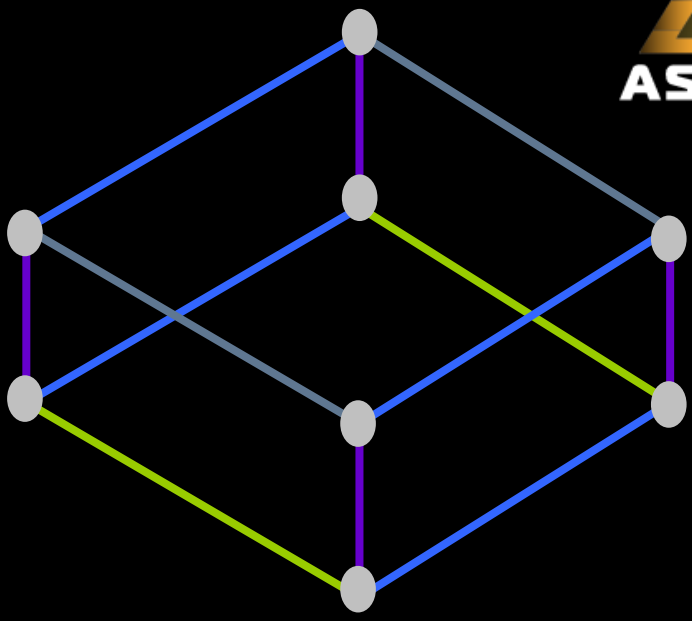
Actes de la Conférence Maghrébine sur les Avancées des Systèmes Décisionnels

ASD'2017



LES SYSTÈMES DÉCISIONNELS

Fondements et Applications

Éditeurs

Sami ZGHAL

Omar BOUSSAID

11ième édition

Conférence sur

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

ASD 2017

ASD 2017

Actes de la 11ième édition

Conférence sur

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

Edités par

Sami ZGAHL et Omar Boussaid

27-29 avril 2017

Tabarka, Tunisie

Préface

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

Forte de son succès graduel et dans le prolongement des éditions précédentes (Agadir–Maroc 2006, Sousse-Tunisie 2007, Mohammedia-Maroc 2008, Jijel-Algérie 2009, Sfax-Tunisie 2010, Blida-Algérie 2012 et Marrakech-Maroc 2013, Hammamet-Tunisie 2014, Tanger-Maroc 2015, Annaba-Algérie 2016), 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 2017, la onzième de son rang, est accueillie cette année par la Tunisie.

ASD 2017 ambitionne de consolider les expériences conduites par les chercheurs, industriels et utilisateurs issus de communautés travaillant sur les systèmes décisionnels. L'objectif de cette onzième 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 2017 a reçu 21 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 2017 est organisée par l'université Jendouba, 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
Sami ZGHAL, Omar BOUSSAID

Comité de pilotage

- BADIR Hassan, (ENSA, Université de Tanger-Tétouan, Maroc)
- BEN ABDALLAH Hanêne (MIRACL, King Abdulaziz University, KSA)
- 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, University of Jeddah, KSA)
- GARGOURI Faiez (MIRACL, Université de Sfax, Tunisie)
- HARBI Nouria (ERIC, Université Lumière Lyon 2, France)

Comité scientifique

- ABDI Mustapha K., Université d'Oran, Algérie
- AHMED OUAMER Rachid, Université Tizi Ouzou, Algérie
- ASFARI Ounas, Université Lyon2, France
- ATMANI Baghdad, Université d'Oran, Algérie
- AYACHI Sonia, ISG, Sousse, Tunisie
- BAAZIZ Abdelhalim, Université Badji Mokhtar, Annaba, Algérie
- BADARD Thierry, Université Laval, Canada
- BADIR Hassan, ENSAT, Maroc
- BADRI Abdelmajid, Université Hassan II, Maroc
- BELLAFKIH Mostafa, INPT Rabat, Maroc
- BELLATRECHE Ladjel, ENSMA Poitiers, France
- BENABES Farouk, Université Badji Mokhtar, Annaba, Algérie
- BEN ABDALLAH Hanene, Université de Sfax, Tunisie
- BENBLIDIA Nadjia, Université de Blida Algérie
- BENHARKAT Nabila, INSA de Lyon, France
- BENSLIMANE Djamel, Université de Lyon1, France
- BENTAYEB Fadila, Université Lumière Lyon 2, 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
- BOUSSAID Omar, Université Lumière Lyon 2, France

- DARMONT Jérôme, Université Lumière Lyon 2, France
- DERRAR Hacene , Université de Blida, Algérie
- FAVRE Cécile, Université Lumière Lyon 2, France
- FEKI Jamel, Université de Sfax, Tunisie
- FERRAG Mohamed Amine, LRS, Université du 8 mai 1945, gUELMA, Algérie
- GARGOURI Faiez, Université de Sfax, Tunisie
- GHOZZI Faiza, Université de Sfax, Tunisie
- HARBI Nouria, Université Lumière Lyon 2, France
- HIDOUCI Walid, ESI Alger, Algérie
- IDRISSI Abdellah, Université Mohammed V, Rabat, Maroc
- KABACHI Nadia, Université Lyon1, France
- KAZAR Okba, Université Biskra, Algérie
- LEMIRE Daniel, Université du Québec à Montréal, Canada
- MAHDAOUIi Latifa, USTHB, Alger, Algérie
- MELIT Ali, Université de Jijel, Algérie
- MEROUANI Hayet Farida, Université Badji Mokhtar, Annaba, Algérie
- MEZIANE Abdelkrim , CERIST, Algérie
- MEZNI Haithem, Université de Jendouba, Tunisie
- 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
- 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
- SASSI Salma, FSJEGJ, Tunisie
- SERIDI Hassina, Université Badji Mokhtar, Annaba, Algérie
- SIDHOM Sahbi, Université de Nancy, France
- TERRISSA Labib, Université Med Khider, Bskra, Algérie
- TESTE Olivier, Université de Toulouse, France
- TISSAOUI Anis, Université de Jendouba, Tunisie
- ZAROUR Nasreddine, Université de Constantine, Algérie
- ZEGOUR Djamel Eddine, ESI Alger, Algérie
- ZGHAL Sami, Université de Jendouba, Tunisie
- ZURFLUH Gille, Université Toulouse Capitole

Comité d'organisation

- ZGHAL Sami, Université de Jendouba, Tunisie
- MOALLA Mohamed Sahbi, ISET Sfax - Tunisie
- SASSI Salma, Université de Jendouba, Tunisie
- TISSAOUI Anis, Université de Jendouba, Tunisie
- MEZNI Haithem, Université de Jendouba, Tunisie

ASD'2017

Conférence sur les Avancées des Systèmes Décisionnels

27-29 avril 2017, Tabarka, Tunisie

Sommaire

Vers une approche pour la prise en compte de l'usager dans l'analyse OLAP

Sadek MENACEUR*, Makhlouf DERDOUR *
Abdelkrim BOURAMOUL **

* Laboratory of Mathematics, Informatics and Systems
Larbi Tebessi University,Tebessa, Algeria
Email : menaceursaddek@gmail.com

* Computer sciences departement
Larbi Tebessi University Tebessa, Algeria
Email: m.derdour@yahoo.fr

** MISC Laboratory, Fundamental Computer Science and its Applications Department
Constantine 2 University, Algeria
Email: abdelkrim.bouramoul@univ-constantine2.dz

Résumé. Les systèmes OLAP sont devenus parmi les solutions prometteuses pour améliorer le processus de prise de décision, d'autant plus que nous voyons une augmentation énorme de volume de données (Big data). Mais de nos jours, ces systèmes s'avèrent inadaptés aux besoins et aux contextes d'analyses des décideurs vus la diversification des particularités des utilisateurs. Cet article présente une nouvelle approche d'analyse en ligne qui prend en compte le concept de personnalisation utilisé pour expliquer comment recevoir à partir d'une grande quantité d'informations uniquement la partie qui intéresse un usager et qui reflète son besoin réel, et ceci à la base d'un ensemble de facteurs définis au départ tels que les besoins et les exigences de l'usager, le profil usager et le contexte de requête. Ces facteurs dirigent la formulation des besoins et les exigences de l'usager sous forme de requêtes fonctionnelles et non fonctionnelles. Les requêtes non fonctionnelles sont utilisées pour réduire ou personnaliser l'espace de recherche et minimiser toutes sortes de masse d'informations inutiles ou bruitées, Alors que l'exploitation des requêtes fonctionnelles réunies au profil de l'usager et du contexte de sa requête conduit à la création d'un cube de données pour une éventuelle session d'analyse OLAP portant sur les informations personnalisées déjà obtenues

Mots clés
Les systèmes OLAP, Big data, profil usager, contextes de requêtes, personnalisation.

1 Introduction

De nos jours, le Big Data est l'un des domaines de recherche les plus sollicités au monde. Ce terme se traduit littéralement par « Grosses Données » ou « Masse de Données ». En date du 22 Août 2015 la commission générale de terminologie et de néologie a associé une définition officielle au terme Big data qui est la suivante « données structurées ou non dont le très

grand volume requiert des outils d'analyse adaptés ». En effet, ces outils d'analyse permettent aux décideurs d'aller plus loin dans l'analyse de données dans le but de prédire de nouvelles valeurs ou de classifier les données. Selon Aubay. (2015), analyser consiste à déterminer les corrélations entre les données dans le but d'extraire des valeurs utiles, des suggestions ou des décisions qui reflètent les besoins réels de l'usager.

Ces besoins suscités sont difficiles à exprimer dès le départ par une requête précise quand il s'agit d'une grande masse de données. Un usager dans un processus d'analyse en ligne OLAP (On Line Analytical Processing) fait recoure à l'utilisation de requêtes ayant des structures complexes afin de trouver l'information pertinente à son besoin. Ce type d'analyse rend la recherche fastidieuse. Par conséquent, Il n'y a pas mieux que de spécifier ce que l'usager cherche exac-tement, et ce, par l'expression de ses préférences. Ces dernières sont liées aux données comportementales de l'usager et représentent ce qu'on appelle profil usager si l'on ajoute les intérêts, les usages, les contraintes, le contexte, etc.

Dans cet article, nous proposons une nouvelle approche d'analyse en ligne, qui s'inscrit dans le domaine de la personnalisation des entrepôts de données, impliquant les connaissances de l'usager tel que ses besoins, son profil, etc. Dans le but de fournir des analyses personnalisées pour recevoir à partir d'une grande quantité d'informations uniquement la partie intéressante. Notre approche est fondée sur trois concepts complémentaires: ***profil, contexte et préférences*** pour minimiser toutes sortes de masse d'informations inutiles ou bruitées et conduire ainsi à une session d'analyse OLAP personnalisée.

Le reste de cet article est organisé comme suite; dans la section 2, nous introduisons nos motivations et le contexte de recherche. Puis, nous présentons brièvement dans la section 3 les différents aspects sur lesquels se base notre proposition, à savoir le profil usager, le contexte système et les préférences usager. Un état de l'art relatif au problème de personnalisation est présenté dans la section 4, Par la suite nous détaillons dans la section 5 l'approche que nous proposons pour la prise en compte de l'usager dans l'analyses OLAP. Enfin, nous concluons cet article et nous évoquons les perspectives de ce travail dans la section 6.

2 Motivations et contexte de travail

Ces dernières années, le sujet de profilisation et contextualisation dans le processus d'analyse en ligne (OLAP) occupe une grande part dans le domaine de l'informatique décisionnelle (Business Intelligence), d'autant plus que le monde aujourd'hui vit une croissance massive dans le volume des données (Big Data) qui a été décrite très souvent comme des données qui dépassent les capacités de l'organisation à stocker ou à analyser dans le but de prendre une décision précise et opportune (Kulkarni, 2013). Dans la littérature le terme Big data a été caractérisé comme ayant une ou plusieurs des quatre dimensions: *le volume, la vitesse, la variété et la véracité* (Laney, 2001 ; IBM, 2014 ; Goes, 2014). Ces dimensions présentent en réalité les grands défis lorsqu'il s'agit de l'analyse des données. Partant de ce constat, nous souhaitons répondre à la problématique suivante : *Parmi le volume important des données stockées, comment formuler une requête qui répond au mieux aux besoins de l'usager et comment renvoyer ensuite un résultat pertinent ?*

3 Etat de l'art

L'analyse en ligne OLAP consiste à exploiter intuitivement les entrepôts de données, néanmoins actuellement ces systèmes ont peu des connaissances sur les usagers. Cela va conduire indirectement à une dégradation dans la performance du processus décisionnel. De ce fait, l'intégration de l'usager dans l'analyse OLAP a fait l'objet de nombreuses recherches (Koutrika et Ioannidis, 2004), (Stefanidis, Pitoura, 2008), (Bentayeb, Favre, Boussaid, 2008), (Jerbi, Ravat, Teste, Zurfluh, 2008), ce type d'intégration permet d'afficher le contenu informationnel pertinent vis-à-vis aux intérêts de l'usager. Cependant, cette pertinence est définie par des éléments **contextuels** directement liés à l'usager, tels que ses centres d'intérêts, ses **préférences** de recherche, etc, l'ensemble de ces éléments est stocké dans une structure appelée *profil usager*.

3.1 Profilisation

Un profil usager apparait comme étant à la base de la personnalisation, il subit des définitions selon le domaine d'utilisation de la personnalisation, Par exemple, dans le domaine de la recherche d'information, les profils sont généralement représentés sous forme de mots clés pondérés (Ferreira et Silva, 2001; Soltysiak et Crabtree, 1998) ou d'un ensemble de fonctions d'utilité sur un domaine d'intérêt (Cherniack et al., 2003), tandis qu'en interaction homme-machine (IHM), les profils contiennent des informations qui vont permettre au système d'adapter l'affichage des résultats selon les préférences de l'usager. Dans le domaine des bases de données, les profils peuvent contenir des conditions de sélection ou de jointure des requêtes SQL (Koutrika et Ioannidis, 2004, 2005a).

D'après le travail de (Bentayeb, Boussaid, 2009), un profil est caractérisé par deux perspectives :

1. La première porte sur *l'implication de l'usager.* Elle peut être explicite, ou implicite. Lorsqu'il s'agit d'une implication explicite, l'usager doit effectuer des interactions directes avec le système, tandis que lors d'une implication implicite le système s'adapte automatiquement à l'usager.

   Ce que nous observons ici, c'est que le mode d'acquisition de l'implication explicite est le plus facile à mettre en œuvre et permet à l'usager de saisir manuellement des informations utilisées dans la construction de son profil. En revanche, le mode d'acquisition implicite repose sur des techniques d'extraction des informations basées sur des mesures de pertinence implicite appliquée sur l'historique d'interactions de l'usager.
2. La deuxième perspective concerne *les fonctions systèmes liées au profil*, elle consiste à définir le profil comme un premier pas, et ensuite, exploiter ce dernier pour une meilleure prise en compte de l'usager.

