupations majeures du Cloud Computing. L'auteur a présenté les questions de sécurité qui peuvent aider les chercheurs à trouver des pistes de solutions ou des mécanismes d'atténuation. Certaines de ces études seront abordées dans cette section.

2.1 Protection et Sécurité des Données dans le Cloud

Les travaux de recherche réalisés dans le domaine de la protection et la sécurité des données entreposées dans le Cloud ont fait ressortir de bons rapports. Dans Xiao Xiao (2013), les auteurs discutent les questions de sécurité du Cloud de manière systématique sur la base d'un ensemble d'attributs. Les auteurs utilisent des attributs tels que la confidentialité, la disponibilité, l'intégrité et d'autres. Pour chaque attribut, certaines menaces étaient examinées et des techniques d'atténuation possibles et solutions ont été proposées. L'étude menée par Sun et al. (2011) porte autour des questions de sécurité, de confidentialité et aide les utilisateurs à reconnaître ces menaces, analyser des méthodes permettant d'éliminer ces risques et de fournir un environnement hautement sécurisé. Les auteurs invoquent les solutions traditionnelles telles que le cryptage et l'autorisation d'accès qui peuvent s'affaiblir devant la scalabitité et propose la confiance comme philosophie pour lutter contre ces menaces de sécurité. Dans Abdul Alsahib S.Aldeend et al. (2015), les auteurs ont présenté les vulnérabilités et les attaques et identifié les directives de solutions possibles pour renforcer la sécurité. En termes d'architectures, Lombardi (2011) propose une architecture innovatrice appelée « Advanced Cloud Protection System (ACPS) ». Dans Sadeghi (2010), de nombreuses architectures et modèles possibles offrant intégrité ainsi que la confidentialité sont présentées se basant sur le chiffrement homomorphe. Dans Zissis (2012), les auteurs présentent la cryptographie comme solution capable d'assurer l'intégrité, la confidentialité et l'authentification de données complexes. Les auteurs dans Lombardi (2010) ont discuté le problème d'intégrité et de protection des données dans le Cloud et ainsi proposent une nouvelle architecture appelée « Transparent Couverture Protection System (TCPS) » pour améliorer la Sécurité dans le Cloud. Aussi bien

que le «VPN sécurisé » proposé par Mathew (2012) comme solution de sécurité. Et enfin, la méthode de « Cryptographie sur les courbes elliptiques » a été proposée par Kumar (2012), qui permet le stockage et le partage sécurisé des données en préservant les fichiers de données et en donnant un accès à l'utilisateur. Un des modèles proposés également, est le Database-as-a-Service (DAS) qui a été introduit par les auteurs de Hacigumus (2002) qui suggèrent une architecture pour le stockage et l'exploitation des données cryptées dans le cloud, depuis lors, cette proposition a reçu beaucoup d'attention de la communauté de recherche.

L'étude de ces travaux de recherche, a permis de conclure qu'il est communément admis que le chiffrement des données, du côté client, est une bonne solution garantissant la confidentialité des données, voir Kamara (2010) et Chow (2009). Nous allons dans la suite présenter les travaux de recherche effectués dans le domaine du chiffrement des données dans le Cloud.

2.2 Chiffrement et Déchiffrement des Données

Dans K. Jasim et al. (2013), les auteurs ont recensé les différents algorithmes de chiffrement et déchiffrement utilisés dans le Cloud ainsi que le mécanisme de la génération et la gestion des clés. Les auteurs proposent une nouvelle architecture de sécurité qui considère les diverses lacunes de sécurité autant que possible. L'environnement propose une nouvelle technique hybride qui combine à la fois l'algorithme « Advanced Encryption Standard (AES) » et le « Quantum Key Distribution (QKD) », comme algorithme principal de sécurité utilisé pour le chiffrement et le déchiffrement. Cette architecture est considérée comme la première technique hybride dans le domaine du Cloud Computing. Les auteurs présentent aussi le « Cloud Data Encryption Based Quantum (CDEQ) » et le « Cloud Encryption Model (CEM) », comme étant les modèles les plus populaires utilisés dans le chiffrement des données dans le Cloud. Quant au « Cryptographic Cloud Storage », les auteurs dans Kamara et al. (2012) proposent des services de stockage privés qui satisferaient les exigences de confidentialité, intégrité, authentification et autres. La plupart des opérations sont faites par le chiffrement des documents stockés dans le Cloud. Cependant, un tel cryptage conduit à une difficulté à la fois dans les processus de recherche dans les documents mais aussi le processus du « Real-time Collaborative Editing (RTCE) ». Au-delà des travaux de recherche en cryptographie cités, il est utile dans cette section de parler aussi des techniques de chiffrement, dites homomorphiques, ceci a été proposé dans Hacigumus (2004) pour résoudre le problème du traitement au niveau du serveur Cloud, notamment les opérations arithmétiques, de comparaison et d'agrégation des données chiffrées. Cette technique a été critiquée par les auteurs dans Einar (2006), d'où ils proposent une approche très simple pour manipuler les requêtes au niveau du serveur, qui ne comporte pas de fonctions de chiffrement et/ou déchiffrement. Ils décrivent en outre des protocoles pour la formulation et l'exécution de requêtes. Ils se concentrent donc sur une variante du DAS qui n'a pas été exploré auparavant : le modèle DAS mixte, où certains attributs (Colonnes) sont sensibles, et donc nécessitent un cryptage, tandis que d'autres ne le sont pas, et sont donc laissés en clair.

2.3 Les Systèmes Multi-agents et la Sécurité des Données dans le Cloud

Dans Talib et al. (2010), les auteurs ont présentés une revue de littérature concernant une utilisation possible des systèmes multi-agents, qui peuvent être bénéfiques dans les plateformes Cloud en facilitant la sécurité des données qui y sont stockées. Ce papier décrit aussi

comment les systèmes multi-agents pourraient être utilisés dans une plateforme Cloud en utilisant un environnement collaboratif du « Java Agent DEvelopment (JADE) ». L'étude réalisée a examiné un certain nombre de thèmes, frameworks, architectures et approches. L'étude a permis aussi de conclure que l'utilisation des systèmes multi-agents pour favoriser la sécurité des données entreposées dans le Cloud est très complexe et se produit dans de multiples niveaux. À la connaissance des auteurs, aucun travail n'a été réalisé dans le domaine. Dans Talib et al. (2012), les auteurs ont proposé une architecture à base de systèmes multi-agents qui a été citée dans leurs précédents travaux. Cette architecture permet de faciliter la confidentialité, l'assurance, la disponibilité et l'intégrité des données La solution proposée se compose de deux principales couches, une couche agent et une couche Cloud et comprend cinq principaux types d'agents. Afin de vérifier le framework de sécurité proposé, une étude pilote a été réalisée. Plus récemment, dans Talib (2014), les auteurs ont discuté la possibilité de jointure entre le Cloud et les « MAS-Based CBR » et il a été spécifié dans le papier comment cela peut être réalisé. Dans Zhou (2014), les auteurs ont proposé un framework avec l'aide de la flexibilité, la commodité, la bonne interaction et la forte capacité d'apprentissage des systèmes multi-agents. Cette architecture comprend 4 agents, à savoir le « Confidentiality Agent », le « Correctness Agent», le « Availability Agent » et le « Integrity Agent ». Ses sous framework réalisés sous forme d'agents permettent d'assurer la sécurité de l'ensemble du framework proposé. Une simulation par 15 utilisateurs a été réalisée pour tester la faisabilité de l'architecture proposée, et a montré une bonne performance.

2.4 Synthèse et Discussion

L'étude de ces travaux de recherche, a permis de conclure qu'il est admis que l'utilisation du chiffrement des données, du côté client, est une bonne solution qui permet de garantir la confidentialité des données. Aussi, chaque étude met l'accent sur un aspect plus que d'autres, certaines se concentrent sur l'architecture, d'autres sur les systèmes cryptographiques et les modèles, ou encore sur les problèmes liées à l'exploitation des données chiffrées.

De notre côté, nous allons essayer de proposer une architecture qui permet de stocker les données chiffrées dans le Cloud, de tenir les clés de chiffrement hors porté des administrateurs Cloud, et de partager la charge d'exécution des requêtes entre le client et le serveur.

Nous avons proposé un processus de chiffrement/déchiffrement de données, où le stockage des clés et leur utilisation pour le déchiffrement se font au niveau client. Le système cryptographique, proprement dit et la politique d'échange de clés n'est pas abordé dans cet article et font partie des perspectives.

L'utilisation des systèmes multi-agents dans la sécurité des données dans le Cloud Computing reste une approche nouvelle et n'a été abordée que dans des travaux limités. L'originalité de notre approche est l'utilisation des systèmes multi agents pour leurs caractéristiques, qui va permettre d'une façon optimale de garantir la confidentialité, la disponibilité et l'intégrité des données au niveau stockage et exploitation. En effet, le système est composé de trois agents qui communiquent entre eux, chacun avec un rôle prédéfini.

3 Solution Proposée

Afin d'atteindre les objectifs que nous avons cités et à travers l'étude des différents travaux mentionnés dans la partie état de l'art ainsi que les solutions existantes sur le marché,

nous avons vu la nécessité de proposer une solution capable de répondre aux exigences et défis de la protection des données dans le Cloud Computing.

La solution qu'on a élaboré vise à réaliser une architecture adaptative à tout type de SGBD utilisé par le fournisseur Cloud, qui permet le chiffrement sélectif des données au niveau colonne et qui peut être gérée et configurée par un administrateur de confiance , à savoir le propriétaire des données et non pas les administrateurs du fournisseur Cloud. Cette solution implémentée en JAVA sera aussi basée sur une approche qui utilise un système cryptographique à base de clés de chiffrement stockées au niveau du propriétaire de données et qui garde, en permanence, les données chiffrées au niveau serveur et n'effectue le déchiffrement qu'au niveau client. D'autre part, l'architecture proposée supporte une large gamme de requêtes SQL, peut être déployée sur un réseau étendu et elle est compatible avec tous les systèmes d'exploitation.

3.1 Architecture du Système

La solution proposée (Voir figure 1), du point de vue architecture, et selon la section comparative des niveaux de chiffrement, se positionne à un niveau intermédiaire entre l'application et le serveur de la base de données. Ceci qui permettra de bénéficier des avantages et échapper aux inconvénients des niveaux de chiffrement existants. Parmi les différentes possibilités de granularité de chiffrement, nous avons choisi la colonne comme la plus petite entité à protéger, vu que le chiffrement de niveau colonne est plus souple et spécifique. Notre architecture présente un ensemble de quatre acteurs : Les administrateurs propriétaires de données (A) ; le Client (B); le serveur en Cloud (C) et le système multi-agents (D) comprenant le Main Agent, le Query Translator et le Query Executor. Nous allons présenter de manière détaillée le rôle de chaque acteur ainsi qu'une description de la structure et du fonctionnement de la solution.

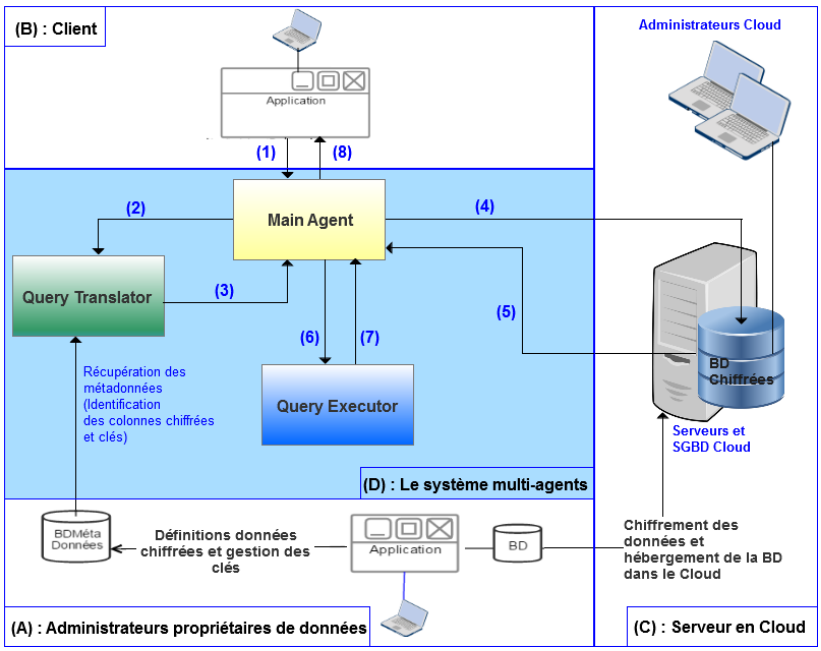


FIG. 1 – *Architecture du système proposé*

3.1.1 Les Acteurs

(A) Les administrateurs propriétaires de données : Ils sont chargés de la définition des données chiffrées (Voir figure 2), ainsi que les clés utilisées pour le chiffrement (métadonnées de chiffrement). Cette métadonnée est utilisée par la suite, d'un côté, pour chiffrer les données avant de les stocker sur la base de données hébergée chez le fournisseur Cloud, et d'un autre côté, par l'outil pour l'exploitation de la base de données. Les administrateurs peuvent mettre à jour la métadonnée et effectuer le chiffrement et le déploiement de la base de données dans le Cloud.
(B) Le Client : Il exploite la base de données, à travers son application, il peut envoyer la requête au « Main agent » et recevoir les résultats sous forme de données déchiffrées.
(C) Serveur en Cloud : Il reçoit la requête version serveur, l'exécute et renvoi le résultat au « Main agent ».
(D) Le système multi-agents :

- **L'agent principal (Main Agent).** Il joue le rôle d'intermédiaire entre le client et la base de données, et coordonne l'envoi et la réception des messages entre le client, les agents et le serveur de base de données dans le Cloud ;
- **Query Translator Agent.** Il reçoit la requête d'origine envoyé par le « Main Agent » et conçoit la requête version serveur, puis renvoie la requête version serveur au « Main Agent » ;
- **Query Executor Agent.** Il reçoit la requête d'origine et les résultats chiffrées envoyés par le « Main Agent » et déchiffre ces résultats d'exécution rendus par le serveur Cloud, fait les calculs et applique les conditions de restriction sur les données déchiffrées et renvoi ensuite les résultats (données déchiffrées) au « Main Agent ».

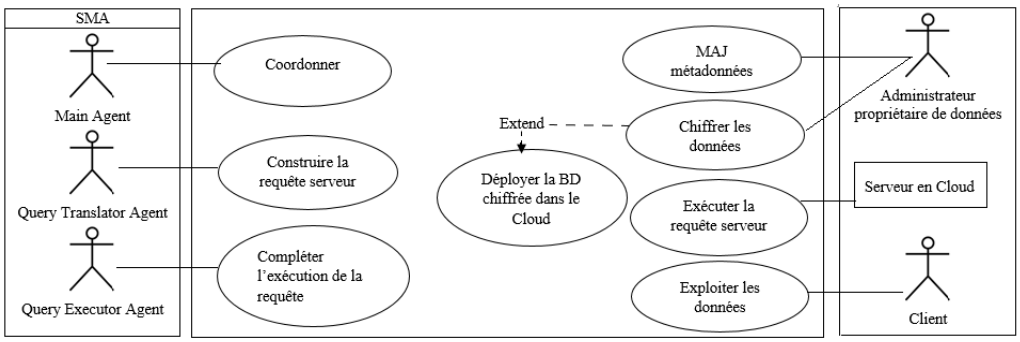


FIG. 2 – *Diagramme de cas d'utilisation des différents acteurs du système*

3.1.2 Fonctionnement

Dans la phase du déploiement initial (Voir figure 1), le propriétaire chiffre sa base de données avant de l'héberger sur la base de données du Cloud. Cette opération doit être toujours possible, même après le premier déploiement ; le chiffrement se fait toujours à base d'une ou plusieurs clés qui sont stockés coté client et gérées par les administrateurs propriétaires de données.

Le système multi-agents, qui joue le rôle de médiation entre les applications utilisateurs et le serveur, doit intercepter et analyser la requête envoyée par le client (1) à l'aide du « Main Agent ». Il doit identifier les colonnes chiffrées et envoyer la requête au « Query Translator Agent » (2), qui se charge de la traduire en requête version serveur. Le « Main Agent » reçoit ensuite la requête version serveur, renvoyée par le « Query Translator Agent » (3). Le « Main Agent », envoi la requête version serveur au serveur, qui héberge la base de données dans le Cloud, pour exécution (4). Il reçoit par la suite les résultats chiffrés de l'exécution de la requête par le serveur (5). Le « Main Agent » envoi la requête d'origine et les résultats chiffrés au « Query Executor Agent » qui complète l'exécution de la requête en se chargeant du déchiffrement (6). Le Main Agent reçoit par la suite les résultats du déchiffrement (données déchiffrées) (7) et envoie finalement ces résultats au client (8).

Les taches de translation et d'exécution des requêtes sont effectuées à base de pseudo-algorithmes spécifiques (Voir Algorithme 1). Le chiffrement et déchiffrement est pris en charge par une fonction simple basée sur une clé définie pour chaque colonne. Cette fonction permet d'augmenter la valeur de la donnée avec un nombre qui dépend de la clé et de sa longueur, elles sont basées sur une valeur calculée par la fonction *Valeur_clé()* (Voir Algorithme 2). Etant donné la donnée *D* et la clé de chiffrement *K* d'une valeur *V*; le chiffré de *D* est *C*, comme suit :

C = **Chiffrer** (D) = ((D+V)*V)/D

Déchiffrer (C) = **Déchiffrer** (C) = V^2 / (C-V) = D

Algorithme 1 Algorithme de translation des requêtes

Entrée : Req (Requête d'origine)

Sortie : ReqS (Requête version serveur)

1: Analyse de la clause « **select** » de la requête :

Extraction des colonnes (séparées par des virgules)

- Identification des colonnes du format **Colonne**[1]
- Identification des colonnes du format **table.colonne**[2]
- Identification des colonnes dans les **expressions arithmétique** (addition, soustraction …), d'agrégation de type f(**Colonne**)[3]**;**
- Construction de la **Liste_Colonnes_Select**

2: Analyse de la clause « **where** » et « **having** »de la requête :

- Extraction des conditions séparées par des opérateur logiques (and,or..)
- Extraction des opérandes séparés par (<,>,=,<>,>=,<=)
- Extraction des colonnes de la même manière que la clause « select » (colonne, Table.colonne, expression arithmétique,fonction)

[1] Exemple : ventes dans *select **ventes** from compta*

[2] Exemple : compta.ventes dans *select **compta.ventes** from compta*

[3] Exemple *f()* : count(*), sum(ventes), max(achats)…

- Construction de la **Liste_Colonnes_Select**

3: Analyse de la clause « **order by** »de la requête

- Extraction des colonnes de la même manière que la clause « select » (colonne, Table.colonne, expression arithmétique,fonction)
- Construction de la **Liste_Colonnes_OrderBy**

4: Construction de la liste globale des colonnes (sans doublants) : **Liste_Colonnes**

5: Constitution de ReqS (Requête version serveur) :

- La clause « Select » contient toutes les colonnes de Liste_Colonne
- Les clauses « where » et « having » contiennent toutes les conditions qui ne concernent pas au moins une colonne chiffrées[4].

Algorithme 2 Algorithme de la fonction Valeur_clé(K)

Entrée: K

Sortie: Numérique

1: V=0 ;

2: pour i de 1 à longeur(K) **faire**

3: V=V+ valeur_ascii(K[i])

4 : fin pour

5 : retourner(V)

3.2 Résultats et Validation

Afin de tester la faisabilité de notre solution, nous avons effectué un ensemble de tests sur des données brutes en utilisant différents exemples de requêtes (Requête simple avec clauses « select » et « where », requête avec colonnes sous forme Table.colonne, expression arithmétique simple, requête de jointure). La validation du système a été faite par l'analyse des résultats fournis et via son fonctionnement global : l'échange des messages, à partir du premier message (requête SQL utilisateur à exécuter) jusqu'au renvoi des données en clair (résultat de l'exécution de la requête). Le test a pris en compte également, la construction de la requête version serveur générée par l'agent Query Translator Agent « QTransAgent » et l'exécution de la requête (déchiffrement de données) par l'agent Query Executor Agent « QExecAgent ».

Nous présentons dans ce qui suit un exemple d'application qui a été réalisé lors du déroulement du prototype.

Soit le schéma suivant (Voir figure 3), les deux colonnes « ventes » et « achats » de la table « Compta », et le « nom_responsable » de la table « Région », sont supposées chiffrées. Cette information se trouve au niveau de la métadonnée, sous forme « table, colonne, clé » (Voir tableau 1). Les colonnes peuvent êtres chiffrées par la même clé ou par des clés différents. Le tableau 2 donne une vue sur les données brutes ainsi que les valeurs des données chiffrées stockées dans la BD de la table « Compta ».

[4] Les conditions contenants au moins une colonne chiffrée seront ignorées, car le serveur en cloud, ne peut pas les évaluer.

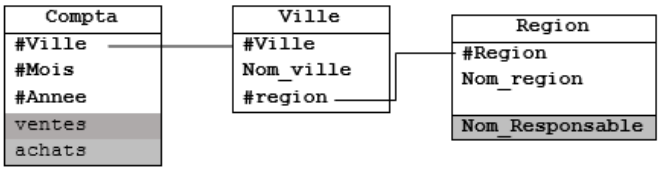


FIG. 3– *Exemple du modèle de données*

Considérons la requête d'origine (envoyée par le client) suivante :

Req :

```
SELECT ville, annee, ventes FROM compta
WHERE ville ='Lyon' AND achats >35000 ;
```

Table	modèle 1	Clé
Compta	Ventes	k1
Compta	Achats	K2
Region	Nom_responsable	K1

TAB. 1 – *Clés et métadonnées selon les tables du modèle*

Ville	Année	Mois	Ventes	Achats
Lyon	2012	01	70000	38000
Lyon	2012	02	55000	36000
Lyon	2012	03	45000	35000
Paris	2012	01	90000	20000

Ville	Année	Mois	Ventes	Achats
Lyon	2012	01	332.1	465
Lyon	2012	02	85.36	37.89
Lyon	2012	03	105.36	37.89
Paris	2012	01	15.23	220

TAB. 2 – *Données brutes (à gauche) et données chiffrées stockées dans la BD (à droite) de la table « Compta »*

Il est évident que la valeur de la colonne « ventes » est chiffrée au niveau de la BD et doit être déchiffrée au niveau client et que la vérification de la condition « ventes >50000» est impossible, car la valeur des ventes est chiffrée au niveau de la base de données, et donc la résolution ne doit être réalisée qu'après déchiffrement des données. La translation de la requête (Req), par le « QTransAgent » donne la requête partielle version serveur **ReqS** suivante :

ReqS :

```
SELECT ville, annee ,ventes,achats FROM compta
WHERE ville ='Lyon' AND achats >35000 ;
```

Le tableau 4 (à gauche) présente le résultat (ResCh) retourné par le serveur :

Ville	Année	Mois	Ventes	Achats
Lyon	2012	01	332.1	465
Lyon	2012	02	85.36	37.89
Lyon	2012	03	105.36	37.89

Ville	Année	Mois	Ventes	Achats
Lyon	2012	01	70000	38000
Lyon	2012	02	55000	36000
Lyon	2012	03	45000	35000

TAB. 4 – *Résultat de la requête retourné par le serveur (à gauche) et le résultat de l'exécution de la fonction de déchiffrement sur la requête ResCh retournée par le serveur (à droite)*

Le « QExecAgent » applique la fonction de déchiffrement pour les colonnes « ventes » et « achats » (Voir tableau 4, à droite) et puis applique la condition de restriction « ventes>50000 », pour obtenir le résultat final (Voir tableau 6). Nous montrons aussi dans la figure 4, les résultats d'exécution de l'outil pour chaque exemple de requête où figure l'ensemble des étapes définies dans l'architecture.

Ville	Année	Mois	Ventes	Achats
Lyon	2012	01	70000	38000
Lyon	2012	02	55000	36000

TAB. 6 – *Résultat de l'exécution de la requête de restriction sur ResCh*

```
SMA Début d'exécution.. [MainAgent] en attente des requêtes Req....
1-- [MainAgent]:Reception message de Requête de  type Req
2-- QTrans: Reception Message Requête de type Req
          Constitution de la requête version server ReqS
          Req  Origine: select ville,mois,annee,achats,ventes from compta
          ReqS générée: SELECT compta.ville ,compta.mois ,compta.annee ,compta.achats
                        ,compta.ventes  from  compta
3-- [Qtrans]: Renvoi ReqS à MainAgent
3-- [MainAgent]:Reception message de Requête de  type ReqS
4-- [MainAgent]:Envoi ReqS au serveur pour execution ) ....
          Traitement de ReqS par le serveur ...........
5-- [MainAgent]:Reception Resultat chiffré ResCh résultat s'exécution par le serveur
      Resultat:
      ville | mois | annee | achats | ventes |
      PARIS |  2   |  2014 | 3080.0 | 2845.0 |
      PARIS |  1   |  2015 |12582.0 | 3390.0 |
      LYON  |  1   |  2014 | 3059.0 | 2768.0 |
      LYON  |  2   |  2014 | 2734.0 | 2725.0 |
      LYON  |  1   |  2015 | 2740.0 | 2716.0 |
6-- [MainAgent] Envoi ResCh [QexecAgent]
6-- [QExecAgent] Reception ResCh + Req .. début execution (déchiffrement)
7-- [QExecAgent] Traitement finalisé.. Envoi Résultat données en claire ResClr à
[MainAgent]
7-- [MainAgent]:Reception Resultat en claire ResClr
8-- [MainAgent]:Envoi ResClr à l'application : Affichage ResClr
      Resultat:
      ville | mois | annee | achats  | ventes  |
      PARIS | 2    | 2014  | 15383.0 | 32157.0 |
      PARIS | 1    | 2015  |  695.0  |  9105.0 |
      LYON  | 1    | 2014  | 16128.0 | 50158.0 |
      LYON  | 2    | 2014  | 66258.0 | 72954.0 |
      LYON  | 1    | 2015  | 62816.0 | 79979.0 |
```

FIG. 4– Résultats d'exécution de l'outil pour chaque exemple de requête

4 Conclusion et Perspectives

Dans le cadre de ce travail, nous avons proposé un prototype préliminaire à base d'agents pour la protection des données entreposées dans le Cloud. Nous avons testé l'interopérabilité des systèmes multi-agents et du Cloud Computing.

Ce travail a été abordé sur trois axes principaux, l'environnement Cloud, le chiffrement des données et le développement d'un système multi-agents. Nous avons été dans l'obligation de concentrer nos efforts sur la conception et l'architecture du système, ainsi que le développement d'un SMA opérationnel. Néanmoins d'autres travaux complèteraient ce travail, à savoir la conception d'un système cryptographique, basé sur une fonction de chiffrement qui répond complètement aux exigences de sécurité, renforcer le SMA par l'exploitation de toutes ses caractéristiques. Notamment, celle de l'intelligence qui permet d'éviter d'exécuter la tâche d'analyse qui donne le même résultat pour les requêtes similaires, penser à la distribution des agents, dans le sens de cloner les agents chargés de l'exécution des requête sur plusieurs machines et nœuds cluster, et implémenter un mécanisme de distribution des messages qui leurs sont envoyés par le « Main Agent ». D'autre part, nous envisageons rendre le SMA plus paramétrable, notamment au niveau de la connexion aux différentes bases de données : adresse du serveur, type de base de données et données de connexion (utilisateur et mot de passe) et finalement compléter l'algorithme de translation et d'exécution des requêtes afin de prendre en charge une large gamme de requête SQL.

Références

Abdul Alsahib S.Aldeen, Y. (2015), *State Of The Art Survey On Security Issue In Cloud Computing Architectures, Approaches And Methods*, Journal of Theoretical and Applied Information Technology, Vol.75. No.1.

Armbrust, M. (2009), *Above the clouds: A Berkeley view of cloud computing*, University of California, Berkeley, Tech. Rep.

Bouganim, L. , Y. Guo (2010), *Database encryption. In Encyclopedia of Cryptography and Security*, Springer, 2e édition.

C.C.A, *CipherCloud Gateway Architecture*, www.ciphercloud.net.

Chow, R. (2009), *Controlling data in the cloud: outsourcing computation without outsourcing control*, In Proceedings of the ACM workshop on Cloud.

Demazeau, Y. (1995) *From interactions to collective behaviour in agent-based systems*, In: Proceedings of the 1st. European Conference on Cognitive Science. Saint-Malo, 117-132.

Ferber, J. (1995), *Les Systèmes multi-agents, vers une intelligence collective*, InterEdition.

Hacigumus, H. (2002), *Providing database as a service*, In International Conference on Data Engineering.

Hacigumus, H., B. Iyer, S. Mehrotra (2004), *Efficient execution of aggregation queries over encrypted databases*, International Conference on Database.

Jasim, K. (2013), *Cloud Computing Cryptography "State-of-theArt"*, World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:7, No:8.

Kamara, S., G. Ateniese, J. Katz (2012), *Proofs of storage from homomorphic identification protocols. In Advances in Cryptology*, -ASIACRYPT '09, volume 5912 of Lecture Notes in Computer Science, pages 319, Springer.

Kamara, S., K. Lauter (2010), *Cryptographic cloud storage*, In Proceedings of the 14th international conference on Financial cryptograpy and data security.

Kumar, A. (2012), *Secure storage and access of data in cloud computing*. International Conference on ICT Convergence (ICTC), 336–339. doi:10.1109/ICTC.2012.6386854

Lo, H., H. F. Chau (1992), *Unconditional security of quantum key distribution over arbitrary long distances*. Science 1999; 283(5410): 2050-2056.

Lombardi, F. (2011), *Secure virtualization for cloud computing. Journal of Network and Computer Applications*, 34(4), 1113–1122.

Lombardi, F. R. Di Pietro (2010), *Transparent Security for Cloud*, 414–415.

Mathew, A. (2012), *Security and Privacy Issues Of Cloud Computing*; *Solutions and Secure Framework*, International Journal of Multidisciplinary Research,Vol.2 Issue 4, ISSN 2231 5780.

Mykletun, E., G. Tsudik (2006), *Aggregation Queries in the Database-As-a-Service Model* ,20th IFIP WG 11.3 working conference on Data and Applications Security.

Sadeghi, A. (2010), *Token-Based Cloud Computing*, 2, 417–429.

Singh, S. (2015), *State-of-the-art Survey on Security Issues in Cloud Computing Environment*, IEEE.

Sun, D. (2011), : *Surveying and Analysing Security, Privacy and Trust Issues in Cloud Computing Environments*, Advanced in Control Engineering and Information Science, Procedia Engineering 15, pp. 2852- 2856.

Talib, A., R. Atan, R. Abdullah, M. Azmi (2012), *Security Framework of Cloud Data Storage Based on Multi Agent System Architecture - A Pilot Study*, 978-1-4673-1090-1/12/, IEEE.

Talib, A., R. Atan, R. Abdullah, M. Azrifah, A. Murad (2010), *Security Framework of Cloud Data Storage Based on Multi Agent System Architecture: Semantic Literature Review*, Computer and Information Science Vol. 3, No. 4.

Talib, A. , N. Elshaiekh (2014), *Multi Agent System-Based on Case Based Reasoning for Cloud Computing System*, APJES II-II, 34-38.

Xiao, Z., Y. Xiao (2013): *Security and privacy in Cloud Computing*, IEEE, Communication Surveys and Tutorials, pp. 843-859.

Zhou, H., S. Qin (2014), *Security framework for cloud data storage based on multi-agent system*, Computer Modelling & New Technologies, 18(12b) 548-553.

Zissis, D., D. Lekkas (2012), *Addressing cloud computing security issues. Future Generation Computer Systems*, 28(3), 583–592. doi:10.1016/j.future.2010.12.006.

Summary

One of the services provided by Cloud Computing is the Database-as-service, which involves hosting data by cloud providers that support the storage and performance requirements. However, these promising services, raise the issue of data protection in terms of confidentiality, availability and integrity. In this work, we exploit the characteristics of multi agents systems to deliver an optimal and secure solution for data storage and exploration in the Cloud. Only sensitive data will be affected by an encryption process making data lose their integrity before storage. Multi profiles, adaptive and cloneable agents will also be used in our architecture to optimize the exploration of the encrypted data in the Cloud environment.

A New Approach for Privacy preserving in Big Data through Access Control

Abdelmalek AMINE, Amine RAHMANI, Reda Mohamed HAMOU, Mohamed El-hadi RAHMANI, Mohamed Amine Boudia

GeCoDe Laboratory, Department of computer sciences, Dr. TAHAR Moulay university of Saida, Algeria

Abstract. Nowadays, the concept of big data grows incessantly; recent researches proved that 90% of the whole data existed on the web had been created in last two years. However, this growing bumped by many critical challenges resides generally in security level; the users care about how could providers protect their privacy on their data. Access control, cryptography, and de-identification are the main search areas grouped under a specific domain known as Privacy Preserving Data Publishing. In this paper, we bring in suggestion a new model for access control over big data using digital signature and confidence interval; we first introduce our work by presenting some general concepts used to build our approach then presenting the idea of this report and finally we evaluate our system by conducting several experiments and showing and discussing the results that we got.

Keywords: Access control, standard deviation, privacy preserving, big data, numeric signature, confidence interval

1 Introduction

Privacy, timeless, scalability of data is the most important problems that big data recognize starting from the first step of data acquisition; in fact, one of the most disturbed principle that are used in big data is the fact of losing control on data. This concept led to a lot of criticism from clients, losing control on your own data means losing everything related to the control even the access control.

Before the coming of the concept big data, controlling access on such data was done locally using the known models such as mandatory models (MAC), discriminatory models (DAC) or role based models (RBAC) but those last cannot be used because of some impediments; in case of DAC models the users defines the right access by himself while in the use of big data the user lose the entire control on his data; in case of MAC models the right access are defined by a major entity like military direction and this does not satisfy the users wishes in big data; moreover, in case of RBAC models, the right access are defines in form of roles where the can have the right from a major entity and can also give the rights on his own data to others which is bumped into reality of losing control. For that many works and propositions are passed by many researches such as in [20] and [7] using cryptography concepts and also in some of the works basing on users' identities.

In this report, we suggest a new model using some complex mathematical concepts such as standard deviation, confidence interval and primitive root to protect access control using users' identities and groups; for that we first introduce some backgrounds and definitions of the mathematical concepts that we practice, and so we introduce the main hypothesis under our good example, talking about its theoretical efficiency and carrying a set of experimentations on a set of information.