L'exploitation du profil peut ; soit nécessiter l'intervention explicite de l'utilisateur qui transforme le système par des choix de recommandations, soit induire une transformation automatique du système. La figure suivante décrit le processus de prise en compte de l'usager dans les entrepôts de données, Mais, *Est-ce que la définition du profil seulement est suffisante pour personnaliser un entrepôt de données ?* Cette question et d'autres seront répondues dans la section 3.2.

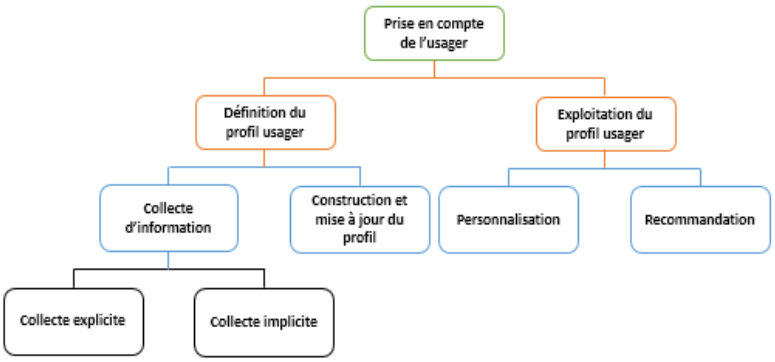


FIG. 1 – *Le processus de prise de compte de l'utilisateur, (Khemiri, 2015)*

3.2 Contextualisation

Définir le profil de l'usager est un critère indispensable, mais n'est pas assez suffisant pour la personnalisation dans les entrepôts de données, il est souvent liée à d'autre critères tel que les ***préférences*** et le ***contexte***. En ce qui concerne les préférences elles sont liées fortement à son profil et on ne peut en aucun cas séparer les uns des autres, mais, leur description à la fois peut changer en fonction du contexte. Une préférence peut être associée à un contexte, dans ce cas, elle est dite contextuelle (ou conditionnelle). Le contexte d'une préférence définit sa portée, c'est-à-dire l'environnement dans lequel elle doit être prise en compte. Donc on peut noter qu'une préférence contextuelle est un couple (P, C), où P est une préférence et C est un contexte.

La partie contexte spécifie les conditions sous lesquelles la préférence P sera activée, où P peut être formulée selon une approche quantitative ou qualitative. Dans la littérature Plusieurs définitions du contexte ont été proposées (Brown et al. 1997; Schmidt et al. 1999). Une définition générale est la suivante:

« Le contexte est toute information susceptible de caractériser la situation d'une entité. Une entité est une personne, un lieu ou un objet qui est considéré pertinent pour l'interaction entre l'usager et l'application, incluant l'usager et l'application » (Dey, 2001).

En fin, il y a une ambiguïté autour des trois concepts : ***profil, contexte et préférences***. Le sens qu'on leur donne change d'une approche à une l'autre et il arrive souvent que l'un d'entre eux soit utilisé à la place des deux autres ou des trois à la fois. Cette ambiguïté de la terminologie rend difficile l'étude et la compréhension de la problématique liée à la personnalisation.

3.3 Personnalisation dans les entrepôts de données

La problématique mentionnée dans la section 2 est connue sous le nom de la personnalisation qui consiste à impliquer les connaissances sur l'usager (ses besoins, son profil, etc.) dans le processus de l'entreposage, que ce soit au niveau de la phase de conception de l'entrepôt, de création ou de l'exploitation des cubes de données, pour permettre à la fin à cet usager d'obtenir des informations pertinentes relatives à son besoins (Domshlak et Joachims, 2007).

Korfhage (1997), dans son agenda de recherche affirme que la personnalisation d'un système consiste à définir, puis à exploiter un profil usager qui ne peut être définie de façon standard, il regroupe souvent un ensemble de caractéristiques servant à configurer ou à adapter le système à l'usager.

Alors, la personnalisation dans les entrepôts des données (ED) a fait l'objet de très nombreux travaux de recherche, et peut se situer à plusieurs niveaux dans un système OLAP; elle peut porter sur le schéma de l'entrepôt de données, la visualisation des données et/ou sur l'interrogation. Tous ces niveaux sont principalement basés sur les profils des usagers et les techniques de recommandation.

3.3.1 Personnaliser un entrepôt de données à base d'un profil usager

Le système de personnalisation dans ce cas repose sur les besoins, les préférences et les caractéristiques des usagers (Ioannidis et Koutrika, 2005), et généralement sur des profils usagers définis (Korfhage, 1997). Il est mentionné précédemment qu'il n'existe pas de consensus pour la définition d'un profil usager, mais un profil comprend généralement un ensemble de fonctionnalités utilisé pour configurer ou adapter le système à l'usager. Ainsi, le système fournit des résultats personnalisés et efficaces (Domshlak et Joachims, 2007) adaptés à un profil usager.

D'autres recherches utilisent les préférences des usagers définies dans leurs profils (Bellatreche et Giacometti, 2005 ; Golfarelli, 2010; Jerbi, Ravat, Teste, Zurfluh, 2008) pour configurer ou adapter le système de personnalisation. Ces préférences peuvent aussi être liées à leurs contextes définissant les cadres d'application des dites préférences (Jerbi, Ravat, Teste, Zurfluh, 2008 ; Garrigos et al, 2009).

Les auteurs (Jerbi, Pujolle, Ravat, et Teste, 2011) distinguent trois objectifs principaux des recherches de personnalisation dans l'entrepôt de données:

1. Personnaliser le schéma des sources de données (Garrigos et al, 2009 ; Bentayeb, Boussaid, 2009), en adaptant les structures de données à des besoins spécifiques des usagers.
2. Personnalisation de la visualisation des requêtes (Bellatreche et Giacometti, 2005), ou représentation (Golfarelli, 2010 ; Jerbi, Ravat, Teste, Zurfluh, 2008).
3. Recommandation de requêtes OLAP (Giacometti, Marcel, Negre, 2009) pour aider à l'exploration des entrepôts des données.

Les deux premiers objectifs semblent affecter la personnalisation centrée sur les données, dans le premier cas en personnalisant le schéma et dans le second cas en représentant des résultats de requêtes personnalisées. Le troisième objectif concerne la recommandation d'une nouvelle méthode de traitement des données, des requêtes.

3.3.2 Personnaliser un entrepôt de données par recommandation ou transformation

La personnalisation par recommandation est l'axe le plus émergent dans la personnalisation des entrepôts des données, il est traitée par divers travaux tels que (Bentayeb, Boussaid, 2009 ; Giacometti, Marcel, et Negre, 2008 ; Giacometti, Marcel, Negre, 2009 ; Chatzopoulou, Eirinaki, Polyzotis, 2009). Dans ces travaux, nous distinguons deux catégories de méthodes de recommandation, les méthodes basées sur le contenu recommandent des objets similaires qui sont basée sur des actions antérieures de l'usager, tandis que les méthodes de

recommandation basées sur le filtrage collaboratif recommandent des éléments en fonction de l'intérêt et de l'usager similaire.

La personnalisation d'un entrepôt des données par transformation est mentionnée par les auteurs dans (Bellatreche et Giacometti, 2005) qui traite la visualisation personnalisée des requêtes OLAP. Les auteurs (Favre, C. Bentayeb, F. et Boussaid, 2007) proposent une solution pour faire évoluer le schéma de l'entrepôt de données en fonction des besoins des usagers. Cette méthode est basée sur des règles "si-alors". Enfin, le travail de recherche dans (Thalhammer, Schrefl, et Mohania, 2001) propose une solution pour développer l'architecture de l'entrepôt de données avec des règles d'événement / condition / action.

4 Une approche d'analyse orientée usager

L'architecture globale de l'approche que nous proposons s'inscrit dans le domaine de la personnalisation de l'analyse en ligne des entrepôts de données, elle comprend trois phases (Figure 2). Dans la première on s'intéresse à la *formulation des requêtes* sur la base des besoins et des préférences usager, où un module de formulation des requêtes permet de spécifier l'ensemble de requêtes fonctionnelles et d'autres non fonctionnelles. La deuxième c'est une phase de *réduction de l'espace de recherche*, elle permet d'exploiter l'ensemble des requêtes non fonctionnelles pour personnaliser le schéma de l'entrepôt de données et éviter toutes sortes de masse d'informations inutiles ou bruitées. Enfin une phase *d'analyse OLAP* sera déclenchée en introduisant l'ensemble des requêtes fonctionnelles précédemment définies. Un cube de données devient disponible pour le décideur pour une éventuelle phase d'exploitation.

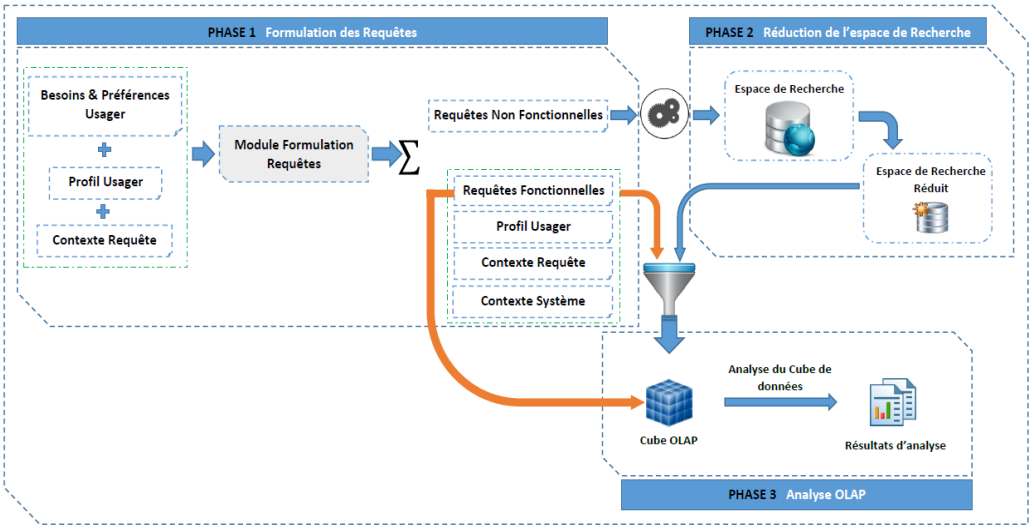


FIG. 2 – *Approche pour la prise en compte de l'usager dans l'analyse en ligne des entrepôts de données*

5 Conclusion et travaux futurs

Le travail présenté dans ce papier se situe dans le contexte de la personnalisation de l'analyse en ligne des entrepôts de données, et plus particulièrement dans le cadre de la personnalisation à base des profils usagers. Notre contribution consiste à présenter un état de l'art dans lequel on décrit les différents concepts de base définissant ce type de personnalisation, tel que, *le profil, le contexte et les préférences*. Ensuite on propose une nouvelle approche fondée sur trois phases, où le profil de l'usager joue un rôle très important dans chacune d'elles. Notre approche présente plusieurs avantages: (1) l'usager devient alors un réel acteur du processus décisionnel, puisque c'est lui qui va gérer le processus de la personnalisation depuis la formulation des requêtes jusqu'à l'obtention de l'information pertinente à son besoins, (2) personnaliser le schéma de l'entrepôt de données avant l'exécution de requête de l'usager donnera des bons résultats lors de la création de cube de données. Nous envisageons dans de travaux de recherches futures de réaliser cette approche, et de la valider à l'aide d'un modèle d'entrepôt de données complexe pour bien présenter les résultats.

Références

Aubay. (2015) Le Big Data. Retrieved from http://www.aubay.com/wp-content/uploads/2015/03/Regard-Aubay-Big-Data-Web.pdf

Bellatreche, L. Giacometti, A. Marcel, P. Mouloudi, H. et Laurent, D. A personalization framework for OLAP queries, The Eighth International Workshop on Data Warehousing and OLAP (DOLAP 2005), 2005, pp. 9-18.

Bentayeb, F., Favre, C., Boussaid, O. (2008). A user-driven data warehouse evolution approach for concurrent personalized analysis needs. Integrated Computer-Aided Engineering, 15(1) :21_36.

Bentayeb, F., Boussaid, O., Favre,C,. Ravat,F,. Teste, O.2009. Personnalisation dans les entrepôts de données : bilan et perspectives, 5eme journées sur les Entrepôts de Données et l'Analyse en ligne (EDA'09), Revue des Nouvelles Technologies de l'Information, RNTI-B-5, Cepadues Editions.

Brown, P., Bovey, J., Chen, X. (1997). Context-aware applications: From the laboratory to the marketplace. IEEE Personal Communications, Vol. 4, No. 5, pages 58–64.

Cherniack, M., Galvez, E.F., Franklin, M.J., Zdonik, S.B. (2003). Profile-Driven Cache Management. Intl. Conf. on Data Engineering (ICDE), IEEE Computer Society, pages 645–656.

Chatzopoulou, G. Eirinaki, M. and Polyzotis, N. "Query recommendations for interactive database exploration," The 21st International Conference on Scientific and Statistical Database Management (SSDBM 2009), 2009, pp. 3-18.

Dey, A. K. (2001). Understanding and using context. Personal and Ubiquitous Computing, Vol. 5, No. 1, pages 4–7.

Domshlak C., T. Joachims (2007). Efficient and Non-Parametric Reasoning over User Preferences. User Modeling and User-Adapted Interaction 17(1-2), 41–69.

Favre, C. Bentayeb, F. et Boussaid, O. Evolution et personnalisation des analyses dans les entrepôts de données. Une approche orientée utilisateur, The 25th Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID 2007), 2007, pp. 308-323.

Ferreira J., Silva A. (2001). MySDI: A Generic Architecture to Develop SDI Personalised Services. Intl. Conf. on Enterprise Information Systems (ICEIS), pages 262–270.

Garrigos, I. Pardillo, J. Mazon, J.-N. et Trujillo, J. "A conceptual modeling approach for OLAP personalization," in Conceptual Modeling-ER Verlag Berlin Heidelberg, 2009, pp. 401-414.

Giacometti, A. Marcel, P. et Negre, E. "A framework for recommending OLAP queries," Proc. The eleventh international workshop on Data warehousing and OLAP (DOLAP 2008) ACM, Pages 73-80, doi:10.1145/1458432.1458446.

Giacometti, A. Marcel, P. et Negre, E. "Recommending multidimensional queries," The 16th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2009), 2009, pp. 453-466.

Golfarelli, M. "From user requirements to conceptual design in data warehouse design – a survey," Data Warehousing Design and Advanced Engineering Applications: Methods for Complex Construction, IGI Global, 2010.

Golfarelli, M. et Rizzi, S. Expressing olap preferences. In SSDBM, pages 83_91, New Orleans, Louisiana USA, 2009.

Goes, P. (2014). Editor's comments: Big data and IS research. MIS Quarterly, 38(3), iii-viii.

IBM. (2014). The four V's of big data. Retrieved from http://www.ibmbigdatahub.com/infographic/four-vsbig-data

Ioannidis. Y, et Koutrika. G, "Personalized systems: models and methods from an ir and db perspective," The 31st conference in the series of the Very Large Data Bases conferences (VLDB 2005), pp. 1365–1365.

Jerbi, H., Ravat, F., Teste, O., Zurfluh, G. "Management of context-aware preferences in multidimensional databases," the Third International Conference on Digital Information Management (ICDIM 2008), pp. 669-675.

Jerbi, H., Ravat, F., Teste, O., Zurfluh, G. (2009). Preference-based recommendations for olap analysis. In DaWaK, pages 467_478.

Jerbi, H. Pujolle, G. Ravat, F. et Teste, O. "Recommandation de requêtes dans les bases de données multidimensionnelles annotées," Revue des Sciences et Technologies de l'Information, Ingénierie des Systèmes d'Information, vol. 16, no. 1, pp. 133-138, 2011.

Khemiri, R. (2015) Vers l'OLAP Collaboratif pour la recommandation des analyses en ligne personnalisée, Thèse Doctorat, Univ Lyon2 France, 23 Septembre.

Korfhage, R. Information Storage and Retrieval, John Wiley & Sons, 1997

Kulkarni, R. (2013). Transforming the data deluge into data-driven insights: Analytics that drive business. Keynote speech presented at the 44th Annual Decision Sciences Institute Meeting, Baltimore, MD

Koutrika, G., Ioannidis, Y. E. (2004). Personalization of queries in database systems. In ICDE, pages 597_608.

Koutrika, G. Ioannidis, Y. E. (2004). Personalization of queries in database systems. Intl. Conf. on Data Engineering (ICDE), IEEE Computer Society, pages 597–608.

Koutrika, G., Ioannidis, Y. E. (2005a). Personalized queries under a generalized preference model. Intl. Conf. on Data Engineering (ICDE), IEEE Computer Society, pages 841–852.

Laney, D. (2001). 3-D data management: Controlling data volume, velocity and variety. Gartner. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Searby, S. (2003) Personalisation - an overview of its use and potential. BT Technology Journal, 21(1):13–19.

Stefanidis, K., Pitoura, E. (2008). Fast contextual preference scoring of database tuples. In EDBT, pages 344_355.

Soltysiak S., Crabtree B. (1998). Automatic learning of user profiles - Towards the personalisation of agent services. BT Technology Journal, Vol 16, No 3, pages 110–117.

Schmidt, A., Aidoo, A. K., Takaluoma, A., Tuomela, U., Laerhoven, K., and de Velde, M. (1999). Advanced interaction in context. Intl. Symposium on Handheld and Ubiquitous Computing, pages 89–101.

Thalhammer, T. Schrefl, M. et Mohania, M. “Active data warehouses: complementing OLAP with analysis rules,” Data & Knowledge Engineering, vol. 39, pp. 241-269, 2001.

Summary

OLAP systems have become among the promising solutions to improve the decision-making process, especially as we see an enormous increase in data volume (Big data). But nowadays, these systems are not adapted to the needs and analytical contexts of decision-makers due to the diversification of the peculiarities of the users. This article presents a new online analysis approach that takes into account the concept of personalization used to explain how to receive from a large amount of information only the part that interests a user and that reflects his real need, On the basis of a set of factors defined at the outset, such as the needs and requirements of the user, the user profile and the request context. These factors guide the formulation of needs and user requirements in the form of functional and non-functional requests. Non-functional queries are used to reduce or personalize the search space and minimize all kinds of mass of unnecessary or noisy information. While the use of functional queries combined with the user profile, and the context of its request leads to the creation of a data cube for a possible OLAP analysis session on the personalized information already obtained.

Keywords

OLAP systems, Big data, user profile, query contexts, personalization

Modélisation de préférences à base de croyance du décideur

LARBI A.[1,2] , MALKI M. [1,2], BEN AHMED A.[3], Boukhalfa K.[4] , LarbiO.[3]
[1]Ecole Supérieure Informatique – Sidi Bel Abbes
[2] Laboratoire LabRI, UDL Sidi Bel-Abbes
3Laboratoire ENERGARID , Université de TAHRI Mohammed Béchar – Algérie
[4]Laboratoire ISL, USTHB Algérie

Résumé

Dans un contexte OLAP, une grande quantité d'informations est fournie à l'utilisateur. La majorité de ces informations n'est pas intéressante ni utile à l'analyse. Afin de régler ce problème, on a pensé à personnaliser les préférences et les requêtes MDX pour une meilleure satisfaction des besoins décisionnels.

Dans cet article une approche quantitative représentée par l'emploi des outils de la théorie de croyance est adoptée pour ce cas en exploitant l'historique de la session utilisateur représentant le fichier log des requêtes entrées précédemment par celui-ci afin de former une expression de préférences.

- *Mots-clés : OLAP, personnalisation, profil, entrepôts de données, théorie d'évidence.*

I. Introduction

Les systèmes OLAP facilitent l'analyse en offrant un espace multidimensionnel des données que les décideurs explorent interactivement par une succession d'opérations OLAP.

Ces systèmes fournissent à l'utilisateur une grande quantité des informations qui peuvent être résumées ou détaillées. Dans la majorité des cas, l'utilisateur ne sera pas intéressé par tous ces informations, ce qui rend la recherche d'une information pertinente plus difficile. Pour cela les chercheurs pensent à personnaliser la requête ou le besoin décisionnel.

La personnalisation de requête consiste à fournir à un utilisateur une information pertinente correspondant à ses préférences et à ses besoins. Les données décrivant ces besoins des utilisateurs sont souvent regroupées sous forme d'un profil**.** Pour construire le profil utilisateur : on peut soit exploiter le fichier log d'usager, soit utiliser les techniques de data mining pour cela et soit compter sur l'utilisateur pour le définir. D'une autre part, le besoin utilisateur varie au cours du temps, ainsi on peut dire qu'un utilisateur peut avoir plusieurs profils en fonction du temps.

Ce papier présente en section 2 un état d'art sur la personnalisation de requêtes puis en section 3 une présentation de l'approche proposée à base de croyance du décideur suivi d'un exemple illustratif de cette proposition et représentant un test préliminaire et enfin une conclusion.

II. Etat d'art

Beaucoup de chercheurs étaient intéressés par l'étude de la personnalisation dans différentes domaines. L'auteur dans [13], analyse l'impact des facteurs de qualité dans la personnalisation de l'information, puis il détaille son incorporation dans un méta modèle de

profil. L'auteur dans [14] décrit et évalue deux approches de reformulation de requêtes sur la base de deux métriques qui sont la couverture et la précision des prédicats utilisés dans l'enrichissement de la requête utilisateur. Le cœur de cette problématique est la reformulation de requêtes. La solution proposée tient compte à la fois du profil utilisateur et de la description des sources de données.

L'auteur dans [8] a signalé l'applicabilité de la personnalisation sur les éléments du schéma concernant cinq approches existantes: Constructeurs de préférence, la personnalisation dynamique, L'OLAP visuel, La recommandation en analysant les sessions et la recommandation en analysant le profil utilisateur.

Plusieurs solutions ont été mises en place pour satisfaire les besoins des décideurs. On peut distinguer plusieurs critères pour classer ces solutions [1]:

Selon la manière d'obtention des profils utilisateurs, certains travaux préfèrent une définition explicite du profil tandis que les autres le génèrent selon le contexte et l'historique de l'utilisateur,

Un autre critère prend en charge la clause où l'on exprime les préférences, certains travaux utilisent la clause WHERE qui est considérée comme une contrainte stricte cependant que les restes préfèrent la clause PREFERRING qui est une contrainte légère, c'est-à-dire que le résultat de la requête doit satisfaire le plus que possible les préférences d'usager.

Selon le but final du processus, certain travaux consistent à transformer la requête cependant que d'autres choisissent de recommander d'autres requêtes à l'utilisateur.

Selon la formulation du profil, la majorité des travaux consistent à définir un profil utilisateur d'une façon explicite : l'utilisateur a la main d'exprimer ses vœux et ses préférences sur les éléments du schéma c'est-à-dire de faire basculer l'un des éléments de l'autre. Deux approches sont utilisées pour cela, une approche qualitative qui est basée sur une comparaison binaire entre les éléments, une autre approche dite quantitative qui consiste à affecter des poids aux éléments. Le reste des solutions évite l'intervention de l'utilisateur. Au lieu d'exprimer les préférences explicitement, le système exploite les fichiers log, l'historique de sessions ouvertes par l'utilisateur et extrait les éléments requêtés précédemment pour former un profil utilisateur, on parle donc d'un profil défini d'une façon implicite.

L'absence de l'intervention de l'utilisateur concernant son profil est montrée clairement dans le travail [2] qui traite la recommandation des requêtes d'une façon que l'on prévoit la prochaine requête entrée par l'utilisateur. L'application de la première approche quantitative est montrée dans [3]. Ce travail utilise un mécanisme basé sur les règles (ECA Evénement Condition Action). La solution [4] consiste à exprimer les préférences du décideur au niveau de la clause GROUP BY en utilisant une algèbre prédéfinie. Dans le domaine de web, l'auteur dans [9] propose un système adaptable qui prend en considération le comportement de l'utilisateur (les clics sur les liens, la séquence des clics, la durée de navigation dans la page …etc.) pour personnaliser le site en ajoutant par exemple les liens souvent visités à la page d'accueil. Notons que le système est basé sur les règles ECA. L'auteur dans [6] propose un modèle pour la recommandation des requêtes en se basant sur des préférences exprimées

à priori par l'utilisateur. L'auteur dans [15], décrit un processus de personnalisation de requêtes décisionnelles à travers une approche d'extraction de règles d'association triadiques.
Il existe d'autres tentatives de personnalisation qui prennent en considération la manière d'affichage d'informations. Ces solutions appelées Visuel OLAP exonèrent l'utilisateur d'écrire la requête MDX, en interagissant avec une interface utilisateur qui lui donne la main de transporter des éléments (dimensions, attributs, mesures …) de navigateur vers l'aire de visualisation, une requête MDX est générée par conséquence [7].
La solution prouvée dans [5] consiste à définir à priori un pré-ordre sur les membres des attributs, un attribut pour chaque dimension afin d'exprimer les préférences d'usager. Par suite, on exploite cet ordre pour chercher un cube dit optimal d'un ensemble de cubes équivalents et on va l'afficher suivant une contrainte de visualisation définie par l'utilisateur pour une meilleure représentation en prenant en considération les dispositifs utilisés (PCA, Mobile, …etc.).
La proposition [5] est une des approches qui comptent sur l'utilisateur pour définir ce profil. Cette méthode a plusieurs limites. On cite qu'elle suppose que l'utilisateur est un connaisseur soit concernant la structure du cube de données soit respectant les membres de ce dernier. La plus part des cas, les décideurs ne sont pas des experts, parfois ils ignorent le modèle de sortie des résultats retenus. Dans ce cas, on ne peut pas compter sur eux pour définir des profils mais on doit les extraire en exploitant l'historique des requêtes de l'utilisateur. D'une autre part, Cette méthode traite seulement les dimensions avec un seul attribut – ce n'est pas toujours le cas – défait c'est jamais le cas ! Comme elle est préoccupée par seulement les attributs catégoriaux et non numérique (les mesures).

III. Préférences à base de croyance

Notre but est de développer un système adaptable. Pour cela, on a proposé une méthode pour extraire les préférences d'utilisateur à partir de l'historique de la session en extrayant les références des requêtes enregistrées dans le log, calculant la masse de chaque référence, puis on les ordonne pour obtenir un ordre sur les membres. Pour cela on utilise la théorie de l'évidence (la croyance) qui contient des outils nous permettent de calculer la masse d'un membre mentionné dans plusieurs requêtes et que d'autres membres partageant la même requête.
La théorie des fonctions de croyance a été développée par Shafer en 1976 à la suite des travaux de Dempster sur les probabilités inférieures et supérieures. Philippe Smets a ensuite énormément contribué au développement de cette théorie grâce à son modèle des croyances transférables (appelé aussi TBM pour Transferable Belief Model), dérivé de celui de Shafer.
La théorie des fonctions de croyance peut être adoptée pour de nombreuses raisons :
- ✓ elle permet de prendre en compte et de modéliser à la fois l'imprécision, l'incertitude et l'incomplétude.
- ✓ elle permet de représenter plusieurs types de connaissance, ce qui offre un cadre riche et flexible,

✓ elle permet de mettre en évidence et de gérer le conflit entre les connaissances,
✓ elle possède des outils qui permettent la prise de la décision.

Notre application de la théorie de la croyance dans le contexte des systèmes d'information n'est pas la première dans la littérature, l'auteur dans [11] s'intéressa à la gestion des imperfections de données spatiales agro environnementales, et à la fusion de résultats de classifieurs crédibilistes dans le cadre de la théorie des fonctions de croyance afin de fournir aux utilisateurs du système d'information un résultat qualifié par un degré de confiance. [12] d'une autre part traite les données géographiques par un fusionnement des données multiples couvrant la même aire en traitant l'imperfection des données par la théorie d'évidence.

A notre connaissance, il n y a pas eu d'études sur la personnalisation de requêtes décisionnelles utilisant la théorie de l'évidence.

1. **Déroulement :** Les références d'une requête : Etant donné un cube « C » et une requête MDX « q » qui interroge ce cube. L'ensemble des références de la requête « q » est défini comme suit : $q_{Refs} = \{Ref_1,..,Ref_i,..,Ref_n,M_1,\ldots,M_m\}$ tel que :

Ref_i est un ensemble de membres de la dimension D_i . Ces références sont extraites à partir des axes de la requête et l'axe de secteur.

M_i est une mesure analysée dans la requête « q ».

Contrairement au langage SQL qui ne donne que les résultats que l'utilisateur a demandé dans sa requête, Le langage MDX considère le niveau d'hiérarchie le plus abstrait (All) concernant les axes non mentionnés dans une requête en agrégeant les faits.

2. **Exemple illustratif :** On Considère la requête suivante

```
SELECT
NonEmptyCrossJoin ([DimProductAndSub].[English Product Category Name].Children,
{[Order Date].[Calendar Year].&[2013],[Order Date].[Calendar Year].&[2014]})
ONROWS,
{[DimGeography].[Sales Territory Group].[Europe],[DimGeography].[Sales Territory
Group].[Pacific]}
ONCOLUMNS
FROM [AdventureWorks]
WHERE [Measures].[Sales Amount]
```

Le cube interrogé est « CubeAdventureWorksDW2014 » , si on considère l'ordonnancement suivant des dimensions :

D_1 est la dimension « DimProductAndSub » ; D_2 est la dimension « DimCustomer »

D_3 est la dimension « DimGeography » ; D_4 est la dimension « Order Date »

On aura : q_{Refs} = { (Accessories,Bikes,Clothing,Components,Unknown) , (All) , (Europe,Pacific), (2013,2014) , [Sales Amount] }

Puisque l'argument « children » est utilisé pour la dimension des produits ça veut dire que tous ces catégories sont préférées, et depuis que la dimension « DimCustomer » n'était pas mentionnée alors sa référence devient (All).