2 State of the art

The access control presents a sensitive domain in informatics security where it consists of defining such policy that allows or not for such user to get the access to such object; with the coming of concepts of big data and data sharing, this domain became a real challenge in research area. Many works are done within this highly active topic where the most of these works use a promising technique called Attribute Based Encryption such as in [32] [26] [29] and [17]; in [7] the author presented his approach of controlling hierarchical access using multiple key assignment in cryptography where he proposed four schemes, in other world four extensions of his work: bounded, unbounded, synchronous and asynchronous in order to give the general idea under temporal access control; in [2] the authors show their new approach of controlling access on resource-deprived environment in sensor data by integrating the Ladon Security Protocol that offers a secure access using end-to-end authentication, authorisation and key establishment mechanisms in PrivaKERB user privacy framework of KERBEROS environment; in [27] the authors introduced a purpose of using Elliptic Curve Cryptography (ECC) to control the access to data over sensor networks so that they presented their implementation of ECC in TelosB sensor network platform and evaluated their results by comparing it with the results of [18] and [19]; in [25] the paper is addressed to introduce the idea of SafeShare that consists of controlling the access by encapsulation of shared data so that their point of view consists of using the ABE to encrypt, encapsulate, audit and log the data in order to define a perform access control policy; other works go to the fact of using data content to control the access such as it is pointed out in [30] and [33].

Before going far in our work, we like to give you a complete grasp about some general concepts that we used in this report

2.1 Standard deviation

It is a mathematical concept that gives the measure of dispersion of a specific population starting from its mean which can be regarded as the average of the population's values, however, the standard deviation is linearly related to the multiplication of the individuals over the population space; the more the individuals are spread the higher is the deflection; the following formula is the used one to calculate the standard deviation of such population of size n.

$$S= \sqrt{\frac{\sum_{i=0}^{n}(X_i- \bar{X})^2}{n-1}} \quad (1)$$

Where the $\bar{X}$ presents the mean that is calculated using the following formula

$$\bar{X} = \frac{\sum_{i=0}^{n} X_i}{n} \quad (2)$$

And X_i is an individual of the population.

the standard deviation that is used in many cases such as in [16] where the authors proposed a new approach for selection of best threshold where the goal is to obtain better results for image segmentation and evaluated their results by comparing it with other conventional methods in term of several criterions such as the number of misclassified pixels; in [8] the authors proposed and evaluated a new query performance predictor for retrieval models using the standard deviation by testing several confidence levels; another use of standard deviation in information sciences is presented in [34] where the authors presented a standard deviation model to answer the problem of failure data in software reliability that presents a major problems in money costs and costumer satisfactions.

2.2 Confidence interval

It is an inferential statistical measurement that represents an interval of probability that such population individual can fall in basing on three essential parameters: the population's mean, the standard deviation and a specific percentage called confidence level. The confidence interval is calculated as follows:

$$\text{CI= mean} \pm \text{marge\_error} \quad (3)$$

Where the merge error presents the remainder between the mean and the extremities of the interval, the equation used to compute the merge error has two different cases: the case of a sample which has a size less than 30 and the one which accepts a size more than 30, the initial difference resides in special value called t-value in the first case and z-value in the second one, these two values are pulled from two different tables known as T-table and Z-table

However, the extraction of values from these two tables is different, meanwhile, in our work we use z-table because of our population has 2 000 individuals in which are divided into 10 groups each one has more than 150 individuals, the computation of error marge passes by two steps:

- Extracting z-value from the table; for that, we must compute the α/2 value where the α is the confidence level, let's get an example of confidence level of 90%, the α/2 value is 0.90/2= 0.45, after that we search the closest value in the table, we find 0.4495 and 0.4505, then for each one of these values we calculate the corresponding row + the corresponding column and we get 1.64 and 1.65, finally the z-value equals to (1.64+1.65) /2= 1.645
- Now we have the z-value, the merge error is calculated using the following formula:

$$\text{Error\_marge= z-value x } \frac{S}{\sqrt{n}} \quad (4)$$

Where the S is the standard deviation and the n is the size of the sample

- Another value could be derived from the standard deviation called the standard error that represents the distribution of the sample and it is figured using the formula 5 as follows:

$$\text{Standard error} = \frac{S}{\sqrt{n}} \quad (5)$$

2.3 Primitive root

In informatics security the primitive root is an important concept used in several cases, especially in the case of sharing the keys in public key cryptography schemes; formally a primitive origin of a number P is the number that satisfies the following attribute:

r is a primitive root of P => $\forall$ i, j $\in \mathbb{N}$, if i $\neq$j than ri mod P $\neq$ rj mod P

Nevertheless, in mathematics there is no accurate way to compute a primitive root of a number, instead, there is a method to verify if such number r is a primitive origin of a number P as shown the following code:

```
Procedure isPrimitiveRoot (number r; number P)
Begin
Compute ℓ (P);
Decompose ℓ (p) to a set of prime factors
For each prime factor fi do
Compute mi= r^ℓ(P)/fi mod P
If all mi ≢ 1 mod P & mi ≢ -1 mod P then r is primitive
root of P
End.
```

The most known algorithm which uses the primitive root is the famous Deffie-Helman algorithm for sharing secrete keys because of his special characteristic that is known as discrete logarithm problem where it is proved that for a number r being primitive root of a number P, if we know r, P, and the result of ra mod P we could never conclude the number a

2.4 Hierarchical identification

The big data knows comes with real evolution not only in term of data volume, but also in term of number of users which makes the identification of them a crucial problem and implies the search for new technics of choosing identities; one of these technics is a promising and new method called Hierarchical Identification that aims to benefit from different information concerning the users such as their groups so that the identities are depending on these information using the concatenation process as shows the figure 1 bellow:

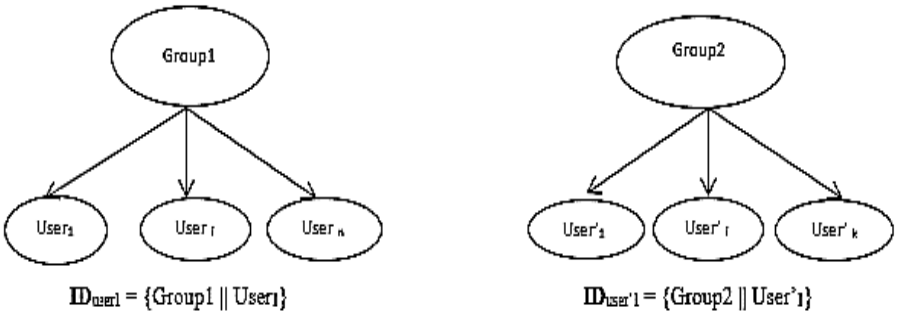


Fig. 1. Hierarchical Identification

This method has a major advantage resides in the ability of using the same identifier for multi users in different groups which allow the identification of big number of users with small size of identities and that can be useful in many cases that are related to the identification such as authentication mechanisms

3 Our approach

Our advance is founded on three independent processes: defining access policy by computing the access control matrix and process of sharing the access rights.

3.1 Computing the access control matrix

This process is based, as the figure 3 shows above, on five steps: identification, normalization of identities, calculation of confidence interval for each group, calculation of digital signature for each user and ultimately determine the access rights by defining the matrix of access rights; in the remainder of this section we will detail each one of the stairs:

3.1.1 Identification**.**

In this step, we target to get the identities of users utilizing the hierarchical identification mechanism in society to afford a standard configuration and size of the identities, we pass an address range of 10000 identities for each group using a concatenation operation between the group's ID where the user belongs and the genuine identity of the user, for example, let's consider a user with IDu= 0001 who belongs to a group which has the IDg= 01, the used identity of the user in our organization will be IDfinal= Edge || IDu= 010001= 10001.

3.1.2 Normalization**.**

We can notice from the formula 4 that the standard deviation has a role in the process of computing the error marge, from its position in the formula, it's clearly shown that the error merge is linearly related to the standard deviation; otherwise, the more bigger

is the standard deviation the more bigger is the error marge, by consequence, the more larger is the confidence interval, for that we suggest a normalisation of identities of the groups in order to create a less propagation rate in the range corresponded to each group so that instead of a maximum difference of 10000 between the extremities of a values of group, we diminish this value to 1 by dividing the identities by 10000. Getting hold of the example of a group in which the ID values go from 00001 to 10000, the normalized values go from 0.0001 to 1.0000; another normalization is used at the level of groups' sizes in order to preclude the influence of the great number of users in the group in the process.

3.1.3 Computing confidence interval**.**

Our system defines for each group a specific confidence interval within the identity range of the group, this interval is estimated using various parameters, leading off from the standard deviation, and so setting the confidence layer, after that computing the merge error, finally utilizing the group means to get the final confidence level; all the computations in this step use the normalized values instead of the real ones.

3.1.4 Computing the digital signatures**.**

This step is independent from the two precedent steps so that it can be executed in parallel with them, however, this step uses the concept of primitive root defined in the section 2.3 in such way that guarantees the unicity of signature for each user; to do that our system, firstly, generates for each group a big prime number P and find one of his primitive roots R, after that, for each user the generated signature equals to RIDfinal mod P; we choose this formula for two reasons: first, since R is primitive root of P then for two different users we will never get the same signature; and the second reason is that to protect the IDfinal of the user because of its major property known as Discrete Logarithm Problem cited in section 2.3, the IDfinal used in this step is in the real form and not the normalised one, thus, our system consists of generating the P as big prime because of two reasons too: first, there is no exact mathematical way to compute the primitive root but instead of that there is a way to verify if such number is primitive root as the procedure in section 2.3 shows using $\ell(P)$, so that we choose the P as prime to optimise the computation because $\ell(P)$ in this case equals to P-1; the second reason of choosing P as prime is also an optimisation reason because the researches proved that for P prime, we have a probability of 0.50 to generate a primitive root between 0 and $\ell(P)$.

3.1.5 Generation of access control matrix**.**

This is the most important step of this process where the access rights are defined starting from the confidence interval of each group, the normalised identity of the user and his own signature; to do that our system conducts a set of tests for each group and each user by verifying if the normalised identity of a user belongs to the confidence interval of the group in order to know in which group the user belongs; once the system defines that, it start comparing the signatures of the data with the one of the user so that if are equals, the user will have full access and all the rights on the data that is considered as his own; else the user will have access on read only on the data that is considered in

this case as shared data with him; for the other groups that the user doesn't belong, he will get no access right to their corresponding data; at the end of this process a matrix user x data is generated that resumes the access control policy.

3.2 Process of sharing the access rights

This procedure has been summed in order to answer some other problem that came to mind; what if such user decide to grant some other user to have write access right to his own data?, Otherwise, we aim by adding this process to allow for users of our system to share the same rights on same datum, to do that, our system uses Deffie-Helman algorithm of sharing cryptographic keys, but in our case to share the signatures between users; first of all, our system generates randomly a big prime number Q and a primitive root r then computed a primary signature for each of the users that will have the same access right S = $r^{ID_{user1}}$ mod Q that corresponds to user1, finally the system computes the final signature S_{final} = $S^{ID_{user2}}$ mod Q and signs the data with it. In this process, our system generates a new big prime Q and his primitive root and utilize them in the stead of the P that corresponds to the group where user1 belongs in order to protect the original signatures of the both users because it is used to sign other data that the users don't want to share rights with each other. The following process resumes this step:

Algorithm RightsSharing ()
Input ID_{user1}, ID_{user2}*: users identities*
Output S_{final}*: final shared signature*
Begin
Generate randomly a big prime Q and a number r < (Q-1);
While (r is not a primitive root of Q) do
Generate randomly a new r < (Q-1);
End while;
Compute S ← $r^{ID_{user1}}$ *mod Q;*
Compute S_{final} *←* $S^{ID_{user2}}$ *mod Q;*
Sign data with S_{final}*;*
End.

However, in this approach we choose to sign the data independently of its content, unlike the work presented in [30] because of two reasons: first, is to protect the privacy of data by perturbing the name in order to hide the real extension of the documents; and secondly, is to prevent problems of distrust like the famous one that Dropbox had recently because of its policy against violation of copyrights where a client claimed to the company from reading the content of his own data via Tweeter after the company prevent him from storing a document because of copyright violation[1].

[1] http://assoquebecois.com/2014/04/01/dropbox-clarifie-sa-politique-sur-lexamen-des-dossiers-partages-pour-les-questions-dmca/

4 experiments and Results

We take a set of experiments by building up a framework consists of 2000 users where each one has ten files stored in our system with a total of 20000 documents which gives an access control matrix of 2000 x 20000 that equals to 40 million right; the users are divided into 10 groups; this section is reserved for the introduction of a set of results using various parameters. But before going to the results, we will present the details of our dataset as shown in table 1

Group	Number of users	Range of identities	Range of normalised identities	Corresponding normalised Mean	Corresponding normalised standard deviation
1	181	[00001 …09999]	[0,0001 …0,9999]	0.096	2.511
2	234	[10000 …19999]	[1,0000 …1,9999]	1.100	3.377
3	205	[20000 …29999]	[2,0000 …2,9999]	2.081	11.140
4	209	[30000 …39999]	[3,0000 …3,9999]	3.020	23.819
5	190	[40000 …49999]	[4,0000 …4,9999]	4.101	25.117
6	184	[50000 …59999]	[5,0000 …5,9999]	5.098	25.649
7	191	[60000 …69999]	[6,0000 …6,9999]	6.101	25.311
8	221	[70000 …79999]	[7,0000 …7,9999]	7.096	23.672
9	193	[80000 …89999]	[8,0000 …8,9999]	8.067	34.671
10	193	[90000 …99999]	[9,0000 …9,9999]	9.098	34.774

Table 1. Dataset details used in our system

As we notice in table 1, the mean is entirely related to the distribution of the values in their specific range and does not necessarily show the core of the range, and we notice also that the standard deviation is always out of the range of the sample because of the use of the ability of two during his computation.

We carry on a set of comparisons organized in two steps: first, we confront a comparison between domains in order to study the influence of the distribution of the sample in our approach, secondly, we study the effect of choosing the confidence level in our approach, and finally, we evaluate our system by comparing it with other conventional work.

The next table shows the result of average of the access rate by domain in many experiments on normalized identities and real ones.

In table 4 below, we will detail the results of the admission rate by group in term of the chosen confidence level in which we used many confidence levels and we take the ones

that give an excited results in some of our groups, the following board indicates the chosen confidence levels with the corresponding z-value of each ace.

Confidence level / Group	20%	28%	28%<<29%	29%	30%	31%
1	48.62	68.51	70.72	71.82	72.36	76.24
2	55.55	78.20	79.91	81.20	84.19	85.90
3	98.54	98.54	98.54	98.54	98.54	98.54
3	97.13	97.13	97.13	97.13	97.13	97.13
5	100.00	100.00	100.00	100.00	100.00	100.00
6	100.00	100.00	100.00	100.00	100.00	100.00
7	100.00	100.00	100.00	100.00	100.00	100.00
8	100.00	100.00	100.00	100.00	100.00	100.00
9	99.48	111.39	122.80	140.41	165.80	193.78
10	100.00	100.00	108.81	124.87	136.79	146.63

Table 2. Access rate by group in term of chosen confidence level

As the table 2 shows, each one of the confidence levels that we choose presents some good results in some groups and in the same time bad results in other groups; the best confidence level for groups group 01 and group 02 is 31% with rate of access of 76.24% in group 01 (138 user from 181) and 85.90% for group 02 (202 users from 234) while this level presents the worst results in programing group with 193.78% of access rate (374 users from 193 authorised), instead of that, the groups programming and security gives best results with less level of confidence using 20% of confidence level with 99.48% for programing (192 users from 193 authorised) and full access rate without error for security; meanwhile, the other groups such as data mining and natural sciences gives excited results without been influenced of the value of confidence level.

The following table presents the results of average of access rate and error rate between domains in term of variation of confidence level in normalised values where the positive value of error rate means that there is less users have access than the authorised ones and negative value means that there is more users that have access than the authorised ones, otherwise, the positive error rate means that there are users who must have access but our system doesn't allow to them to get access while the negative value means that there are some users must not access to data but our system allows to them to have access.

Confidence level (%)	Average of access rate (%)	Error rate (%)
20	89.92	10.08
28	95.37	4.63
28<<29	97.73	2.27
29	101.39	- 1.39
30	105.47	- 5.47
31	109.81	- 9.81

Table 3. Results of average of access rate and error rate in term of confidence level variation

From the table 3 we can clearly notice that the confidence level between 28% and 29% gives better results with an average rate of access about 97.73% even if it represents some weaknesses in the last two domains where the access rate exceeds the 100 % (122.79% for 9 domain (44 unauthorized users), and 108.29% for 10 (16 unauthorized users)), the reason of why we didn't determine the exact value between 28 and 29 is that because all values within this range gives the same z-value which is about 0.365. The use of confidence level equals to 20% presents a major advantage because of all the values of access rate doesn't exceed the 100%, which means for all data there is no unauthorized access while it presents the worst result in term of error rate with more than 10% of authorized users could not access to data that must get access to because of the less access rate in the first domains where only 48.61% (only 88 users) of authorized users could access in 1 domain and 55.55% (only 130) could access in 2 domain. So, as the table shows, once we defined a confidence level starting from 29% the results became more badly (average of 28 of unauthorized users could access to data for 29%, 110 to 30%, and 197 to 31%).

After introducing a set of outcomes using a variation of parameters, we put our system in confrontation with a set of conventional works in the image of the system presented in [18] named TinyECC, and the one shown in [27] under the name ECC-AC in term of time of generating a signature and time of verifying the signature, however, our system generates a signature of average of size of 128 bits because of the role of a prime number of sizes of 2048 bits and primitive root of 1024 chips; the following table introduces the effects of time of generation and verification of signatures

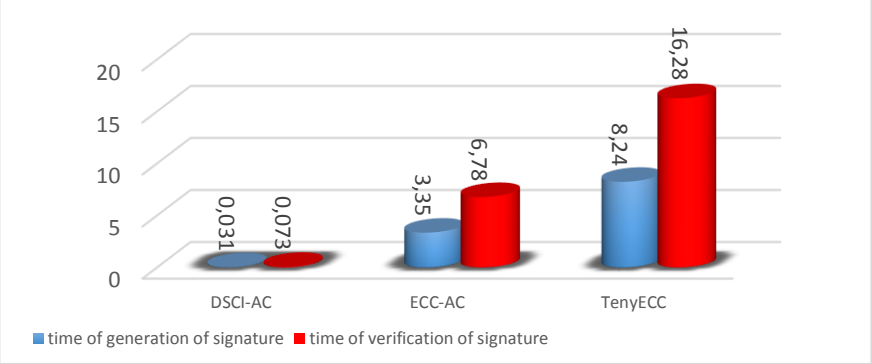


Figure 2. Results of comparison of time of generation and verification of signatures

As figure 2 indicates, our system doesn't take much time for generating or verifying the signatures and that's due to the fact of using a simple computations to generate signatures and also the signature are used to perturb the data via their names which make its verification low costs because the system doesn't need to treat the data content to have signature, thus, in our approach only the server is the responsible of computing and verifying the signatures which leads us to eliminate the time of connection for users. In the other hand; we notice clearly that the time of generation of a signature is less than the one of its verification because the generation is based only on one and one only operation while the verification takes more time because of the number of operations resides on searching if the user belongs to the group in order to define if he has already the access or not then compare the two signatures to define the right that he has.

5 Conclusion

In this paper, we ushered in a new glide path of applying digital signature and the confidence interval in order to answer three essential questions: how could we control the approach to information that we don't hurt even the control on?, To answer it, we first divided the users into groups by their domains then compute for each group its own confidence interval that we used in our system in parliamentary procedure to ascertain who has access to data and who doesn't, after determining which user has access to the data, another question came to mind; for the users who have access to data, which right should they have, is that full access or read only access? We answered this question by using the digital signature generated using another mathematical concept called primitive root basing on prime numbers and random theories in order to precisely which access right each user must take; then by assisting these two questions we could define the last access control matrix; the final question that we answered in this study is that if such user decide to afford full access on his data for another user, how could we ensure that? To respond that we offered the use of Deffie-Hellman algorithm of sharing cryptographic keys in order to permit users to partake in the same signature by consequence have the same access right on the same data.

As future work, we will usher in new models of using meta-heuristics technics to improve the results of this work by searching for the appropriate assurance level for each

group we also will give other models using cryptography whose purpose is to prevent the server from recognizing the genuine signatures of the users. In the final stage, it only remains to mention that the security in Big Data is all grounded on trust then that no trust no security.

6 References

Arunkumar, S., Raghavendra, A., Weerasinghe, D., Patel, D., & Rajarajan, M. (2010, October). Policy extension for data access control. In Secure Network Protocols (NPSec), 2010 6th IEEE Workshop on (pp. 55-60). IEEE.

Astorga, J., Jacob, E., Huarte, M., & Higuero, M. (2012). Ladon 1: end-to-end authorisation support for resource-deprived environments. Information Security, IET, 6(2), 93-101.

Altman, D. G., & Bland, J. M. (2005). Standard deviations and standard errors.Bmj, 331(7521), 903.

Bagheri, E., Babaei, S., & Khayyambashi, M. R. (2009, August). A new method for consistency of access control in web services. In Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference on (pp. 567-569). IEEE.

Camenisch, J., Mödersheim, S., Neven, G., Preiss, F. S., & Sommer, D. (2010, June). A card requirements language enabling privacy-preserving access control. In Proceedings of the 15th ACM symposium on Access control models and technologies (pp. 119-128). ACM.

Chen, Y. R., Chu, C. K., Tzeng, W. G., & Zhou, J. (2013, January). Cloudhka: A cryptographic approach for hierarchical access control in cloud computing. InApplied Cryptography and Network Security (pp. 37-52). Springer Berlin Heidelberg.

Crampton, J. (2009). Cryptographically-enforced hierarchical access control with multiple keys. The Journal of Logic and Algebraic Programming, 78(8), 690-700.

Cummins, R., Jose, J., & O'Riordan, C. (2011, July). Improved query performance prediction using standard deviation. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (pp. 1089-1090). ACM.

Curé, O., Kerdjoudj, F., Le Duc, C., Lamolle, M., & Faye, D. (2012, September). On the potential integration of an ontology-based data access approach in NoSQL stores. In Emerging Intelligent Data and Web Technologies (EIDWT), 2012 Third International Conference on (pp. 166-173). IEEE.

Di Vimercati, S. D. C., Foresti, S., Jajodia, S., Paraboschi, S., & Samarati, P. (2007, September). Over-encryption: management of access control evolution on outsourced data. In Proceedings of the 33rd international conference on Very large data bases (pp. 123-134). VLDB endowment.

Guo, J., Baugh, J. P., & Wang, S. (2007). A group signature based secure and privacy-preserving vehicular communication framework. Mobile Networking for Vehicular Environments, 2007, 103-108.

Goyal, V., Pandey, O., Sahai, A., & Waters, B. (2006, October). Attribute-based encryption for fine-grained access control of encrypted data. InProceedings of the 13th ACM conference on Computer and communications security (pp. 89-98). ACM.

Keathley, E. F. (2014). Big Data and Bigger Control Issues. In Digital Asset Management (pp. 99-115). Apress.

Kelani Bandara, K. B. P. L. M., Wikramanayake, G. N., & Goonethillake, J. S. (2007, August). Optimal selection of failure data for reliability estimation based on a standard deviation method. In Industrial and Information Systems, 2007. ICIIS 2007. International Conference on (pp. 245-248). IEEE.

Khalil, I., Khreishah, A., & Azeem, M. (2014). Consolidated Identity Management System for secure mobile cloud computing. Computer Networks,65, 99-110.

Li, Z., Cheng, Y., Liu, C., & Zhao, C. (2010, March). Minimum Standard Deviation Difference-Based Thresholding. In Measuring Technology and Mechatronics Automation (ICMTMA), 2010 International Conference on (Vol. 2, pp. 664-667). IEEE.

Li, M., Yu, S., Zheng, Y., Ren, K., & Lou, W. (2013). Scalable and secure sharing of personal health records in cloud computing using attribute-based encryption. Parallel and Distributed Systems, IEEE Transactions on, 24(1), 131-143.

Liu, A., & Ning, P. (2008, April). TinyECC: A configurable library for elliptic curve cryptography in wireless sensor networks. In Information Processing in Sensor Networks, 2008. IPSN'08. International Conference on (pp. 245-256). IEEE.

Malan, D. J., Welsh, M., & Smith, M. D. (2004, October). A public-key infrastructure for key distribution in TinyOS based on elliptic curve cryptography. In Sensor and Ad Hoc Communications and Networks, 2004. IEEE SECON 2004. 2004 First Annual IEEE Communications Society Conference on (pp. 71-80). IEEE.

Miklau, G., & Suciu, D. (2003, September). Controlling access to published data using cryptography. In Proceedings of the 29th international conference on Very large data bases-Volume 29 (pp. 898-909). VLDB Endowment.

Ortiz, P., Lazaro, O., Uriarte, M., & Carnerero, M. (2013, June). Enhanced multi-domain access control for secure mobile collaboration through Linked Data cloud in manufacturing. In World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2013 IEEE 14th International Symposium and Workshops on a (pp. 1-9). IEEE.

Perera, C., Zaslavsky, A., Christen, P., & Georgakopoulos, D. (2014). Sensing as a service model for smart cities supported by internet of things. Transactions on Emerging Telecommunications Technologies, 25(1), 81-93.

Shtok, A., Kurland, O., & Carmel, D. (2009). Predicting query performance by query-drift estimation. In Advances in Information Retrieval Theory (pp. 305-312). Springer Berlin Heidelberg.

Stevens, G., & Wulf, V. (2009). Computer-supported access control. ACM Transactions on Computer-Human Interaction (TOCHI), 16(3), 12.

Thilakanathan, D., Calvo, R., Chen, S., & Nepal, S. (2013, December). Secure and Controlled Sharing of Data in Distributed Computing. In Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on(pp. 825-832). IEEE.

Tu, S. S., Niu, S. Z., Li, H., Xiao-ming, Y., & Li, M. J. (2012, May). Fine-grained access control and revocation for sharing data on clouds. In Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW), 2012 IEEE 26th International (pp. 2146-2155). IEEE.

Wang, H., Sheng, B., & Li, Q. (2006). Elliptic curve cryptography-based access control in sensor networks. International Journal of Security and Networks, 1(3), 127-137.

Wang, Z. H., Zhi, S. S., & Liu, H. M. (2012, July). MSHS: The mean-standard deviation curve matching algorithm in HSV space. In Machine Learning and Cybernetics (ICMLC), 2012 International Conference on (Vol. 3, pp. 1064-1069). IEEE.Yang, Y., & Zhang, Y. (2011, September). A generic scheme for secure data sharing in cloud. In Parallel Processing Workshops (ICPPW), 2011 40th International Conference on (pp. 145-153). IEEE.

Yu, S., Wang, C., Ren, K., & Lou, W. (2010, March). Achieving secure, scalable, and fine-grained data access control in cloud computing. InINFOCOM, 2010 Proceedings IEEE (pp. 1-9). Ieee.

Zeng, W., Yang, Y., & Luo, B. (2013, October). Access control for big data using data content. In Big Data, 2013 IEEE International Conference on (pp. 45-47). IEEE.

Zhang, X., Liu, C., Nepal, S., Dou, W., & Chen, J. (2012, November). Privacy-Preserving Layer over MapReduce on Cloud. In Cloud and Green Computing (CGC), 2012 Second International Conference on (pp. 304-310). IEEE.

Yang, Y., & Zhang, Y. (2011, September). A generic scheme for secure data sharing in cloud. In Parallel Processing Workshops (ICPPW), 2011 40th International Conference on (pp. 145-153). IEEE.

Nabeel, M., Bertino, E., Kantarcioglu, M., & Thuraisingham, B. (2011, October). Towards privacy preserving access control in the cloud. InCollaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2011 7th International Conference on (pp. 172-180). IEEE.

Bandara, K., Wikramanayake, G. N., & Goonethillake, J. S. (2007, August). Optimal selection of failure data for reliability estimation based on a standard deviation method. In Industrial and Information Systems, 2007. ICIIS 2007. International Conference on (pp. 245-248). IEEE.

Smart Approach for the solar irradiation estimation based on Multi-Agent System

Mohamed Amir Abbas*, Nadjia Benblidia**
Nour El Islam Bachari***

*Faculty of Science, Saad Dahleb University - Blida, Algeria
m.amir.abbas@gmail.com
** Faculty of Science, Saad Dahleb University - Blida, Algeria
benblidia@gmail.com
*** Faculty of Biological Sciences, Houari Boumediene University - Algiers, Algeria
bachari10@yahoo.fr

Abstract. In this paper, we present a new approach for a smart estimator of the solar irradiation based on multi-agent system. This approach is adaptive with the evolution of the meteorological data provided from several sources by the use of intelligent techniques (neural networks). The solar irradiation data could be used to facilitate the taking of decision in the choice of the right geographical place for implementing solar energy equipment. One of our main cares is to be able to classify those geographical sites, from where the data is collecting, in clusters representing the potential of the solar energy production. The smart part in our approach is assured by the use of multi-agent architecture which allowing the intelligent agents to collaborate autonomously in providing the best estimation through superposed levels. Each level regroups a set of specialized agents working independently and communicating with the agents of the other levels under the supervision of one controller agent in order to reach one goal: deliver the best data to make decision.
Keywords: Meteorological data, Solar irradiation, Multi-Agents System, Decision Application.

1 Introduction

The development of the renewal energy sources is becoming more attractive and beneficial for the most of the economical societies in the world, further to their inexhaustible provisioning and its positive ecological effects. As a main source, the solar energy is been explored in thermic systems (heat energy) and photovoltaic systems (electricity energy).

The measurement of solar irradiation is very important in the designing and implementation decision of the energetic systems, but the measured data is not always available especially in isolated areas, due to the cost and unavailability of measurement instruments.

The availability of several other meteorological data types in the most weather station, have motivated researchers to think about the integration of those parameters for estimating solar radiation hoping to find correlations that lead the estimation of the solar irradiation.

Recently, several empirical hybrid formulas using some meteorological data types have been tested. The meteorological parameters include the duration of sunshine, temperature, cloudiness, humidity, vapor pressure etc.

Through this work, we wish to design and develop an intelligent estimator for the global solar irradiation using an evolutionary and multimodal data, such: temperature, insolation and humidity, where the most of this data are available along the year.

The complexity of the present approach is not located at the models of global solar irradiation estimation, but in the coordination of tasks between agents and the monitoring of their numbers.

This article is organizing as below: the second section presents one used method of AI (Artificial Intelligence) which is Neural Networks, followed by a short resume on the concepts of multi-agent systems. In section three, we develop the global architecture of our approach by a definition of the agent's entities and their levels of interaction. The fourth section describes briefly our implementation roadmap using the database (Weather-Temperature/insolation) followed by the proposed evaluation metrics, then a conclusion and perspectives.

2 State of the art

These last years, the intelligent techniques (*neural networks, fuzzy logic, decision trees*) are been mostly used to design human mental models where the traditional methods are inefficient. The solar irradiation prediction is one of those domains where the intelligent methods were applied successfully, exceptionally with the neural networks, Yacef and al. (2012), Inman and al. (2013), Martí and Gasque (2011).

2.1 Neural networks

It is a matrix of elementary units (neurons) interconnected between them, grouped into several groups (layers). Fig 1. The S neurons of the same layer are connected to the R inputs, where a weight wi,j is associated for each connection which connects neuron i to its input j, so the global weights of a layer define a matrix W of dimension S × R.

i: indicates the range of the neuron on the layer.

j: indicates the number of the input neuron.

A neural network regroups three types of layers. The first layer is named "Input Layer", the last layer is named "Output Layer" and between these two layers, we have the "Hidden layers".

One of the most powerful properties of a neural network model is its capacity to learn from its environment, to improve its performance through a process of experimental learning. Generally, the learning stage involves the modification of the weights values that connect the neurons of a layer to other neurons, Habiboulaye (2006).

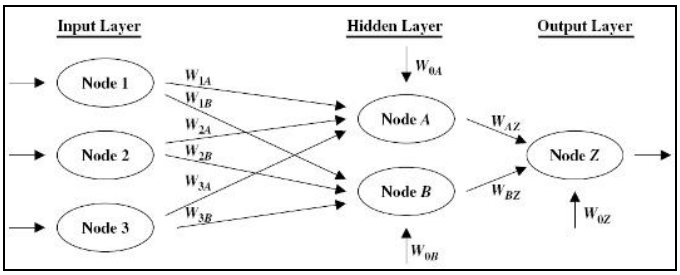


FIG. 1 – *Neural Network Architecture.*

The network design starts with learning stage on representative samples in order to define the first weights of connections between neurons, Blansche (2006). In the literature we found two types of learning method:

- Supervised: when it is possible to provide a desired output. The weights of the neural network are adjusted based on the error signal which equals to the difference between outputs provided by the network and the desired output;
- Unsupervised: when the adapting of the weights depends only on the internal criteria of the network. The adaptation is done only with the input signals. No error signal and no desired output are defined.

We have two main types of neural networks:

1) Static neural networks: Called also feed-forward networks, the information is propagating in one direction from the left to the right. In the literature, we can find:
   (a) Multilayer Perceptron network (MLP): it is an assembly of several layers connected between them, from the left to the right, the output of the left layer will be the input of the next layer,
   (b) Radial Base Function network (RBF): it constitutes from only three layers. Entry layer, RBF hidden layer and the output layer which has a linear function,
   (c) Extreme Learning Machine (ELM): it could be improving through a learning process in short time.
2) Dynamic neural networks: Here the information is propagating in duplex directions. Also in the literature, we can find:
   (a) Neural Network Auto-Regressive with Exogenous Inputs (NNARX): its layers have recurring connections (retroaction of the outputs onto the inputs of the same layer),
   (b) Elman Network: it contains only two layers, with the recurring connection on its first layer.

2.2 Multi-Agent Systems

A MAS is a collection of autonomous intelligent agents able to communicate, collaborate and act in a common environment to perform several common or individual tasks, in order to solve a complex problem, Occello (2003).

The application of multi-agent systems has expanded greatly, reaching a wide range of technical fields related to artificial intelligence, distributed systems and software engineering.

The principle of classical artificial intelligence is defined by the development of autonomous program for the execution of complex tasks based on expert knowledge and centralized in a single system. Reverse to a distributed system which is constituted from a set of independent programs that appears to the user as a single system, Saidane and al. (2005).

The advantage to use MAS is to take advantage of the two disciplines (AI & Distributed System), and simplify the development, implementation and execution control of this complex system.

In the context of intelligent predictions, the multi-agent systems are very effective to apply several specialized process simultaneously on an autonomous basis, with a possibility of adaptation with the evolutionary of data.