- Exemple : Soit un log qui contient les requêtes suivantes :

q_1 = {(Mountain-Bike-38, Mountain-Bike-39), All,(France,Canada), All, [Sales Amount]}

q_2 = {(Mountain-Bike-38), All, (Europe), (2013,2014), [Sales Amount]}

q_3 = {(Touring Bikes, Accessoiries), All, (Europe), (2014), [Sales Amount]}
q_4 = {(Bikes, Accessoiries), All, (France, Germany), All, [Sales Amount]}
q_5 = {(Mountain Bikes, Touring Bikes), All, (NA), (2013, 2014), [Sales Amount]}
q_6 = {(Bikes, Clothing), All, All,(2013, 2014), [Freight]}
q_7 = {(Mountain-Bike-38), All, (Europe), (2014), [Freight]}
q_8 = {(Bikes, Accessories), All, (Canada,France,Germany), (2013), [Freight]}
q_9 = {(Socks), All, (NA,Europe), (2014), [Freight]}

On considère le contexte « Sales Amount », l'utilisateur a entré cinq requêtes avec différentes références, on peut donc calculer la masse de chaque référence en prenant en considération son nombre d'apparition. La masse affectée à chaque requête est trouvée à partir de sa fréquence calculée au niveau d'une classe (contexte) c'est-à-dire : la masse de la requête « q_i » est égale à 1 divisé par le nombre total de requêtes de même classe (contexte) que « qi » c'est-à-dire : $m(q_1) = 1/5 = 0.2$; $m(q_8) = 1/4 = 0.25$

Le nombre de contextes trouvé dans le log détermine le nombre de fonctions de masses allouées, on écrit :

$m_1(q_1) = m_1(q_2) = m_1(q_3) = m_1(q_4) = m_1(q_5) = 0.2$
$m_2(q_6) = m_2(q_7) = m_2(q_8) = m_2(q_9) = 0.25$

Cette masse correspond à seulement les références liées aux dimensions et non les mesures depuis que les requêtes soient déjà classées selon leurs contextes. On néglige par suite les références triviales des requêtes qui ne donnent aucune addition capable d'expliquer les préférences d'utilisateur comme la référence « All ». On aura donc concernant les requêtes « q_1 » et « q_3 » les masses comme suit :

$m_1(q_1) = m_1$(Mountain-Bike-38, Mountain-Bike-39, France, Canada) = 0.2
$m_1(q_3) = m_1$(Touring Bikes, Accessoiries, Europe, 2014) = 0.2

Une multiple source des préférences expliquées par la pluralité des contextes nous obligent de les combiner pour avoir une seule expression générale des préférences.

Ce qui nous conduit à la règle de combinaison originale, connue en tant que règle de combinaison de Dempster, qu'est une généralisation du théorème de Bayes [10]. Ce théorème met clairement en valeur l'accord entre des sources multiples et ignore les conflits grâce à un facteur de normalisation en calculant la masse du conflit entre les sources. Cette masse est trouvée en sommant les produits des masses des intersections vides entre les requêtes des deux classes : Dans l'exemple précédent, on a 4 intersections vides ($\{q_1 \cap q_6\}, \{q_1 \cap q_9\}, \{q_4 \cap q_7\}, \{q_4 \cap q_9\}$) donc: $m(\emptyset)=4x(0.2x0.25)=0.2$. Ce conflit est utilisé par suite dans l'opérateur de Dempster pour normaliser la fonction de masses composée en divisant toutes ces masses sur $1 - m(\emptyset) = 0.8$, on aura par exemple m(2013,2014)= 0.125. A la fin de cette phase, on obtient les masses des sous-ensembles de l'ensemble général **Ω** l'ensemble de tous les membres du cube. On doit par suite calculer la croyance de ces sous-ensembles dont les redondants vont avoir une préférence élevée expliquée par une masse élevée.

a. Le calcul de la fonction de crédibilité (croyance) : La fonction de croyance, *bel*(A), mesure la force avec laquelle on croit en la véracité de la proposition A et elle est définie

comme étant la somme de toutes les masses de croyance des éléments focaux B $\in 2^{\Omega}$ inclus dans A. Formellement :

$$bel(A) = \sum_{B \subseteq A, B \neq \emptyset} m(B) \qquad \forall\, A \subseteq \Omega.$$

Remarque : Pour les cas où l'on a un seul contexte concernant les requêtes dans le log, il suffit de passer directement à calculer les crédibilités des références de ces requêtes.

La dernière étape nous permet d'obtenir des masses corresponds aux membres singuliers (des sous-ensembles de **Ω** de cardinalité égal à 1). Le but est de trouver le poids de chaque membre ce qui nous conduit à trouver l'ordre partiel des membres selon la préférence d'utilisateur. Pour cela on doit calculer la probabilité pignistique.

b. La probabilité pignistique : Après avoir trouvé la crédibilité de chaque référence, on doit calculer la probabilité pignistique de chaque singleton dans cette référence en divisant la crédibilité de la référence sur sa cardinalité. La dernière étape du processus consiste à trouver la masse totale pour chaque singleton en sommant ces probabilités et par suite de les trier dans l'ordre décroissant afin d'obtenir un pré-ordre sur les membres du cube. Les membres n'ayant pas une masse tangible sont classés dans la fin de l'ordre avec une préférence équivalente.

IV. Conclusion

Cette approche proposée nous a permis en premier lieu d'améliorer l'approche de l'article [5], comme elle a simplifié la tâche aux décideurs dont la majorité ne sont pas informaticiens et n'ont pas forcément de connaissance sur la structure d'entrepôts de données utilisée. On peut étaler l'étude à d'autres travaux pour montrer l'utilité et l'efficacité de cette proposition. Les premières expérimentations sont très encourageantes.

Bibliographie :

[1] Patrick Marcel : OLAP Query Personalisation and Recommendation: An Introduction. M.-A. Aufaure and E. Zim´anyi (Eds.): eBISS 2011, LNBIP 96, pp. 63–83, 2012. Springer-Verlag Berlin Heidelberg, (2012)

[2] Giacometti, A., Marcel, P., Negre, E., Soulet, A.: Query recommendations for OLAP discovery driven analysis. Proceedings of ACM 12th International Workshop on Data Warehousing and OLAP, Hong Kong, China, November 6, pp. 81–88 (2009)

[3] F. Ravat, O. Teste: Personalization and OLAP Databases. Springer US, Annals of Information Systems, vol. 3, New Trends in Data Warehousing and Data Analysis, (2009) , pp. 1-22.

[4] M. Golfarelli, S. Rizzi, and P. Biondi : MYOLAP: An Approach to Express and Evaluate OLAP Preferences. IEEE transactions on knowledge and data engineering, VOL. 23, NO. 7, pp 1050-1064. JULY (2011)

[5] Bellatreche, L. Giacometti, A. Marcel, P. Mouloudi, H. and Laurent, D. A personalization framework for OLAP queries. In DOLAP '05: Proceedings of the 8th ACM international workshop on Data warehousing and OLAP, pages 9–18, New York, NY, USA, (2005). ACM.

[6] Jerbi, H., Ravat, F., Teste, O., Zurfluh, G.: Preference-Based Recommendations for OLAP Analysis. In: Pedersen, T.B., Mohania, M.K., Tjoa, A.M. (eds.) DaWaK 2009. LNCS, vol. 5691, pp. 467–478. Springer, Heidelberg (2009)
[7] Mansmann, S., Scholl, M. H.: Visual OLAP: A New Paradigm for Exploring Multidimensonal Aggregates. In Proceedings of IADIS International Conference on Computer Graphics and Visualization (MCCSIS'08), Amsterdam, The Netherlands, 24 - 26 July, (2008), pp. 59-66.
[8] N. Kozmina, L. Niedrite : OLAP Personalizaton with User-describing Profiles. 2011
[9] Garrigós, I., Gómez, J.: Modeling User Behaviour Aware WebSites with PRML. In Proceedings of the CAISE'06 Third International Workshop on Web Information Systems Modeling (WISM '06), Luxemburg, June 5-9, (2006), pp. 1087-1101.
[10] Patrick Vannoorenberghe « Un état de l'art sur les fonctions de croyance appliquées au traitement de l'information » Revue I3, vol. XX, num. XX. (2004)
[11] Karima Zayrit « Fusion de données imparfaites dans un système d'information agro-environnemental - Une approche basée croyance », (2012)
[12] Ana-Maria OLTEANU « Fusion de connaissances imparfaites pour l'appariement de données géographiques - Proposition d'une approche s'appuyant sur la théorie des fonctions de croyance » Thèse, Université Paris-Est, (2008)
[13] Dimitre Kostadinov, Mokrane Bouzeghoub, Stéphane Lopes, Accès personnalisé à des sources de données multiples : évaluation de deux approches de reformulation de requêtes, INFORSID, (2008)
[14] Mokrane Bouzeghoub, Sylvie Calabretto, Nathalie Denos, Rami Harrathi, Dimitre Kostadinov, AnTe Nguyen, Veronika Peralta, Accès personnalisé aux informations : approche dirigée par la qualité INFORSID, (2007)
[15] Sid Ali Selmane, Fadila Bentayeb Omar Boussaid , Règles d'Association Triadiques pour la personnalisation de requêtes décisionnelles, RNTI. (2014)

Abstract

In the OLAP context, a large amount of information is provided to the user. In general, the majority of those information is not interesting and inadequate from the user's viewpoint. In order to solve this problem, we thought at personalizing the decisional system and the MDX queries.

In this paper, a quantitative approach represented by using the Dempster Shafer theory's tools was adopted. The log of the past queries issued by the user is exploited to form a preference expression.

***Keywords** : OLAP, personalization, profile, Data Warehouse, Dempster–Shafer theory.*

Répondre aux Questions « *Why* » pour les Applications BI : Modélisation et Approche

Meriem Amel Guessoum, Rahma Djiroun, Kamel Boukhalfa

Laboratoire LSI/USTHB, Alger, Algérie

{mguessoum,rdjiroun,kboukhalfa}@usthb.dz

Résumé. Les applications de Business Intelligence fournissent au décideur de l'entreprise un ensemble d'informations agrégées et d'indicateurs décisionnels lui facilitant la compréhension du fonctionnement actuel de l'entreprise et l'identification des tendances du marché pour la prise de décision. Pour accéder aux informations décisionnelles, le décideur exprime généralement ses besoins en langage naturel plutôt qu'en langage formel. La question exprimée en langage naturel peut être formulée via un ensemble de mots clés ou bien en une phrase commençant par : *What, Who, Why*, etc.(« WH-questions »). Dans la littérature, plusieurs travaux se sont intéressés aux questions « What » et « Who », alors que la question « *Why* » omniprésente dans le processus de prise de décision, n'a pas été abordée dans un contexte de Business Intelligence. Nous nous intéressons dans ce papier aux questions de type « *Why* » où nous proposons une modélisation ainsi qu'une approche basée sur l'analyse de tendances pour traiter ce type de questions. Nous avons développé un outil nommé *Why-Question analyzer* pour montrer la faisabilité de notre approche.

1 Introduction

Les interfaces naturelles (Hearst (2011)) sont devenues populaires et ont fait l'objet de plusieurs travaux de recherche dans différents domaines : moteurs de recherche, bases de données, web sémantique, etc. Ces dernières années plusieurs travaux dans le domaine du Bussiness Intelligence (BI) spécialement les entrepôts de données (ED) ont été menés pour intégrer ce genre d'interfaces (Naeem et al. (2012), Popowich et al. (2012), Saias et al. (2012), Kuchmann-Beauger et Aufaure (2011, 2012), Kuchmann-Beauger (2013), Bargui et al. (2008)).
Pour accéder aux informations décisionnelles de l'ED, l'intervention du concepteur (*IT designer*) est nécessaire pour transformer les besoins du *décideur* exprimés en langage naturel en une ou plusieurs requêtes formelles exprimées dans un langage spécifique (SQL, MDX, etc.). Cette démarche présente certaines limites : incompréhension du besoin décisionnel, perte d'information et dépendance du décideur du *IT designer*.
Pour pallier les lacunes de ce scénario d'interrogation des ED, le remplacement de l'intervention du *IT designer* par une interface naturelle semble un moyen efficace afin de supporter et traiter des questions exprimées en Langage Naturel (LN).
Les questions exprimées en LN peuvent être classées selon plusieurs catégories tels que les

questions de types « WH » [1] (*What*, *Who*, *Where*, etc.), les questions constituées d'un ensemble de mots clés, questions de définition et les question d'opinions, etc.
Dans un contexte de BI, les questions en LN qui ont attiré l'attention des chercheurs sont les questions "*What*"(Naeem et al. (2012), Saias et al. (2012),Kuchmann-Beauger et Aufaure (2011, 2012)) et les questions de type mots clés (Kuchmann-Beauger et Aufaure (2012) Kuchmann-Beauger (2013)). Généralement, ces questions ne correspondent pas complètement aux besoins des décideurs. En effet, dans les applications BI, les décideurs cherchent en général à connaitre l'origine de phénomènes observés par rapport à une certaine activité(diminution des ventes, augmentation des recours, etc.). Ce besoin décisionnel peut prendre la forme d'une question de type « *Why* ». Par exemple dans le domaine d'accidentologie, un décideur peut poser la question : *" pourquoi le nombre d'accidents a augmenté cette année ?"*. Une réponse à cette question pourrai être la mise en évidence d'un ensemble d'indicateurs fournis au décideur lui permettant de comprendre l'origine du phénomène tel que le facteur humain, plus précisément l'âge du conducteur. En conséquence, une décision pourrai être prise par rapport aux conditions d'obtention du permis de conduire pour les plus jeunes.
Dans la littérature, la question « *Why* » a été qualifiée de complexe par Moriceau et al. (2010), car les réponses attendues sont en général des explications nécessitant des techniques particulières pour les produire. Ce type de questions a été largement traité dans le domaine de Recherche d'Informations (RI) comme dans (Girju (2003),Verberne (2006, 2007), Verberne et al. (2007), Moriceau et al. (2010), Oh et al. (2012), Baral et al. (2012), Oh et al. (2013)). Néanmoins, les modèles proposés dans ces travaux ne sont pas adaptés pour les applications BI, car ils ne prennent pas en compte les concepts multidimensionnels caractérisant les ED (faits, mesures, dimensions, hiérarchie, etc.) alors que ces concepts (indicateurs décisionnels) sont importants dans un système décisionnel.
À notre connaissance aucun travail n'a été proposé pour prendre en charge les questions de type « *Why* » pour les applications BI. Notre objectif consiste à fournir au décideur un moyen pour détecter le/ les facteur(s) influant sur un phénomène pour une aide efficace à la prise de décision. Nous proposons dans ce papier, une modélisation ainsi qu'une approche basée sur l'analyse de tendances pour traiter et répondre aux questions « *Why* ».
Le papier est structuré comme suit : nous analysons dans la section 2 un ensemble de travaux connexes. Nous présentons dans la section 3 un exemple de motivation. Dans la section 4, nous proposons une modélisation des questions décisionnelles de type « *Why* ». L'approche que nous proposons est présentée dans la section 5. L'outil développé ainsi qu'une première évaluation de l'approche sont présentés dans la section 6. Enfin, dans la section 7, nous concluons ce papier et nous abordons quelques perspectives de travail.

2 Travaux Connexes

Les décideurs, dans les applications BI, cherchent généralement à connaitre l'origine de phénomènes observés sur l'activité de l'entreprise. Ce besoin peut être exprimé sous forme d'une question « WH » (*What, Who, Where*, etc.) ou une question constituée d'un ensemble de mots clés. Ce type de questions a été largement traité dans le domaine du RI. À notre connaissance aucun travail n'a abordé la question de type « Why » dans le contexte du BI. En plus des

1. pronoms interrogatifs utilisés pour poser une question, comme : who, when, where, etc. .Ils sont parfois appelés Wh-words

travaux dans le domaine du BI, nous avons intégré dans notre étude bibliographique les travaux traitant la question « Why » dans le domaine du RI (voir tableau 1). Nous analysons ces travaux selon les critères suivants : Entrées (corpus et type de la question), Sorties (le résultat de l'approche) et objectif (le but de l'approche proposée).

Pour la plupart des travaux abordant la question « *Why* » dans le domaine du RI, il a été question de proposer des approches pour mettre au point des systèmes de Questions réponses afin de répondre à la question « *Why* ». Ces approches cherchent à identifier des réponses dans des sources de documents comme dans (Moriceau et al. (2010); Oh et al. (2012, 2013); Verberne (2006); Girju (2003)) et (Verberne (2007); Verberne et al. (2007)). Néanmoins, dans (Baral et al. (2012)), l'auteur a traité une question de type « *Why* » par rapport à une base de connaissances définie selon le domaine de la Biologie. Les réponses à une question « *Why* » sont en général des explications basées sur le principe de *causalité*. Nous présentons brièvement, quelques différentes méthodes traitant la question « Why » : (1) dans (Oh et al. (2013)), les auteurs ont proposé une approche qui explore l'existence d'expressions causales dans des documents japonais (e.g Tsunamis are caused by the sudden displacement of huge placement of water) ; (2) selon une méthode basée sur un ensemble de modèles lexico-syntaxiques se référant à la causalité dans des textes anglais (Girju (2003)) et (3) par rapport à des patrons spécifiques inspirés par l'observation qu'une question « *Why* » et ses réponses suivent souvent le fait que si quelque chose de désirable ou indésirable se produit alors ses raisons sont respectivement désirables ou indésirables (Oh et al. (2012)).

Dans le contexte du BI, l'objectif principal des approches proposées était de procurer au décideur un moyen intuitif et flexible pour interagir avec l'ED. Que ce soit pour remplacer le langage formel (SQL, MDX) par une requête exprimée en LN dans le processus d'interrogation des ED comme dans (Naeem et al. (2012), Saias et al. (2012), Kuchmann-Beauger et Aufaure (2011, 2012), (Kuchmann-Beauger (2013)) ou bien pour faciliter le processus de génération du schéma conceptuel d'un magasin de données à partir d'un ensemble de requêtes introduites en LN comme dans (Bargui et al. (2008)). Ces approches ont toutes tenu compte de façon impérative de l'aspect de la modélisation multidimensionnelle d'un ED. Nous avons observé que les requêtes exprimées en LN les plus traitées étaient les questions de type « *What* » et les questions de type mots clés. Alors que dans le contexte du BI, il serait intéressant de pouvoir poser concrètement en LN une question de type « *Why* », pour laquelle la production d'une ou plusieurs requêtes formelles comme réponses seront probablement insuffisantes, nécessitant des techniques particulières pour les traiter tel que : *" pourquoi Le nombre d'accidents a augmenté cette année ?"*.

En formulant une question « *Why* », le décideur cherche à avoir, vis-à-vis d'une représentation multidimensionnelle, des indicateurs décisionnels lui permettant d'avoir une réponse à sa question ou bien une aide pour affiner sa question par rapport à d'autres éléments multidimensionnels. Par contre, les réponses à une question « *Why* » traitée dans le RI sont en général des explications produites dans des clauses, des phrases (Kato et al. (2005)) ou bien des paragraphes (Verberne (2007), Verberne et al. (2007)), localisées dans des corpus de documents. Nous soulevons donc, le fait que les approches proposées dans le domaine du RI ne peuvent pas être adoptées pour traiter une question décisionnelle de type « *Why* ». Par conséquent, modéliser une question décisionnelle de type « *Why* » est nécessaire, dans la perspective de proposer une approche qui traite ce type de question.

À notre connaissance, l'approche que nous proposons est la première qui permet à un décideur

Critères							Travaux connexes	Recherche d'Information								Business Intelligence								
								Moriceau et al. (2010)	Baral et al. (2012)	Oh et al. (2012)	Oh et al. (2013)	Verbeme (2006)	Girju (2003)	Verbeme (2007)	Verbeme et al. (2007)	Naeem et al. (2012)	Saias et al. (2012)	Popowich et al. (2012)	Kuchmann-Beauger et Aufaure (2011)	Kuchmann-Beauger et Aufaure (2012)	Kuchmann-Beauger (2013)	Bargui et al. (2008)	Notre approche
Entrée	Question utilisateur	Type	Langage Naturel	WH exprimés	Question *What*											X	X		X	X	X		
					Question « *Why* »	Formalisée			X														
						Non formalisée		X		X	X	X	X	X	X								X
				WH non exprimés																		X	
			Ensemble de termes	Mots clés																X	X		
				Sélection de termes assistée.														X					
		Modélisation																					X
	Corpus	Entrepôts de sonnées														X	X	X	X	X	X		X
		Un ensemble de documents.						X		X	X	X	X	X	X								
		Base de connaissances.							X														
		Sources de données.																				X	
Sortie	Requêtes formelle	SQL														X		X	X	X	X		
		MDX															X			X	X		
	Textuelle	Réponses formalisées selon	Une base de connaissance						X														
			Modèles en Langage Naturel																				X
		Réponses non formalisées	Paragraphes					X				X		X	X								
			Clause ou des phrases							X	X		X										
	Schéma conceptuel d'un magasin de données.																					X	
	Résultats graphiques.																X	X		X	X		X
Objectif	Pertinence.							X	X	X	X	X	X	X	X	X	X	X	X	X		X	X
	Performance.									X	X					X		X		X	X	X	X

TAB. 1 – *Comparaison des Travaux connexes.*

d'interroger concrètement un ED avec une question « *Why* » exprimée en LN et de générer des réponses en Langage Naturel.
Avant d'entamer notre approche, nous présentons dans la section suivante, un exemple de motivation nous permettant d'expliquer les étapes de l'approche.

3 Exemple de Motivation

Afin d'expliquer notre approche, nous considérons un modèle simplifié de l'ED issu de l'étude du risque routier dans le domaine de l'accidentologie (Derbal et al. (2016)). L'entrepôt de données est modélisé par un schéma en flocon de neige (voir figure 1). La table de faits « *Incident* » est composée des mesures : *nombre d'accidents*, *nombre de morts*, *nombre de blessés*, *nombre de véhicules impliqués*. Le modèle se compose d'un ensemble d'axes d'analyse (dimensions) organisés en hiérarchie comme la dimension *Temps*, *Wilaya*, *Conducteur* et *Véhicule*. Un décideur en accidentologie peut avoir des besoins décisionnels exprimés en LN sous forme d'une question « *Why* » par rapport au nombre d'accidents, de morts ou de blésées. Soit l'exemple de la question « *Why* » (Q_1) que nous utilisons tout au long de ce papier *"Pourquoi le nombre d'accidents a augmenté ?"*. Pour répondre à cette question d'une manière naïve, le

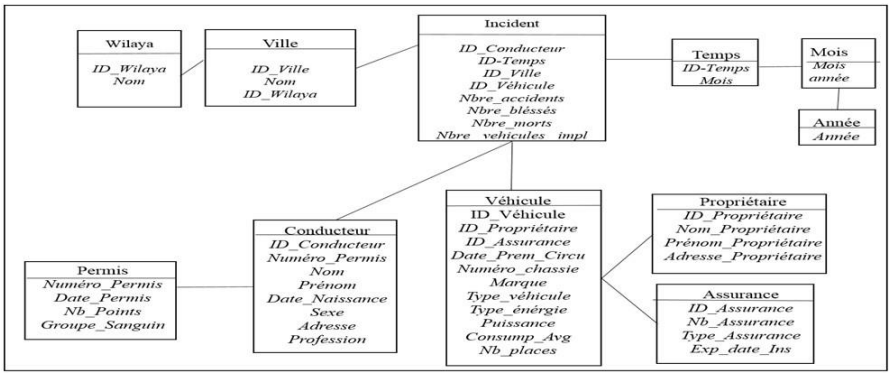


FIG. 1 – *Modèle conceptuel d'un ED reflétant le phénomène du risque routier.*

décideur fait intervenir le *IT designer* pour exprimer son besoin. Ce dernier, transforme ce besoin en une série de requêtes exécutables (SQL, MDX, etc) pour interroger l'ED. Les résultats obtenus seraient des instances générées par rapport aux mesures selon tous les axes d'analyse de l'ED. Le décideur doit par la suite analyser chaque doublet (<*mesure, dimension*>) à part, de multiples aller retours entre les différentes dimensions sont possibles pour faire une synthèse complète. Les efforts d'analyse sont considérables et posent des problèmes de rapidité, d'efficacité et une perte de temps, engendrant parfois des analyses erronées dues au volume important de données et par conséquent le décideur pourrai diverger du but décisionnel initial. Une solution à ce problème, serai de retourner au décideur des réponses en LN, lui permettant de diminuer considérablement sa charge de travail en termes de temps et d'efforts tout en réduisant la dépendance du décideur au *IT designer*. Notre approche est proposée dans ce contexte. Avant de présenter cette dernière, nous proposons, dans la section suivante, une modélisation pour les questions décisionnelles de type « *Why* ».

4 Modélisation des questions-décisionnelles de type « *Why* »

Nous considérons un ED modélisé par un schéma en flocon de neige, composé d'une table de fait (F) comportant un ensemble de mesures (M) tel que $M = \{m_1, ..m_i..m_n\}/ i = 1..n$, un ensemble de dimensions (D) tel que $D = \{D_1, .D_j., Dt..D_m\}/j = 1..m$ où Dt référence la dimension temporelle. Chaque D_j est décrite par un ensemble d'attributs (A) tel que $A = \{a_1, ..a_k..a_p\}/k = 1..p$. Une dimension D_j est munie ou non de niveaux d'hiérarchie (N) tel que $N = \{n_1, ..n_t..n_s\}/t = 1..s$, nous notons donc une dimension : $D_j[n^*[a_k]]$.

Nous modélisons une question décisionnelle de type « *Why* » selon les éléments multidimensionnels qu'elle référence tels que les mesures, les dimensions, les niveaux, les membres, etc. et des tendances observées sur une activité durant une *période* donnée tels que : diminution, augmentation, baisse, hausse, stagnation, changement, stabilité, etc.

Nous proposons dans la figure 2, une modélisation UML d'une question décisionnelle de type « *Why* ». Cette dernière peut être composée ou non de *mesures* M. Les dimensions qu'elle peut référencer : la *dimension temporelle* Dt tel que la date pour spécifier le temps, *autres dimensions* D_j telles que client, produit, etc. et d'un indicateur de tendance (augmentation, di-

minution, etc.). Nous considérons que la question « *Why* » peut comprendre ou non des filtres (*f*). Un *filtre* ***f*** consiste à appliquer une restriction sur les valeurs (V) des attributs d'une dimension tel que $V = \{v_1, ...v_e, ...v_r\}/e = 1..r$ ou par rapport aux valeurs (V^l) d'une mesure m_i tel que $V^l = \{v^l_1, ...v^l_g, ...v^l_z\}/g = 1..z$. Un filtre ***f*** est défini en fonction d'un ensemble d'opérateurs (OP) tel que $OP \in$ { *égale, où, entre, supérieur à, inférieur à, sauf, comprise entre, etc.*}. Nous notons donc un filtre $\boldsymbol{f} : (\boldsymbol{f}[OP][D_j[n^*[[a_k[v_e]]]]$ ou $\boldsymbol{f}[OP][m_i[v^l_g]])$. Par exemple *"pourquoi le nombre d'accidents a augmenté durant les années comprises entre 2010 et 2016"* où OP = "comprises entre" et "2010, 2016" est un filtre à appliquer à un attribut de la dimension année.

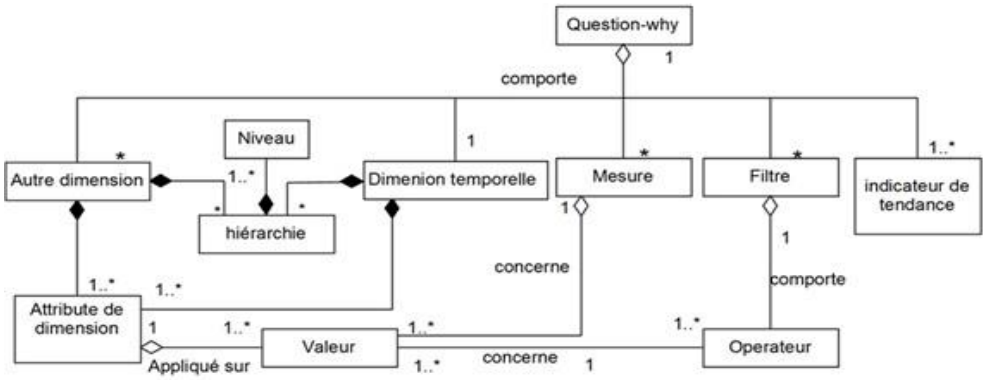


FIG. 2 – *Modélisation de la question « Why ».*

5 Notre Approche pour Analyser et Répondre aux questions « *Why* »

Dans une question décisionnelle « *Why* », le décideur s'intéresse aux tendances d'activités évaluées par rapport à un ensemble de mesures analysées selon plusieurs dimensions. Par conséquent, nous considérons, deux catégories de questions « *Why* » selon la façon dont les mesures sont référencées : (1) questions « *Why* » avec mesures tel que *"pourquoi le nombre d'accident a augmenté"* et (2) questions « *Why* » sans mesures tel que *"pourquoi le risque routier ne cesse de s'accroitre"*. Dans chaque catégorie, les dimensions peuvent être exprimées ou non par le décideur. Par exemple, dans la question *"pourquoi le risque routier a connu un pic en 2016"*, la dimension temporelle est référencée explicitement tandis que les mesures (par exemple, le nombre de morts) sont implicitement citées.

Dans le cadre de notre approche, nous nous intéressons uniquement aux questions avec mesures (la deuxième catégorie fera objet de nos travaux futurs). En effet, le traitement des questions sans mesures est plus compliqué nécessitant une étude sémantique approfondie que nous sommes entrain de faire.

L'objectif de l'approche, que nous proposons dans le présent papier, est de permettre au décideur d'introduire en entrée une question « *Why* » afin de produire en sortie une ou plusieurs réponses en LN avec des résultats graphiques.