Our conception is based on the GAIA method, Zambonelli and al. (2003). We began by defining the entities of agents with their properties (name, roles, responsibilities, knowledge, protocols, group), then we have modeled the interaction between these agents using conceptual charts.

3 Architecture of the proposed approach

Our approach is the result of a merger between four phases of data processing, each onewill concerns a specialized agents regrouped in one level. They communicate with the agents of the other levels through several connections.

The choice of the four levels was inspired from the main standard functionality of a business intelligent system based on the data warehouse source.

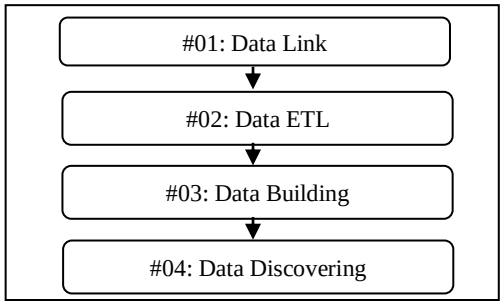


FIG. 2 – *Architecture of the proposed approach.*

We start always by a data link step. At this level, our system will assure only the availability of connections and access authorizations to the different data sources. For each data source connection a specific agent will be deployed as a responsible of that link.

In the second level, we find the ETL data engine (***Extract-Transform-Load***). It concerns the definition of queries structures and plugins by three specialized agents, everyone will assume one of the below functions:

- Extraction and selection of the raw data from the sources through the data links;
- Transformation and adaptation of the selected data as per the integrity and aggregation rules;
- Loading of the transformed new data in the global database (Data warehouse).

The agents of the third level will be in charge to execute the algorithms for the solar irradiation estimation. Here, the number of agents is not fixed, since it depends on the number of the selected calculation methods. In our case we have only one agent which is using a neural network method.

For the interaction and communication between our system and the final user, we defined the last level named "Data discovering". Specialized agents could interact with the user through a unique display interface, where every agent is responsible to display only a part of the global data (Data-Mart) as sub-source. In our case, we need only one agent for displaying the solar irradiation data.

The approach will involve the below reactive agents' types:

- ***AgS***: the Supervisor agent acts as the leader of the group of agents. He is responsible of coordination of tasks between the system's agents. His main task is to make an adaptive selection of the right agent that fit the nature of data treatment;
- ***AgK***: Linker agent, his main task is to assure and initiate the connection link between the system and the data sources. He has the needed authorizations and parameters to execute the queries of authentications in order to connect safely to the required data source;
- ***AgE***: Extractor agent. His mission is to extract the raw data from the sources based on the initiated connections, which are created by the linker agent. Generally, the extraction actions are planned in fixed moments, and the selected data is the new records not yet selected in the past. For that, this agent is referring to the supervisor agent where the extraction schedules are defined;
- ***AgT***: Transformer agent. Based on the application of predefined aggregations and transformation rules, this agent will interact on the new extracted raw data in order to clean and adapt them as per the formal data formats before its loading in the global database;
- ***AgL***: the Loader agent is the responsible of loading the new adapted data in the global database of the system. He can differentiate between the natures of the data and allocate it in the specific tables (fact or dimension table);
- ***AgB***: the Builder agent adopts an intelligent technique for the data processing. He will use the new transformed loaded data in the estimation of the solar irradiation factor. Regarding the importance of this building step in our approach, we can define several agents with different techniques, and they will contribute in the estimation process in order to strengthen the results;
- ***AgD***: The Discoverer agent provides and displays the result of the final building data to the user, and that after the confirmation of the user authentication. According to the user requests and allowed privileges, the discoverer agent will deliver the data as a part of the global database (data Mart).

In resume, our approach could contain at least 7 agents or more, it depends on the number of the data sources and applied estimation methods.

As it is showing in the table.1, our agents in different stages are simulating some procedures based on neural networks and algorithmic methods.

Stage	Agent	Procedure	Method
1	AgK1	Link to Databases	SQL queries plugin
	AgK2	Link to Files Directories	LDAP queries plugin
2	AgE	Data extration	SQL & LDAP queries plugin
	AgT	Data transformation	PL SQL algorithm
	AgL	Data loading	PL SQL algorithm
3	AgB	Data building	Neural network algorithm
4	AdG	Data discovering	User interface plugin

TAB. 1 – *Applied agents with methods.*

Let's present now the main stages processing in detail, after this list of presupposed hypothesis regarding the agent's nature:

- ***hyp01***: *We suppose that all agents are synchronized and know well their positions in the environment, and we have unique defined communication protocol between them;*
- ***hyp02****: Each agent is oriented to proceed one and only one of the below tasks: link data-source, extract data, transform data, load data, build data, discover data;*
- ***hyp03****: We have only one supervisor agent who manage and control all other agents, and knows well their status;*
- ***hyp04****: We suppose that the supervisor agent could detect all environment changes successfully. We avoid the treatment of expected errors and incidents in the detections;*
- ***hyp05****: We suppose that all agents are initially in standby status (excepting the supervisor agent), and they do not start any process only if they will get the instruction from the supervisor agent.*

3.1 Link Data Source Stage

Once a new data source is detected by the supervisor agent in the system environment, he will request the linker agent to identify the data source connection.

The wake-up message, attached with detected data source properties, will be sent from the supervisor agent to the linker agent, where he should build and test the specific query for the link connection to the new data source based on the approval of the data source administrator (owner of the data source). If the connection test is successful, he stores the new link query in his own local database, and replies to the supervisor agent with completion message. After that, the linker agent status will be changed to standby.

We have set this stage (definition of the link connection with the data source), because a major of decision systems are provisioning the data from several sources simultaneously, through a specific driver's connectors and authentication properties. It was preferable to design an agent who will be in charge to manage all that links properties, and facilitate the data access to the other agents.

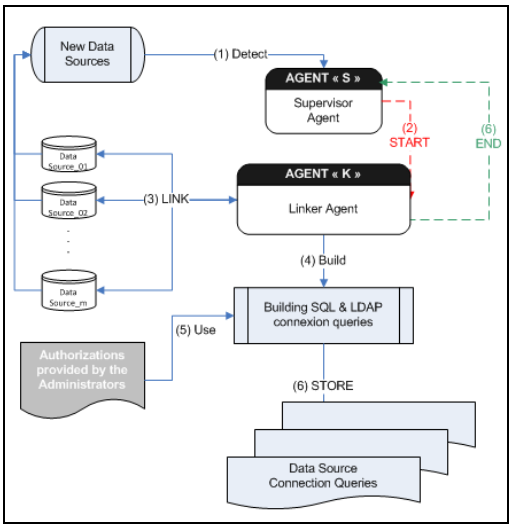


FIG. 3 – *Operating diagram of the data link stage.*

3.2 Extract-Transform-Load Data Stage

As one of the main part of any intelligent business system, we have those three sequential actions of data treatment, starting by the extraction then transformation and completing with the loading of transformed and adapted data to the global database (data warehouse).

In the irradiation solar estimation, the data is extracting from several sources (temperature, insolation, humidity, wind…), by the extractor agent. He is using the SQL transaction to extract the data from the sources, which are in relational type, through external joining queries in order to avoid a duplicated data.

The extraction periods could be planned by the supervisor agent, who is synchronizing the extractions on scheduled periods or in cases of new data source link setup by the linker agent.

The extracted data will be automatically transformed by the transformer agent, by applying the integrity rules and aggregation of data, using PL-SQL algorithms adapted and adjusted periodically as per the evolution of the data.

To finalize this stage, the supervisor agent will instruct the loader agent to load the transformed data into the global database, within its right place (destination tables) using PL-SQL algorithms.

In the irradiation solar estimation, we store the meteorological data in the fact tables. The dimension tables will contain: solar site, data type, period, timing (date, hour, day, month, year, season…).

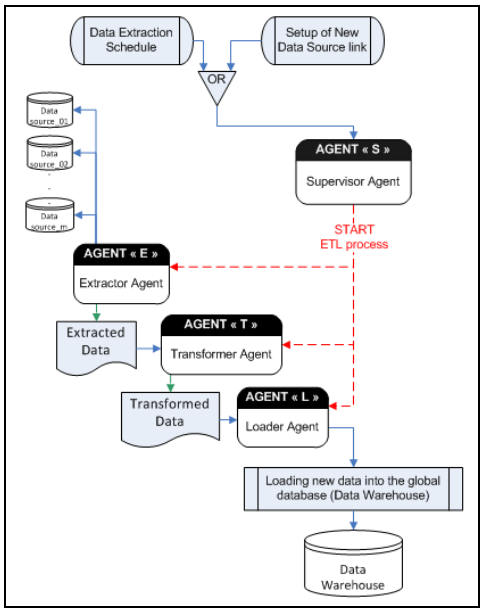


FIG. 4 – *Operating diagram of the ETL data stage.*

3.3 Building Data Stage

It involves a collection of several agents. Everyone will apply an intelligent technique to estimate the global solar irradiation factor using the collected meteorological data.

We have choicen the application of one static neural network method (MLP) associated to one builder agent, which had given good results in the solar irradiation estimation as per the existent research works, Yacef and al. (2012), Martí and M. Gasque (2011), Benghanem and Mellit (2010).

This network will contain three layers. The neurons of the first layer (input layer) depends on the number of the used meteorological data type, the second layer (hidden layer) will contain the possible combination models of the meteorological data types (if we have two inputs data ex: minimum and maximum temperature; we will get three combination models: [T_{min}], [T_{max}] and [T_{min} , T_{max}]). In the last layer (output layer) we have only one neuron of the global solar irradiation value.

This part of data processing could be developed and detailed in separate and complete paper. Where, we present the possible applied functions for the weight network calculation, and the error signal estimation. In addition, the enumeration of possible methods and algorithms for the network learning. However, in this paper, we have summarized this stage through a general presentation of its mains flow steps.

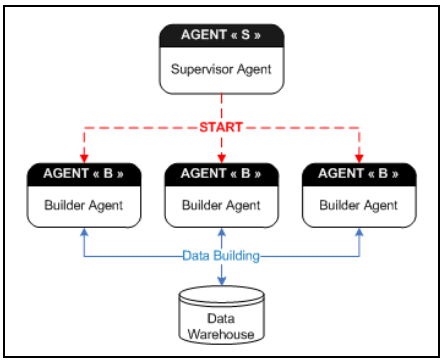


FIG. 5 – *Operating diagram of the Building data stage.*

3.4 Discovering Data Stage

The final value of building a data warehouse or any kind of data processing is to restore the data later to end users, using analysis tools and dashboards well adapted according to access rights, assigned initially to the data requester.

The discoverer agent is able to evaluate the user access rights under the authority of the supervisor agent. He only retrieves the part of the data from the global database (data warehouse) as a data-Mart. It will be delivered to the requesting user, who can do his analysis job conveniently through a data source with less voluminous, and easier to explore compared to the use of the global database.

In the end, we can get a several final users accessing simultaneously to the system, working with same or different data independently from the global database. This feature of data access segmentation is transparent for the user, also, the system background process (ETL, Building), which don't affect the quality of the data exploration by the user, because a dedicated agent is in charge to manage the requests of the users, while the other agents are doing their tasks independently and in the same time.

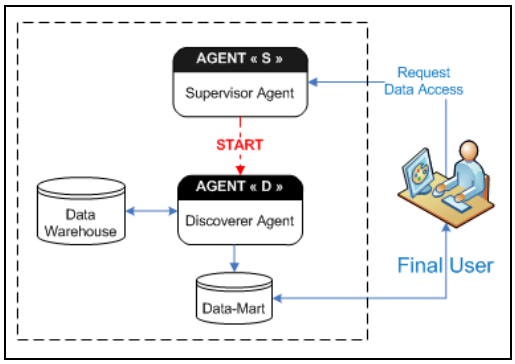


FIG. 6 – *Operating diagram of the Discovering data stage.*

4 Implementation

In this section, we present our roadmap in the implementation of the above approach for the global solar irradiation estimation. We will use the platform Jade (Java Agent Development Framework) developed in Java according to the FIPA specifications in the laboratory TILAB.

The implementation concerns three important databases containing meteorological data. The first one represents the duration of daily sunshine/insolation in (hour) unit. The second database contains the daily temperature measurements in (°C) unit, with T_{max} and T_{min} values. The third database contains the humidity values in (%) unit, with RH_{max} value. Those databases are providing to us by the biology laboratory of the Houari Boumediene University (Algiers, Algeria). They concern one of the Algerian weather stations managed by the MNO (Meteorology National Office). The station' properties are as below:

- **Station Name** : ORAN SENNIA ;
- **Altitude**: 90 m \ **Latitude**: 35°38 N \ **Longitude**: 00°36 W ;
- **Collection Period**: 1980 – 2010.

The aim was to estimate the daily global solar irradiation in (hW/m²) based on the **Ojosu Komolafe** model, Ojosu and Komolafe (1987), depending on the insolation fraction, temperature, and the maximum relative humidity. The fraction of insolation (S/S_0) equals to the fraction of sunshine duration measured gross S (daily duration measurement in the station) by the theoretical sunshine duration S_0 (length of the astronomical day from sunrise to sunset).

To estimate the global solar irradiation (G), Ojosu and Komolafe had proposed the following formula:

$$G = G_0.(a + b.(S/S_0) + c.(T_{min}/T_{max}) + d.(RH/RH_{max})) \quad (1)$$

With:

G_0 is the extraterrestrial solar irradiation.

a, b, c and d are the estimated regression coefficients for the weather site.

The proper distribution of tasks and the exchange of messages communicated between agents allow us to monitor the proper implementation of the global decision-making.

In the context of estimating global solar irradiation, there are many methods to check whether an estimator or predictor is effective, Martí and Gasque (2011), Kalogirou (2001), Mellit and Kalogirou (2008). During this work, we perform a statistical test for comparison the presented approach. The metrics that we can use are described as below on daily basis:

- *The mean absolute error (MAE)*: it is a quantity frequently used to measure the difference between the predictions and measurements. It is giving by:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |(G_{mi} - G_{ei})| \quad (2)$$

  With:

  G_{mi} is the ith measured solar irradiation value, G_{ei} is the ith estimated solar irradiation value, n is the number of values.

- *The mean percentage error (MPE):* it is given by:

$$MPE = \left(\frac{1}{n}\sum_{i=1}^{n} \frac{|(G_{mi} - G_{ei})|}{G_{mi}}\right) \times 100 \quad (3)$$

- *The mean bias error (MBE):* it is defined as the average algebraic difference between simulation and measurement, its formula is given as follows:

$$MBE = \frac{\sum_{i=1}^{n} (G_{mi} - G_{ei})}{n} \quad (4)$$

5 Conclusion

Through this work, we wanted to introduce a new smart approach based on a multi-agent system, in order to estimate the global solar irradiation using a meteorological data. With the MAS architecture, the prediction system is becoming independent and divided to several sub-modules.

We are convinced that experiments are mandatory, and must been applied for any research work before its validation. The thing that remains for us, to be applying as next step, since this work is a part of preparing a doctoral thesis, where experimentations and simulations are not yet been completed.

Our approach could be deployed and simulated on several areas, such the prediction of intelligent application, exploring a multimodal and dynamic data. Its conception is adaptive with any intelligent business model.

After the completion of the first implementations, we planned in the future to involve the satellite pictures in the estimation process, collected form the METEOSAT satellite. The deployment of new agent for the image processing is a must thing, and we are confident that it will not affect our first conception, since we are using distribution intelligent architecture, and it could be improved and extended easily.

References

A. Blansche (2006). *Unsupervised classification with attribute weighting by evolutionary methods*. Doctoral Thesis, Louis Pasteur – Strasbourg I University, France.

AB. Habiboulaye (2006). *Dynamics Classification of Non-Stationary Data Learning And Monitoring Of Evolutionary classes*. Doctoral Thesis, Sciences and Technologies Lille University, France.

A. Mellit, SA. Kalogirou (2008). *Artificial intelligence techniques for photovoltaic applications: Areview*. Progress in Energy and Combustion Science n°1-1, 52-76.

A. Saidane, H. Akdag, I. Truck (2005). *MSA Aggregation Approach and Cooperation of classifiers*. 3rd International Conference: Sciences of Electronic, Technologies of Information and Telecommunications (SETIT 2005), Tunisia.

F. Zambonelli, NR. Jenningsy, M. Wooldridge (2003). *Developing Multiagent Systems: The Gaia Methodology*. Journal ACM Transactions on Software Engineering and Methodology (TOSEM) Volume 12 Issue 3.

JO. Ojosu, LK. Komolafe (1987). *Models for estimating solar radiation availability in South Western Nigeria*. Nigerian Journal of Solar Energy, Vol.6, 69-7.

M. Benghanem, A. Mellit (2010). *Radial Basis Function Network-based prediction of global solar radiation data: Application for sizing of a stand-alone photovoltaic system Al-Madinah*. Saudi Arabia. Energy ,Vol.35, 3751-3762.

M. Occello (2003). *Methodology and architectures for the conception of multi-agent systems*. Doctoral Thesis, Joseph Fourier University, Grenoble, France.

P. Martí, M. Gasque (2011). *Improvement of temperature-based ANN models for solar radiation estimation through exogenous data assistance*. Energy Conversion and Management ,Vol.52, 990-1003.

R. Yacef, M. Benghanem, A. Melli (2012). *Prediction of daily global solar irradiation data using Bayesian neural networks: A comparative study*. Renewable Energy, vol.48, 146-154.

RH. Inman, HTC. Pedro, CFM Coimbra (2013). *Solar forecasting methods for renewable energyintegration*. Progress in Energy and Combustion Science, article in press, 1-42.

S. Kalogirou (2001). *Artificial neural networks in renewable energy systems applications: a review*. Renewable and Sustainable Energy Reviews, n°12; 5(4), 373-401.

Résumé

Dans cet article, nous présentons une nouvelle approche pour un estimateur intelligente de l'irradiation solaire basée sur système multi-agents. Cette approche est adaptative avec l'évolution des données météorologiques recueillies à partir de plusieurs sources de données en utilisant des techniques intelligentes (réseaux de neurones). Les mesures d'irradiation solaire pourraient être utilisées dans la prise de décision sur le bon choix du lieu géographique pour l'installation des équipements d'exploitation de l'énergie solaire. Un de nos principaux objectifs est de pouvoir classifier ces sites géographiques, d'où les données sont recueillent, en clusters représentant les niveaux du potentiel de la production d'énergie solaire. La partie intelligente dans notre approche est assurée par l'utilisation de l'architecture multi-agent qui permet aux agents intelligents de collaborer de manière autonome dans la fourniture de la meilleure estimation à travers des niveaux de travail superposés. Chaque niveau regroupe un ensemble d'agents spécialisés travaillant de façon autonome et communiquant avec les agents des autres niveaux sous la supervision d'un agent superviseur afin d'atteindre un but commun: fournir les meilleures données pour prendre des décisions.

Mots clés: Données météorologiques, Irradiation solaire, Système Multi-Agents, Application décisionnelle.

Un environnement sémantique à base d'agents pour la formation à distance (E-Learning)

Samir Bourekkache*, Okba Kazar*
Laid Kahloul*, Faiez Gargouri**, Aicha-Nabila Benharkat***

*Laboratoire de l'informatique intelligente, université de Biskra, Algérie
s.bourekkache@gmail.com
kazarokba@yahoo.fr
kahloul2006@yahoo.fr
**Laboratoire MIRACL
Institut supérieur d'informatique et du multimédia de Sfax
BP 3030 - 3018 Sfax TUNISIE
faiez.gargouri@fsegs.rnu.tn
***Laboratoire LIRIS, département d'informatique, Université de Lyon, France
nabila.benharkat@insa-lyon.fr

Résumé. Aujourd'hui, les établissements d'enseignement, tels que les universités, de plus en plus offrent des contenus d'E -Learning. Certains de ces cours sont utilisés avec l'enseignement traditionnel (face à face ou présentiel), tandis que d'autres sont utilisés entièrement en ligne. La création de contenu d'apprentissage est une tâche principale dans tous les environnements d'apprentissage en ligne. Les contraintes de réduire au minimum le temps nécessaire pour développer un contenu d'apprentissage, d'augmenter sa qualité scientifique et de l'adapter à de nombreuses situations, ont été un principal objectif et donc plusieurs approches et méthodes ont été proposées. En outre, les caractéristiques intellectuelles et sociales, ainsi que les styles d'apprentissage des individus, peuvent être très différents. Dans ce travail, nous développons un système collaboratif pour créer et annoter le contenu éducatif en utilisant le système multi-agents. La contribution de notre système est l'hybridation des techniques d'adaptation avec celles de la collaboration et du Web sémantique (ontologie, annotation). Nous représentons les profils des apprenants et le contenu d'apprentissage en utilisant des ontologies et des annotations pour répondre à la diversité et aux besoins individuelles des apprenants. Nous utilisons le paradigme agent, dans la phase d'implémentation de notre système, pour bénéficier des points forts de ce paradigme tels que la modularité, autonomie, flexibilité... etc.

Mots-clés : E-learning, apprentissage adaptatif, contenu éducatif, système collaboratif, Web sémantique, ontologie, système multi-agent …etc

1 Introduction

Le E-Learning doit être plus adaptable et flexible pour aider les apprenants à acquérir la connaissance. Typiquement, les systèmes d'E-learning traditionnels ignorent les fonctions de personnalisations telles que la différence dans les styles d'apprentissage, les capacités, les préférences…etc. Le fait qu'il existe un manque de connaissances sur chaque individu, le processus d'apprentissage n'est pas adapté aux besoins spécifiques des apprenants. Ainsi, ces

systèmes utilisent le même processus, les mêmes styles et les mêmes documents pédagogiques pour tous les apprenants. Afin de concevoir un système éducatif pour l'apprentissage adaptatif, nous devons permettre la livraison de contenu d'apprentissage selon plusieurs critères tel que : le niveau d'intelligence, les préférences, les besoins de l'apprenant, et les styles préférés. En outre, les développements récents des technologies du web sémantique ont montré une tendance à l'utilisation d'ontologie de promouvoir le processus éducatif et adaptatif.

Plusieurs travaux qui utilisent la technologie de web sémantique, l'annotation des contenus pédagogiques et les ontologies Tatyana (2011), Hyun-Sook (2012), Sylvain (2007), Boyce (2007), Bremgartner (2012), ont indiqué que l'utilisation de ces technologies ont beaucoup de succès dans la conception et le développement du contenu éducatifs. Elles fournissent un support d'aide pour les enseignants à créer le contenu éducatif, à accéder facilement à ce contenu et à livrer les cours de façon personnalisée Bourekkache (2009, 2014).

Dans ce travail, nous proposons une nouvelle approche pour le développement de systèmes d'E-Learning personnalisé. L'objectif est de développer les contenus éducatifs en utilisant les ontologies et l'annotation. Aussi, on exploite les documents pédagogiques annotés pour offrir un apprentissage personnalisé. La force de cette étude est de créer le modèle de l'utilisateur ontologique, modèle de contenu et le modèle pédagogique séparément pour accroître la flexibilité et la réutilisabilité du système.

Pour exploiter les ontologies réalisées dans notre approche on développe une plateforme de formation à distance qui par un nombre d'agents afin de profiter des points fort de ce paradigme tel que modularité, autonomie … etc Bourekkache (2014).

2 Fondements théoriques

2.1 Web sémantique et E-learning

L'utilisation du Web sémantique dans l'E-Learning offre un sens commun et des métadonnées traitables par les machines. Aussi, le contenu éducatif est sémantiquement annoté, cette annotation facilite la recherche de contenu adéquat dans chaque cas et la combinaison des nouveaux objets pédagogiques. Le processus d'adaptation selon les caractéristiques de l'apprenant, est basé sur les requêtes web Sémantique et la navigation à travers le contenu d'apprentissage activé par un background ontologique.

2.2 Apprentissage personnalisé

Les apprenants qui participent dans un environnement éducatif sont généralement hétérogènes c'est à dire ils ont différentes préférences, différent background, objectifs et style d'apprentissages…etc. Si on utilise la même séquence des parties de cours pour tous les apprenants, alors la qualité et le niveau de compréhension seront très faibles. Ainsi, ces apprenants ne peuvent jamais apprendre en utilisant le même contenu pédagogique et le même style d'apprentissage. Donc, l'une des exigences des systèmes d'apprentissage à distance est l'apprentissage personnalisé. Le principe de l'apprentissage personnalisé est de livrer pour chaque apprenant le document pertinent en fonction de ses préférences, besoins, objectifs et ses caractéristiques et styles d'apprentissage Bourekkache (2009), Bourekkache (2014).

2.2.1 Adaptation de contenu éducatif

Dans une formation présentielle, l'enseignant explique le cours en utilisant plusieurs exemples, détails et manières d'explications selon le niveau des étudiants. Dans notre système nous utilisons l'adaptation de cours pour fournir une plateforme qui répond aux besoins de plusieurs niveaux des apprenants (faible, moyen, fort...etc.). Nous adaptons le contenu selon les résultats des tests pour chaque apprenant. L'adaptation de contenu éducatif dans notre système est la recherche des documents de référence (RefDoc) pertinents pour chaque apprenant.

2.2.2 Styles d'apprentissage

Le domaine de styles d'apprentissage est complexe à définir et affecté par plusieurs aspects. Le style d'apprentissage peut être défini : une description des attitudes et des comportements qui déterminent la manière d'apprentissage préférée pour une personne. Felder (1988) a défini le style d'apprentissage : "les forces de caractéristiques et préférences dans la manière dont les apprenants prennent les informations et le processus d'apprentissage". Une autre définition Sabine (2007) est : "la manière complexe et les conditions dans lesquelles, les apprenants perçoivent, traitent et stockent le plus efficacement possible, et rappellent ce qu'ils tentent d'apprendre".

3 Modélisation de notre approche

D'un côté, notre approche est conçue pour répondre aux nouveaux besoins d'enseignants en termes d'assistance des travaux collaboratifs et à distance en donnant l'opportunité à un groupe d'auteurs pour la construction et l'annotation d'un contenu éducatif qui reflètent leurs points de vue et qui enrichit la connaissance sur ce contenu éducatif. D'un autre côté, notre système éducatif et adaptatif offre l'opportunité à adapter les cours et les stratégies pour assurer un apprentissage personnalisé pour chaque apprenant. Notre système contient trois acteurs qui agissent dans une plateforme de E-learning : l'enseignant, l'apprenant et l'administrateur. L'enseignant fournit les informations pour l'élaboration et l'indexation du contenu éducatif. L'apprenant acquiert le savoir et enrichit ces connaissances personnelles. Chaque apprenant doit bénéficier d'une formation personnalisée en fonction de ses caractéristiques particulières. L'administrateur d'une institution de formation assure la gestion de la formation, gestion d'auteurs, gestion d'apprenants, gestion des groupes, …etc. La figure suivante présente l'architecture de notre approche.

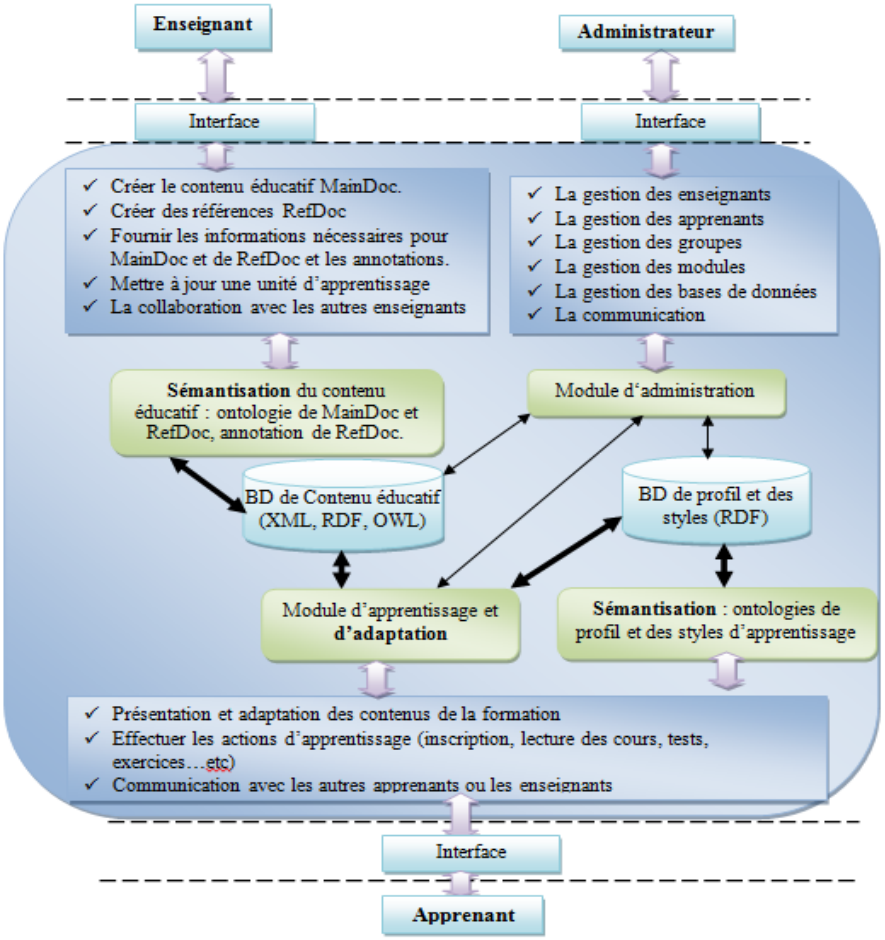


FIG. 1 – *Architecture générale de notre approches.*

3.1 Ontologie de profil

Dans les systèmes E-learning adaptatif, plusieurs problèmes sont posés. Tout d'abord, c'est la question de savoir comment créer des profils d'utilisateurs précis et complets et comment ils peuvent être utilisés pour reconnaître les apprenants et décrire ses comportements. Aussi, comment on peut construire les relations hiérarchiques entre les différentes parties du contenu éducatif et comment on peut reconnaitre la prochaine étape d'apprentissage en fonction de profil d'apprenant. Un remède possible à ces problèmes est la conception ontologique de profil de l'utilisateur et du contenu éducatif.

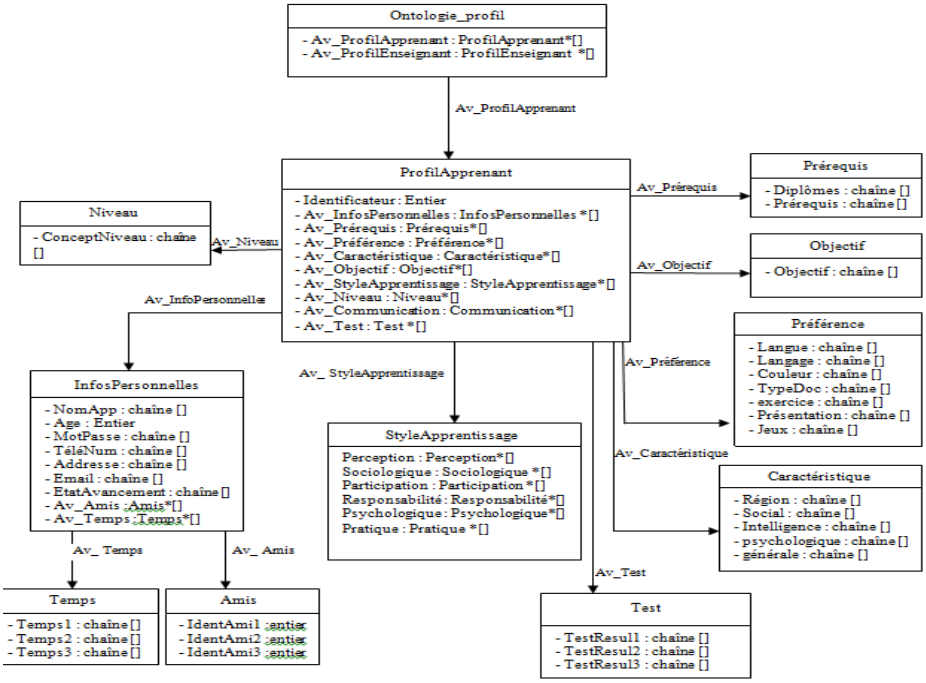


FIG. 2 – Ontologie de profil d'apprenant.

Cette ontologie contient toutes les informations nécessaires pour une vue précise sur l'apprenant et ses préférences et caractéristiques (fournies par l'apprenant et systèmes).

Exemple de liste sans numérotation :

- Identificateur : un nombre entier considéré comme une clé de chaque apprenant ;
- Prérequis : une classe qui représente le background (connaissances acquises) ;
- Niveau : est une classe qui représente les résultats de l'apprenant pour chaque concept de cours. Test : contient tous les tests et leurs résultats ;
- La classe InfosPersonnelles : détermine toutes les informations personnelles de l'apprenant (nom, prénom, âge, téléphone, email, adresse, état d'avancement). Cette classe contient : la class Temps et la classe Amis.
- Préférence : exprime toutes les préférences de l'apprenant tel que : (couleur, langue, langage, préférer les exercices ou non, préférer les jeux ou non ...etc).
- Caractéristique : qui contient les caractéristiques sociales, psychologique, …etc).
- StyleApprentissage : contient plusieurs sous classes qui présentent les stratégies ou les styles préférés pour chaque apprenant (on les utilise pour l'adaptation de stratégies).

3.2 Contenu éducatif

On construit le contenu éducatif en utilisant la technologie du Web sémantique afin d'organiser ce contenu et de facilité la recherche de document pertinent. Un groupe d'enseignants essayent de rédiger et annoter le contenu pédagogique, et former les relations sémantiques entre ces documents. Puis, on exploite ce contenu annoté sémantiquement pour

guider l'apprenant à mieux comprendre l'objet éducatif en utilisant les techniques d'adaptation. On a deux types de documents éducatifs :

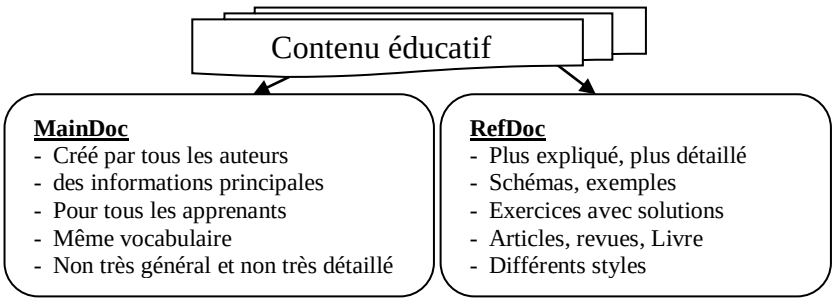


FIG. 3 – Le contenu éducatif.

Le contenu éducatif est un élément critique dans E-Learning puisque les apprenants vont apprendre et manipuler ce contenu éducatif et il aide les enseignants peuvent transférer leurs connaissances aux apprenants.

- MainDoc : est le document principal, comme le cours donné par l'enseignant. Ce contenu pédagogique est créé par tous les auteurs de même cours. Il contient les informations nécessaires pour comprendre les concepts de contenu éducatif.
- RefDoc : est le document de référence utilisé dans des cas particuliers (apprenant échouant, excellent, ou une requête de recherche). Les enseignants fournissent un ensemble de références : pour plus d'explications, plus de détails, exercices avec solutions, livre, thèse, schémas explicatifs, …etc.

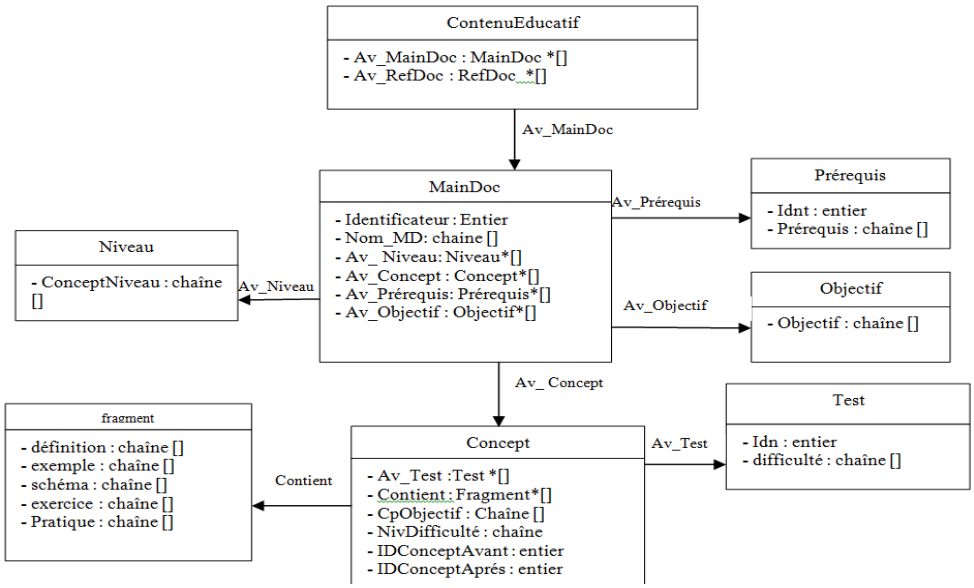


FIG. 4 – Ontologie de MainDoc.

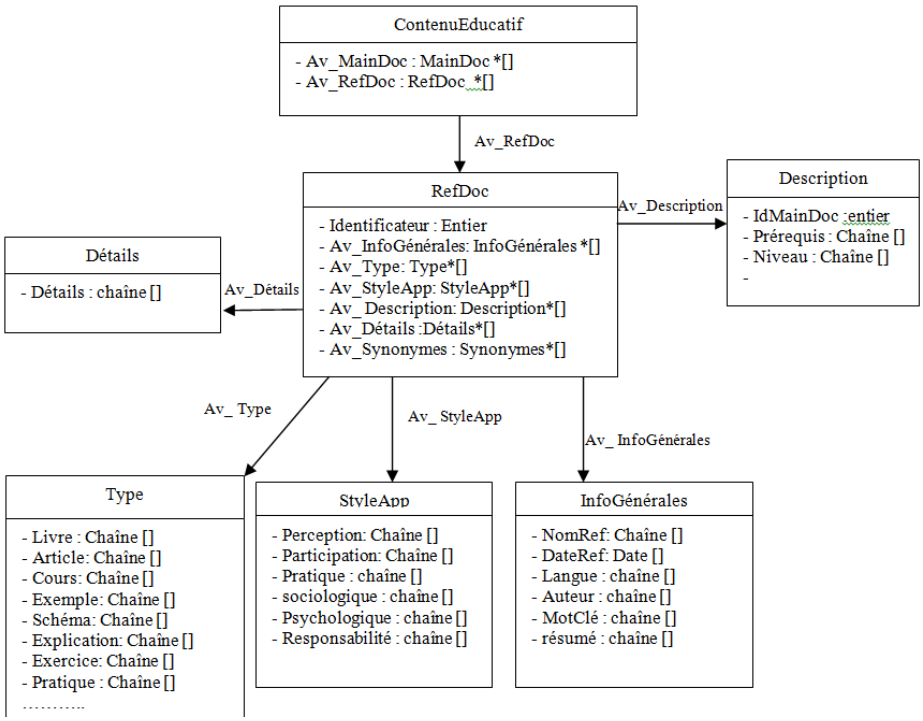


FIG. 5 – Ontologie de RefDoc.

RefDoc est le deuxième type de contenu d'apprentissage, il peut être des exemples explicatifs, des schémas explicatifs, documents contiennent des explications détaillées, des résumés, livres, articles ...etc. Nous allons utiliser la principale annotation pour le RefDoc (en utilisant LOM) : Le titre, mots clés, les mots les plus répétés dans le document et le degré de pertinence. Aussi, l'enseignant doit fournir des informations supplémentaires sur le Refdoc pour compléter l'annotation de Refdoc, par exemple:

- Av_type : type de RefDoc (exemple, schéma, livre, …etc),
- Av_format : le format de RefDoc.
- Av_Style : style d'apprentissage (textuel, auditif, ou visuel, Collaboratif, compétitif, global, séquentiel, détectif, intuitif, dépendant, indépendant …etc).
- Exigé_par : important pour la compréhension de quel MainDoc ou RefDoc.
- Patie_de : partie de quel RefDoc.
- Réference_de : référence pour quel MainDoc, Av_version : la version de RefDoc.
- Av_prérequis : des explications des prérequis,
- Est_exercice : Des exercices avec solutions, Exige : Exige la lecture de quel Doc.
- Niv_détail : détaillée ou non, contient des informations supplémentaires ou non,
- Av_niveau : Orientés vers les apprenants excellents ou à ceux qui ont échoué, quiz …

Ces annotations sont réalisées par Protégé 4.3 en utilisant SubClass pour la hiérarchie des classes et AnnotationProperties pour les propriétés.

4 Fonctionnement du notre système

Dans cette section, nous présentons le fonctionnement de notre approche pour atteindre l'objectif de créer et d'annoter les documents éducatif.

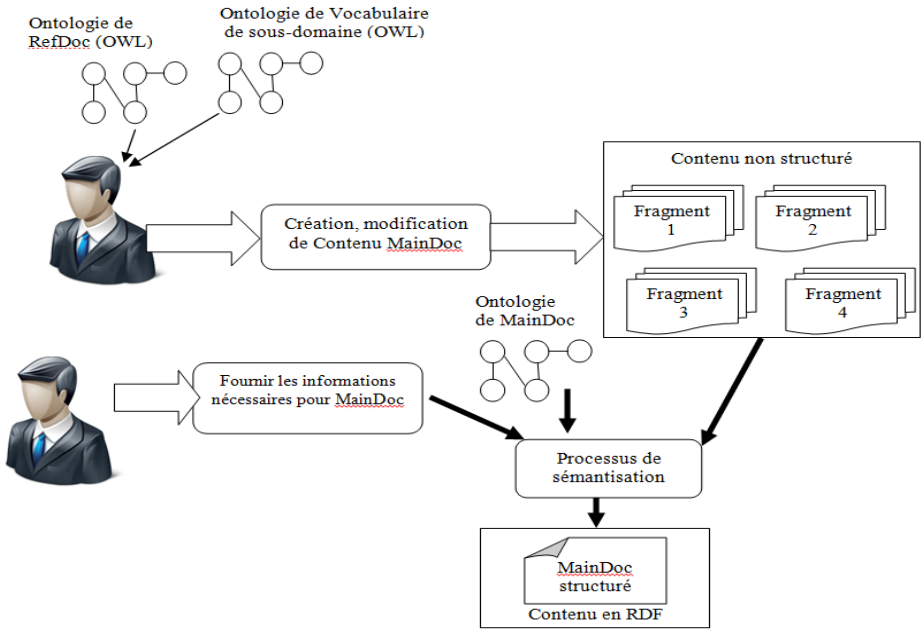


FIG. 5 – Le processus de développement du MainDoc.

Le processus de création du MainDoc Commence par la rédaction des fragments (concepts) de MainDoc par l'enseignant. Le vocabulaire utilisé dans le concept créé doit suivre l'ontologie de vocabulaire. Ensuite, l'enseignant utilise l'interface de notre système pour donner les concepts et les informations de chaque concept selon l'ontologie de MainDoc. Le processus de sémantisation (structuration de MainDoc et ses informations générales) est démarré. Comme résultat, nous avons un document structuré en utilisant RDF pour la description du MainDoc.

La figure suivante explique la démarche pour créer et annoter le RefDoc en utilisant notre approche.

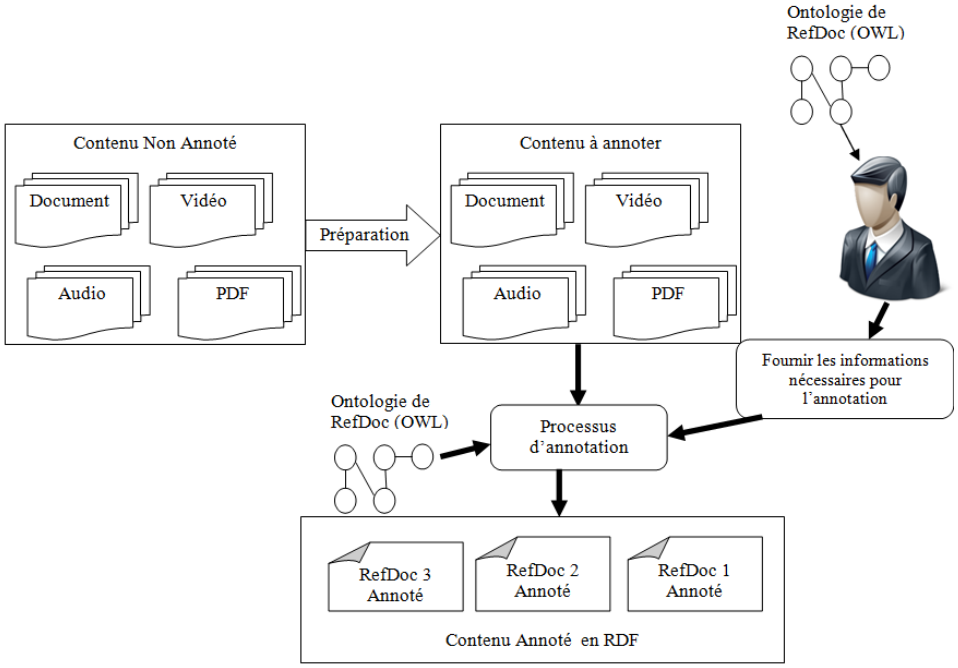


FIG. 6 – Le processus de développement du RefDoc.

La figure précédente explique le processus d'annotation des ressources d'apprentissage. Ce processus représente la partie de sémantisation du contenu éducatif. Pour cela, nous avons trois étapes : 1) la préparation de RefDoc à annoter c'est à dire la préparation des descriptions et des informations utilisées pour l'annotation. L'enseignant peut utiliser tous les types de documents existants : PDF, DOC, AVI, MP3, …etc, aussi il peut prendre une partie de document (partie de livre, article, thèse …etc) comme il peut créer (écrire) un RefDoc (explication, exercices, …etc). L'enseignant doit avoir une vue détaillée sur le document (titre, sous titres, niveau, contenu, l'auteur, version, date …etc). 2) En se basant sur l'ontologie de RefDoc et les annotations utilisées, l'enseignant fournit, à travers l'agent interface, toutes les informations et les descriptions nécessaires pour l'annotation de RefDoc. 3) Notre système exploite ces informations fournies par l'enseignant et l'ontologie de RefDoc pour le stockage de RefDoc avec ses annotations en RDF qui référence (décrit) le RefDoc.

Voici par exemple une déclaration RDF qui décrit une ressource (schedule.pdf) par les propriétés : NameRefDoc, TitleRefDoc, et StylePerception ; et la relation PartOf

```
<rdf : Description rdf : about="&RefDoc;schedule.pdf ">
        <NameRefDoc> schedule </NameRefDoc>
         <TitleRefDoc> schedule _definition</TitleRefDoc>
         <StylePerception> visual</StylePerception>
       < PartOf rdf : resource="&RefDoc;OperatingSystem.pdf "/>
</rdf : Description>
```

Le Refdoc annoté est utilisé dans des cas particuliers (Pour enrichir les règles de l'adaptation il suffit d'ajouter d'autres règles dans la base de règles. Donc, on présente quelques règles:

- L'apprenant échoue (moins de 50%): ici nous cherchons RefDoc qui a des annotations: des exemples, des schémas et des exercices avec solutions, utiliser le style auditif.
- L'apprenant échoue (moins de 25%): nous cherchons RefDoc qui contient des annotations: plus de détails, explication simple, utiliser le style auditif et visuel... etc. S'il ne suit pas les conseils des enseignants et du système nous cherchons des RefDoc pour les apprenants indépendants (style indépendant).
- L'apprenant échoue (moins de 10%): nous cherchons Refdoc qui a des annotations: les explications des prérequis, quiz, utiliser le style active. S'il fait des communications avec ses amis : on utilise des RefDoc qui exigent un travail collectif (style collaboratif).
- Pour les excellents apprenants : nous cherchons Refdoc qui a des annotations: plus d'informations, résumés, articles, …etc.
- Quand un apprenant fait une recherche en utilisant des mots-clés qui existent dans l'ontologie cours et leurs synonymes dans le processus de recherche.
- Un apprenant qui pose plusieurs questions : nous cherchons Refdoc qui a des annotations: plus de détails, utiliser le style collaboratif.
- Un apprenant qui a échoué et qui manipule plusieurs activités et exercices (pratique) : on cherche les documents de référence qui ont le style détectif et on change le style de son profil par ce nouveau style.
- Si un apprenant a échoué, et il participe dans forum : on change son profil par le style actif et on cherche des RefDoc qui ont l'annotation pour Actifs.
- Si un apprenant échoue et qui a le style séquentiel dans le profil, on essaye de le changer au style global (cherchons des RefDoc qui ont comme annotation style global).
- Dans toutes les situations précédentes (particulièrement dans le deuxième échec dans le même concept) : on utilise des questionnaires qui contiennent plusieurs questions pour la détection automatique de style pertinent pour chaque apprenant (questions psychologiques et sociologiques, question sur le style préféré …etc).

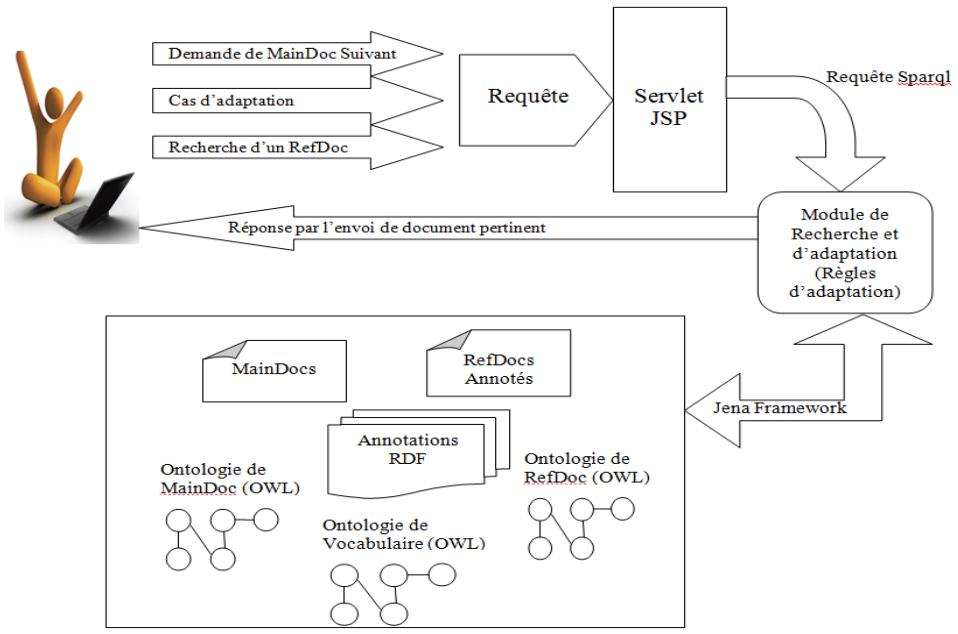


FIG. 6 – Le processus d'adaptation et fournir un contenu éducatif.

L'apprenant demande un cours (concept suivant du MainDoc), ou il demande une opération de recherche, ou il est dans un cas qui exige une adaptation de contenu ou de style d'apprentissage. Donc, Une requête va être transmise via servlet (JSP) vers l'agent Gestion cours (dans le cas de demande d'un concept de MainDoc), ou vers l'agent adaptation (dans les deux autres cas). Ce dernier utilise Framework Jena pour extraire ou décrire dedans la base de données cours en RDF (en utilisant Sparql pour le langage de requête pour RDF). En appliquant les règles d'adaptation on forme une requête adéquate pour obtenir le RefDoc pertinent dans chaque cas. Dans le cas de recherche, on utilise les mots donnés et leurs synonymes pour trouver les documents pertinents.

5 Conclusion

Nous avons exploité la technologie de Web sémantique pour améliorer l'apprentissage des apprenants. La représentation sémantique des contenus pédagogiques, en utilisant l'ontologie cela facilite la recherche et la réutilisation des documents éducatifs. Ainsi, elle aide la machine à comprendre et manipule les documents d'apprentissage. Les apprenants ont des caractéristiques hétérogènes, donc il faut assurer un apprentissage personnalisé pour chaque apprenant. L'adaptation de contenu et de stratégies (style d'apprentissages) s'avère très pertinente pour cet objectif. Finalement nous avons présenté une architecture pour la formation à distance. Notre approche assure la création et l'annotation des documents éducatifs, en utilisant la technologie de Web sémantique, et le bon guidage de l'apprenant durant sa formation et essaie de satisfaire les besoin des apprenants selon les niveaux et selon ses préférences.

Références

Bourekkache S., O. Kazar (2009). Agent-Based Approach for E-Learning". *International Journal of Emerging Technologies in Learning (iJET)*, Vol 4, No 4.

Bourekkache S., O. Kazar , N. Benharkat, L. Kahloul (2014). A cooperative multi-agent approach for the creation and annotation of adaptive content for e-learning. *Journal of e-Learning and Knowledge Society (Je-LKS), Vol 10, No 1.*

Boyce S., C. Pahl (2007). Developping domain ontologies for course content. *Educational Technology & Society*, vol. 10, pp. 275-288.

Bremgartner V., J. F. de Magalhães Netto (2012). Improving Collaborative Learning by Personalization in Virtual Learning Environments Using Agents and Competency- Based Ontology. *Frontiers in Education Conference (FIE), IEEE.*

Felder R. M., L. K. Silverman 1988). Learning and Teaching Styles in Engineering Education. *Engineering Education*, 78 (7), 674–681.

Hyun-Sook Ch., K. Jung-Min (2012). *Ontology Design for Creating Adaptive Learning Path in e-Learning Environment.* Proceedings of the International Multi-Conference of Engineers and Computer Scientists, Hong Kong.

Sabine G., (2007). *Adaptivity in Learning Management Systems Focussing on Learning Styles*, Thèse Pour l'obtention du titre de Docteur de l'université de technologie, Vienna.

Sylvain D., (2007) *Exploiting Semantic Web and Knowledge Management Technologies for E-learning.* thèse pour obtenir le titre de Docteur en Sciences, université de Nice-Sophia Antipolis.

Tatyana I., (2011). *Adaptive Open Corpus E-Learning and Authoring, Using Collaborative Ontology Learning*", 9th IEEE International Conference on Emerging eLearning Technologies and Applications, Stará Lesná, Slovaquie.

Summary

Nowadays, educational institutions, such as universities, more and more offer E-Learning contents. The creation of learning content is a main task in every E-learning environment. The constraints of minimizing the time required for developing a learning content, for increasing its scientific quality and to adapt according to the preferences and the individual needs of the learners have been a principal aim. We develop a collaborative system to create and annotate educational materials. The contribution of our system is the hybridization of adaptation techniques with those of collaboration and Semantic Web (ontology, annotation).

Médiation Sémantique dans MedPeer : Un Système d'Intégration de Sources de Données Hétérogènes Basé sur les Ontologies

Naïma Souâd Ougouti*, Haféda Belbachir*, Youssef Amghar**

*University des Sciences et de la Technologie d'Oran-Mohamed Boudiaf (USTO-MB)
Souad.ougouti@univ-usto.dz; s_ougouti@yahoo.fr; h_belbach@yahoo.fr;
**LIRIS UMR 5205, Insa of Lyon,
youssef.amghar@insa-lyon.fr

Résumé. Avec l'avènement du web sémantique et dans le but d'un partage efficace des sources d'information présentes sur le Web, de nouvelles possibilités d'intégration de données multi-sources voient le jour. Le processus de médiation sémantique est devenu une tâche incontournable dans cette nouvelle génération de systèmes qui utilisent des ontologies. Nous présentons dans cet article, les mesures de similarités utilisées pour trouver les correspondances entre les ontologies locales représentant les sources de données présentes au niveau des pairs et l'ontologie globale de domaine présente au niveau des super-pairs qui forment MedPeer: notre nouveau système d'intégration de sources de données hétérogènes et distribuées dans un environnement P2P.
Mots-clés: Web sémantique – Ontologies –Similarité sémantique – Voisinage.

1. Introduction

Nous nous intéressons dans cet article au problème de la médiation sémantique dans MedPeer, notre système d'intégration de données hétérogènes dans un environnement P2P, Ougouti et al. (2011). Nous présentons donc une mesure de similarité globale entre les concepts d'une ontologie de domaine et ceux des ontologies locales des sources de données présentes sur les pairs du système. Dans ce travail, nous faisons l'hypothèse de pairs contenant des bases de données relationnelles. En première étape, des ontologies locales doivent être générées pour décrire les schémas de ces bases de données, ceci est réalisé par le biais d'une nouvelle méthode que nous avons proposée nommée Relationnel.OWL2E qui permet à partir d'un schéma relationnel de générer automatiquement son ontologie correspondante basée sur le langage OWL2, Ougouti et al. (2013, 2015). En deuxième étape, le processus de médiation sémantique peut commencer, en comparant les concepts des deux ontologies dans le but de trouver des correspondances, qui seront stockées et utilisées dans le routage sémantique et la réécriture des requêtes.

Cet article est organisé comme suit : dans la section 2, nous présentons un état de l'art des principales approches d'alignement d'ontologies. En section 3, nous introduisons la mesure de similarité globale utilisée dans le cadre de notre travail. Dans la section 4 nous mettons l'accent sur la similarité de voisinage qui constitue la contribution essentielle de ce travail. La section 5 est consacrée à la présentation et à la discussion des résultats. Enfin nous terminons par une conclusion.

2. Etat de l'art

Pour pouvoir comparer un ensemble de concepts, il est nécessaire de disposer d'une mesure de similarité qui permet de trancher sur la similitude ou la dissimilitude de ces concepts. Dans le domaine de l'alignement des ontologies, plusieurs travaux sur l'état de l'art ont été proposés, on peut notamment citer ceux de Shvaiko, et Euzenat (2013), Bernstein et al. (2011) et Rahm (2011). Certains travaux sont basés sur une mesure de similarité globale qui est une somme pondérée de plusieurs caractéristiques tels que Les éléments lexicaux, les relations structurelles, la structure interne, les relations sémantiques et enfin les extensions (instances de classes et valeurs des propriétés). Parmi ces approches d'alignement d'ontologies, nous citerons par exemple les systèmes : OLA (OWL Lite Alignment) Euzenat et al. (2004), ASMOV (Automated Semantic Mapping of Ontologies with Validation), Jean-Mary et Kabuka (2007), H-MATCH, Castano et al. (2003), COMA++ (COmbining MAtching) et Aumueller, et al. (2005). D'autres systèmes plus récents correspondent à HurTUDA, Hertling (2012), LogMap, Jimenez-Ruiz et Cuenca Grau (2011), LYAM++, Tigrine et al. (2015), et enfin S-Match1, Giunchiglia (2012).

Enfin, pour plus d'informations, il existe une campagne annuelle d'évaluation des outils d'alignement, appelée L'OAEI (The Ontology Alignment Evaluation Initiative) qui permet de comparer les résultats obtenus par les méthodes d'alignement participantes sur différents jeux d'ontologies, et dont le dernier rapport est disponible dans Cheatham et al. (2015).

3. Médiation sémantique

Pour calculer la similarité globale entre deux concepts, nous avons fait le choix de nous baser sur la mesure de similarité introduite dans Senpeer Faye (2007) qui elle-même est basée sur d'autres méthodes telles que Rodriguez et al. (2003) et Castano et al. (2003).

La similarité que nous avons choisie d'utiliser repose sur une méthode qui combine plusieurs techniques d'appariement et ou le score global est une somme pondérée des scores d'appariement partiels, elle est basée sur des techniques linguistiques et structurelles.

Etant donné deux ontologies Od (ontologie de domaine) et Ol (ontologie locale), l'alignement de ces deux ontologies consiste à trouver $|Od|\times|Ol|$ éléments de correspondances <IDij, ci(od),cj(ol),γij>, avec IDij identifiant unique de la correspondance, le concept ci(od) $\in$ Od , le concept cj(ol) $\in$ Ol et γij le degré de similarité entre les deux concepts. L'affinité sémantique entre les concepts est établie si leur similarité est supérieure à un seuil minimum de similarité. Notons aussi que la qualité de l'appariement est évaluée dans l'intervalle [0,1] par souci de normalisation, et que nous ne considérons que les correspondances sémantiques exactes et directes (équivalence).

La similarité entre deux concepts est fonction de :

- Leur similarité linguistique (similarité de leurs ensembles de synonymes, des textes les décrivant (les étiquettes) et de leurs types).
- La similarité de leurs voisinages sémantiques.

La similarité globale entre deux concepts (ci(od),cj(ol)) est donc calculée comme suit :

$$Sim_g(ci(od),cj(ol)) = \lambda . Sim_l(ci(od),cj(ol)) + (1-\lambda) Sim_v(ci(od),cj(ol)) \quad (1)$$

Avec Sim_l : la similarité linguistique des deux concepts, Sim_v la similarité de leur voisinage sémantique. Ces deux mesures sont calculées sur la base d'un autre type de similarité qu'est la similarité lexicale entre deux concepts et que nous introduisons dans la prochaine section.

3.1 Similarité lexicale

L'appariement entre deux termes A et B est calculé en utilisant la similarité lexicale SL(A,B) proposée par Maedche et Staab (2002), elle-même basée sur la distance dl de Levenshtein (1966) qui est égale au nombre minimal de caractères qu'il faut supprimer, insérer ou remplacer pour passer d'une chaîne à l'autre. SL(A,B) est donnée par la formule suivante :

$$SL(A,B) = max(0, \frac{min(|A|,|B|) - dl(A,B)}{min(|A|,|B|)}) \in [0,1]. \qquad (2)$$

3.2 Similarité linguistique

La similarité linguistique de deux concepts est une somme pondérée de la similarité de leurs deux ensembles de synonymes $Sim_{syn}(ci(od),cj(ol))$, de la similarité de leurs types $Sim_{type}(ci(od),cj(ol))$ et de la similarité des commentaires $Sim_{com}(ci(od),cj(ol))$

$$Sim_l(ci(od),cj(ol)) = \omega . Sim_{syn}(ci(od),cj(ol)) + \alpha . Sim_{type}(ci(od),cj(ol)) + \mu . Sim_{com}(ci(od),cj(ol)) \qquad (3)$$

avec $\omega, \alpha, \mu \geq 0$ *et* $\omega + \alpha + \mu = 1$

3.2.1 Similarité des types

Elle est basée sur la similarité des types fournit par le système Cupid Madhavan et al. (2001) qui utilise une table fournissant des coefficients de similarité dans l'intervalle [0, 1] entre des pairs de types de données.

3.2.2 Similarité des synonymes

Pour calculer la similarité entre deux ensembles de synonymes on se base sur la mesure de Tversky (1977) qui calcule la similarité de deux objets en comparant leurs caractéristiques communes et distinctives : plus les objets partagent des caractéristiques, et moins ils ont de caractéristiques distinctives, plus ils sont similaires.

$$Sim_{syn}(c_i(od), c_j(ol)) = \frac{|syn(c_i(od)) \cap syn(cj(ol))|}{|syn(c_i(od)) \cap syn(cj(ol))| + \alpha|syn(c_i(od)) \setminus syn(cj(ol))| + (1-\alpha)\,|\,syn(\,cj(ol)) \setminus syn(c_i(od))|} \qquad (4)$$

Avec :

- syn(ci(od)) : ensemble des synonymes du concept i de l'ontologie de domaine.
- syn(cj(ol)) : ensemble des synonymes du concept j de l'ontologie locale.
- $0 \leq \alpha \leq 1$, « ∩ » représente l'intersection et « \ » représente la différence.

3.2.3 Similarité des étiquettes et des commentaires

La propriété d'annotation *rdfs :comment* permet de fournir une description textuelle de la ressource. Pour calculer la similarité des textes contenus dans les commentaires nous utilisons la fonction cosinus qui utilise la représentation vectorielle complète (fréquence des mots) et qui quantifie la similarité entre deux vecteurs comme le cosinus de l'angle entre eux. Quand les commentaires sont identiques, l'angle entre les vecteurs est nul et le cosinus vaut 1 et à l'opposé des commentaires entièrement différents sont représentés par des vecteurs orthogonaux donc leur similarité est nulle.

$$\text{similarité}(\mathbf{A},\mathbf{B}) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} {A_i}^2} \times \sqrt{\sum_{i=1}^{n} {B_i}^2}} \qquad (5)$$

Puisque les poids ne peuvent pas être négatifs, alors on aura $0 \leq$ similarité(A,B) ≤ 1.

4. La similarité du voisinage sémantique

Le contexte d'un concept est très important dans le calcul de la similarité sémantique, nous partons de l'intuition que deux concepts sont similaires s'ils sont liés à d'autres concepts eux même similaires. Nous considérons le voisinage sémantique d'un concept i noté V(ci) comme étant l'ensemble des concepts ayant un lien sémantique direct ou indirect avec ce concept.

Soit G le graphe correspondant à l'ontologie O, G={N, R} ou N est l'ensemble des nœuds (concepts) de ce graphe et R l'ensemble des arcs (liens sémantiques) entre ces concepts. Nous introduisons les notions suivantes :

- Nœud(Ci) : le nœud correspondant au concept ci.
- Père(n) : le nœud père de n
- Fils(n) : Le nœud fils de n

Soit n1=nœud(c1) et n2=nœud(c2), r=le lien entre n1 et n2, les opérations sont formulées comme suit :

- Arrivant_à : n2 ϵ arrivant_à (n1; r) ssi n2 = Père(n1) et l[(n2; n1)] = r:
- Sortant_de : n2 ϵ sortant_de(n1; r) ssi n2 = fils(n1) ^ l[(n1; n2)] = r:

Nous définissons le voisinage d'ordre 1 (liens directs) du nœud ni comme suit :

V1(ni)={nj/ nj ϵ arrivant_à (ni; r1) ou nj ϵ sortant_de (ni; r2)}

Dans notre travail nous considérons le voisinage indirect d'ordre *n*, définit comme suit :

V(ni)= V1(ni) ∪ ***V2(ni)*** ∪ ***V3(ni)*** ∪ **…..**∪ ***Vn(ni)***

$V_k(n_i)$ contient les voisins directs des concepts présents dans $V_{k-1}(n_i)$:

$$\forall\, C_i \in V_{k-1}(n_i)\ \ V_k(n_i) = \cup V1(C_i)$$

Ce voisinage peut être vu comme l'ensemble des concepts qui se trouvent à un rayon *n* du concept n_i dans le graphe. Ce rayon sera défini selon les besoins (expérimentation) et selon la taille du graphe des ontologies. Pour un rayon assez grand, les voisinages peuvent différer de manière significative et la similarité des voisinages sera faible ; à l'opposé, un rayon faible augmente la probabilité d'avoir un voisinage contenant plus d'éléments commun.

La similarité de voisinage entre deux concepts est le rapport entre le nombre de correspondances (similarités linguistiques) entre les deux ensembles de voisinage et le nombre total des concepts des deux voisinages. Elle est définie comme suit :

$$Sim_v(c_i(od), c_j(ol)) = \frac{|\{x \in V(c_i(od))/\exists y \in V(c_j(ol)) \wedge Sim_l(x,y) > \varepsilon_{acc}\}|}{|V(c_i(od)) \cup V(c_j(ol))|} \qquad (6)$$

5. Présentation des résultats et Discussion

Pour évaluer et mesurer la qualité des résultats de la fonction de similarité introduite dans la section précédente, nous utilisons les mesures utilisées dans la recherche d'information à savoir *le rappel* , *la précision*, leur moyenne harmonique la *F-mesure* et la

métrique *Global*. Puisque notre méthode de recherche des correspondances est une extension de la méthode utilisée dans Senpeer, nous avons essayé de savoir si les améliorations proposées ont été efficaces. Pour cela, nous avons comparé nos résultats avec ceux de Senpeer pour deux jeux de tests différents :

Le premier jeu concerne l'exemple utilisé par Senpeer. En effet, nous avons voulu savoir, si notre méthode donnait de bons résultats comparés à ceux de Senpeer. Les résultats obtenus sont les suivants :

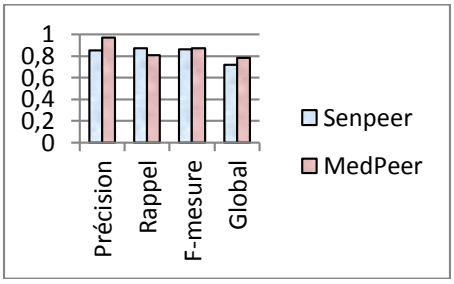


FIG. 1 - *Comparaison entre SenPeer et MedPeer par métrique*

Ces différents résultats montrent que notre approche se comporte bien et qu'elle est aussi efficace que Senpeer pour cet exemple. Rappelons que Senpeer ne considère que le voisinage d'ordre 1, c'est-à-dire que les liens directs.