L'architecture de l'approche est telle que illustrée dans la figure 3. Elle est constituée de quatre phases : (1) Analyse de la question « *Why* », (2) Génération de la requête formelle, (3) Analyse de tendances et (4) Génération des résultats. Les détails de chaque phase sont présentés dans la suite de cette section.

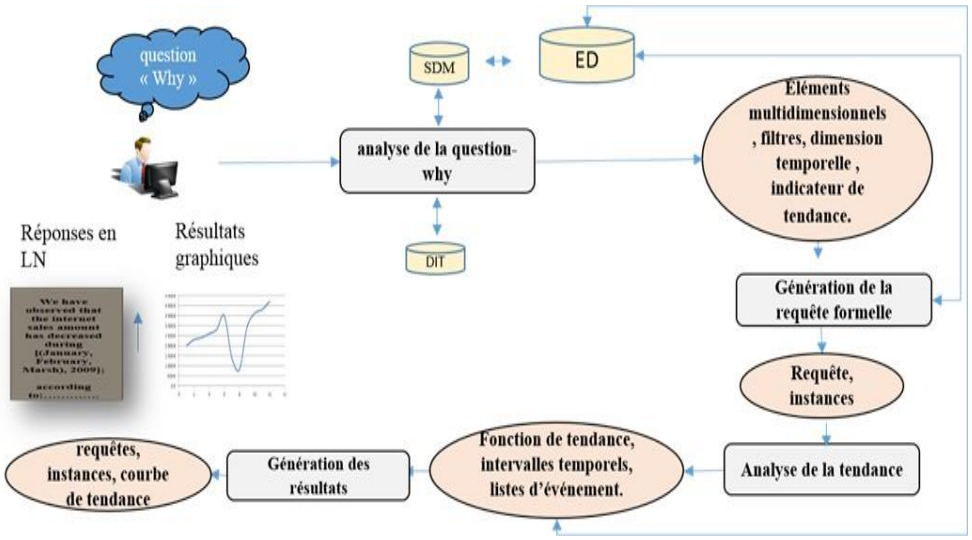


FIG. 3 – *Architecture de l'approche proposée.*

5.1 Analyse de la question « *Why* »

Une phase de traitement automatique du langage s'impose pour analyser la question « *Why* » afin d'obtenir l'ensemble d'informations caractérisant la question. Cette analyse comporte deux étapes : (1) l'analyse lexico-syntaxique et (2) l'identification des éléments multidimensionnels, des filtres et des indicateurs de tendance.

1. *L'analyse lexico-syntaxique :* dans cette étape, nous identifions les différentes unités lexicales ainsi que les relations syntaxiques existantes entre les termes de la question utilisateur. Ceci peut être effectué par des analyseurs lexicaux et syntaxiques existants comme *StandFord Pos Tagger* (Manning (2011)) et *Stanford parsers* (Bajwa et al. (2012)).
2. *Identification des éléments multidimensionnels, des filtres et de l'indicateur de tendance :* pour identifier les éléments multidimensionnels, nous nous appuyons sur une structure de données multidimensionnelles (SDM) chargée automatiquement à partir du schéma de l'ED (métadonnées). Nous associons des étiquettes pour spécifier chaque information chargée dans la SDM tels que mesure, dimension, attributs de dimension, niveau d'une hiérarchie de dimensions ainsi que la dimension temporelle. La SDM vise également à identifier les dimensions qui sont en relation avec la mesure antérieurement identifiée.
   Les indicateurs de tendances sont identifiés selon un dictionnaire d'indicateurs de tendance (DIT) que nous avons construit. Ce dernier comporte un ensemble d'indicateurs

(mots) faisant référence à un changement de tendance comme : diminution, augmentation, hausse, élevé, baisse, etc.
Nous identifions les filtres contenus dans la question en se basant sur un ensemble d'indicateurs de filtre tels que : où, égal, comprise entre, cet, supérieur à, inférieur à, sauf, etc. Une donnée temporelle (comme *Année 2016*, *mois de Janvier*) introduite au niveau de la question « *Why* » est considérée comme un filtre à appliquer sur la dimension temporelle. Nous considérons que les termes qui sont syntaxiquement liés aux éléments multidimensionnels préalablement identifiés comme filtres, par exemple "*ville d'alger*", ville et alger sont syntaxiquement liés et alger est une valeur de l'attribut de la dimension ville.

Comme exemple, l'analyse de la question Q_1 présentée dans la section 3 produit les informations suivantes : { indicateur de tendance (IDT)= *a augmenté* }, { mesure m_{Q_1}= *nbre-accidents* }, { Dimension temporelle Dt_{Q_1} = *Temps [mois, année]* }, { les dimensions D liées à la mesure m_{Q_1} : D_{Q_1}= {*Wilaya[ville]*, *véhicule [propriétaire, assurance]*, *conducteur[permis]*}.

5.2 Génération de la requête formelle

Afin de générer une requête exécutable (Q) équivalente aux informations issues de la phase d'analyse de la question « *Why* », indépendamment du modèle logique et physique de l'ED, nous générons une requête formelle (R) à partir de la question « *Why* » (q) tel que :
R =< description des mesures (DM), description de la dimension temporelle (DDT), description des autres dimensions (DD) > tel que :
$DM =< m_q, (f[OP][m_q[v_q]])^* >$.
$DDT =< Dt_q[n^*[a_q]], f[OP][Dt_q[n^*[a_q[v_q]]] >$.
$DD =< (D_q[n^*[a_q]])^*, (f[OP][D_q[n^*[[a_q[v_q]]])^* >$.
Par rapport à la requête formelle R, nous avons défini un ensemble de templates de requêtes exécutables conformes aux langages d'interrogation des ED comme SQL, MDX, XQuery, etc. Une fois le template rempli avec les informations appropriées et la requête Q adéquate est exécutée, nous récupérons les instances résultantes afin de procéder à la phase d'analyse de tendances. En appliquant cette phase par rapport à la question Q_1, nous obtenons :

```
Select sum(nbre-accident) as total-accidents , année, mois
From Incident, Temps, Mois, Année
Where temps.ID-Temps== Incident.ID-Temps
et Temps.mois==Mois.mois et Mois.année== Année.année et Année.année= "2016"
Group by année, mois ;
```

TAB. 2 – *Requête exécutable.*

$R_{Q_1} =< nbre-accidents, année = 2016, mois = \{janvier, ..., décembre\} >$. À partir de R_{Q_1}, nous pouvons générer une requête exprimée en SQL tel que illustré dans le tableau 2. Les instances obtenues suite à l'exécution de la requête SQL sont présentées dans le tableau 3.

Valeurs de la dimension temporelle *Temps*	Janv-16	Févr-16	Mars-16	Avr-16	Mai-16	Juin-16	Juil-16	Aout-16	Sept-16	Oct-16	Nov-16	Déc-16
valeurs de la mesure *nbre-accidents*	2324	3873	2711	2905	1162	1937	1743	2704	3096	2329	2130	1942

TAB. 3 – *Instances.*

5.3 Analyse de Tendances

L'objectif de cette phase est de synthétiser une perception qualitative des données numériques des mesures ($M[V^l]$) observées selon une période basée sur la dimension temporelle Dt, visant à détecter les tendances globales telles qu'une augmentation, une diminution. Pour y parvenir, nous utilisons un modèle mathématique basé sur des méthodes d'analyse numérique : *fonction de tendance* (Dufour (2003)). Cette fonction permet de décrire mathématiquement les données $(Xi, Y\,i)$ où X est la période définie selon Dt et Y est $M[V^l]$ défini selon les mesures M et les dimensions D spécifiées ou non dans la question « *Why* ». La fonction de tendance est basée sur le principe de la régression non-linéaire (Mellac (2013)), pour laquelle la courbe ne passe pas nécessairement par toutes les coordonnées (xi, yi), mais les approche le plus possible. Elle permet d'effectuer une bonne approche descriptive et d'atteindre la précision souhaitée sans être encombré par les informations induites par les multiples oscillations locales de $Y\,i$.
La fonction de tendance est construite en fonction des $M[V^l]$ et Dt. Cette dernière peut prendre la forme d'une de ces fonctions : polynomiale, logarithmique, exponentielle, sinusoïdale [2] : $f(Xi) = Y\,i$.
Par la suite, nous devons chercher la valeur de R (l'erreur relative), définie comme suit :

$$R = \sqrt{\frac{1}{k}\sum_{i=0}^{k}(Y_i - f(Xi))^2} \tag{1}$$

Où k est le nombre de coordonnées $(Xi, Y\,i)$.
Soit la forme polynomiale : $f(x) = P^n(x) = \sum_{j=0}^{n} a_j x^j$.

La fonction idéale $f(x)$ est obtenue lorsque R atteint sa valeur minimale. Ceci est effectué, lorsque les dérivées partielles de R s'annulent simultanément :

$$\frac{\partial R}{\partial a_0} = 0, \ldots, \frac{\partial R}{\partial a_j} = 0, \ldots, \frac{\partial R}{\partial a_n} = 0 \tag{2}$$

Ce système d'équations, nous conduit donc à déterminer les paramètres $\{a_j\}$.
Une fois la fonction de tendance déterminée, il devient désormais possible de procéder à une étude standard de fonction mettant en relief les différents aspects tels que : point haut, point bas, diminution et augmentation majeures avec les intervalles respectifs et l'amplitude des

2. https://onlinehelp.tableau.com/current/pro/desktop/fr-fr/trendlines-model.html

variations ΔY. Chacun de ces relief constitue un événement que nous classons dans une liste, pour récupérer par la suite l'intervalle (I) de la tendance recherchée la plus importante. Suite

Événement	Données temporelles	variations ΔY
Augmentation	0	1401.1
Point haut	-0.49473	0
Diminution	0	-2169.1
Point bas	-0.15252	0
Augmentation	0	1264.7
Point haut	0.20053	0
Diminution	0	-971.46
Point bas	0.46536	0
Augmentation	0	84.359

TAB. 4 – *Un exemple d'événements identifiés.*

à cette phase, la fonction de tendance et R obtenues pour la question Q_1 sont : $f(x) = 1.0e+05 * (0.0208 + 0.0545x + 0.0393x^2 - 0.6768x^3 - 0.0361x^4 + 1.5486x^5)$ et $R = 327.8737$. La liste d'événements identifiés est telle que présentée dans le tableau 4.

5.4 Génération de résultats

Les résultats d'une question « *Why* » sont obtenus suite à une opération de projection de l'intervalle temporel I par rapport aux dimensions D. Une projection consiste à générer des requêtes exécutables, selon lesquelles les instances résultantes révèlent des éléments de réponses à la question « *Why* ». Cette projection est effectuée selon une requête formelle (R^I) tel que $R^I =< D_j[n^*[a_k]], Dt[n^*[a_k]], f[OP[Dt[n^*[a_k[v_e]]]]] >$ tel que I correspond à $f[OP[Dt[n^*[a_k[v_e]]]]]$. Nous interprétons ensuite les instances obtenues sous forme de ré-

1. Nous avons observé que le nombre d'accidents a augmenté pendant [Février 2016] selon les wilaya: M'sila, Na'ama, Laghouat.

2. Nous avons observe que le nombre d'accidents a augmenté pendant [Février 2016] selon les vehicules: camion poids lourd, transport commun

.

3. Nous avons observé que le nombre d'accidents a augmenté pendant [Février 2016] selon les conducteurs: hommes.

FIG. 4 – *Réponses en Langage Naturel.*

ponses en LN. Pour ce papier, nous avons pré-défini un template (T) qui capture les résultats des dimensions D concernées par la tendance recherchée tel que :
T = "Nous avons observé que" $< m_q ><IDT >$ "pendant" $< I >$ "selon" $< D >$.
Comme exemple de réponses par rapport à la question Q_1, nous exposons la figure 4.

Pour appuyer les réponses générés en LN, nous produisons un résultat graphique représentant la courbe de tendance tel que présentée dans la figure 5. Cette dernière illustre clairement la période de la tendance recherchée ainsi que les variations telles que l'augmentation et la diminution.

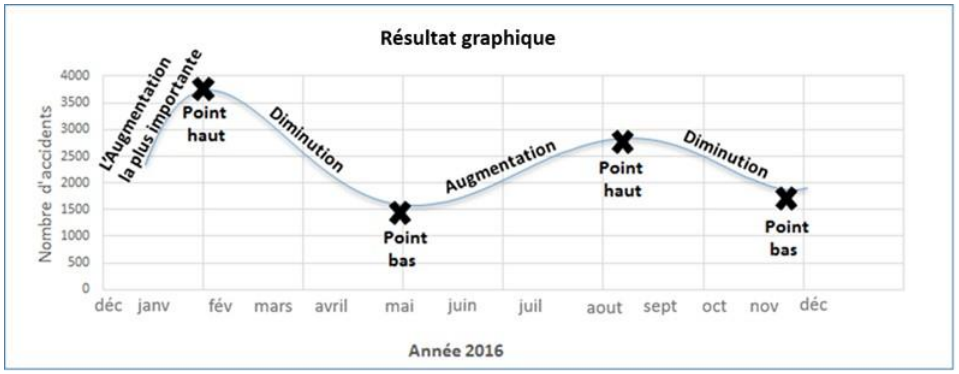


FIG. 5 – *Résultats graphiques.*

6 Implémentation et Evaluation

Pour prendre en compte un jeu de données réel, nous avons considéré l'ED *Microsoft Adventure-Works DW 2012* [3], portant sur les ventes, les achats, la gestion des produits, la gestion des contacts et des ressources humaines. À cet effet, nous avons constitué une base d'un minimum de 50 questions « *Why* » [4] relatives au contexte de l'ED. Nous avons défini cette base de questions selon des combinaisons effectuées entre les différents éléments multidimensionnels existants dans l'ED et par rapport à des questions qui peuvent être posées dans un environnement d'entreprise tel que *"pourquoi l'entreprise n'évolue pas ?"*.

Nous avons implémenté notre approche sur une machine Pentium(R) Dual Core CPU T4500 2.30GHz avec 2G de mémoire RAM. Nous avons utilisé les langages *MATLAB 2014* et *JAVA* dans l'environnement *NetBeans IDE 8.0.2* ainsi que *Microsoft SQL Server* pour exploiter l'ED *Microsoft Adventure-Works DW 2012*.

Nous avons développé un outil nommé *Why-Question analyzer*, offrant au décideur une interface graphique lui permettant d'exprimer son besoin sous forme d'une question « *Why* » en LN et de visualiser des résultats en LN et sous forme graphique (voir figure 6).