Le deuxième jeu concerne deux exemples d'ontologies locales correspondantes aux bases de données suivantes :

Base de données 1 : *Bdmed1*

Patient(numordre, nom, prenom, datenaiss, lieunaiss, profession, adresse, NSS)
Diagnostic (numordre, code_diag, codemed, date_maladie, observation)

Base de données 2 : *Bdmed2*

Malade (code,nom_malade,prenom_malade,datenaissance,ville_naissance,N_rue, rue,codepostal, ville, NSS)
Dossier-medical (code,id_dossier,date-creation)
Maladie (code,id_dossier,code_diag,date_diag,observation)

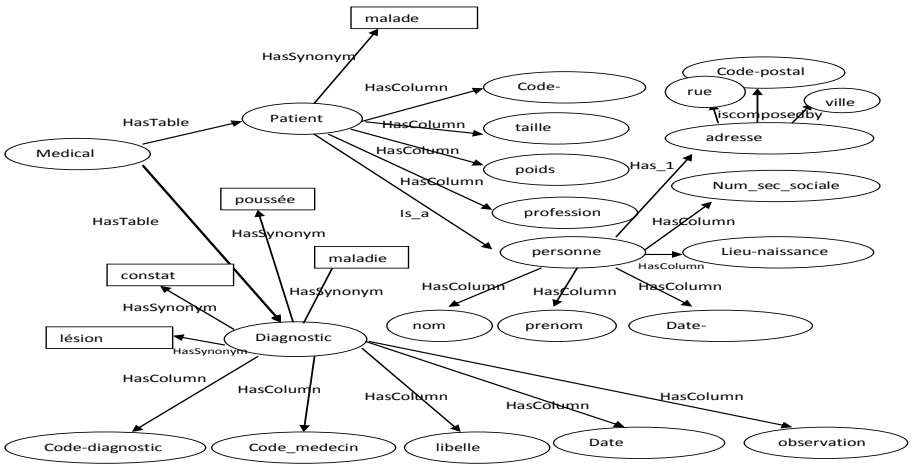


FIG. 2– *Une Partie de l'ontologie de domaine*

Dans cet exemple nous allons montrer que la méthode de SenPeer ne se comporte pas bien puisque le taux des réponses *False Négative* sera très élevé à cause de certains liens non pris en compte dans le voisinage. En effet, la méthode SenPeer ne prend en compte dans le voisinage que les liens père et fils.

En réalisant les tests sur les ontologies locales bdmed1 , bdmed2 et l'ontologie de domaine (FIG. 2) les résultats ont été peu concluants au regard des valeurs des métriques retournées et données par le graphique suivant :

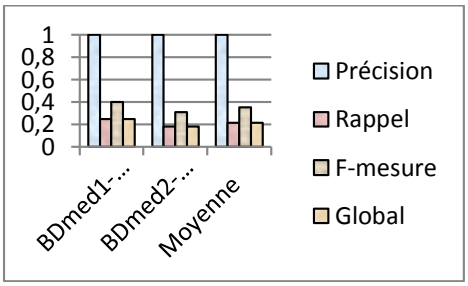


FIG. 3 - *Résultats de l'application de SenPeer sur les ontologies du domaine Médical.*

Nous avons appliqué notre méthode sur les mêmes ontologies. Nous sommes partis au début de notre intuition que l'exploration de tous les liens sémantiques pour enrichir le voisinage donnerait de meilleurs résultats. Nous avons remarqué que pour certaines correspondances, ceci s'est révélé très intéressant par contre pour d'autres correspondances qui étaient déjà établies, la similarité n'était plus établie. Nous avons fait la constatation suivante : Il y a certains cas ou en élargissant le voisinage à plusieurs liens (grand rayon), le nombre de voisins augmentent et la similarité de voisinage chute. Exemple de « patient.numordre » de *bdmed1* avec « patient.codepatient » de l'ontologie de domaine, la similarité de voisinage était à 0,5 et la similarité globale établie alors qu'en élargissant le voisinage la similarité de voisinage a chuté à 0,07 ce qui a conduit à un résultat *False Négative.* En faisant ce constat, nous avons décidé de ne pas explorer d'emblée tous les liens dés le début mais de calculer à chaque niveau d'exploration la similarité de voisinage, puis de prendre la valeur maximale trouvée. Ceci nous a permis de personnaliser le calcul de la similarité de voisinage à chaque cas de similarité. Certaines sont établies uniquement en explorant les liens *is-a*, d'autres ont eu besoin de pousser la recherche en élargissant le rayon.

Nous remarquons une nette amélioration des indices de *Précision, rappel, F-mesure* et *Global*, il y a beaucoup plus de similarités retournées par rapport à la méthode de SenPeer. Comme le montre l'histogramme suivant :

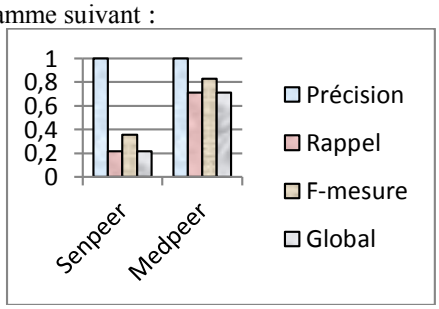


FIG. 4 - *Histogramme comparatif entre notre approche Medpeer et celle de Senpeer*

6. Conclusion

Nous avons présenté dans cet article une technique de découverte des correspondances sémantiques entre les concepts de différentes ontologies implémentées au sein de notre nouveau système d'intégration de données hétérogènes nommé MedPeer. Les modèles combinant plusieurs aspects des concepts (propriétés, position dans la hiérarchie, contexte) ont pour avantage d'être complets et de prendre en compte le maximum d'information contenue dans l'ontologie, c'est pour cette raison que la mesure de similarité globale que nous avons introduite est basée sur plusieurs mesures dont la plus importante est la mesure du voisinage sémantique qui doit explorer un certain nombre de liens sémantiques. Ceci est facilité par le choix d'utiliser des ontologies comme format de description des sources locales, car il exploite toute la richesse des relations sémantiques offertes par celles-ci. Le modèle de similarité sémantique proposé constitue donc une amélioration générale d'un modèle de similarité sémantique pouvant être adopté dans d'autres domaines.

Références

Aumueller, D., Do, H. H., Massmann, S., Rahm, E. (2005). *Schema and Ontology Matching with COMA++*. Proc. SIGMOD Conf., ACM Press, 906–908.

Bernstein P.A., Madhavan J., Rahm E. (2011*). Generic schema matching, ten years later*. In PVLDB 4(11), 695-701.

Castano, S., Ferrara, A., Montanelli, S. (2003). *H-match: an algorithm for dynamically matching ontologies in peer-based systems.* Proceedings of the 1st VLDB International Workshop on Semantic Web and Databases (SWDB 2003) Berlin, Germany, 231–250.

Cheatham M., Dragisic Z., Euzenat J., Faria D., Ferrara A., Flouris G., Fundulaki I., Granada R., Ivanova V., Jimenez-Ruiz E., Lambrix P., Montanelli S., Pesquita C., Saveta T., Shvaiko P., Solimando A., Trojahn C., Zamazal O. (2015). *Results of the Ontology Alignment Evaluation Initiative 2015*. In Proc. 20th ISWC ontology matching workshop (OM), Boston (MA US), 60-115.

Euzenat, J., Loup, D. , Touzani, M., Valtchev, P. (2004). *Ontology alignment with OLA.* Proceedings of Third International Semantic Web Conference 2004.

Faye, D. C. (2007). *Médiation de données sémantique dans SenPeer, un système pair-à-pair de gestion de données*. PHD thesis, Nantes University.

Giunchiglia F., Autayeu A., Pane J. (2012). *S-Match: An Open Source Framework for Matching Lightweight Ontologies*. In Semantic Web J., vol. 3, no. 3, 307—317.

Hertling S. (2012). *Hertuda: Results for OEAI 2012*. In Seventh International Workshop on Ontology Matching.

Jean-Mary, Y., Kabuka, M. R. (2007). *ASMOV: Ontology Alignment with Semantic Validation.* SWDB-ODBIS Workshop, Vienna, Austria, 15-20.

Jimenez-Ruiz E. and Cuenca Grau B. (2011). *Logmap: Logic-based and scalable on-ontology matching*. In The Semantic Web–ISWC 2011, Springer, 273–288.,

Levenshtein, V. (1966). *Binary codes capable of correcting deletions, insertions, and reversals.* Soviet Physics Doklady, 10(8), 707-710

Madhavan J., Bernstein P. A. et Rahm E. (2001). Generic schema matching with cupid. In VLDB ’01: Proceedings of the 27th International Conference on Very Large Data Bases, pages 49–58, San Francisco, CA, USA.

Maedche A. and Staab S. (2002). *Measuring similarity between ontologies*. In EKAW ’02: Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, pages 251–263.

Ougouti, N. S., Belbachir, H., Amghar, Y., Benharkat, N. (2011). *Architecture Of MedPeer:* A New P2P-based System for Integration of Heterogeneous Data Sources. Proceedings of the International Conference on Knowledge Management and Information Sharing (KMIS), Paris, 351-354.

Ougouti, N. S., Belbachir, H., Some F., Ouattara I. (2013). *Relational.OWL2E: Une nouvelle approche de représentation du schéma d’une base de données relationnelle basée sur OWL*. International Journal of Information Technology and Computer Science, 48-53.

Ougouti, N. S., Belbachir, H., Amghar, Y. (2015). *A New OWL2 Based Approach for Relational Database Description*. International Journal of Information Technology and Computer Science, 48-53.

Rahm E. (2011). *Towards large-scale schema and ontology matching*. In: Schema Matching and Mapping, 3-27.

Rodriguez, A. M., Egenhofer, M. J. (2003). *Determining semantic similarity among entity classes from different ontologies*. IEEE Transactions on Knowledge and Data Engineering, 15(2), 442–456.

Shvaiko P. and Euzenat J. (2013). Ontology matching : state of the art and future challenges. In IEEE Trans. Knowl. Data Eng. 25(1), 158-176.

Tigrine A., Bellahsene Z., Todorov K. (2015). *LYAM++ Results for OAEI 2015*. In Proc. 20th ISWC ontology matching workshop (OM), Boston (MA US), 176-180.

Tversky A. (1977). *Features of similarity*. Psychological Review, (84), 327–352.

Summary

With the advent of Semantic Web and for an effective sharing of information sources present on the Web, new multi-source data integration possibilities emerge. The semantic mediation process has become an essential task in this new generation of systems that use ontologies. We present here the similarity measures used to find correspondences between local ontologies representing data sources and domain ontology present on super-peers that form MedPeer: our new heterogeneous and distributed data sources integration system in a P2P environment.

Ontologie générique des concepts des Ahadiths El Nabawia El Charifa

Meftah Dahmouni*, Hassina Aliane*
Kamel Boukhalfa**

*Ecole nationale Supérieure d'Informatique (ESI), Alger, Algerie
dmeftah@gmail.com
**CERIST, Alger, Algerie
haliane@hotmail.com
*** LSI, Dept Informatique, Faculté d'Informatique et D'electronique, USTHB, Alger, Algerie.
boukhalk@gmail.com l

Résumé. Science is now an important part of knowledge in Arabic that has now wants to put available on the Internet. In this context, the field of hadith knows a few works. In this article, we offer work related computerization of Islamic sciences in general and in particular Semantic work ahaadeeth.

Mots clés. Ontologie, représentation sémantique, hadith, Matn, concept.

1 Introduction

Les connaissances islamiques sont représentées par le coran, el ahadiths el nabawia et les différents livres islamiques. En effet, les ahadiths représentent la deuxième source de législation chez les musulmans.

Plusieurs axes d'informatisation peuvent toucher ces connaissances, la recherche d'information, la classification des ahadiths et le traitement automatique naturel des textes des ahadiths, …etc.

Nous avons choisi à concevoir une ontologie générique qui représente un index globale pour les sujets (thèmes ou concepts) qu'inclus le corpus des ahadiths, ainsi que pour les différentes livres islamiques (doctrine : العقيدة, jurisprudence : الفقه, …).

Cet article s'articule autour des points suivants : après une brève définition de la science du hadith en générale et la science de la terminologie du hadith en particulier, nous présentons une synthèse des travaux reliés à notre étude (articles et applications) après la représentation de l'ontologie comme modèle de représentation de connaissances. En suite, nous donnons une synthèse sur les difficultés de classification des ahadiths, et nous proposons notre ontologie qui est fondée sur la classification à base de thèmes. Enfin nous présentons l'exploitation de cette ontologie dans la section implémentation.

2 Les sciences du hadith

Les deux auteurs الطحان (1405هجرية) et أبو شهبة (2010) indique que les sciences du hadith concerne l'étude, l'authentification et la transmission des paroles du prophète Mohamed. Cette étude prend en compte tout ce qui a été rapporté non seulement comme paroles mais également les consentements (acquiescements) ou descriptions (caractéristiques) physiques ou morales du prophète. Ces sciences concerne l'étude des deux parties du hadith qui sont : l'énoncé ou le contenu du hadith qui est appelé Matn (متن), et la chaine de transmission des ahadiths (énoncé) qui est appelée Saned (سند). Cette chaine comporte un ensemble de narrateurs, qui commence par la couche des compagnons (الصحابة) jusqu'à la dernière couche qui est représentée par le compilateur (collecteur des ahadiths ; exemple El Boukhari : البخاري).

3 Les ontologies comme modèle de représentation des connaissances

Psyché et al. (2004) définissent l'ontologie comme suit : *"An ontology is an explicit specification of a conceptualization"*. Donc, l'ontologie c'est une spécification explicite de conceptualisation. Autrement, une ontologie selon Gomez Pérez et Benjamins (1999), représente un ensemble de : 1) Concepts : un concept c'est une entité. 2) Relations : les relations relient les différents concepts. 3) Fonctions : une fonction est un cas particulier d'une relation, où un élément de la relation est définit sur la base des autres éléments (précédents). 4) Axiomes : se sont les règles définies comme vrais, qui sont utilisées pour inférer d'autres connaissances. 5) Instances : représentent l'extension d'un domaine (autres objets de l'ontologie).

Dans cet article, nous proposons les ontologies comme : 1) forme de représentation et de capitalisation des ahadiths 2) outil pour le partage de connaissances et communication**,** via l'interopérabilité entre différents systèmes (collaboration) et échange de connaissance entre ces systèmes (communication). 3) support pour la réutilisation des connaissances, où un utilisateur peut utiliser une ontologie existante d'un autre utilisateur, afin d'élargir son domaine de connaissances. 4) support pour le développement d'outils pour les différents utilisateurs qui s'intéressent à l'étude de ce domaine (hadith).

4 Travaux reliés

Nous distinguons deux catégories de travaux qui ont des liens avec notre étude. Premièrement, les travaux de représentation sémantique des autres domaines des connaissances islamiques (coran et autres corpus des connaissances islamiques). Deuxièmement, les travaux relatifs à notre étude, qui s'intéresse aussi bien à notre domaine de connaissance (hadith) et notre domaine de représentation de connaissances (ontologie).

A propos des ontologies des autres domaines de connaissances, nous avons : Baqai et al. (2009) ont présenté une ontologie du coran, où la majorité des classes sont des classes primitives, à savoir : « verset : آية », « chapitre :سورة », « division :جزء », …etc. Saad et al.

(2011) ont décrit une ontologie de prière, où ces informations sont extraites du coran, des ahadiths et des livres de l'école de Chafia (مذهب الشافعية) uniquement.

Ta'a et al. (2013) ont présenté la construction d'une ontologie du coran à base de thèmes, en définissants tous les thèmes (concepts et sous concepts) et en les reliant aux versets, aux chapitres et aux divisions. Cette ontologie est validée par des experts.

Saad et al. (2010, 1) et Saad et al. (2010, 2) ont décrit un cadre générale de représentation des connaissances islamiques via une ontologie, avec comme cas particulier les concepts liés à la prière.

Concernant les travaux relatifs à l'ontologie du hadith, nous constatons qu'uniquement les trois articles des auteurs Baqai et al. (2009), Azmi et Bin Badia (2010) et Harrag et al. (2013), ont présenté une ontologie du hadith relative au Matn. Baqai et al. (2009) ne décrivent pas l'ontologie en détail. Par contre, Azmi et Bin Badia (2010) présentent en détail l'ontologie, où les classes de cette ontologie sont : une classe « personne » qui a deux sous classes « narrateur » et « auteur », les deux classes « livre » et « chapitre » et enfin la classe « hadith ». Cette dernière, comporte les propriétés: « titre » du hadith, « contenu » du hadith (Matn) et le « sujet » du hadith. De plus, cet article dessine l'arbre relatif aux narrateurs d'un hadith donné. Cette ontologie, est enrichie par Dalloul (2013), afin d'être utilisée pour juger le Saned d'un hadith. Harrag et al (2013) présentent la construction automatique d'une ontologie sur la base des règles d'association présentées dans les ahadiths du livre Sahih El Boukhari (صحيح البخاري). Pour cela, cette ontologie considère le chapitre (كتاب) comme une classe et la section (باب) comme une sous classes. En conséquence, l'ontologie comporte 99 classes (عدد الكتب). Les relations qui sont présentées dans l'article sont de type "part of", "kind of" and "synonym of". L'auteur Al-Masri présente une ontologie des concepts d'une seule partie des ahadiths, concernant les ahadiths de médecine prophétique, avec 6 classes. Les relations de cette ontologie, sont : la relation « synonyme » et les relations liées à la médicine, comme la relation « traiter : يعالج», …etc.

En conséquence, nous constatons l'absence d'une ontologie générique détaillée relative à tous les thèmes des ahadiths.

5 Les ontologies comme modèle de représentation de connaissances.

5.1 Difficulté de classification des ahadiths

Vu la diversité et l'hétérogénéité de classification des ahadiths, nous avons rencontrés énormément de difficultés pour construire l'ontologie. Nous citons uniquement trois types des classifications différentes (présenté dans le site web1(2016) : islamweb), afin de montrer la difficulté rencontrée pour opter à une seule classification. Ces trois classifications sont :

1. Classification basée sur des chapitres et des sections : par exemple sahih El Boukhari (صحيح البخاري), comportant les ahadiths très bon (أحاديث صحيحة).
2. Classification basée sur les narrateurs : comme exemple Mus'ned El Imam Ahmed (مسند الإمام احمد).
3. Classification alphabétique basée sur la première lettre du premier mot hadith : nous citons El Djamiâ El Kabir d'El Imam El Soyouti (الجامع الكبير للإمام السيوطي).

Alors pour chercher des ahadiths d'un sujet donné, il faut procéder à lire tous les ahadiths des livres, du début à la fin, des deux derniers types de classifications sus cités. En revanche, le premier type de classification (par chapitre), permet au lecteur d'accéder directement au chapitre ou section concernant le sujet adéquat, néanmoins, ce livre ne contenant pas tous les types des sujets, par exemple le sujet de l'éducation des enfants (تربية الأولاد)[1] n'existe pas dans ce livre et dans n'importe quel livre des ahadiths, ainsi qu' il y a des titres des sujets qui ne corresponds pas à la terminologie utilisée actuellement, qui est liée à la vie contemporaine. par exemple, le chapitre El Far'â oua El Atira (كتاب الفرع و العتيرة [2]) qui signifie El dhabiha (الذبيحة).

Pour cela nous avons opté à une classification à base de thèmes que comportant le Matn de chaque hadith. Notons qu'il est possible de classifier les ahadiths selon leurs authenticités (très bon : صحيح, Bon : حسن, Faible : ضعيف), mais cette classification existe dans les livres des ahadiths, et en conséquence il suffit de les collecter et les organiser uniquement dans les classes adéquates.

5.2 Proposition d'une ontologie des thèmes des ahadiths

Alors, pour la construction de notre ontologie, nous nous sommes basé sur les livres des ahadiths, en tenant compte de leurs diversités de classification, afin d'extraire les concepts relatifs aux différents thèmes. Dans ce contexte, parfois, nous avons utilisé les concepts cités dans ces livres (exemple, doctrine : عقيدة, jurisprudence : فقه, …) et parfois nous avons utilisé nos propres concepts, ainsi que notre propre regroupement des différents sujets en se basant sur la terminologie des sciences actuelles, après consultation des experts du domaine des ahadiths (chercheurs, Imams, …).

Nous avons suit la méthode stanford pour construire notre ontologie, qui est décrite dans le guide de Noy Natalya et McGuiness (2002).

A la fin, notre ontologie proposée comporte 176 classes (concepts) dont quatorze concepts principaux les plus abstraits du premier niveau, 15 propriétés et 9 relations. Pour le développement, nous avons utilisé protégé 2000. Nous donnons une synthèse générale sur les différentes classes :

1. العقيدة (doctrine): contient «pilier de foi : أركان الإيمان» qui sont : croyance au Ellah :الإيمان بالله, croyance aux anges : الإيمان بالملائكة, croyance aux livres (Coran, Bible, …) : الإيمان بالكتب, croyance aux prophètes : الإيمان بالأنبياء و الرسل, croyance à la résurrection : الإيمان باليوم الأخر, croyance au destin : الإيمان بالقضاء و القدر et croyance aux invisibles : الإيمان بالغيبيات (djinn et tombe : الجن و القبر).

2. الفقه (jurisprudence) : concerne les rituels religieux : العبادات (prière : الصلاة, dépense : الزكاة, jeûne : الصوم et pèlerinage : الحج), funérailles : الجنائز, jurement : الأيمان, gage : النذر et expiation : الكفارات, chasse : الصيد, animal abattu : الذبيحة, obéissance au dieu : النسك, ablution :الطهارة, relations : المعاملات. Ce dernier concept concerne les transactions financières : المعاملات المالية (vente, achat, …), relations entre les personnes : الأحوال شخصية (mariage : الزواج, divorce : الطلاق,

[1] Présenté dans le site web2 (2016) : http://www.ahlalhdeeth.com

[2]El Far'â oua El Atira (كتاب الفرع و العتيرة) : signifie un animal abattu.

manumission : العتق et héritage : الميراث), justice : القضاء (jugements et punitions : concerne les ahadiths des règles de jugement, les punitions, …), pouvoir et autorité : الإمارة و الحكم (concerne les concepts liés aux pouvoir, comme élection : البيعة, condition du choix du président : بيعة الخليفة, …), bonnes éthiques : الأخلاق الحسنة (vérité, fidélité, …), mauvais éthiques : الأخلاق السيئة (mensonge, trahison, …) et la relation du musulman avec les autres musulmans, avec les non musulmans, avec l'environnement, avec sa famille, … etc.

3. سير و تاريخ (comportement et histoire): contient les concepts liés à la vie du prophète, leurs expéditions et compagnies, et les paroles du prophète envers leurs compagnons.
4. القران (coran) : concerne les concepts des méthodes de lecture du coran, l'explication du coran (histoire des prophètes : قصص الأنبياء و الرسل et autres histoires : قصص أخرى…) et vertus des corans : فضائل القران.
5. أماكن و أوقات (places et périodes): comporte les concepts qui ont une relation avec les ahadiths qui parlent des places (مكة, المدينة, …) et des périodes (mois de jeûne, la journée du vendredi, …).
6. ذكر و دعاء و آداب (dire, prière, politesse): contient les concepts des ahadiths de dire et prière et les concepts liés aux ahadiths des différentes bonnes règles dans la vie quotidienne (politesse de manger, politesse de marcher dans la route, politesse de dialogue, politesse de voisinage, …).
7. وصايا دنيوية (conseils de la vie) : se sont les ahadiths relatifs aux conseils de vie données par le prophète. Par exemple : « appreniez vos enfants la natation, le tire et la montée des chevaux ».
8. البداية و النهاية (début et fin) : comporte les concepts qui ont une relation avec des ahadiths de création et de la fin de l'être humain et de l'univers (الكون). Par exemple, naissance de l'être humain, sa vie et sa mort, ainsi que la création de l'univers, son mouvement, la fin du monde, … etc.
9. الأعمال (travaux): comporte les concepts liés aux travaux (أعمال) et les règles de travaux (l'intension : النية, charité et bonne maitrise du travail : الإحسان, bonne intension : الإخلاص, …).
10. كرامات، بشارات و وعود (miracles, bonnes annonces et menaces) : concerne les concepts liés à la miracle (كرامة), bonne annonce (بشارة) et ouâid (وعيد).
11. المنهيات (interdictions): comporte les concepts relatifs aux différents interdictions.
12. الفضائل (préférences: vertus): concerne les préférences en général et les préférences de la science en particulier.
13. طب و غذاء (médecine et nourriture) : comporte les concepts relatifs à la médecine (maladie, médicament, …) ainsi qu'aux différentes nourritures (fruits, grain, …).
14. متفرقات (divers) : concerne les ahadiths qui ne sont pas liés aux concepts cités avant.

De plus, notre ontologie comporte : 1) les propriétés suivantes : la clé, le sujet, définition du sujet, type1 du hadith (très bon, Bon, Faible), compilateur du hadith (exemple, El Boukhari : البخاري) , type2 du hadith (Abrogatif et abrogé :ناسخ و منسوخ), type3 du hadith

(important et moin important : راجح و مرجوح), type4 du hadith (Sens unique ou non : محكم و متشابه) origine du hadith (saint[3] : قدسي , Relevé[4] : مرفوع, …), vérificateur du hadith[5] (exemple, El Albani : الألباني), type de senna (parole, acte, …), type du livre, Contenu du hadith : Saned + Matn. 2) les types des relations suivantes : " speciality of", "kind of" , "rule of", "rule_tartile_oua_tadjwide of", "fadhaile of", "explication of" et "is a". Sachant que, le type de relation dans le premier niveau (Figure 1), entre le concept « hadith » et tous les concepts les plus généraux est « speciality of ».

6 Implémentation

Nous avons implémenté notre ontologie avec l'éditeur protégé 2000. Le remplissage de l'ontologie se fait manuellement à travers une interface permettant l'insertion, la modification et la suppression des ahadiths (instances de l'ontologie).

Pour l'exploitation de l'ontologie (figure 1), nous avons réalisé deux interfaces de navigation, à savoir : 1) la première est une interface de personnalisation, destinée aux experts du domaine des sciences des ahadiths, leur permettant l'ajout, la suppression et la modification des concepts (thèmes) en cas de nécessité.

L'exemple suivant montre l'utilisation de cette interface : en utilisant le terme « l'âme : الروح » lors de la recherche, nous extrayons tous les ahadiths contenant le lemme de ce mot. Pour les ahadiths contenant aussi, le lemme « décédé : مات » et/ou «décès : ميت », alors ces ahadiths seront ajouté à la sous classe « décès : الموت » qui appartienne à la classe principale « début et fin : البداية و النهاية ». Mais si les ahadiths contenant uniquement le lemme « Ame : روح », comme le hadith suivant : « الأرواح جنود مجندة ما تعارف منها ائتلف و ما تناكر منها اختلف », sera ajouté à la classe principale « divers : متفرقات » s'ils n'existent pas beaucoup des ahadiths de ce genre, sinon dans le cas contraire nous procédons à créer une nouvelle sous classe « l'âme : الروح » comme fils de la sous classe « croyance des invisibles : الإيمان بالغيبيات » qui appartienne à la classe principale « doctrine : عقيدة». 2) la deuxième est une interface de recherche des ahadiths sur la base d'un concept donnée, destinée aussi bien aux utilisateurs publiques qu'aux experts.

La recherche se fait au moyen du langage SPARQL, qui est développé par W3C et est utilisé pour interroger le langage OWL de l'ontologie.

Par exemple, si nous utilisons le mot «prière : الصلاة», lors de la recherche alors le résultat contient, tous les instances (les ahadiths) contenant ce mot, ainsi que toutes les prières de type obligatoire « الصلوات المفروضة » comme « El Sob'h : الصبح », « El Dhoh'r : الظهر », « El As'r : العصر », « El Maghrib : المغرب » et « El Icha : العشاء » et toutes les prières de type facultatives « صلوات النافلة », comme « El Chaf'a : الشفع », « El Khosouf : الخسوف », « El Raouatib : الرواتب », …etc. Ces deux types de prières sont extraits via l'utilisation de la relation « Kind of » entre la sous classe père « El Salat : الصلاة» et la sous classe « Prières

[3] Saint : son source de paroles c'est Ellah

[4] Relevé: ce qui ajouté au prophète à savoir : parole, action, souci, acquiescement ou caractéristique du prophète.

[5] Vérificateur de hadith : indique est ce que el hadith (très bon, Bon, Faible)

obligatoires :الصلوات المفروضة » d'une part, et entre la sous classe père « El Salat : الصلاة» et la sous classe « Prières facultatives :صلوات النافلة » d'une autre part[6].

Egalement, à travers l'opération de lemmatisation qu'offre le traitement automatique de la langue naturelle, nous pouvons extraire tous les ahadiths contenant les mots de la même famille. Par exemple, (صليت، صلى، الصلاة، الصلوات، يصلون، صلي، ...) avec un seul mot clé utilisé pour la recherche (exemple, j'ai prié : صليت).

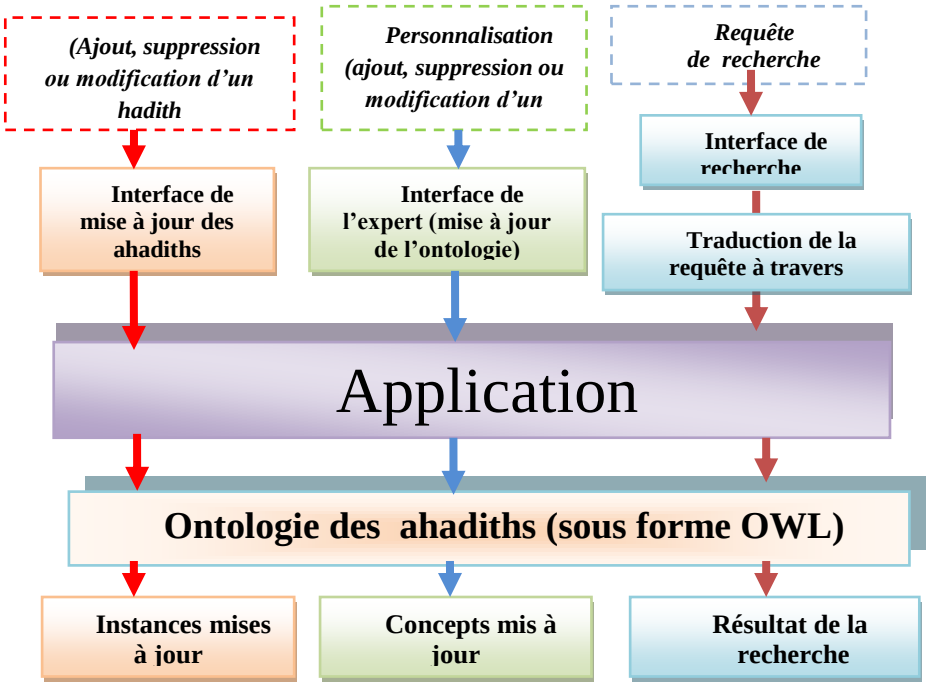


FIG. 1 – *Interface de navigation.*

7 Conclusion

Nous avons implémenté une ontologie sur la base d'une classification des thèmes. L'ajout des instances se fait d'une manière manuelle. Néanmoins, son exploitation est assurée par une interface de personnalisation de l'ontologie pour les experts pour un éventuel ajout, modification ou suppression d'un concept, ainsi qu'une autre interface concerne la recherche des ahadiths comportant un concept donné. Cette ontologie est utilisée pour la recherche sémantique ou indexation sémantique des corpus des ahadiths, ainsi que pour la classification automatiques des ahadiths.

Notre futur travail, consiste à réaliser un système de classification automatique des ahadiths, qui nous permet le peuplement automatique de l'ontologie.

[6] El dhoh'r est un type « Kind of » des prières obligatoires et El Rouatib est un type des prières facultatives (الظهر هو نوع من الصلوات المفروضة و الرواتب هو نوع من صلوات النافلة).

Références

Al-Masri, M. G. (2015). *An Ontology Based Approach to Enhance Information Retrieval from Al-Shamelah Digital Library*. Thèse magistère, May.

Azmi, A., et N. Bin Badia (2010) . *e-Narrator: an application for creating an ontology of hadiths narration tree semantically and graphically*. The Arabian Journal of Science and Technology 35(2C), 86–91.

Baqai, S., A. Basharat, H. Khalid, A. Hassan, and S. Zafar (2009) . *Leveraging semantic web technologies for standardized knowledge modeling and retrieval from the Holy Qur'an and religious texts*. FIT.

Dalloul, Y. M. (2013). *An Ontology-Based Approach to Support the Process of Judging Hadith Isnad*. Degree of Master in Information Technology, March.

Gomez Pérez, A., et V.R Benjamins (1999). *Overview of Knowledge Sharing and Reuse Components: Ontologies and problem-Solving Methods*. Proceeding of the IJCAI-99 workshop on Ontologies and problem-Solving Methods (KRR5), Stockholm (Suède), pp. 1.1-1.15.

Harrag, F., A. Alothaim, A. Abanmy, F. Alomaigan, et S. Alsalehi (2013) . *Ontology Extraction Approach for Prophetic Narration (Hadith) using Association Rules*. IJIACST, Vol. 1, Issue 2, September, 17-26.

NOY Natalya, F., D. MCGUINESS (2002). *Développement d'une ontologie 101 : Guide pour la création de votre première ontologie*. Université de Stanford.

Psyché, V., O. Mendes, et J. Bourdeau (2004) . *Apport de l'ingénierie ontologique aux environnements de formation à distance*, In STICEF, Vol. 10.

Saad, S., N. Salim, H. Zainal, S. Azman, M. Noah (2010) . *A Framework for Islamic Knowledge via Ontology Representation*. International Conference on Information Retrieval and Knowledge Management, 16-18, UiTM, Shah Alam, Selangor.

Saad, S., N. Salim, H. Zainal (2010). *Towards Context-Sensitive Domain of Islamic Knowledge Ontology Extraction*. International Journal for Infonomics (IJI).

Saad, S., N. Salim, H. Zainal, and Z. Muda (2011). *A process for building domain ontology: An experience in developing Solat ontology*. In Proceedings of IEEE International Conference on Electrical Engineering and Informatics, 1-5.

Site web1 (2016). *http://articles.islamweb.net/media/index.php?page=article&lang=A&id=16725*

Site web2 (2016). *http://www.ahlalhdeeth.com/vb/showthread.php?t=274600*

Ta'a, A., S. Z. Abidin, M. S. Abdullah, B. Abdul Bashah, M. Ali, M. Ahmad (2013) . *AL-QURAN THEMES CLASSIFICATION USING ONTOLOGY*. Proceedings of the 4 th International Conference on Computing and Informatics, ICOCI

أبو شهبة، م. م. (2010). *الوسيط في علوم ومصطلح الحديث*. الطبعة الثانية، منشورات عالم المعرفة.