Pour montrer l'efficacité de notre approche, nous avons calculé le temps de génération des réponses d'une question « *Why* ». Faute d'approche similaire dans la littérature, nous avons comparé notre approche avec une approche naïve. Cette dernière génère des instances sous forme de tableaux pour chaque dimension et chaque mesure de l'ED. Le décideur analyse manuellement et séparément ces différents résultats tandis que notre approche génère automatiquement des résultats graphiques et des réponses en LN dans une seule interface. Nous avons

3. /msftdbprodsamples.codeplex.com
4. Les 50 questions sont accessibles à : //wq-bi.jimdo.com/

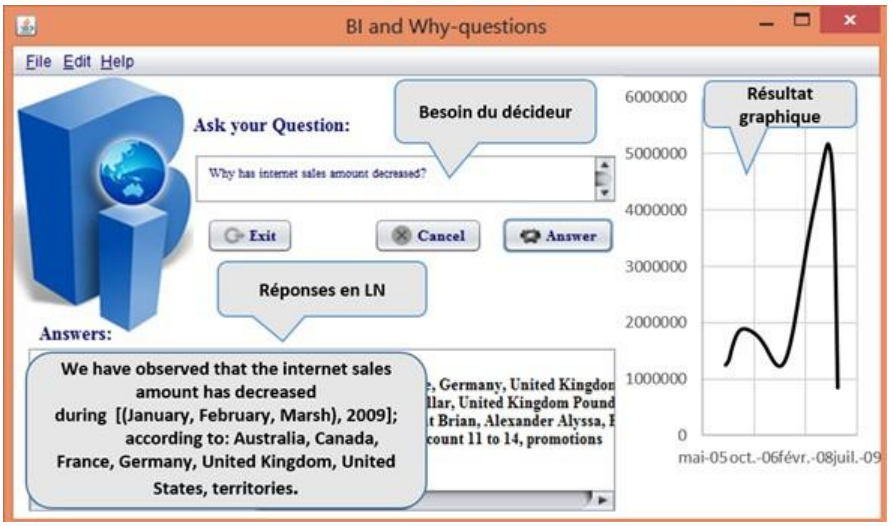


FIG. 6 – *Capture d'écran du Why-Question analyzer.*

calculé le temps de réponse de chaque approche pour un ED comportant une mesure et neuf dimensions (y compris la dimension temporelle).
Pour comparer les deux approches dans un cas d'utilisation réelle, nous incluons le temps nécessaire pour qu'un décideur analyse les résultats. Nous avons estimé que le temps d'analyse est d'une minute quand il s'agit d'un doublet < mesure, dimension>. Le temps d'analyse n'est pas linéaire, par conséquent, nous avons classé les dimensions à partir de la dimension qui génère un minimum de résultats à celle qui produit plus de résultats. La figure 7 montre le temps de réponse des deux approches en intégrant le temps d'analyse du décideur. Les résultats montrent que notre approche, par rapport à l'approche naïve, permet au décideur d'avoir des résultats à ses questions « *Why* » en un temps très raisonnable.

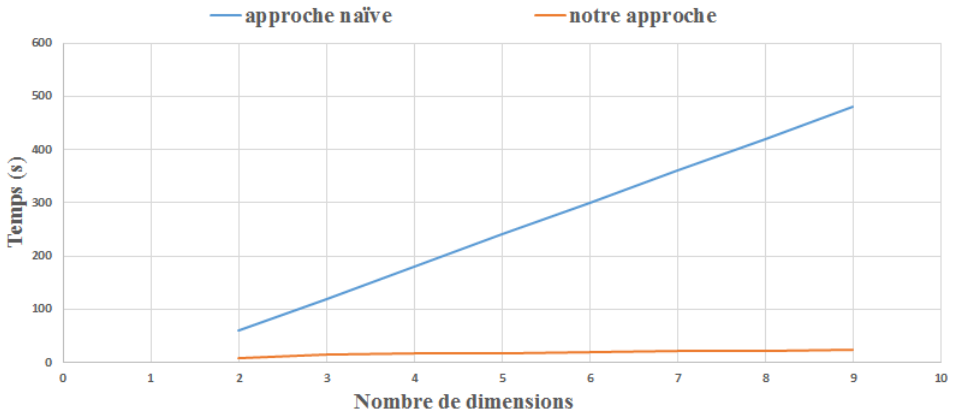


FIG. 7 – *Résultats de l'évaluation préliminaire.*

7 Conclusion et Perspectives

Dans cet article, nous avons proposé une approche abordant les questions décisionnelles de type « *Why* » ainsi qu'une modélisation pour ce type de questions. Notre approche est basée sur un modèle mathématique permettant de synthétiser une perception qualitative des données numériques afin de détecter les tendances d'une mesures observées selon les différents axes d'analyse de l'ED. Nous avons développé l'outil *Why-Question Analyser* qui permet à un décideur d'exprimer son besoin sous forme d'une question « *Why* » et lui fournir un ensemble de réponses en LN et des résultats graphiques.

Dans notre approche, les réponses à une question « *Why* » sont générées uniquement à partir de l'ensemble des dimensions de l'ED or que les réponses à une question « *Why* » peuvent être liées à d'autres mesures de l'ED. Nous travaillons actuellement sur cette partie de l'approche pour exploiter les éventuelles relations de cause à effet entre les mesures pour améliorer les réponses générées.

L'approche proposée explore uniquement l'ED pour extraire des réponses alors que si la réponse est partiellement ou totalement externe à l'ED (données météorologiques, événements externes, etc.), l'approche reste limitée. Nous comptons intégrer des sources de données externes à l'ED pour apporter des éléments de réponse aux questions « *Why* » quand les données de l'ED ne nous permettent pas de les avoir.

Références

Bajwa, I. S., M. Lee, et B. Bordbar (2012). Translating natural language constraints to ocl. *Journal of King Saud University-Computer and Information Sciences 24*(2), 117–128.

Baral, C., N. Ha Vo, et S. Liang (2012). Answering why and how questions with respect to a frame-based knowledge base : a preliminary report. In *LIPIcs-Leibniz International Proceedings in Informatics*, Volume 17. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Bargui, F., J. Feki, et H. Ben-Abdallah (2008). A natural language approach for data mart schema design. In *Proc. of 9th Int. Arabic Conference on Information Technology (ACIT), Hammamet-Tunisia.*

Derbal, K. A., Z. Tahar, K. Boukhalfa, I. Frihi, et Z. Alimazighi (2016). From spatial data warehouse and decision-making tool to SOLAP generalisation approach for efficient road risk analysis. *IJITM 15*(4), 364–386.

Dufour, J.-M. (2003). Ajustement de courbes de tendance par des méthodes de régression.

Girju, R. (2003). Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12*, pp. 76–83. Association for Computational Linguistics.

Hearst, M. A. (2011). 'natural'search user interfaces. *Communications of the ACM 54*(11), 60–67.

Kato, T., J. Fukumoto, F. Masui, et N. Kando (2005). Are open-domain question answering technologies useful for information access dialogues ?—an empirical study and a proposal of a novel challenge. *ACM Transactions on Asian Language Information Processing (TALIP) 4*(3), 243–262.

Kuchmann-Beauger, N. (2013). *Question Answering System in a Business Intelligence Context.* Ph. D. thesis, Ecole Centrale Paris.

Kuchmann-Beauger, N. et M.-A. Aufaure (2011). A natural language interface for data warehouse question answering. In *Natural Language Processing and Information Systems*, pp. 201–208. Springer.

Kuchmann-Beauger, N. et M.-A. Aufaure (2012). Natural language interfaces for datawarehouses. In *8èmes Journées francophones sur les Entrepôts de Données et l'Analyse en ligne*.

Manning, C. D. (2011). Part-of-speech tagging from 97% to 100% : is it time for some linguistics ? In *Computational Linguistics and Intelligent Text Processing*, pp. 171–189. Springer.

Mellac, K. (2013). Méthodes d'analyse de données en régression non linéaire. méthodologie. Technical report, Méthodologie [stat.ME].

Moriceau, V., X. Tannier, et M. Falco (2010). Une étude des questions "complexes" en question-réponse. In *Actes de la Conférence Traitement Automatique des Langues Naturelles (TALN 2010, article court), Montréal, Canada*.

Naeem, M. A., S. Ullah, et I. S. Bajwa (2012). Interacting with data warehouse by using a natural language interface. In *Natural Language Processing and Information Systems*, pp. 372–377. Springer.

Oh, J.-H., K. Torisawa, C. Hashimoto, T. Kawada, S. De Saeger, J. Kazama, et Y. Wang (2012). Why question answering using sentiment analysis and word classes. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 368–378. Association for Computational Linguistics.

Oh, J.-H., K. Torisawa, C. Hashimoto, M. Sano, S. De Saeger, et K. Ohtake (2013). Why-question answering using intra-and inter-sentential causal relations. In *ACL (1)*, pp. 1733–1743.

Popowich, F., M. Mosny, et D. Lindberg (2012). Interactive natural language query construction for report generation. In *Proceedings of the Seventh International Natural Language Generation Conference*, pp. 115–119. Association for Computational Linguistics.

Saias, J., P. Quaresma, P. Salgueiro, et T. Santos (2012). Binli : An ontology-based natural language interface for multidimensional data analysis. *Intelligent Information Management 4*(5).

Verberne, S. (2006). Developing an approach for why-question answering. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics : Student Research Workshop*, pp. 39–46. Association for Computational Linguistics.

Verberne, S. (2007). Paragraph retrieval for why-question answering. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 922–922. ACM.

Verberne, S., L. Boves, N. Oostdijk, et P.-A. Coppen (2007). Evaluating discourse-based answer extraction for why-question answering. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 735–736. ACM.

Summary

Business Intelligence applications provide the company's decision maker a set of aggregated information and decision-making indicators to facilitate the understanding of the current company's functioning and the identification of market trends for decision making process. To access decisional information, the decision maker should expresses his / her requirements in natural language rather than in formal language. The question of the decision-maker expressed in natural language can be formulated according to a set of key words or a sentence starting with: *What, Who, Why,* etc. (Known as "WH-questions"). In the literature, several works have focused on the questions "What" and "Who", while the question "*Why*" omnipresent in the decision-making process, has not been not addressed in a context of Business Intelligence. In this paper, we are interested in questions of type "*Why*" for which we propose a modeling and an approach based on trend's analysis. We have developed a tool named *Why-Question analyzer* to show the feasibility of our approach.

Une extension du standard XACML basée sur ARBAC pour contrôler l'accès à différents niveaux de données hébergées dans un environnement de Cloud

Sara Namane*, Nouria Harbi**, Nacira Ghoualmi*
*Laboratoire Réseaux et Systèmes, Université Badji Mokhtar, Annaba, Algérie
naamanesara2005@yahoo.fr, ghoualmi@yahoo.fr
** Laboratoire ERIC, Université Lyon II, Lyon, France
nouria.harbi@univ-lyon2.fr

Résumé. Le stockage des données chez un fournisseur de services info-nuage public est un service qui a atteint un développement sans précédent. La confidentialité des données reste la préoccupation majeure des propriétaires, c'est ce qui a empêché l'émergence généralisée vers une solution Cloud. Pour faire face à cette problématique plusieurs solutions utilisant la cryptographie ont été proposées. Lorsque les données sont chiffrées, l'exécution des requêtes devient plus difficile, d'autre part le propriétaire n'a pas la possibilité de spécifier la partie de données accessible. Dans cet article nous présentons un modèle de contrôle d'accès qui est une extension XACML basée sur ARBAC. Les données hébergées sont considérées selon les niveaux hiérarchiques de la structure d'une base de données : niveau base de données, niveau table ou niveau colonne. La proposition optimise la décision d'accès en minimisant le nombre de politique de sécurité à vérifier et permet aux propriétaires de spécifier la partie de données accessible par un utilisateur.

1 Introduction

Le cloud computing est un modèle permettant d'accéder à un réseau partagé de ressources informatiques configurables (réseaux, serveurs, stockage, applications et services), qui peuvent être rapidement provisionnés et libérés avec un minimum d'effort de gestion [Mell P. et Grance T.]. Le stockage des données au sein du Cloud est un service qui a atteint un développement sans précédent. C'est ce qui a permis aux propriétaires de données de différentes organisations de stocker leurs données locales sur différentes zones de stockage virtuelles hébergées par le Cloud avec la possibilité d'atteindre leur contenu dès le besoin. Cependant, le paradigme de stockage de données introduira également quelques problèmes de sécurité tout en offrant beaucoup de commodités. Il est évident que les propriétaires de données s'inquiètent que leurs données soient mal utilisées ou accédées par des utilisateurs non autorisés. La confidentialité des données, la préservation de la vie privée ainsi que l'efficacité entravent l'expansion rapide du cloud. Un mécanisme de contrôle d'accès efficace et sécurisé devient un moyen pour faire face à ce dilemme. En outre, les politiques de sécurité peuvent utiliser différents formats de données allant d'une simple donnée à d'énormes bases, comment gérer l'accès à plusieurs données demandées par une requête quelque soit sa complexité tout en garantissant la souplesse et l'efficacité. Dans cet article, une extension du standard XACML basée sur ARBAC (Attribute and Role Based Access control) est proposée à fin

de garantir la confidentialité des données hébergées au Cloud tout en fournissant la possibilité d'accéder à différents niveaux de la structure d'une base de données (niveau base de données, niveau tables ou niveau colonnes) et omettre l'accès à d'autres. Le reste de cet article est organisé comme suit : un état de l'art sur le contrôle d'accès au Cloud computing est présenté dans la Section 2, la Section 3 présentera l'architecture proposée. La conclusion et les perspectives seront discutées en Section 4.

2 Etat de l'art sur le contrôle d'accès dans le cloud computing

Récemment, un grand nombre de modèles de contrôle d'accès au cloud computing ont été proposés. Dans cette section, ces travaux de recherche vont être décrits selon les trois modèles les plus connus : RBAC, ABAC et ARBAC

2.1 Les contrôles d'accès basés sur RBAC

Zhu et. al. (2011) ont proposé un modèle de contrôle d'accès (CoRBAC) basé sur RBAC, plus précisément sur le RBAC distribué (dRBAC), le modèle proposé fusionne les services d'authentification distribuée de (dRBAC) et étend la fonction du CA, en offrant l'authentification inter-domaines et l'affectation de rôles inter-domaines. En outre le modèle proposé a amélioré l'efficacité du contrôle d'accès en ajoutant des caches hiérarchiques.
W. Chunlei et al. (2012) ont introduit une valeur de permission, un rôle quantifié et une valeur de comportement pour construire un modèle de contrôle d'accès basé sur un rôle quantifié. Ce modèle a été validé dans un prototype du cloud computing, ce qui a donné une réduction du nombre de rôles, une amélioration du processus d'autorisation et une implémentation dynamique des permissions.
L. Sun et al. (2012) ont proposé un modèle de contrôle d'accès sémantique basé sur RBAC. Des vocabulaires structurés et hétérogènes ont été utilisé avec les ontologies dans le système e-healthcare, ce qui a permis de résoudre le problème d'un contrôle d'accès distribué dans un environnement dynamique tel que le cloud computing. Ce modèle reste un modèle purement théorique car aucune implémentation n'a encore été proposée.
F. Yue-qui et al. (2015) ont proposé un modèle de contrôle d'accès spécifique au cloud computing qui se repose sur RBAC ainsi qu'un modèle d'accès basé sur les tâches. Ce qui a permis d'avoir les avantages des deux modèles en intégrant une valeur de réputation. Cette valeur a pour but de diminuer le nombre d'accès non autorisés. Le point faible de cet article est que la valeur de réputation n'a pas été précisée.