الطحان، م. (1405هجرية). *تيسير مصطلح الحديث*. الطبعة السابعة، منشورات مركز الهدى للدراسات.

Context Aware Recommender System for access to adapted the Web Information System

BELKHIR Fatiha*,Rezoug Nachida**
University Saad Dahleb-Blida
Department of Computer science, University Saad Dahleb, Blida, Algeria
*fatihabelkhir@hotmail.com.**; n_rezoug@esi.dz

Abstract. With advancements of mobile environment technology and the occurrence of different types of devices, through wireless networks, the users can access and exchange information where they are at any time with any type of device. These new requirements are that the information requested and research should not only answer the need for the user but also should be supported on any type of device. Indeed, the mobile user (which changes all the time, the localization or device) can obtain a lot of information that is mostly useless and if they are, they might be unsupported on his device.

In this paper we propose a context aware recommender system enabling the nomadic user has access to a web information system through its mobile device this will be done by providing information adapted to its contextual information such as the user profile and context of use.

Keywords: context awareness, Web Information Systems, context adaptation CARS

1 Introduction

The mobility is deeply inscribed in human nature; the concept of mobile computing refers to the possibility for the users that have a mobile devices or mobile computer access to services and advanced applications through a dividing network infrastructure independently of their physical localization or attitude movement (Samuel Pierre, 2011).

The nomadic users can access the web-based information system, through their mobile device such as mobile phones to get the de want information account to their characteristics and those of their mobile device, WIS developer must rely on mechanisms able to offer to the nomadic users and adapted information to the context of use which is based on a representation of various elements such as the activities of the user, the preferences, characteristics of the device used, localization or the time (Benazzouz Yazid, 2011).

In such a system, adaptation of information becomes a major factor enabling to provide and adaptive services for the user, the adapting information to the nomadic user to take account the preferences of the user and context of use (the localization of the user, the characteristics of DM).

To facilitate the development of context aware applications, we are interested to context aware recommender system, the aim is to establish an access adapting to the mobile WIS to result a data relevant according to the context of use and the user profile.

The paper is organized as follows, we present in Section1 the introduction. Section 2 we present the notion of context, the definition of context awareness., in section 3 we present state of the art, then we concentrate on the discussion of the work of ubiquitous WIS, Context in recommender systems, Mobile and ubiquitous context- aware systems .In section 4 , we present the architecture of our proposed system for access to the mobile WIS detailing the different phases of each layers .In section 5 we present the implementation and realization of our work . Section 6 we conclude our proposed work.

3 State of the art

In terms of representation of the context of use and context aware. According to (Soukkarieh Bouchra, 2010) a summary of the work is detailed in the following table..

Research made by:	Discussing work of WIS context aware
(Virgilio R. D , 2005)	-was used to model WIS context aware through an architecture spreading a set of adaptation requirement, adaptation is assured taking into account the profile (used to represent an abstract configuration of web pages) and the context of use(used to represent these web pages taking into consideration the characteristic of the device, user preferences, and location).
(Kirsch-Pinheiro M. 2006)	- She focuses on the nomadic user and formalizes the context with the object model. - Takes into account the physical aspects (focusing or the device) also aspects related to cooperative process in which the user is involved (group, role, activitie.)
(Hinz and al 2004)	Presenting a new architecture named AMACONT The latter aims to dynamic generation of web presentations according to user preferences and its device capabilities.

TAB. 1 -Discussing work

A more recent work that really adapts access to web information account the profile and context of the user is proposed by (CARRILLO RAMOS, 2007):

This research was conducted in aim to provide for nomadic users access to information according to their device and adapt the information to the context of use and the user profile. To achieve this goal a solution was realized "a Framework called PUMAS" The approach that was chosen is that of the agents. Indeed the architecture of PUMAS is composed of four Multi-Agent Systems (MAS) later in this work a Contextual Profile Management System (CPMS) has been developed that contributes to the adaptation of the information delivered to a nomadic user on three aspects: I) formalization of the concept of user preference, ii) a contextual matching algorithm iii) a mechanism that manages conflicts may arise between the user preferences. At the end of this work CPMS was integrated into PUMAS within MAS dedicated to the adaptation of information.

4 Approach proposed for a mobile WIS

To facilitate the development of context-aware applications, we propose an access architecture mobile web information systems, the objective of this architecture is the adaptation of the information in mobile environments with the aim is to establish access to mobile WIS, why we interested to context aware recommender systems in order to achieve relevant data depending on the context of use and user's profile. The architecture of access to mobile WIS proposed is based on three layers: user layer, semantic layer and context aware recommender system layer as shown in Fig-1.

At the user layer, we retrieve the user profile information and its context, that information is collected for use in the next step.

At the semantic layer, we use ontology to model the contextual profile to write the ontology of contextual profile.

We offer a contextual profile recommendation algorithm, which allows providing the professional user "teaching" or "architecture" the most relevant information either by its current location or his favorite activity or only against its location.

At the context aware recommender system layer, we choose the contextual modeling method , One of the techniques of this method is the heuristic approach, this last it allowed us to represent a predictive model .

In Order to recommend to the user a WIS appropriate against to its activities and items already consulted by him in the past, We used two filtering methods "content-based" and "collaborative".

4.1 Architecture of access to the mobile WIS

Architecture of access to the mobile WIS is composed of three principal layers, The user layer, the semantic layer and context aware recommender system layer. see figure1

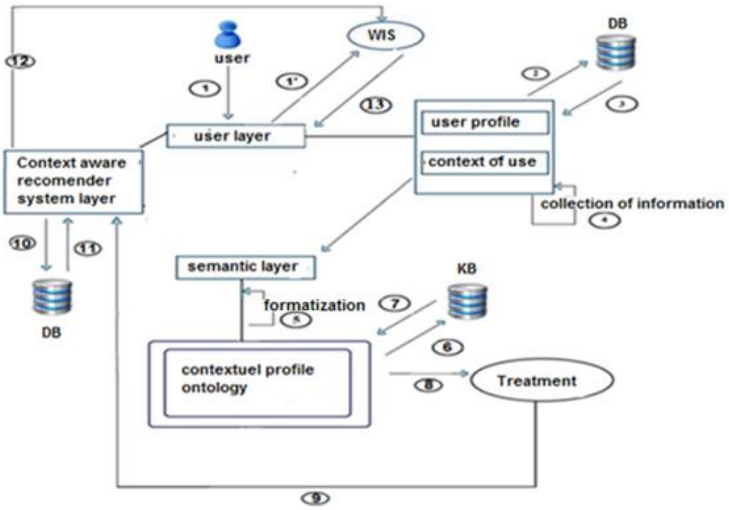


FIG. 1 – Architecture of access to the mobile WIS

4.2 General description of the proposed architecture

Numéro	Description
1+1’	User Authentication to access WIS by its profile
2	Sending data entered by the user to the database
3	Establish access to WIS
4	Collection of information
5	Formalization of the matrix
6	Sending data appears on the knowledge base
7	Sending Results
8	Sending data for treated
9	Sending the treaties result
10	Sending the recommended data
11	Result of the recommendation
12	Sending the results of the recommendation

13	Displaying the final result of the relevant data to the appropriate WIS to a user

TAB. 3- Architecture Description

4.3 Different phases of the architecture for adapted access mobile WIS

4.3.1 Collection of information

Collecting phase information is the phase of the user layer, the latter is constituted of a set of user profiles and a set of context of use, the aim of user profile is to store all data potentially useful on a user (Romain Picot-Clémente, 2011). The profile is composed of the static characteristics and evolving characteristics, whereas the only factor of the context of use considered as part of our theme is the localization, This phase permit to collect the contextual information for a user. The collection of contextual information doing with the manner explicitly and implicitly.

Example:

Implicitly: providing to a new user a web form asking him to choose his current activity.

Explicitly: detected the localization of the user through a commercial API (Google maps) helps us to recover the current activity of the user.

In our proposal, the collection of information is based on the user profile and its context, to detect the user's activity preferences, to extract its activity from their profile, and its localization from the context of the user to define the current localization of the user.

4.3.2 Formalization of the matrix

Consider a matrix VI with size N*M simplified manner representing the ontology of purpose. N is the number of concepts purposes and M is the number of individuals Of the domain. The matrix connects each individual and each purpose by a weight.

6.3.3 Modeling of contextual profile

The contextual profile modeling phase is the phase of the semantic layer, the later its aim to model the context of use and the user profile. We chose an ontology modeling.

4.3.3.1 Ontology of contextual profile

We chose the ontology as a model of representation of all the contextual profile concepts, it allows to model the context and profile, we mentioned mainly, the ontology representation scheme profile and usage context inspired work (CARRILLO RAMOS, 2007):

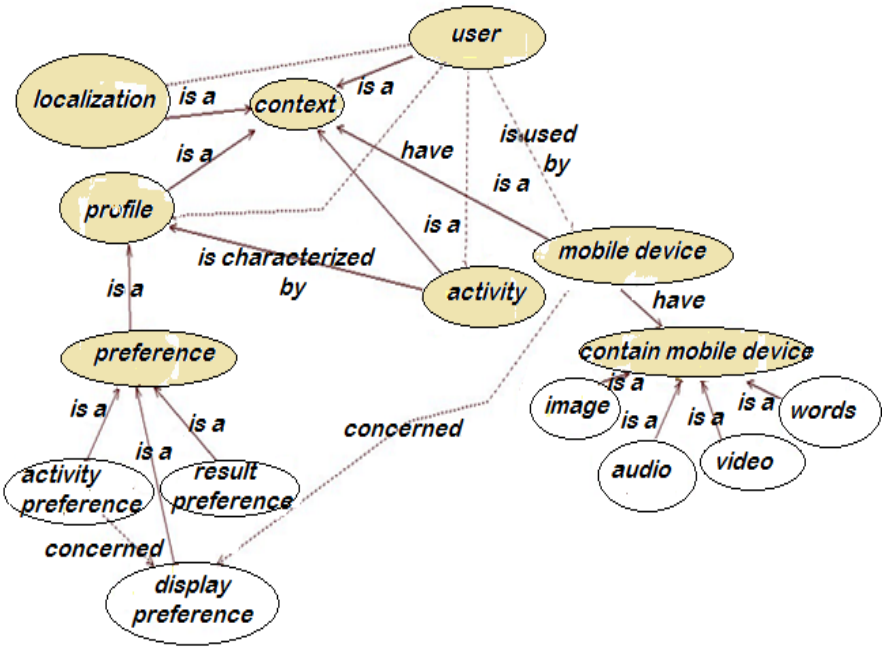


FIG. 2– Ontology of contextual profile

The explanation of the concepts of this ontology is inspired work of (CARRILLO RAMOS, 2007) ,these classes are described as following:

4.3.3.2 Management knowledge base

To represent the knowledge base of our system we define all classes and all classes that relate the property and describe the ontology of contextual profile.

4.3.4 Information treatment

The treatment phase is a phase that allows the treatment of concepts obtains from the semantic layer to this, we proposed a pseudo contextual user profile algorithm, this algorithm allows to deduct a favorite activity user, we focus on its activity and its context of use.

4.3.4.1 Pseudo algorithme of contextual profile

Step 1 : Authentification () ;
Step 2 : retrieve the location of the user
Var localization= get_Localization() ; // taking location
Step 3 : recommend information (by localization or by activity preference and localization)

```
If  (User is new)
          {
          var activity_ prefer =get_ Activity(localization) ;
          }  // take the activities corresponding to its detected location
Else
{
Var Activity []=get_Activity(id_ user) ;
Var Preference-A [] ;
}
For (int i=-1 ;i< Activite.taille();i++)
{
                  // browse the tab activity until his final size
                Preference-A [ i ]=getRepititionActivity(Activity[i]) ; // take the value of
activity and search function by the number of activity that has been repeated
}

Var IndicMaxi =Max (Preference-A) ; //take the indication of the most used activity

Var activity_ prefere =Activity [IndicMaxi] ; // removed the activity that corresponds to the
indication found

if(activity_ preferer . getLocalization()==localization)
{
printf("recommend this as this activity prefer") ;
}
Else {  Printf("make the recommended by locazation") ;}
```

4.3.5 The recommendation

The recommendation is a context aware recommender system layers, The recommendation is realized when the information is collected (context of use, profile) , the context of use is constituted by information on the current localization of the user, the user localization can be obtained using a GPS device or by another method, we use a commercial API (Google maps) which permit to pass for a coordinate based location to an address, however, the profile constituted of information on the activities of the user.

On the other hand, the pairing of all the information collected account the classes ontology are made at the contextual modeling phase .In order to deduce the preferred activity of a user, the phase treatment is based on user activity and location.

4.3.5.1. Heuristic approach

For the use of contextual information, we used the context modeling approach that is predictive models to assess the class of an object or to assess the value or range of an attribute, for this we have presented an example of this heuristic approach to modeling.

Example recommendation user contextual information (teacher, researcher or architect):

Model: User Treatment that has as a research activity, teacher or architect

UL: user location.
US: user situation.
UP: user preference

if UL. Location == "university" then
US. Activity = "researcher"; UP. Preference = " consult schedule conference";
else if UL.Lieu == "Department" then
US.Activité = "teacher"; UP. Preference = "do appointment";
if UL. Location == " office research" then
US. Activity = "architect"; UP. Preference = "consult plan".

Heuristic M1

M1: "consult schedule conference"
AU: university.
UC: user researcher

if AU. Program == "conference presentation"
then if date = 19/07/2014
then if UC.coopére == " assist conference"
Then valueOf (M1) = "yes";
else if UC.coopére == "does not assist conference"
Then valueOf (M1) = "no";
else if UA.Programme == "conference presentation" then
if date> 07/19/2014
then if US.coopére == "see conference" then valueOf (M1) = "yes".

Heuristic M2

M2: do appointment
LE: list of end of study students
EU: teacher user

If UE.Activité = "teacher"; then if LE.cycle == "Master 2"; then if LE.cycle == "license";
then if S.spécialité = "computer" then if T.time = mm / ss / hh Then valueOf (M1) = "yes".

Heuristic M3

M1: Consult plan
AU: office research
UR: user architect

if AU. Program == "plan presentation"
then if US.coopére == "view plan" then valueOf (M1) = "yes".

4.3.5.2. Information filtering

The recommendation for objective to recommend items to users and its context according SIW profile .The filtering methods recommendation system " content-based filtering" and "collaborative filtering" are those used in our proposal.

Collaborative filtering: for the "collaborative filtering" method, we try to recommend to the user a WIS considering the similarity between the activities of a user preferences and those of other users.

For example: user 1 has activity as "researcher", that user usually consults the conferences, the system allows to recommend conferences has already seen or conferences which deals with the same subject.

5 Experience

We have constructed an application of context aware recommendation which adapts the information to mobile users and permit access to WIS that correspond his preferences of activities based on its activities and its current location.

Example of interface: “localization of a user profession «architect», «teacher» or «researcher».

This interface aim to localize a new user of activity (architect or teacher, researcher), once the new user authenticate the system detects its current localization and indicates its activity.

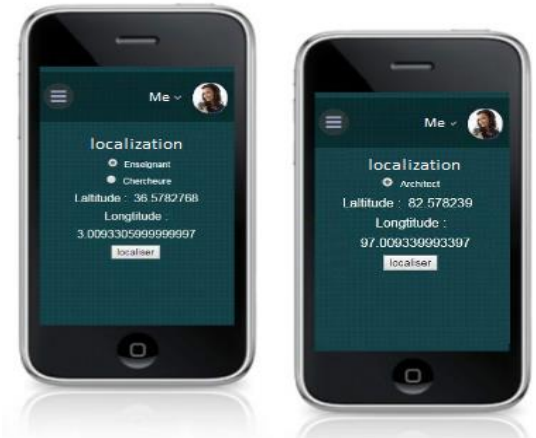


FIG. 3– Interface of localization

Example of interface: recommendation plan

Usually the user consulted the list of plan , the aim of recommendation is to try to recommending it the plan which focus with the same subject which has already consulted this interface gave us the results of the recommendation plan.

FIG. 4–: Interface of recommendation plan

6 Conclusion

Today, new needs in information systems have appeared, for that with the emergence of the technology, the information systems called ubiquitous so that the nomadic user can have information according to when or where it is located.

The development of our work permit to conceive a system which adapts information which permit mobile users to access a WIS taking into account their mobility, the use of different mobile devices and their capabilities As solutions we put the focus on a context aware recommender system for adequate access to WIS contributing pertinent information to nomadic users based on their mobile devices.

References

Samuel pierre, Réseaux et système informatique mobile, Edition revue et augmentée.

Benazzouz Yazid, « Découverte de contexte pour une adaptation automatique de services en intelligence ambiante», Thèse de doctorat, École Nationale Supérieure des Mines de Saint-Étienne, page 29-31,2011.

Anind K. Dey. Understanding and using context. Personal Ubiquitous Computing, 5(1) :4–7, 2001.

Schmidt, A., Beigl, M. et Gellersen, H. (1999). There is more to context than location. Comput Graphics, 23(6):893_901.

Henricksen, K. (2003). A Framework for context-aware pervasive computing applications. Thèse de doctorat, University of Queensland, Queensland, Australia.

Behlouli Belhanafi Nabiha , « Ajout de mécanismes de réactivité au contexte dans les intergiciels pour composants dans le cadre d'utilisateurs nomades», Thèse de doctorat, l'Université d'Évry Val d'Essonne , 2006.

Nachida Rezoug, Fahima Nader, Omar Boussaid, "Implémentation d'OLAP dans les environnements mobiles : aperçu de l'état de l'art". International Conference on Information Systems and Technologies Tebessa, Algeria, ICIST, 24-26 April 2011.

Villanova-Oliver M. (2002). Adaptabilité dans les systèmes d'information sur le web : Modélisation et mise en œuvre de l'accès progressif. Thèse de doctorat, Institut Nationale Polytechnique de Grenoble, France.

CARRILLO RAMOS Angela Cristina, « Agents ubiquitaires pour un accès adapté aux systèmes d'information : Le Framework PUMAS» thèse de doctorat, Université Joseph Fourier, 2007

Jérôme Gensel, Marlène Villanova-Oliver, Manuele Kirsch-Pinheiro , « Modèles de contexte pour l'adaptation à l'utilisateur dans des Systèmes d'Information Web collaboratifs», Université Leuven,2008.

Soukkarieh Bouchra, « Technique de l'internet et ses langages : vers un système d'information Web restituant des services Web sensibles au contexte», thèse de doctorat , l'Université Toulouse III - Paul Sabatier,2010

Bouzeghoub, M., Kostadinov, D. Personnalisation de l'information : aperçu de l'état de l'art et définition d'un modèle flexible de profils. Actes de CORIA 2005 (Grenoble, France, 9-11 mars, 2005), pp. 201-218.

Romain Picot-Clémente, « Une architecture générique de Systèmes de recommandation de combinaison d'items. Application au domaine du tourisme»,
l'Université de Bourgogne, thèse de doctorat, 2011.

geoffary bonnin , « vers des systemes de recommandation robustes pour la navigation Web : inspiration de la modélisation statistique du language»,thése de doctorat ,université Nancy2,2010.

Resnick, P. and Varian, H. R. 1997. Recommender Systems, Communications of the ACM, 40(8):56-58.

Gediminas Adomavicius and Alexander Tuzhilin, « Context-Aware Recommender Systems»,2010.

Virgilio R. D. et Torlone R. (2005). A General Methodology for Context-Aware Data Access. In Proceedings of the 4th ACM International Workshop on Data engineering for wireless and mobile access (MobiDE'05), p. 9-15, Baltimore, Maryland, USA.

Kirsch-Pinheiro M. (2006). Adaptation Contextuelle et Personnalisée de l'Information de Conscience de Groupe au sein des Systèmes d'Information Coopératifs, Thèse de doctorat, Université Joseph Fourier, Grenoble 1, France.

Hinz M. and Fiala Z. (2004). AMACONT: A System Architecture for Adaptive Multimedia Web Applications. In Proceedings of the 11th international conference on 3D web technology, p. 65-74, Germany.

Schilit, B.N., and Theimer, M.M., Disseminating active map information to mobile hosts. IEEE network, 8(5):22–32, 1994.

P. J. Brown, J. D. Bovey, and X. Chen. Context-aware Applications : from the Laboratory to the Marketplace. IEEE Personal Communications, 4(5) :58–64, October 1997.

Ryan, N., Pascoe, J., and Morse, D., Enhanced Reality Fieldwork: the Context-Aware Archaeological Assistant. Gaffney, V., van Leusen, M., Exxon, S.(eds.) Computer Applications in Archaeology. British Archaeological Reports, Oxford, 1997.

Dey A.K., Abowd, G.D., and Salber, D., A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. Human-Computer Interaction, 16(2):97–166, 2001.

Résumé

Avec les progrès de la technologie de l'environnement mobile et l'apparition de différents types d'appareils, et des réseaux sans fil, les utilisateurs peuvent accéder et échanger des informations où ils sont à tout moment avec tout type d'appareil. Ces nouvelles exigences sont que les renseignements demandés et de la recherche ne doivent pas seulement répondre à l'utilisateur mais doit également adapter ces informations à tout type d'appareil. En effet, l'utilisateur mobile (qui change tout le temps, de localisation ou de dispositif) peut obtenir beaucoup d'informations qui sont la plupart du temps inutiles et si elles sont, ils ne pourraient être pris en charge sur son dispositif.

Dans cet article, nous proposons un système de recommandation sensible au contexte permettant à l'utilisateur nomade un accès à un système d'information sur le Web via son appareil mobile cela sera fait en fournissant des informations adaptées à ses informations contextuelles telles que le profil de l'utilisateur et le contexte d'utilisation

Une nouvelle approche basée sur la détection d'opinion par SentiWordNet pour les résumés automatiques de textes par extraction

Mohamed Amine BOUDIA*, Reda Mohamed HAMOU**, Abdelmalek AMINE***, Ishak H.A Meddah ****

*,**,***Laboratoire de Gestion des Connaissances et des Données Complexes(GeCoDe Lab)
Department d'informatique, , Université Dr. Tahar Moulay Saida, Algeria
**** Université Science et Technologie d'Oran USTO.

mamiamounti@yahoo.fr * hamoureda@yahoo.fr ** abd_amine1@yahoo.fr ***
ishak.meddah@yahoo.com****

Résumé. Dans cet article, nous proposons une nouvelle approche basée sur la détection d'opinion par le SentiWordNet pour la production de résumé automatique de textes en utilisant la technique d'extraction par scoring adapté à la détection d'opinion. Les textes sont décomposées en phrases puis représentés par un vecteur de scores d'opinion. Le résumé se fera par élimination des phrases dont l'opinion est différente de celle du texte originale. Cette différence est aussi exprimé par un seuil d' opinion. L'hypothèse suivante : « les unités textuelles qui ne partagent pas la même opinion du texte sont des idées utilisées pour le développement ou une comparaison et que leurs absences n'ont pas la vocation d'atteindre la sémentique du résumé. » a été vérifiées par la mesure statistique du Chi_2. Enfin nous avons déterminé un intervalle de seuil d'opinion qui engendreront les évaluations optimales.

Mots clés : Opinion Mining, SentiWordNet, Résumé automatique, Extraction, Détection d'opinion, N-Gram.

1 Introduction et problématique

A l'heure actuelle , l'un des problème majeur des informaticiens est l'accès au contenu des informations, l'accès en lui-même ou autrement dit les infrastructures software et hardware ne sont plus un obstacle, le problème majeur c'est l'augmentation exponentielle de la quantité d'information textuelle électronique. Cela engendre l'utilisation des outils plus spécifiques autrement dit l'accès au contenu des textes par des moyens rapides et efficaces est devenu une tâche nécessaire.

Résumer un texte s'avère une technique intéressante pour l'accès rapide au contenu des informations textuelles. Un résumé est un texte sous une forme réédité du texte originale en taille plus réduite qui se réalise sous la contrainte de garder la sémantique d'un document autrement dit minimiser l'entropie sémantique. Le but de cette opération est d'aider le lecteur à repérer les informations intéressantes pour lui sans être obligé de lire entièrement le document. L'utilisation des résumés automatiques à pour but de réduire le temps de recherche

pour trouver les documents pertinents ou de réduire le traitement des longs textes en identifiant les informations clées.

Pour faire un résumé automatique, la littérature actuelle présente trois approches : par extraction, par compréhension et par classification.

Un autre axe de recherche qui a pris de l'ampleur ces dernières années, en occurrence l'Opinion Mining ou le fait de détecter une opinion d'une phrase, paragraphe ou texte. Notre travail consiste à l'utilisation de méthodes de détection d'opinion pour produire un résumé. Nous proposons cette hypothèse :

« les unités textuelles qui ne partagent pas la même opinion du texte sont des idées utilisées pour le développement ou une comparaison et que leurs absences n'ont pas la vocation d'atteindre la sémantique du résumé. »

Nous allons construire un résumé a partir des phrases qui ont une opinion proche de celle de texte selon un seuil d'opinion, notre travail répondra au question suivantes :

- Est-ce que notre hypothèse est vérifiable ? si oui est-ce qu'elle est valide ?
- Quel est l' impact du seuil d'opinion sur la qualité du résumé ?
- L'opinion mining peut t-elle rapporter un plus pour le résumé automatique ?:

2 Our proposed approach

Notre approche se base sur l'identification des opinions des unités textuelles (expressions, propositions, phrases, paragraphes), l'identification de l'opinion du texte originale , enfin l'extraction des unités textuelles qui partagent les même opinion que le texte original. Nous partons de l'hypothèse citée dans l'introduction. L'opinion est une expression des sentiments d'une personne envers une entité ou un aspect de l'entité (Liu, 2010). L'entité peut être un produit, une personne, un événement, une organisation ou un sujet.

SentiWordNet sera utilisé pour filtrer les mots porteurs d'opinion et calculer les scores d'opinions, de telle sorte que :

Si (score_opinion (terme i) <0) **alors** l'opinion du (terme i) est négative,

sinon positive

Notre approche suivra alors les étapes suivantes:.

2.1 Prétraitement

Les mots vides ne vont pas être supprimé, car la méthode de résumé automatique par extraction consiste à extraire les phrases les plus informatives sans les modifier: en supprimant des mots vides de sens sans avoir des informations sur leur impact morphosyntaxique dans les phrases, on risque d'avoir un résumé incohérent d'un point de vue morphologique. Le nettoyage consistera alors à supprimer les émoticônes, de remplacer les espaces par « _ » et supprimer les caractères spéciaux (#, \, [,].....).

Dans notre étude on aura besoin de deux représentations : représentation en sac de mots. Et représentation en sac de phrases. Ses deux représentations sont introduites dans le cadre du modèle vectoriel.

La matrice d'occurrence phrases- mot sera générer à l'issue des deux représentations précédentes, cette matrice est de taille égale au nombre de phrases dans le texte x le nombre de mot dans le texte, les poids p_{ik} représente le nombre d'occurrence du mot k dans la phrase i.

2.2 Détection d'opinion par SentiWordNet.

On réduit la matrice « phrase-terme » à une matrice « phrase– terme porteur d'opinion » en filtrant les termes du vecteur vi par le SentiWordNet. Les termes inéxistants dans le dictionnaire d'opinion seront éliminés.

A la fin de cette étape, une matrice M de taille n * p sachant que n est égale aux nombres de phrases et p est égale aux nombres de terme porteur d'opinion. M_{ij} indique l'occurrence du mot (porteur d'opinion) j dans la phrase i.

Une Matrice O est génèrée de même taille que la matrice M tel que :

$$O_{ij} = M_{ij} * score\ (j)$$

Le score (j) représente le score du terme j obtenu par le SentiWordNet.

2.3 Construction de Résumé

Une fois la matrice O prête, On calcule le score des phrases ainsi que celui du texte afin de procèder a l'ultime étape qui représente le résumé final.

Le score d'opinion pour les unités textuelles (phrases, paragraphe ou texte) est égale à la moyenne des scores des termes porteurs d'opinion obtenu par le SentiWordNet .

Le score d'opinion de chaque phrase sera calculé comme suit :

$$Score\_phrase\ [i] = \frac{\sum_{i=0}^{n} O_{ij}}{\sum_{i=0}^{n} M_{ij}}$$

Tel que n=nombre terme porteurs d' opinion dans l' unité textuelle

Enfin on identifie l'opinion de texte qui est la moyenne des score d'opinion des phrases qui le composent

$$Score\_texte = \frac{\sum_{i=0}^{R} Score_{phrase}[i]}{R}$$

Telque R taille de vecteur V' (nombre de phrases)

"La procédure suggérée revendique le principe que les mots de haute fréquence dans un document sont les mots importants" [Luhn 1958], Dans notre cas l'adaptera comme suit « La procédure suggérée revendique le principe que les phrases qui partagent la même opinion que le document (texte) sont des phrases importantes » autrement dit les phrases qui ne partagent pas la même opinion avec le texte sont des phrases qui on été utilisées par l'auteur pour le développement d'idée ou une comparaison, cela veut dire qu'on peut les éliminer sans provoquer une grande perte de sens.

L'étape finale consiste à sélectionner les phrases qui ont les mêmes opinions que le texte pour selon la règle suivante:

Si (valeur absolu(score_texte - seuil_phrase[k]) <= seuil_voisignage)
Alors sélectionné la phrase k

Où on sélectionne les phrases qui on un degré d'opinion proche de celui du texte. Selon un seuil de voisinage : on sélectionnant que les phrases ayant un score proche (positivement ou négativement) à celui score du texte original par le seuil d'opinion.

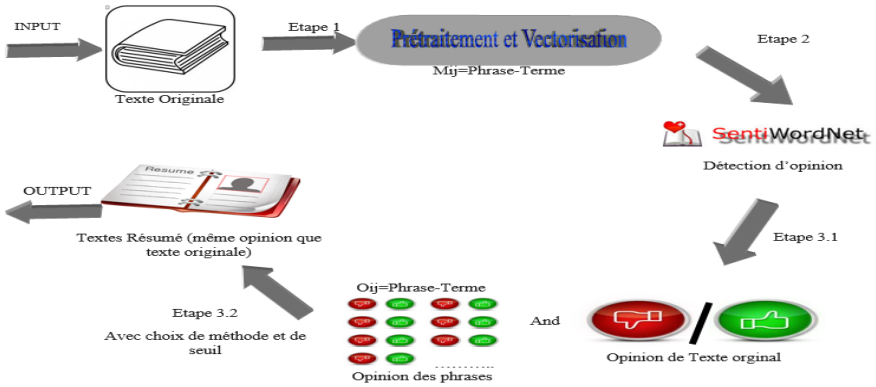


FIGURE 1 : PROCESSUS COMPLET DE L'APPROCHE PROPOSE

3 Expérimentation

Pour vérifier notre hypothèse déjà citée nous utiliserons la mesure Chi2 qui est une mesure statistique bien connue, elle évalue le manque d'indépendance entre une unité textuelle et un texte. Elle utilise les mêmes notions de co-occurrence mot/texte que l'information mutuelle normalisé.

$$X^2\,(ut_k, text_i) = \frac{|T|.[P\,(ut_k, text_i).P\,(\overline{ut_k}, \overline{text_i}) - P\,(ut_k, text_i).P\,(\overline{ut_k}, text_i)]^2}{P(ut_k).P(\overline{ut_k}).P(text_i).P(\overline{text_i})}$$

L'utilisation de la mesure du Chi2 déterminera l'indépendance des phrases éliminées par la détection d'opinion par rapport au texte original et favorise l'absence des termes et les termes les plus fréquents puis prend en considération les informations du texte. Une valeur élevée du Chi-2(k,i) reflète une dépendance entre la phrase k et le texte i.

La validation de notre étude est assurée par la robustesse de résumé, pour cela nous utiliserons deux méthodes d'évaluation en l'occurrence la ROUGE-SU(2) (Recall-Oriented Understudy for Gisting Evaluation –Skip Unit) et la F-Mesure.

3.1 The Data Used

On a utilisé comme corpus le texte « Hurricaine» en langue française qui contient un titre et 20 phrases et 313 mot , après les processus de prétraitement et de vectorisation en sac de mots, nous obtenons 171 token différent.

Les résumés référence : on a pris trois résumés références produit succesivement par le résumeur CORTEX, Essential Summarizer, et un résumé produit par un expert humain.

3.2 Résultats

Seuil	Cortex			Essential Summarizer			Expert Humain		
0,00125	VP=0	FN=74	P=0,2757	VP=3	FN=53	P=0,5893	VP=5	FN=62	P=0,7287
	FP=6	VN=91	R=0,4690	FP=3	VN=112	R=0,5137	FP=1	VN=103	R=0,5325
	F-Mesure 0,3473		ROUGE-SU(2) 0,0	F-Mesure 0,5489		ROUGE-SU(2) 0,00535	F-Mesure 0,6153		ROUGE-SU(2) 0,06172
0,0025	VP=14	FN=60	P=0,6513	VP=5	FN=51	P=0,4561	VP=10	FN=57	P=0,5612
	FP=6	VN=91	R=0,5634	FP=15	VN=100	R=0,4794	FP=10	VN=94	R=0,5265
	F-Mesure 0,6043		ROUGE-SU(2) 0,18918	F-Mesure 0,4674		ROUGE-SU(2) 0,08928	F-Mesure 0,5433		ROUGE-SU(2) 0,12345
0,00375	VP=16	FN=58	P=0,5236	VP=8	FN=48	P=0,4424	VP=13	FN=54	P=0,4960
	FP=18	VN=79	R=0,5153	FP=26	VN=89	R=0,4583	FP=21	VN=83	R=0,4940
	F-Mesure 0,5194		ROUGE-SU(2) 0,21621	F-Mesure 0,4502		ROUGE-SU(2) 0,14285	F-Mesure 0,4950		ROUGE-SU(2) 0,16049
0,005 à 0,0075	VP=39	FN=35	P=0,6885	VP=13	FN=43	P=0,4254	VP=21	FN=46	P=0,4824
	FP=18	VN=79	R=0,6707	FP=44	VN=71	R=0,4247	FP=36	VN=68	R=0,4836
	F-Mesure 0,6707		ROUGE-SU(2) 0,52702	F-Mesure 0,4251		ROUGE-SU(2) 0,23214	F-Mesure 0,4830		ROUGE-SU(2) 0,25925
0,00875	VP=41	FN=33	P=0,6473	VP=14	FN=24	P=0,4025	VP=22	FN=45	P=0,4478
	FP=26	VN=71	R=0,6430	FP=53	VN=62	R=0,3945	FP=45	VN=59	R=0,4478
	F-Mesure 0,6451		ROUGE-SU(2) 0,55405	F-Mesure 0,3951		ROUGE-SU(2) 0,25	F-Mesure 0,4478		ROUGE-SU(2) 0,27160
0,01 à 0,01125	VP=46	FN=28	P=0,6780	VP=15	FN=41	P=0,3970	VP=23	FN=44	P=0,4375
	FP=26	VN=71	R=0,6767	FP=57	VN=58	R=0,3861	FP=49	VN=55	R=0,4360
	F-Mesure 0,6776		ROUGE-SU(2) 0,62162	F-Mesure 0,3915		ROUGE-SU(2) 0,26785	F-Mesure 0,4367		ROUGE-SU(2) 0,28395
0,0125	VP=49	FN=25	P=0,6428	VP=25	FN=31	P=0,4668	VP=30	FN=37	P=0,4613
	FP=36	VN=61	R=0,6455	FP=60	VN=55	R=0,4623	FP=55	VN=49	R=0,4594
	F-Mesure 0,6441		ROUGE-SU(2) 0,66216	F-Mesure 0,4685		ROUGE-SU(2) 0,44642	F-Mesure 0,4604		ROUGE-SU(2) 0,37037
0,01375	VP=49	FN=25	P=0,5934	VP=25	FN=31	P=0,4276	VP=30	FN=37	P=0,4144
	FP=46	VN=51	R=0,5939	FP=70	VN=45	R=0,4188	FP=65	VN=39	R=0,4113
	F-Mesure 0,5936		ROUGE-SU(2) 0,66216	F-Mesure 0,4232		ROUGE-SU(2) 0,44642	F-Mesure 0,4129		ROUGE-SU(2) 0,37037
0,015	VP=50	FN=24	P=0,5612	VP=26	FN=30	P=0,4011	VP=38	FN=29	P=0,4662
	FP=54	VN=43	R=0,5594	FP=78	VN=42	R=0,3930	FP=66	VN=38	R=0,4622
	F-Mesure 0,5603		ROUGE-SU(2) 0,67567	F-Mesure 0,3970		ROUGE-SU(2) 0,46428	F-Mesure 0,4662		ROUGE-SU(2) 0,46913
0,01625 à 0,0175	VP=51	FN=23	P=0,5327	VP=34	FN=22	P=0,4653	VP=42	FN=25	P=0,4756
	FP=61	VN=36	R=0,5301	FP=78	VN=37	R=0,4644	FP=70	VN=34	R=0,4768
	F-Mesure 0,5314		ROUGE-SU(2) 0,68918	F-Mesure 0,4648		ROUGE-SU(2) 0,60714	F-Mesure 0,4762		ROUGE-SU(2) 0,51851
0,01875 à 0,02	VP=53	FN=21	P=0,4903	VP=46	FN=10	P=0,5791	VP=49	FN=18	P=0,5260
	FP=61	VN=36	R=0,4921	FP=78	VN=37	R=0,5715	FP=75	VN=29	R=0,5050
	F-Mesure 0,4912		ROUGE-SU(2) 0,71621	F-Mesure 0,5753		ROUGE-SU(2) 0,82142	F-Mesure 0,5055		ROUGE-SU(2) 0,60493
0,02125	VP=53	FN=21	P=0,4542	VP=46	FN=10	P=0,5592	VP=49	FN=18	P=0,4756
	FP=76	VN=21	R=0,4663	FP=83	VN=32	R=0,5489	FP=80	VN=24	R=0,4812
	F-Mesure 0,4608		ROUGE-SU(2) 0,71621	F-Mesure 0,5545		ROUGE-SU(2) 0,82142	F-Mesure 0,4783		ROUGE-SU(2) 0,60493
0,0225 à 0,02625	VP=64	FN=10	P=0,5672	VP=47	FN=9	P=0,5226	VP=57	FN=10	P=0,5422
	FP=76	VN=21	R=0,5406	FP=93	VN=22	R=0,5152	FP=83	VN=21	R=0,5263
	F-Mesure 0,5536		ROUGE-SU(2) 0,86486	F-Mesure 0,5189		ROUGE-SU(2) 0,83928	F-Mesure 0,5341		ROUGE-SU(2) 0,70370
0,0275 à 0,03	VP=64	FN=10	P=0,5191	VP=47	FN=9	P=0,4809	VP=59	FN=8	P=0,5420
	FP=82	VN=15	R=0,5097	FP=99	VN=16	R=0,4892	FP=87	VN=17	R=0,5220
	F-Mesure 0,5144		ROUGE-SU(2) 0,86486	F-Mesure 0,4850		ROUGE-SU(2) 0,83928	F-Mesure 0,5318		ROUGE-SU(2) 0,72839

TABLE 1. RESULTAT D'EVALUATION DE RESUME PRODUIT (CANDIDAT) ROUGE ET F-MESURE EN UTILISANT 3 RESUME REFERENCE CORTEX, ESSENTIEL SUMMERIES, EXPRET HUMAIN

Le tableau suivant illustre d'une façon explicite les phrase gardé (G) et les phrase élimi-né (E) à chaque seuil d'opinion et donne le valeur chi_2 pour chaque phrase avec le texte originale ainsi que le taux chi2 de chaque résumé par rapport le texte originale.

Phrase / Seuil	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	Taux Chi_2 %	Taux de rédu %
0	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	0	100
0,00125	E	E	E	E	E	E	E	E	E	E	E	E	E	E	G	E	E	E	E	E	E	5,422	95,24
0,0025	E	E	E	E	E	E	E	E	E	G	E	E	E	E	G	E	E	E	E	E	E	9,836	90,48
0,00375	E	E	E	E	E	E	E	E	E	G	E	E	E	E	G	E	E	E	E	E	G	14,24	85,72
0,005 à 0,0075	E	E	E	E	E	E	E	E	E	G	G	E	E	E	G	E	E	E	E	E	G	18,01	80,95
0,00875	E	E	E	E	E	E	E	E	E	G	G	E	E	E	G	G	E	E	E	E	G	22,26	76,2
0,01 à 0,01125	E	E	E	E	E	E	G	E	E	G	G	E	E	E	G	G	E	E	E	E	G	27,17	71,43
0,0125	E	E	E	G	E	E	G	E	G	G	G	E	G	E	G	G	E	G	E	E	G	51,66	52,39
0,01375	E	E	E	G	G	E	G	E	G	G	G	E	G	E	G	G	E	G	E	E	G	55,59	47,62
0,015	E	E	E	G	G	E	G	E	G	G	G	G	G	E	G	G	E	G	E	E	G	59,90	42,86
0,01625 à 0,0175	E	E	E	G	G	E	G	E	G	G	G	G	G	E	G	G	G	G	E	E	G	64,33	38,1
0,01875 à 0,02	E	E	G	G	G	E	G	E	G	G	G	G	G	E	G	G	G	G	E	E	G	68,42	33,34
0,02125	E	E	G	G	G	G	G	E	G	G	G	G	G	E	G	G	G	G	E	E	G	73,01	28,58
0,0225 à 0,02625	G	E	G	G	G	G	G	E	G	G	G	G	G	E	G	G	G	G	E	E	G	76,83	23,81
0,0275 à 0,03	G	E	G	G	G	G	G	E	G	G	G	G	G	E	G	G	G	G	E	G	G	81,33	19,05
>0,03	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	100	0
Chi-2	0,4456	0,5113	0,47671	0,54112	0,48832	0,53389	0,57254	0,68138	0,71500	0,51435	0,43795	0,50251	0,83666	0,47430	0,63190	0,49601	0,51648	0,76084	0,50841	0,52420	0,51382		

TABLE 2. REPARTITION DES PHRASES DANS RESUME PRODUIT (CANDIDAT) POUR CHAQUE SEUIL AVEC TAUX CHI_2 ET TAUX DE REDUCTION, ET VALEUR CHI_2 POUR CHAQUE PHRASE

4 Interprétation et Discussion

Nous avons tester notre approche avec un seuil incrémentale à 0,00125 afin de voir l'impact du seuil d'opinion sur la qualité du résumé ainsi on pourra recommander une fourchette de seuil qui conduit a des bons résultat (bon résumé).

ROUGE est une métrique d'évaluation semi-automatique intrinsèque qui se base sur le nombre de co-occurrence entre un résumé candidat et un ou plusieurs résumés de référence divisé par la taille de ces derniers. Sa faiblesse est qu'elle se base que sur les résumés références et néglige le texte orignal.

La F-Mesure est l'une des métrique les plus robuste utilisée pour l'évaluation des classification ; La F-Mesure est une combinaison de Rappel et Précesion. Pour l'adaptation on ajoute à la force de F-Mesure le faite que nous procèderons à une évaluation éxtrinsèque au début, et on enchaîne avec une évaluation intrinsèque : donc une évaluation hybride. Pour un résumé automatique à taux de réduction réduit, la F-Mesure donne des évaluations meilleures que celle de ROUGE car elle prend en considération l'absence de terme. Mais contrairement

à ROUGE l'évaluation de résumé candidat à taux de réduction élevé peux être faussée puisque la valeur de faux négative sera maximale ce qui donnera de bonnes évaluations pour des résumé généralement médiocre.

Les deux variables taux de réduction de texte et le pourcentage de Chi_2 ont une corrélation inverse, ce qui indique que le nombre de phrases retenu augmente la dépendance au texte orignal. Cette indication est logique et attendu, mais se basant sur la table 2, on constate que le fait de retenir les phrases P4, P13 et P18 le taux de chi-2 est augmenté de plus de 25%, grace a leur forte dépendance par rapport au texte original (égale a à 0.57 pour P4 et superieur à 0.71 pour les phrases P13 et P18).

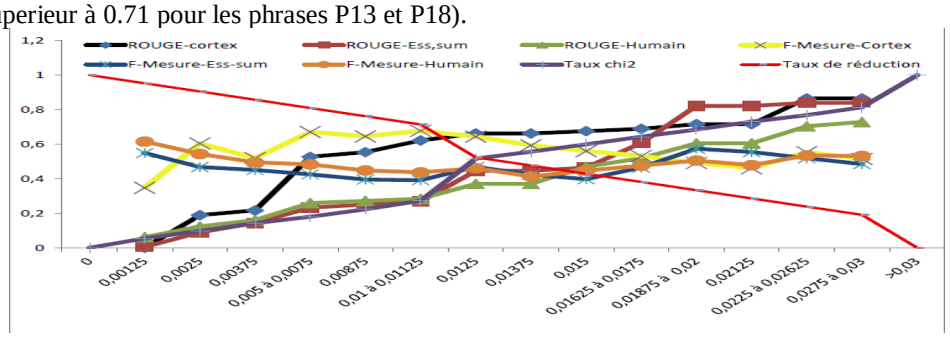


FIGURE 2. ROUGE VS F-MESURE (POUR LES 3 RESUMES REFERENCE) VS TAUX DE CHI2 VS TAUX DE REDUCTION

Enfin on peut remaquer que tous les indices utilisés pour l'évaluation atteignent leur valeurs sont optimale ensemble entre le seuil 0.0125 et 0.0175, dans cet intervalle toutes les valeurs d'évaluation sont bonnes pour le résumé candidat.

On peut voir à partir du figure 2 et table 2 que la sélection des phrases qui sont moins dépendante au texte orginale, n'améliore pas les résulats, par exemple : entre le seuil 0,0125 et 0,01375, la seule différence est la sélection de la cinqième phrase par le deuxième seuil, dans la table 2 on s'appercoit que la valeur Chi_2 de cette phrase est faible ce qui est lisible aussi sur le graphe par la stagnation de ROUGE et une lègére chute de la F-Mesure pour les trois résumés référence. Notre hypothèse est ainsi confirmée.

5 Conclusion et perspective

Dans cet article, nous avons présenté une nouvelle approche pour la production d'un résumé automatique par extraction basée sur la détection d'opinion par SentiWordNet

Dans une première étape , nous avons proposé une hypothèse qui a servi de support de cette approche « les phrases qui ne partagent pas le même opinion sont des phrases qui ont été ajoutées pour le développement et leur suppression ne provoque aucune entropie de donnée » . La deuxième étape , nous avons expliquer notre approche de détection d'opinion et proposer une technique souple afin de sélectioner les phrases : seuil d'opinion .les résultats obtenus, ont démontrés que notre hypothèse est valide ; et par conséquent peut contribuer à résoudre une des problématiques majeure des résumés automatiques : la réduction de l'entropie d'information et la conservation de la sémentique.

En perspective nous tenterons de perfectionner le résumé automatique par extraction basé sur la détection d'opinion ; par l'application d'autres techniques et de méthodes classique telle que la détection de tématique ainsi que le Fuzzy Opinion.

Réference

1. Boudin , F. and Morin, E. (2013). Keyphrase extraction for N -best Reranking in Multi- Sentence Compression . In Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL) .
2. CLAVEAU , V. (2012). Vectorization , Okapi and Calculating Similarity for NLP : For-get for the TF -IDF Finally . In Proceedings of the Joint Conference EHD - NLP - RECITAL 2012 , Volume 2: NLP .
3. Hamou, R. M., Amine, A., & Lokbani, A. C. (2012). The Social Spiders in the Clustering of Texts: Towards an Aspect of Visual Classification. *International Journal of Artificial Life Research (IJALR)*, *3*(3), 1-14.
4. LIN C.-Y. (2004). ROUGE : A Package for Automatic Evaluation of Summaries. In M.-F. MOENS & S. SZPAKOWICZ, Eds., Text Summarization Branches Out : ACL-04 Workshop, p. 74–81, Barcelona.s
5. Amine, A., Hamour, R. M., Simonet, M., (2014) Detecting Opinions in Tweets. CoRR abs/1402.5123
6. MANI I., KLEIN G., HOUSE D., HIRSCHMAN L., FIRMIN T. & SUNDHEIM B. (2002). Summac : a text summarization evaluation. Natural Language Engineering, 8(1), 43–68.
7. OVER P., DANG H. & HARMAN D. (2007). DUC in context. IPM, 43(6), 1506–1520.
8. Baccianella S., Esuli A., Sebastiani F., (2010), SentiWordNet 3.0 : An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining , the conference on Language Resources and Evaluation (LREC'10).
9. Bethard S., Yu H., Thornton A., Hatzivassiloglou V., Jurafsky D., (2004) ,Automatic extraction of opinion propositions and their holders, Working Notes of the AAAI Spring Symposium on Exploring Attitude and Aect in Text : Theories and Applications.
10. Bossard A., Généreux M., Poibeau T., (2008), Description of the LIPN Systems at TAC2008 : Summarizing Information and Opinions, Proceedings of the Text Analysis Conference.
11. Boudin F., Torres-Moreno J.-M., El-Bèze M., (2008), A Scalable MMR Approach to Sentence Scoring for Multi-Document Update Summarization, COLING Conference, p. 21-24.
12. Pak A., Paroubek P., (2010), Twitter as a Corpus for Sentiment Analysis and Opinion Mining, the conference on International Language Resources and Evaluation (LREC'10).
13. BACCIANELLA S., ESULI A., SEBASTIANI F. (2010). SentiWordNet 3.0 : An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In Proceedings of LREC'10.
14. BOUCHLEGHEM R., ELKHLIFI A., AND FAIZ R. (2010). Automatic extraction and classification approach of opinions in texts. ISDA 2010, IEEE Press, 918-922.

Abstract. In this paper, we propose a new approach based on the detection of opinion by the SentiWordNet for the production of text summarization by using the scoring extraction technique adapted to detecting of opinion. The summary will be done by elimination of sentences whose opinion is different from the original text. This difference is expressed by a threshold opinion. The following hypothesis: "textual units that do not share the same opinion of the text are ideas used for the development or comparison and their absences have no vocation to reach the semantics of the abstract" Has been verified by the statistical measure of Chi_2 which we used it to calculate a dependence between the unit textual and the text. Finally we found an opinion threshold interval which generate the optimal assessments.

Keywords: Opinion Mining, SentiWordNet, automatic summarization, Extraction, N-Gram.

Une nouvelle méthode pour le calcul du skyline basée sur le tri

Zekri Lougmiri

Université d'Oran 1 Ahmed Ben-Bella
Faculté des Sciences Exactes & Appliquées
Département d'informatique Laboratoire LAPECI
BP 1524, El-M'naouer, Maraval, Oran 31000, Algérie
lougmiri@gmail.com
zekri.lougmiri@univ-oran.dz

Résumé. Les requêtes skyline sont importantes pour la prise de décisions multicritères. Ces requêtes offrent un ordre partiel puisqu'il est impossible d'avoir un ordre total sur des données contradictoires. Dans ce papier et dans une première étape, nous rappelons notre méthode de réduction de l'espace des points afin de réduire le nombre des tests de dominance. Cette réduction est justifiée car le nombre de points en entrée pourrait être trop élevé. Dans une seconde étape, nous présentons des résultats qui montrent l'importance de la réduction, alors que dans la troisième étape, nous présentons une nouvelle méthode pour le calcul du skyline. Cette méthode est de type Divide & Conquer qui sera comparée à BNL (Bloc Nested Loops).

1 Introduction

Les requêtes skyline retournent un ensemble d'objets, appelés skyline, qui ne sont dominés par aucun autre objet. Ce type de requêtes est utile pour la prise de décisions multicritères, puisqu'un ordre total ne peut être établi sur l'ensemble des attributs. Le skyline a été connu sous le nom de front de Pareto et a été défini dans le contexte des mathématiques à partir des travaux de l'économiste Vilfredo Pareto (Pareto, 1896). En 2001, il fut adapté dans le contexte des bases de données (Börzsonyi et al.,2001). La relation de dominance se définit comme suit : Soit un ensemble de points S dans un espace vectoriel de dimension d (d critères). Soit D l'ensemble de toutes les dimensions $D = \{d_1, d_2, \ldots, d_d\}$, et soient p et q deux points de S. La relation de dominance ($\succ$) suivant D est pour $1 \leq i ; j \leq d$, p domine q, notée $p \succ q$, si et seulement si $\{ \forall d_i \in D, p(i) \leq q(i)\}$ et $\{\exists d_j \in D/p(j) < q(j)\}$. Lorsque $p \not\succ q$ et $q \not\succ p$ alors p et q sont dits non dominés ou concurrents. A partir l'ensemble de données S, l'opérateur skyline renvoie l'ensemble des points qui ne sont dominés par aucun autre, suivant toute les dimensions D : $SkyD(S) = \{p \in S/\nexists\ q \in S/q \succ p\}$.

Les données sur lesquelles s'applique cet opérateur sont soit réelles, soit synthétiques. Les données réelles sont issues de domaines réels comme le sport, les banques, les agences immobilières alors que les données synthétiques sont le produits de programmes qui produisent ces données suivant des lois de statistiques (Börzsonyi et al.,2001) . Les données synthé-

tiques sont de trois types. Le premier définit les données corrélées. Celles-ci représentent un environnement dans lequel les points qui sont bons dans une dimension sont également bons dans les autres dimensions. Le deuxième type contient les données anti-corrélées où un environnement dans lequel les points qui sont bons dans une dimension sont mauvais dans l'une ou toutes les autres dimensions. Les donnes indépendantes forment le troisième type où toutes les valeurs des attributs sont générées indépendamment à l'aide d'une répartition uniforme. Les algorithmes qui traitent le skyline doivent être progressifs (Kossmann et al.2002) ; ie ils doivent retourner les résultats au fur et à mesure qu'ils les détectent.

Dans ce papier, nous détaillons la méthode de réduction et le calcul des points Max_{sys} et Min_{sys} (Zekri et al., 2015). Nous donnons de nouveaux résultats. Ces résultats incluent les statiques concernant les premiers points skyline détectés et les quantités des points éliminés par rapport à l'ensemble de points global. Les expérimentations à base de données réelles et synthétiques sont données.

Le reste de ce papier se présente comme suit. La section 2 donne des travaux reliés. La section 3 présente la méthode de réduction de DCRD pour Divide-and-Conquer for Reduced Data (Zekri et al., 2015) qui produit un espace de candidats réduit, sur lequel nous appliquons un nouvel algorithme de type D&C. Cette nouvelle combinaison est appelée SRDS, pour Sorted and Reduced Data Space. La section 4 présente les résultats des expérimentations. La conclusion et les futurs travaux sont donnés dans la section 5.

2 Travaux liés

(Börzsonyi et al.,2001) était le premier travail ayant adapté l'optimisation au sens de Pareto dans les bases de données. Intuitivement, le calcul du skyline consiste à comparer chaque point *p* avec tous les autres et si aucun point ne le domine alors *p* est un point skyline. L'algorithme BNL, pour Bloc Nested Loops, (Börzsonyi et al.,2001) utilise cette technique directe. Il met en mémoire une liste candidate et teste à chaque fois si un nouveau point p domine un ou plusieurs points déjà insérés. Si c'est le cas, il est inséré et l'ensemble des points dominés est écarté sinon il passe au point suivant. Cette méthode peut être utilisée facilement et ne requiert pas de prétraitement, sauf qu'elle est gourmande en mémoire et en temps de calcul. Divide&Conquer D&C(Börzsonyi et al.,2001) divise l'entrée en plusieurs partitions et détermine le skyline de chaque partition. Par la suite, les skyline sont fusionnés et les points dominés sont écartés. Cette méthode est meilleure que BNL mais souffre des multiples duplications lors de la fusion. Notons que ni BNL ni DC ne fonctionnent progressivement. L'algorithme Bitmap(Tan et al., 2001) consiste à encoder dans des vecteurs bitmap toutes les informations de chaque point selon le nombre de points distincts sur chaque axe. La comparaison des vecteurs bitmap se fait par la suite. Bitmap est progressif, mais nécessite beaucoup d'opérations et de codage en commençant par la détermination des points distincts dans chaque axe. Les tests dupliqués sont un problème qui nuit à cet algorithme. Index(Tan et al., 2001) consiste à trier les données sur chacun des d axes dans un ordre croissant. Afin de déterminer le skyline, les points sont testés de façon circulaire. Le problème est que la récupération des coordonnées des points peut prendre du temps. Cette méthode est bien adaptée pour les applications progressives, elle retourne aussitôt les premiers points, sauf que les auteurs ne déduisent pas les points skyline induits par le tri. Nearest Neighbor NN(Kossmann et al., 2002) est un algorithme qui utilise les R-Trees pour indexer les don-

nées. Il partitionne l'espace sur chaque axe selon le point le proche voisin de l'origine. NN est progressif et est efficace dans un espace `a deux dimensions, mais il souffre du problème de duplications des éliminations pour 3 dimensions et plus. A partir de 4 dimensions, il devient difficile de l'appliquer. Les auteurs proposent différentes techniques pour remédier à ces problèmes. Branch and Bound Skyline BBS (Papadias et al., 2003) exploite les R-Tree, la méthode de Branch and Bound et NN afin de calculer, progressivement, le skyline. Son plus grand problème est qu'il souffre de requêtes redondantes. Le nombre de dimensions élevés est un autre problème qui pénalise BBS. Les auteurs de (Yuan et al., 2005) proposent Skycube. Ils calculent les skyline fils de toutes les combinaisons possibles des points dans le treillis. Lorsqu'ils passent au niveau supérieur ou inférieur du treillis, ils fusionnent ces fils. (Jongwuk et al. 2009) a repris Index (Tan et al.2001) et l'a couplé avec l'algorithme Best Position algorithms (Reza et al.2007) afin d'optimiser les accès aux listes. Mais la technique qu'ils ont présentée est meilleure pour les données uniformes. Vu l'évolution constante des méthodes de calcul du skyline et des différentes projections et utilisations, de nouveau travaux récents sont apparus. De point de vue machine, des travaux ont exploité les multiprocesseurs et les cartes graphiques. (Bøgh et al.,2015) présente SkyAlign qui est une nouvelle stratégie basée sur le tri. Cette stratégie combines le CPU et la GPUpour accélérer le calcul du skyline (Bøgh et al., 2013). (Bøgh et al., 2013) décrit GNL dans le but de calculer le skyline sur GPU. GNL est une parallélisation de BNL sur GPU. Ce travaille présente aussi une méthode naïve pour calculer le skyline. GNL a présenté des performances meilleures que BNL et cette méthode naïve.Avec l'augmentation des volumes des données, (Yuanyuan et al., 2015) propose des calculer une variante du skyline en caclculant les sous-espace skyline afin de satisfaire les utilisateurs.

3 La méthode SRDS

Notre méthode DCRD (Zekri et al., 2015), est une méthode analytique qui se base sur le tri pour la réduction de l'espace des candidats. Ainsi, DCRD s'inscrit dans la famille des algorithmes qui appliquent un prétraitement sur les données. Nous reprenons la même méthode de réduction de DCRD et nous présentons un nouvel algorithme de type D&C dans la section 3.3. Dans (Zekri et al., 2015), nous avons appliqué l'algorithme D&C, tel qu'il était présenté dans (Börzsonyi et al.,2001). DCRD a montré que si on combine D&C avec notre méthode de réduction, les résultats seraient meilleurs que si on l'appliquait sur les données brutes. Notre nouvel algorithme est baptisé SRDS, pour Sorted and Reduced data Space, évite les fusions dupliquées desquelles souffre D&C (Börzsonyi et al.,2001). Dans la partie suivante, nous présentons brièvement la méthode de réduction de DCRD puis, nous enchaînons par la présentation de SRDS.

Soit *S* l'ensemble des points d'une base de données en entrée. Soit *d* le nombre d'attributs de chaque objet de la base de données. Ainsi, chaque point (objet) de *S* possède *d* coordonnées dans un espace vectoriel correspondant. Supposons, pour simplification, la résolution du skyline dans le sens min sur toutes les dimensions.

3.1 Méthode de réduction de DCRD

Il est souhaitable de calculer le skyline d'une manière progressive (Kossmann et al., 2002), c'est à dire retourner les points skyline au fur et à mesure que les algorithmes les

détectent. Afin de respecter ce souhait, nous avons proposé d'utiliser le tri pour déduire les premiers points skyline.

On crée et trie d listes dans l'ordre croissant selon les *d* axes. Nous avons prouvé dans (Zekri et al,. 2015) que si un point A possède une valeur unique minimale sur un axe alors A est un point skyline. Si par contre il existe plusieurs points avec la même valeur minimale sur un axe, alors on procède à tester la dominance entre eux sur le reste des axes.

Si les données sont du type corrélé, alors la plupart, sinon la totalité, des points skyline seront édités. Pour les autres types de données, un nombre important de points sera retourné à l'utilisateur. Dans ce stage, nous avons démontré deux théorèmes utiles (zekri et al,. 2015).

L'ensemble de points permet de déduire les deux points que nous avons baptisés Minsys, l'équivalent du point idéal (Benayoun et al., 1971), et Maxsys, l'équivalent du point nadir. Les d composantes de Minsys (Maxsys) sont les plus petites (grandes) composantes des premiers points skyline déterminés. Afin de réduire l'espace des candidats, nous avons démontré le théorème suivant (zekri et al,. 2015) :

Soit p un point de S. Si p possède d ou (d-1) composantes supérieures aux d ou (d-1) composantes correspondantes du point Max_{sys}*, où d est le nombre de dimensions, alors p est éliminé.*

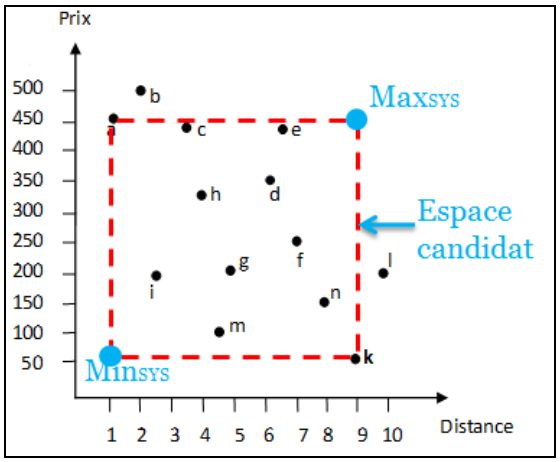


FIG. 1 – *Présentation de l'espace résultat à l'issu de la réduction*

La figure 1 présente l'espace candidat résultat après réduction de l'espace. On remarque alors les points b et l sont éliminés puisque chacun d'eux possède une composante supérieure à celle équivalente de Maxsys.

3.2 Sorted and Reduced Data Space

A l'issue de la phase 2, l'espace de dominance contient tous les points considérés comme candidats. Les autres points sont définitivement éliminés. Il ne reste que de les comparer entre eux et éliminer les points dominés. Cette phase se compose de deux étapes.

Etape1 : Diviser pour régner. Nous utilisons un algorithme de type diviser pour régner qui permet de trouver les points Skyline en un temps très réduit. On divise l'ensemble obtenu à chaque fois en deux sous-ensembles gauche et droit de façon récursive jusqu'à ce que la

taille de la partition soit inférieure ou égale à 3, selon le nombre pair ou impair de tuples de l'espace candidats. Chaque partition est confiée à un thread qui calcule son skyline local.

Etape 2 : Fusion. La fusion se fait à la manière de Round Robin où, à chaque fois, on teste la dominance entre les skyline locaux. A la différence de D&C de (Börzsonyi, et al., 2001), nous ne risquons pas d'avoir des duplications dans les tests.

4 Expérimentation

Les expérimentations ont été réalisées dans une machine dotée d'Intel Core i5 2,50 GHz, et d'une mémoire de 4 Go, sous Windows 7, 64 bits. Le programme est écrit sous Java. MySQL 5.1.41 est utilisé comme système gestion des bases de données. Il est interrogé depuis un serveur Apache 2.2.14. Nous rappelons qu'on résout le skyline dans le sens Min.

BNL est le meilleur algorithme dans le meilleur des cas, ie, quand le nombre de points Skyline est petit alors que D&C est le meilleur dans le pire des cas, ie quand le nombre de points Skyline est important. Dans cette partie, nous comparons SRDS à BNL. La comparaison est faite en variant la dimensionnalité et la cardinalité des points. La comparaison est faite selon le temps global, bien que SRDS retourne des points skyline très tôt alors que BNL n'est pas progressive.

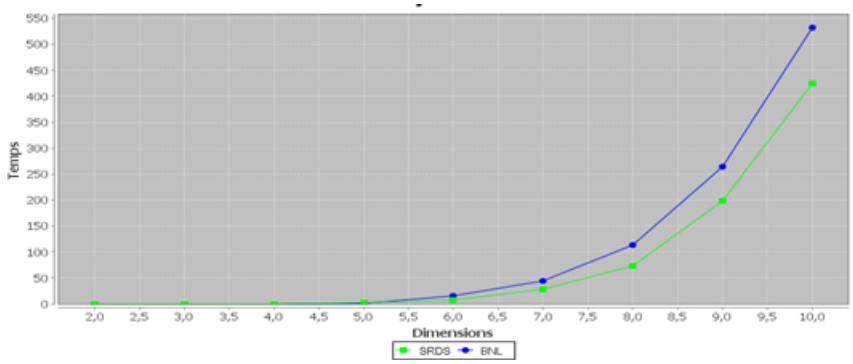


FIG. 2 – *Résultat du changement de dimension dans les données indépendantes*

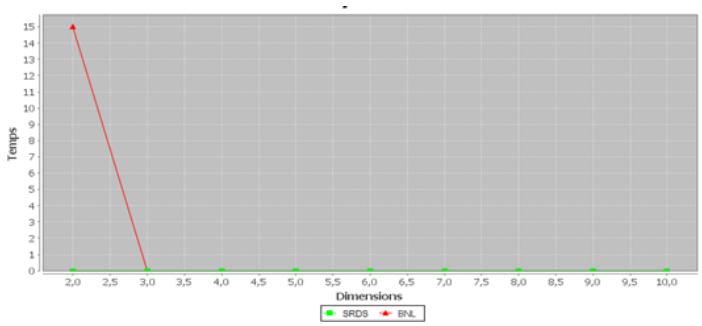


FIG. 3 – *Résultat du changement de dimension dans les données corrélées*

Sur la figure 2, on remarque qu'après 6 dimensions, l'intervalle entre les deux courbes s'élargit car le nombre de points skyline devient important. A la fin, pour 10 dimensions, on obtient une différence de 100 sec entre SRDS et BNL, ce qui est important pour une requête. En réalité entre 2 et 5, il y a une différence mais elle n'est pas montrée car il fallait faire un zoom.

La figure 3 montre qu'au début, quand le résultat est égal à 4 points, BNL fait 15 ms pour les calculer, alors que SRDS les déduits directement depuis son index. A partir de 3 dimensions, on a un seul point comme résultat, donc le temps reste fixe.

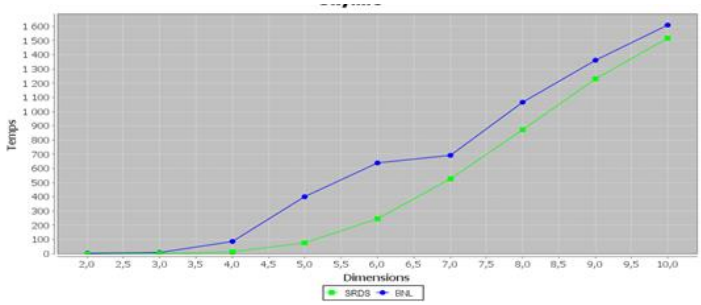


FIG. 4 – *Résultat du changement de dimension dans les données anti-corrélées*

La figure 4 montre que SRDS a réalisé une meilleure performance que BNL sur les données anti-corrélées. Sur les 10000 points et à partir de d=5, l'écart en millisecondes s'est bien creusé entre ces deux méthodes. Cet écart a atteint 350 millisecondes pour d=6 et globalement, SRDS a été plus rapide que BNL.

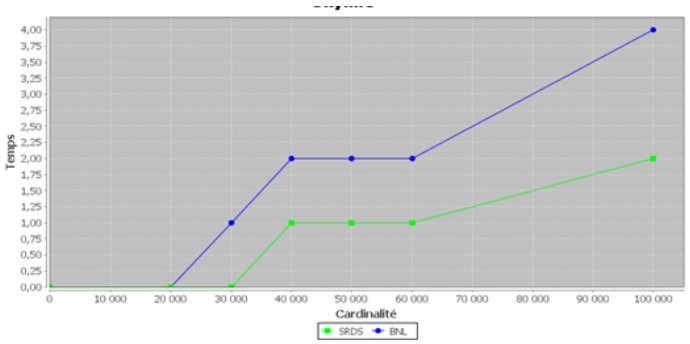


FIG. 5 – *Résultat du changement de la cardinalité dans les données indépendantes*

La figure 5 montre que BNL a mis moins d'une seconde pour trouver les points skyline dans les deux tables qui ont une cardinalité inférieure à 20.000 tuples, alors que SRDS a fait ce même temps jusqu'à 30.000 tuples. A la fin, dans les tables de 100.000 tuples, nous avons une différence de 2 secondes entre les deux algorithmes.

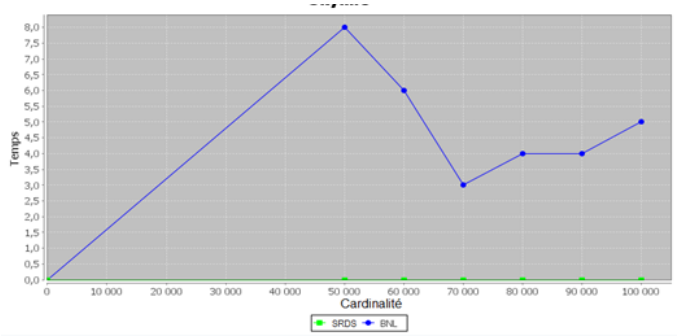


FIG. 6 – *Résultat du changement de la cardinalité dans les données corrélées*

On remarque sur la figure 6, que dans cette catégorie, le SRDS a fourni immédiatement les skyline et reste constant dans le temps, alors que BNL est obligé de passer par toutes les comparaisons possibles induisant un temps de réponse élevé.

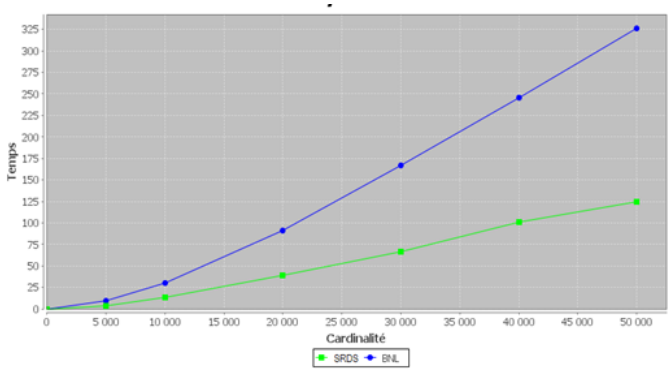


FIG. 7 – *Résultat du changement de la cardinalité dans les données anti-corrélées*

Sur la figure 7, on remarque que SRDS a présenté une meilleure performance que BNL, bien que ce type de données soit difficile à manipuler. Cette figure montre que BNL consomme un temps linéaire en fonction de la cardinalité puisque le nombre de tests de dominance augmente proportionnellement avec elle alors que SRDS a été plus rapide. Pour 50000 points, SRDS est plus rapide de200ms. Cette courbe montre que plus la cardinalité augmente plus l'écart en ms augmente aussi.

5 Conclusion

Dans ce papier, nous avons présenté une nouvelle méthode pour le calcul du skyline ainsi que de nouveaux résultats. Cette méthode analytique permet d'éliminer les points qui ne feront plus part du skyline. L'élimination est importante car elle permet d'éviter un grands nombre de tests de dominance. Nous avons montré que les prétraitements sont importants car ils permettent de gagner un temps considérable.

Ainsi, nous avons donné de nouveaux théorèmes et nous avons expliqué les différentes étapes de SRDS. Les expérimentations menées sur des données réelles et synthétiques ont montré l'importance du tri. Celui-ci permet de déduire un nombre important de réponses instantanées et offre la possibilité aux algorithmes de fonctionner progressivement. Actuellement, nous travaillons sur la distribution de cette approche afin qu'elle puisse fonctionner sur un système pair-à-pair hybride.

Références

R. Benayoun, J. de Montgolfier, J. Tergny, O. Larichev(1971). Linear programming with multiple objective functions: Step method (STEM), *Mathematical Programming*, tm. 1, no. 3, pp. 366-375.

S. Börzsonyi, D. Kossmann, K. Stocker(2001), The skyline operator, *ICDE2001*, pp 421-430.

Bøgh, k., Assents, I., Magnani, M. : Efficient gpu-based skyline computation. In DaMoN'13(2013).

Bøgh K, S. Chester I. Assent, Work-Efficient Parallel Skyline Computation for the GPU. Proceedings of the VLDB Endowment, 2015, Vol. 8, No. 9.

D. Dhaenens-Flipo(2005). *Optimisation Combinatoire Multi-Objectif. Apport des Méthodes Coopératives et Contribution à l'Extraction de Connaissances*. Thèse d'habilitation. USTL Lille.

L. Jongwuk, k. Jinhan, H. Seung-won(2009). Supporting efficient distributed skyline computation using skyline views. *In International Conference on Data Warehousing and Knowledge Discovery (DaWaK)*.

D. Kossmann, F. Ramsak, S. Rost(2002). Shooting stars in the sky: An online algorithm for skyline queries. *In Lochovsky et Shan*, pages 275-286.

D. Papadias, Y. Tao, G. Fu, B. Seeger(2003). An optimal and progressive algorithm for skyline queries. *ACM SIGMOD*. June 9-12, San Diego, California, USA. pp 467-478.

V. Pareto(1896). *Cours d'économie politique*. Volumes 1-2. Lausane, Switzerland.

K. Tan, P. Eng, B. Ooi(2001). Efficient Progressive Skyline Computation. *Proceedings of the 27th International Conference on Very Large Data Bases*, Rome, Italy, September.

Y. Yuan, X. Lin, Q. Liu, W. Wang, J. Xu, J. Yu, Q. Zhang(2005). Efficient computation of the skyline cube. *In VLDB*, pp. 241-252.

Yuanyuan L, ZhiyangLi, Mianxiong Dong,Wenyu Qu,Changqing Ji,Junfeng Wu. Efficient subspace skyline query based on user preference using MapReduce. Ad Hoc Networks, pp 10-115. 2015

L. Zekri, H. Belaicha, Contribution au calcul du skyline par réduction de l'espace candidat. EGC-2015 Luxembourg pp 221-226. 2015

Summary

The skyline queries are important for multi-criteria decision-making. These queries provide a partial order since it is impossible to have a total order on contradictory data. In this paper, as a first step, we recall our point of space reduction method to reduce the number of dominance tests. This reduction is justified because the number of input points may be too high. In a second step, we present results that show the size of the reduction, while in the third stage, we present a new method for calculating the skyline. This method is of type Divide & Conquer, which will be compared with BNL.

Une nouvelle stratégie pour le calcul du skyline sur GPU

Hadjer BELAICHA*, Lougmiri ZEKRI**
Larbi SEKHRI***
* Laboratoire LITIO
belaichahadjer@gmail.com
** Laboratoire LAPECI
lougmiri@gmail.com
*** Laboratoire RIIR,
sekhrilarbi@yahoo.fr
Université d'Oran 1 Ahmed Ben Bella BP 1524, El-M'Naouer, 31000 Oran, Algérie

Résumé. Les processeurs graphiques, ou GPUs Graphics Processing Units, sont devenus un outil important pour le calcul intensif fortement parallèle dans de différents domaines scientifiques et industriels. La raison est que les GPUs offrent la possibilité d'exécuter une instruction sur une grande quantité de données simultanément, réduisant ainsi les temps de traitement. Nous nous intéressons au calcul du skyline d'une base de données multidimensionnelle. Plusieurs travaux se sont intéressés au calcul du skyline sur un ou plusieurs CPUs alors que peu de travaux exploitent les GPUs malgré le parallélisme élevé fourni. Nous présentons une nouvelle stratégie pour le calcul intensif du skyline sur GPU. Nous comparons cette stratégie avec la méthode Divide-and-Conquer qui s'exécute sur le CPU. Les expérimentations menées montrent des gains considérables en termes de temps.

1 Introduction

Les systèmes d'information actuels donnent accès à un grand nombre de sources de données. Lorsque l'utilisateur soumet une requête, il est confronté à des réponses qui sont souvent conflictuelles de façon qu'il ne soit pas facile de décider lesquelles sont les meilleures. Dans ce contexte, le calcul du skyline est devenu un paradigme important afin d'aider l'utilisateur à choisir à partir d'énorme quantité de données disponibles en identifiant un ensemble d'objets de données intéressants. Par exemple, un système de classification des hôtels est interrogé pour trouver les hôtels qui ne sont pas chers et qui sont près de la mer. L'opérateur skyline a été introduit pour résoudre ce type de problème. Il localise l'ensemble d'objets (points) concurrents d'une base de données selon la relation de dominance $\succ$ qui est définie comme suit : Soient D l'ensemble de critères $D = \{1,2,\ldots, d\}$, et p et q sont deux objets de la base de données S, et pour $1< i, j \leq d$, p domine q *noté,* $p \succ q$ *:* si et seulement $\{ \forall d_i \in D / p(i) \leq q(i)\}$ et $\{ \exists d_j / p(j) < q(j) \}$. Lorsque $p \not\succ q$ et $q \not\succ p$ donc p et q sont dits non dominés ou concurrents. A partir l'ensemble de données S, l'opérateur skyline renvoie l'ensemble des points qui ne sont dominés par aucun autre, suivant toutes les dimensions D : Zekri et BELAICHA (2015) $Sky_D(S) = \{p \in S \mid \nexists\, q \in S : q \succ p\}$.

Depuis la définition de l'opérateur skyline en 2001(Börzsonyi et al. 2001), plusieurs travaux ont répondu à ce problème dans de différents systèmes tels que les systèmes centralisés à un seul processeur (CPU), et les systèmes distribués multi-processeurs (Multi-CPU). Bien que les algorithmes proposés aient bien résolu le problème du skyline, leur résolution est

coûteuse en termes de temps de calcul. Elles restent insuffisantes en matière d'exploitation des nouvelles machines parallèles, d'autant plus que les bases de données actuelles sont larges. A fin de remédier à ce problème, les sociétés comme NVIDIA et AMD, ont développé des architectures, permettant le développement et l'exécution de codes généraux sur GPU et d'utiliser les processeurs de la carte graphique comme un outil de calcul puissant. Actuellement, ces capacités sont exploitées pour accélérer la charge de travail dans des domaines multiples. Ainsi, les GPU sont devenus de véritables plateformes de calcul intensif.

Dans ce papier, nous présentons une solution de type BNL qui ouvre plusieurs fenêtres en parallèle. Notre solution exploite au maximum le parallélisme offert par la carte graphique, ce qui permet d'éviter l'oisiveté des processus de laquelle souffrent certains travaux.

Le reste de ce papier se présente comme suit. La section 2 donne un état de l'art du domaine. La section 3 donne notre approche. La section 4 présente les résultats de la comparaison entre notre approche avec DC qui tourne sur le CPU. La section 5 présente la conclusion.

2 Etat de l'art

L'opérateur skyline et son calcul ont attiré beaucoup d'attention depuis 2001 avec l'apparition de (Börzsonyi et al. 2001) qui a adapté le calcul du front de Pareto dans les bases de données. Depuis, plusieurs approches ont été proposées pour rendre le calcul plus efficace. Ces approches peuvent être divisées en deux grandes classes. La première classe regroupe toutes les solutions qui s'exécutent dans le CPU. Ces solutions, elles-mêmes, se divisent en trois sous classes : CPU centralisé, distribué et multicoeurs. Parmi les approches qui s'exécutent dans le CPU centralisé on rencontre; l'algorithme BNL(Block Nested Loop) (Börzsonyi et al. 2001) qui teste chaque point avec tous les autre points. Son problème est qu'il consomme beaucoup de mémoire et du temps d'exécution. L'algorithme DC (Divide and Conquer) (Börzsonyi et al. 2001) divise l'ensemble de données d'entrée en plusieurs partitions. Le skyline partiel de chaque partition est calculé; ensuite, DC fusionne les skylines partiaux pour trouver le skyline final, mais cette fusion cause de multiples testes de duplication. Le NN (Nearset Neighbor) (Kossmann et al. 2002) est une méthode qui utilise les R-Trees comme une structure d'index. NN trouve le point le plus proche à l'origine, ensuite partitionne l'espace selon chaque axe par rapport à ce point. Cet algorithme est bien adapté pour les applications on-line. Il fonctionne efficacement dans un espace à deux dimensions, mais souffre du problème de duplications des éliminations lorsque la dimension dépasse trois Zekri et BELAICHA (2015). (Yuan et al. 2005) ont proposé une nouvelle structure appelée Skycube. Ceci consiste à calculer toutes les combinaisons possibles du treillis afin de générer selon leurs deux algorithmes BUS et TDS le skyline. Aayant remarqué que des points ne feront jamais partie du skyline final, et dans l'objectif de réduire le nombre de tests de dominance, Zekri et BELAICHA (2015) proposent une méthode de filtrage qui permet d'éliminer ce type de points. Plusieurs algorithmes ont été conçus pour les applications distribuées, par exemple (Lee et al. 2009) est un algorithme qui calcule le skyline dans un environnement distribué en utilisant des vues qui stockent un pré-calcul de résultat skyline d'un sous espace de la base de données d'entré. Le but est de minimiser le coût total d'accès en triant l'ensemble des entrées à l'avance et de réduire le nombre de tests de dominance en stockant les points non skyline, mais ces derniers occuperont un espace mémoire inutile vu qu'ils n'ont pas d'importance. Dans les architectures multi-cœurs, les cœurs participent à l'intérieur d'un processeur et communiquent tout simplement en mettant à jour la mémoire principale. APS-

kyline (Linkes et al. 2014) est un algorithme de calcul du skyline dans un système multi-cœurs. Il utilise le modèle de partitionnement à base d'angle. APskyline s'exécute en deux phases comme DC, sauf que pour le partitionnement, un pourcentage configurable de l'ensemble de données est utilisé pour pré-calculer les limites de partitionnement. Dans ce travail, les auteurs proposent une heuristique pour trouver le bon partitionnement qui peut ne pas être la solution optimale.

La deuxième classe regroupe les travaux qui calculent sur GPU. Peu de travaux ont été proposés dans cette classe. (Choi et al. 2012) propose une technique d'optimisation pour la conception d'algorithmes de traitement du skyline qui utilise le modèle SIMD (Single Instruction Multiple Data). Ils ont appliqué l'algorithme BNL dans une version parallèle sur GPU; le but était de minimiser le nombre de threads morts pendant l'exécution. Le GNL se déroule principalement en deux phases. La première consiste à partitionner la base de données en sous ensembles sur les différents blocs où chaque thread teste un point avec tous les autres points d'un même bloc. Dans la deuxième phase, il fusionne les skylines partiaux des différents blocs de la même manière. Cette méthode simple peut être utilisée comme une méthode naïve applicable sur GPU, mais le manque de la communication entre les threads durant l'exécution cause des tests inutiles qui augmentent le temps d'exécution. (Bogh et al. 2013) présentent GGS (GPGPU skyline) qui vise à partager la charge entre CPU et GPU d'une manière à profiter pleinement de leurs capacités. Ils utilisent le CPU pour garder la trace des données à traiter dans le GPU, ainsi que le filtrage des points en éliminant ceux qui sont dominés, tandis que le GPU traite des lots de données d'une taille fixe α où α indique le nombre de points à comparer avec un seul point à chaque itération. GGS trie les données selon une fonction monotone, en suite chaque thread compare chaque point de l'ensemble de données avec les α premiers points du même ensemble, jusqu'à ce que toute la base de données soit traitée. Bien que cet algorithme soit adapté aux applications GPGPU, transfert de données entre CPU et GPU cause un surcoût de communication lorsque la taille de la base de données est considérable, en outre que l'estimation de α dépend du type de données d'entrée.

3 GPU Skyline algorithm

Bien que plusieurs algorithmes aient été proposés pour calculer le skyline, leurs résolutions restent coûteuses en temps d'exécution. Pour réduire ce coût, la plupart des algorithmes skyline pensent à éviter les tests de dominances non nécessaires le plus tôt possible en éliminant les points non skyline ou en partitionnant les données d'entrée (Börzsonyi et al. 2001).

Dans cette partie, nous présentons une nouvelle méthode pour le calcul du skyline sur GPU appelée GSA pour GPU Skyline Algorithm. GSA utilise le modèle SIMD et la fonctionnalité du GPU pour effectuer les tests en parallèle, fusionner les skyline partiaux et enfin, déduire le skyline final. Notre algorithme est conçu pour tenir en compte les limitations de saturation de mémoire et évite l'exécution des tests de dominances inutiles. Nous visons à exploiter les capacités offertes par la carte à savoir, le multithreading, global and shared memories

Au début la base de données est divisée en parties égales, selon le nombre de blocs de la carte graphique. Chaque bloc va charger la partie qui lui a été destinée initialement, jusqu'à son épuisement. Le reste de GSA se déroule en deux phases comme suit

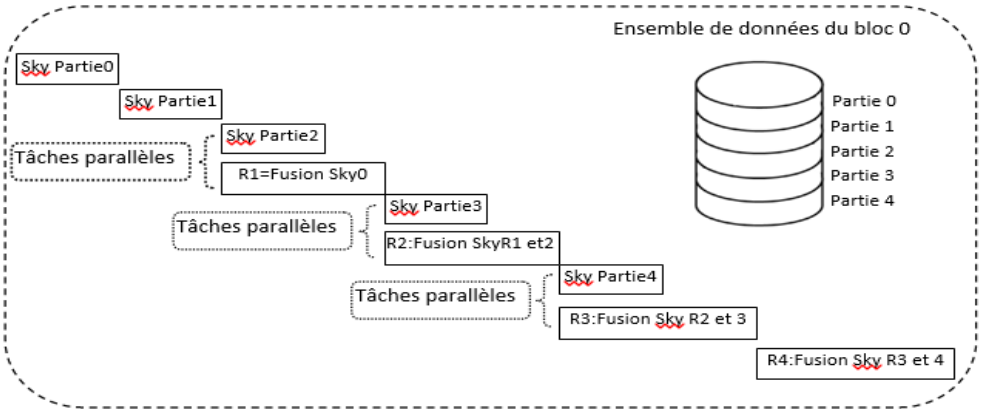


FIG. 1 – *Processus de fusion d'un bloc.*

3.1 Phase 1 : calcul du skyline partiel :

Les données de chaque bloc sont stockées dans la mémoire partagée, où les threads d'un même bloc accèdent librement. Cette phase se déroule en deux étapes parallèles :

1. **Etape 1 : Le calcul du skyline de chaque partie** : Au départ, et pour une raison d'accélération de calcul et gestion de saturation de la mémoire, l'ensemble de données d'entrée initial de chaque bloc est divisé en petites parties. Ensuite, le skyline de chaque partie est calculé de la manière illustrée dans la figure 2 (par multiple de deux) jusqu'à ce que toutes les parties soient parcourues.
2. **Etape 2 : La fusion des résultats des différentes parties** : Une fois le résultat du calcul du skyline des deux premières parties est retourné, la fusion de ces deux derniers commence en parallèle avec le calcul du skyline de la troisième partie.

On répète ces deux étapes jusqu'à ce que toutes les parties soient traitées et le skyline partiel de chaque bloc est retourné. La figure 1 montre la stratégie de calcul dans un bloc.

3.2 Phase 2 : calcul du skyline final :

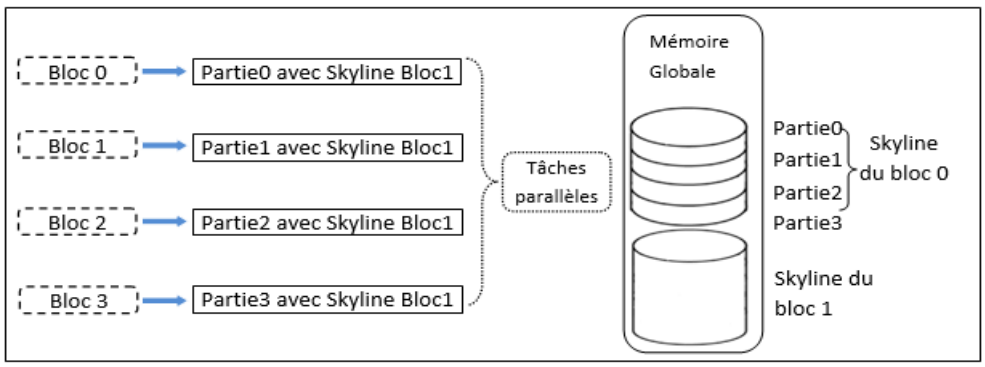


FIG. 2 – *Fusion des skyline des blocs*

Après le calcul du skyline de chaque bloc, des tests de dominance entre les blocs doivent être effectués pour calculer le skyline final. Puisque les threads d'un bloc ne peuvent pas accéder aux données d'un autre bloc, nous avons opté de faire la fusion directement dans la mémoire globale pour éviter la saturation de la mémoire partagée et accélérer le calcul puisque les threads des différents blocs vont partager ce calcul. La fusion est faite Round Robine entre blocs. Ce résultat sera lui-même fusionné avec le skyline du bloc suivant jusqu'à ce qu'on revienne au premier bloc.

Pour accélérer le calcul de la fusion entre les skyline de deux blocs, et une exploitation maximale de la GPU, nous divisons le skyline partiel du premier bloc en N parties (où N est le nombre de blocs) et chaque bloc va être responsable de la partie qu'il va fusionner avec le skyline du deuxième bloc. La figure 2 illustre la fusion du skyline de deux blocs.

	Axe 1	Axe 2	Axe 3
A	1	3	7
B	4	2	8
C	6	2	9
D	11	7	11
E	9	12	9
F	3	10	1
G	7	6	1
H	10	8	3
Y	2	5	6
Z	8	9	4

TAB. 1 – *Exemple de trois critères*

3.3 Exploitation maximale de la GPU :

Réduire le temps d'exécution reste un défi à relever surtout quand le nombre de critères est important. Pour minimiser ce temps, nous chargeons *d* threads pour comparer les points deux à deux. Chaque thread compare *d* critères. Nous disposons aussi d'un thread coordinateur entre les threads d'une seule comparaison. Pour tester la dominance entre deux points de *d* critère il nous faut (*d+1*) threads (voir la figure 3). Ainsi, le traitement sera parallélisé entre les différents threads en faisant une exploitation maximale des ressources de la GPU.

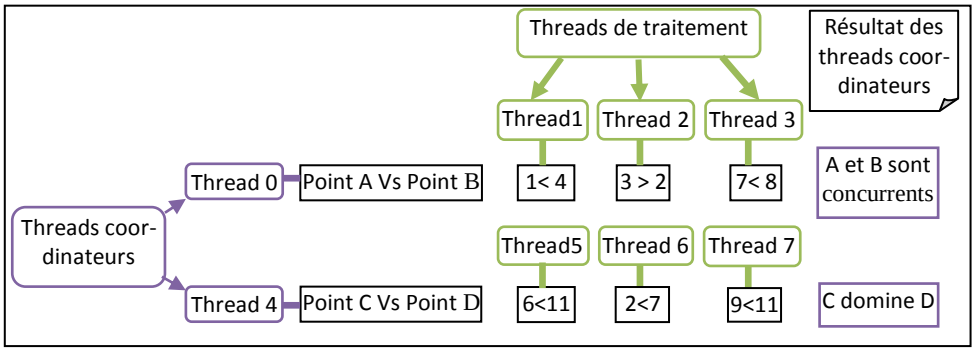


FIG. 3 – *Processus de comparaison des points*

3.4 Gestion de la mémoire:

Le transfert de données entre le CPU et la GPU est coûteux, donc nous réduisons ce trafic en transférant les données une seule fois. Selon l'architecture de la GPU, la latence de lecture de la mémoire globale est relativement lente, donc on utilise la mémoire partagée qui peut être accédée par les threads du même bloc. L'accès à cette mémoire est plus rapide, vu qu'elle est proche des threads, mais elle est d'une capacité limitée. Pour éviter la surcharge de la mémoire partagée, nous avons partitionné les données en plusieurs partitions qui vont être traitées une après l'autre. La communication entre les threads se fait via cette mémoire où la mise à jour se fait rapidement. Ce procédé permet d'éviter les tests inutiles.

4 Expérimentation

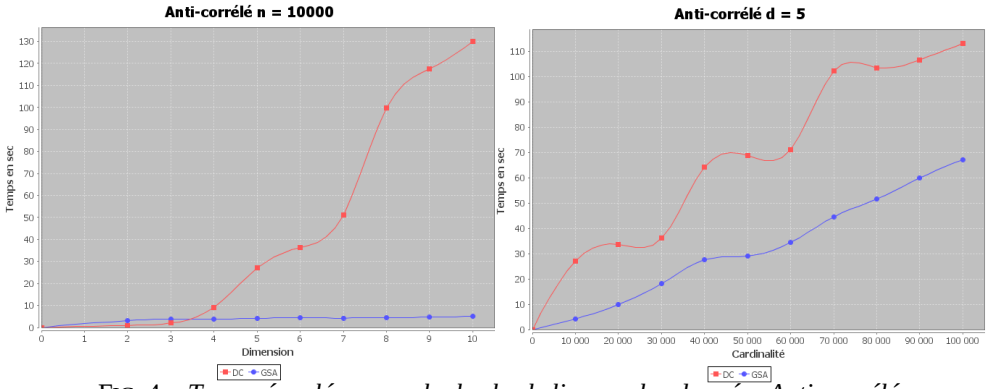


FIG. 4 – *Temps écoulé pour calculer le skyline sur les données Anti-corrélées.*

Les expérimentations ont été réalisées sur une machine dotée d'Intel Core i5 2,50 GHz, et d'une mémoire de 4 Go, équipée d'une carte graphique NVIDIA GeForce GT 525 M avec 2 Go. Sous Ubuntu 12.04. Les mêmes bases de données synthétiques de (Börzsonyi et al. 2001) sont utilisées. Il s'agit de trois types de données : corrélées, anti-corrélées et indépendantes. Nous avons varié la dimension *d* dans l'intervalle [2,10] et la cardinalité *n* dans l'intervalle [10.000,100.000] points. Nous avons comparé GSA et DC selon ces deux paramètres. Nous avons calculé le temps de calcul en millisecondes.

La figure 4 illustre les temps effectués par DC et GSA. Nous remarquons un accroissement des temps effectués par DC, plus la dimension augmente plus le temps augmente. Sur chaque dimension, il calcule la médiane dans un procédé one-way ou multi-way. Ensuite, il calcule le skyline sur dimension. Ce procédé se répète sur toutes les dimensions. Ce qui explique sa montée dans le temps. Les temps effectués par GSA sont stables. Cette stabilité est due au multithreading. Nous proposons d'utiliser autant de threads que de dimensions et chaque thread exécute les tests sur une seule dimension en parallèle à la fois.

En variant la cardinalité, nous remarquons DC a effectué plus temps que GSA. GSA a augmenté son temps, car plus le nombre de points augmente plus il y a des partitions causant plus de swapping entre la shared memory et la global memory. Mais l'utilisation massive du parallélisme a fait que GSA a consommé moins de temps que DC.

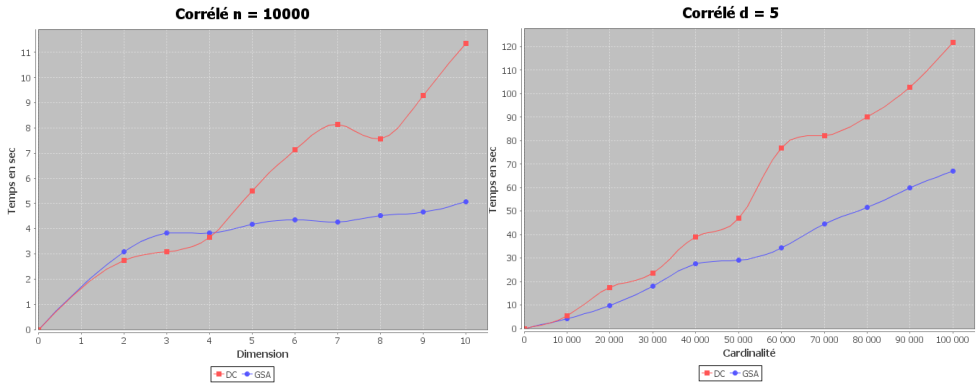


FIG. 5 – *Temps écoulé pour calculer le skyline sur les données corrélées.*

La figure 5 présente l'exécution de DC et GSA sur les données corrélées. Pour les petites dimensions 2, 3 et 4 DC a consommé moins de temps que GSA, mais ce type de problème se fait pour des dimensions larges. On remarque alors que GSA a été plus rapide que DC, suite à la manière avec laquelle on exécute les tests. Les *d* threads comparent les points deux-à-deux sur la même composante.

En variant la dimension, les deux algorithmes ont vu leurs courbes augmentées. Mais GSA reste toujours plus performant. L'exploitation massive du parallélisme est fructueuse. Dans notre, tous les threads sont fonctionnels et se divisent la charge.

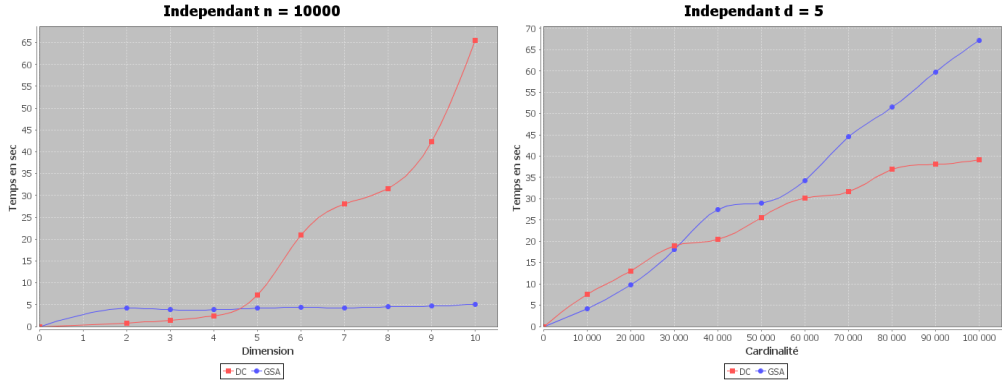


FIG. 6 – *Temps écoulé pour calculer le skyline sur les données indépendantes.*

La figure 6 montre la comparaison entre DC et GSA en termes des temps effectués sur les données indépendantes. En variant la dimension, pour les petites dimensions DC était plus rapide. En réalité, pour un nombre réduit de threads le séquentiel est meilleur car les procédés de synchronisation ralentissent l'exécution. Mais avec l'augmentation des dimensions, GSA devient plus performant et mieux encore fournit des temps stables. Ce constat est le fruit de l'exécution des tests parallèles sur les composantes du même rang des points.

La variation de la cardinalité a fait profiter à DC. En réalité, le type indépendant est difficile à manipuler (Börzsonyi et al 2001), les chercheurs souhaitent éviter ce type de données. Ce type est caractérisé par un nombre important des skyline ; d'où GSA a sauvegardé beau-

coup de point et a fait plus d'accès à la mémoire global qui est lente. C'est pour cette raison, que DC a consommé moins de temps que GSA.

5 Conclusion

Dans ce papier, nous avons traité le problème des requêtes skyline dans la GPU. Ces requêtes sont importantes pour la prise de décisions multicritères. Nous avons défini l'opérateur skyline ainsi qu'état de l'art qui regroupe un ensemble de travaux du domaine.

Nous avons introduit la technique de tests de dominance en utilisant la GPU, qui peut réduire considérablement le coût de ces tests par rapport à la solution CPU. Nous avons aussi fait communiquer les threads pour éviter les tests inutiles en proposant un partitionnement de données pour une bonne gestion de la mémoire. Bien que GSA soit plus performant que DC sur la plupart des données, il est moins de performant sur les données indépendantes.

La solution que nous avons présentée a pris pour but d'éviter que des threads restent oisifs ce qui réduit considérablement les temps d'exécution.

Références

Bogh.k, I.Assents et M.Magnani(2013). Efficient gpu-based skyline computation.In DaMoN'13.

Börzsonyi .S, D.Kossmann et K.Stocker(2001). The skyline operator. In ICDE2001.volume pages 421-430.

Choi.W ,L.Liu et B.Yu(2012). Multi-criteria decision making with skyline computation. In IEEE IRI 2012. Pp 316-323.

Kossmann.D, F.Ramsak et S.Rost (2002).Shooting stars in the sky: an online algorithm for skyline queries. In Lochovsky et Shan.volume pages 475-286.

Lee.J , J.Kim et S.Hwang(2009). Supporting efficient distributed skyline computation using skyline views. In International Conference on Data Ware-housing and Knowledge Discovery (DaWaK). Pp 24-37

Linkes.S, A. Vlachou et C. Doulkeridis (2014). Improved skyline computation for multicore architectures. In DASFAA. Pp 312-326

Yuan.Y, X.Lin et Q.Lui(2005). Efficient computation of the skyline cube. In VLDB, volume pages 241-252.

Zekri.L , H.BELAICHA(2015). Contribution au calcul du skyline par réduction de l'espace candidat. In EGC. Volume pages 221-226.

Summary

Graphics processors units GPU have become an important tool for the highly parallel computing in various scientific and industrial fields. GPUs offer the ability to execute an instruction on a large amount of data simultaneously, reducing processing time. In this paper, we present our solution for computing skyline in GPU. Our method is highly parallel. Compared to D&C, our proposition presents better performance in terms of time execution.

Index des auteurs

A
Abbas M. A., 179
Abdelmalek A., 65, 77, 89, 165, 231
Aliane H., 211
Amghar Y., 203
B
Bachari N. E. I., 179
Badir H., 151
Barr M., 125
Belaicha H., 247
Belbachir H., 203
Belkhir F., 219
Ben Abbes A., 113
Benatia I., 1
Benblidia N., 179
Bencherif K., 101
Bendjenna H., 1
Benharkat A. N., 101
Berrahal S., 101
Bouarara H. A., 89
Bouchabou A., 53
Boudia M. A., 65, 77, 165, 231
Boukhalfa K., 125, 211
Boulkrinat N. H., 53
Bourekkache S., 191
C
Chaibi S., 137
D
Dahmouni E. C. M.,211
Drias H., 53
E
Eom S., 1
F
Farah I. R., 15, 113
G
Gargouri F., 191
Ghazouani F, 15
H
Hamou R. M., 65, 77, 89, 231
Harbi N., 151
Hocine K., 125
I
K
Kabachi N., 151
Kahloul L., 191
Kassimi D., 41
Kazar O., 41, 191
Khellouf H., 53
L
Laouar M. R., 1
M
Malki M., 101
Meddah I.H.A., 231,
Mellah H., 53
Messaoudi W., 15

O
Ougouti N. S., 203
R
Rahmani M. E., 77, 165
Rezoug F., 219
Rhazlane S., 151
S
Saouli H., 41
Sekhri L., 247
Senouci M., 27
Z
Zekri L., 27, 239, 247
Zemani I. M., 27