2.2 Les contrôles d'accès basés sur ABAC

D.R Dos Santos et al. (2013) ont proposé un modèle de contrôle d'accès à base de risque au sein d'une fédération de cloud sans la nécessité d'une fédération d'identité. Le modèle proposé utilise des politiques de risques sous la forme de fichiers XML, lors d'une demande d'accès par un utilisateur du même cloud, la requête est gérée par le modèle ABAC classique, par contre lors d'une demande d'accès par un utilisateur appartenant à un autre cloud,

si il n'y a pas une fédération d'identité entre les deux cloud , le modèle de contrôle d'accès à base de risque est activé.
Riad et al. (2015) ont proposé un modèle de contrôle d'accès AR-ABAC qui utilise le modèle ABAC classique et une nouvelle notion proposée par les auteurs nommée les règles d'attributs (AR). Ces règles d'attributs permettent l'accès aux objets selon leurs degrés de sensibilité, d'autre part elles permettent de déterminer combien d'attributs et quel type d'attributs sont utilisés pour prendre une décision d'accès.
A. Chen et al. (2016) ont proposé un modèle de contrôle d'accès dynamique basé sur le risque, ce modèle repose sur le modèle ABAC ainsi qu'un mécanisme d'évaluation de risques. Les auteurs ont analysé la régression itérative en se basant sur une fenêtre de flux de données, c'est ce qui a permis de calculer efficacement les facteurs de risque environnementaux d'un demandeur.
F. Khan et al. (2016) ont proposé un modèle de contrôle d'accès basé sur le cryptage d'attributs à plusieurs autorités. Le modèle proposé permet au propriétaire de données de spécifier quelle partie de données peut être accéder par un utilisateur en utilisant des attributs cryptés, d'autre part le modèle proposé a pris en considération la réduction du cout du cryptage en diminuant les attributs répétitifs. Ce modèle n'a pas pris en considération le cas où les utilisateurs quittent le système et reste un modèle purement théorique car aucune implémentation n'a encore été proposée.

2.3 Les contrôles d'accès basés sur RBAC et ABAC

E. Mon et al. (2011) ont proposé un modèle de contrôle d'accès qui combine les deux approches RBAC et ABAC pour assurer la confidentialité des données dans un cloud privé. Ce modèle reste un modèle purement théorique car aucune implémentation n'a encore été proposée.

2.4 Comparaison et synthèse sur les travaux existants

L'étude des travaux cités auparavant (TAB.1) sur le contrôle d'accès dans un environnement de cloud, nous a permis de trouver ce qui suit :

Les modèles de contrôle d'accès basés sur ABAC ont été tous validé dans un environnement de cloud, tandis que ceux utilisant RBAC ne l'ont pas tous été, ce qui montre la souplesse du modèle ABAC par rapport à RBAC dans un environnement distribué et dynamique tel que le cloud computing.

La gestion des attributs n'est pas applicable dans les modèles basés sur RBAC, car ceux-ci n'utilisent pas d'attributs. Les travaux basés sur ABAC n'ont pas pris en considération ce point, (Riad et al. 2015) ont utilisé des règles d'attributs pour déterminer le nombre et le type d'attributs à utiliser lors d'une décision d'accès.

(F. Khan et al., 2016) ont proposé un modèle de contrôle d'accès qui a permis de spécifier la partie de données accessible par un utilisateur contrairement aux autres travaux qui n'ont pas tenu compte de ce point important qui pourra faire partie des exigences d'un propriétaire de données.Tous les travaux cités auparavant n'ont pas tenu compte de la minimisation du nombre de politiques de sécurité à vérifier, ce qui pourra réduire d'une manière significative le temps de réponse à une requête d'accès.

(E. Mon et al. , 2011) ont proposé un modèle de contrôle d'accès basé sur les deux modèles RBAC et ABAC à la fois, ce modèle n'a géré aucun des critères cités auparavant.

Modèle de classification	Citation	Approche	Techniques	Validation	Gestion des attributs	décomposition des données	minimisation des politiques à verifier
RBAC	Zhu et. Al (2011)	dRBAC	•Authentification inter-domaine • Affectation de rôles inter-domaine • Caches hiérarchiques	⊕	•	☒	☒
	W. Chunlei et Al (2012)	Rôle quantifié	• Valeur de permission • Role quantifié • Valeur de comportement	⊕	•	☒	☒
	L. Sun et Al (2012)	Un contrôle d'accès sémantique	• Ontologies • Vocabulaires structurés et hétérogènes	☒	•	☒	☒
	F. Yue-qui et Al (2015)	Un contrôle d'accès à base de tâche s	• Valeur de réputation	⊕	•	☒	☒
ABAC	D.R Dos Santos et Al (2013) & A. Chen et Al (2016)	Un contrôle d'accès à base de risque	• Une politique de risque • Une politique ABAC	⊕	☒	☒	☒
	Riad et Al (2015)	Un contrôle d'accès selon la sensibilité des objets	• Politiques ABAC • Règles d'attributs	⊕	⊕	☒	☒
	F. Khan et Al (2016)	Un contrôle d'accès basé sur le cryptage	• Plusieurs autorités d'attributs • Cryptage des attributs	⊕	☒	⊕	☒
ARBAC	E. Mon et Al (2011)	Un contrôle d'accès selon le degré de sensibilité des données	• Niveau de sécurité des données • Niveau d'utilisateur • Un gestionnaire de sécurité	☒	☒	☒	☒

TAB 1 *comparaison des travaux existants*

☒ Non supporté ⊕ Supporté ● Non applicable

3 Contrôle d'accès à des données organisées d'une manière hiérarchique basé sur les modèles RBAC, ABAC et une extension du modèle XACML

3.1 Motivation

Le contrôle d'accès est un mécanisme qui permet d'assurer la sécurité des données hébergées au sein du cloud en spécifiant les permissions acceptées pour chaque utilisateur. Selon les travaux cités dans la section précédente, un modèle de contrôle d'accès est basé sur RBAC, ABAC ou ARBAC. Une requête générée par un utilisateur peut demander l'accès à différents types de données. Ces données sont généralement organisées dans des bases de données composées de plusieurs tables. Chaque table contient plusieurs colonnes qui représentent les attributs de la table. D'un point de vue physique, ces données sont vues comme une hiérarchie de données. D'un autre côté le propriétaire peut spécifier la partie des données accessible pour certains utilisateurs et omettre l'accès à une autre partie au sein d'une même table en utilisant des politiques de sécurité spéciales. Comment peut un modèle de contrôle d'accès gérer des requêtes d'accès à des données appartenant à différents niveaux hiérarchiques de la structure d'une base de données (niveau base de données, niveau table, niveau colonnes) tout en respectant les politiques de sécurité exigées par le propriétaire des données ? Quel langage pouvons-nous utiliser pour spécifier les politiques de sécurité pour des données organisées d'une telle manière tout en assurant un contrôle d'accès efficace ? Comment réduire le nombre de politiques à vérifier afin d'améliorer le processus de contrôle d'accès et diminuer le temps de réponse à des requêtes d'accès dans un environnement distribué et avec un grand nombre d'utilisateurs tel que le cloud ?

3.2 Le modèle proposé

Un environnement de cloud computing est souvent composé d'un fournisseur de service (A) qui fournit le service de stockage et une organisation (B) qui fait appel à ce service pour héberger ses données dans des bases de données (i). Le modèle proposé (FIG. 1) se compose des éléments suivants :

(a) le propriétaire des données, est celui qui héberge les données au sein du cloud, spécifie les règles et les politiques de sécurité.
(b) Les utilisateurs de données : ce sont ceux qui génèrent des requêtes pour manipuler les données.
(c) Le gestionnaire de données : il intercepte les requêtes SQL, extrait les données concernées et les envoie au PEP.
(d) PEP (Policy enforcement point) fait partie du modèle XACML, d'habitude nous le trouvons dans le serveur d'autorisation avec les autres composants, il se charge de créer les requêtes XACML et les envoyer au PDP.
(e) le serveur d'autorisation qui est une extension du modèle XACML, contient le PDP (Policy decision point) : le module qui permet de vérifier les valeurs des attributs des requêtes d'accès avec les valeurs d'attributs des politiques de sécurité et des métapoli-

tiques. Le PIP (Policy information point) est l'élément qui fournit les attributs manquants concernant les utilisateurs de l'organisation. La base de politiques contient toutes les politiques de sécurité concernant les tables et les colonnes. La base des métapolitiques contient les politiques de sécurité concernant les bases de données hébergées.

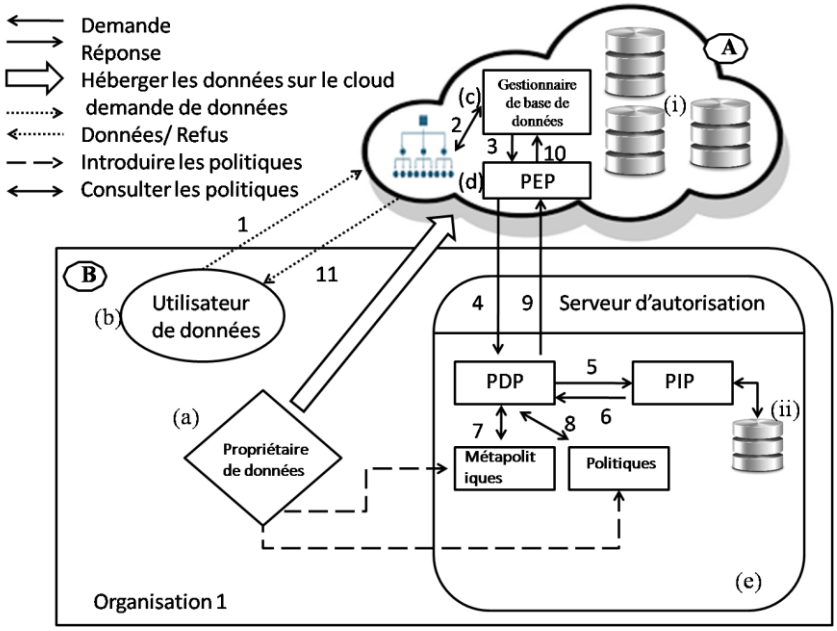


FIG. 1– *Architecture proposée*

Dans cet article, les deux modèles RBAC et ABAC ont été combinés et déployés dans une architecture XACML distribuée, le choix du modèle RBAC est dû à l'importance qu'il donne aux rôles des utilisateurs par rapport à leur identité. Dans une organisation, la notion de rôle est importante pour déterminer le contenu que peut consulter un utilisateur, l'utilisation de RBAC au sein de notre modèle a permis de gérer les données de manière efficace sachant qu'un utilisateur peut occuper plusieurs rôles, en outre un rôle peut être affecté à plusieurs employés ce qui nous permet de gérer les permissions affectées aux rôles et non les permissions affectées à chaque employé. Le modèle ABAC a été utilisé pour sa souplesse et son dynamisme. Ce modèle vérifie les valeurs des différents attributs avec les objets de la politique d'accès, l'un de ses points faibles est lié à la gestion d'attribut en termes de nombre et de type à utiliser pour prendre une décision concernant une demande d'accès. Pour faire face aux inconvénients des deux modèles, nous les avons fusionnés pour tirer profit des avantages de chacun. Le modèle proposé utilise le standard XACML (Rissanen E., 2014) pour spécifier les politiques de sécurité et les métapolitiques pour des données organisées de manière hiérarchique. D'autre part, le PEP a été mis du côté du fournisseur pour permettre la création des requêtes XACML, tandis que les politiques et leur vérification sont du côté propriétaire pour garantir leur intégrité et éviter leur cryptage qui ajoutera des tâches admi-

nistratives supplémentaires au propriétaire. La division des politiques en deux ensembles facilitera la tâche à l'administrateur et évitera la vérification de l'ensemble des politiques de sécurité par le PDP ce qui permet d'avoir un modèle plus efficace.

3.3 Exemple d'un contrôle d'accès à des données d'une organisation hébergées dans un environnement de cloud

Soit l'exemple suivant, un ensemble de données d'une organisation qui sont organisées dans deux bases de données (BDD 1 et BDD 2) avec différentes tables (achats, vente,…), chaque table contient plusieurs attributs (Produit, fournisseur, …) (FIG.2). Les étapes suivantes sont réalisées lors du passage d'une organisation à une solution cloud.

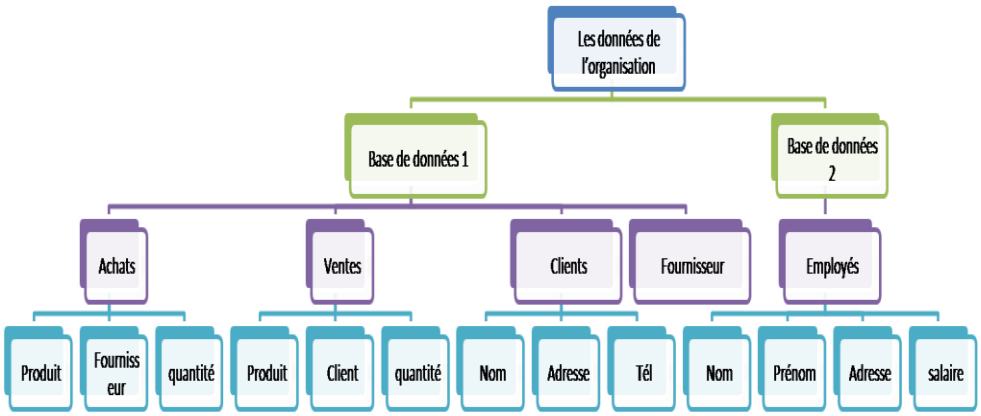


FIG. 2– *Plan des données du propriétaire*

Etape 1 : Hébergement des données par le propriétaire

Les données d'une entreprise ou organisation sont souvent sous la forme d'une base de données. Une base de données est composée de tables et chaque table d'un ensemble de colonnes, l'organisation de ces données est faite de manière hiérarchique car la demande d'accès sera pour une colonne dans une table qui appartient à une base de données. Chaque donnée est identifiée en utilisant son URI (Berners-Lee T. et aL., 2005) car les données peuvent être répétitives, de cette manière l'URI permet de trouver la donnée demandée par l'utilisateur. Lorsqu'un propriétaire décide d'héberger ses données dans un Cloud, il crée les bases de données adéquates, crée les tables ainsi que le plan de ses données (FIG.2). Le propriétaire doit définir les politiques d'accès aux bases de données (Métapolitiques) ainsi que les politiques d'accès aux tables et aux colonnes (politiques) en utilisant le rôle des utilisateurs au sein de l'organisation, la spécification de ces politiques est faite en langage XACML 3 pour ressources hiérarchiques avec décisions multiples (FIG.3). Les données sont ensuite hébergées sur le Cloud, tandis que les métapoli-

tiques et les politiques sont mises dans les deux bases spécifiques aux politiques au sein de l'organisation.

```
<Policy xmlns="urn:oasis:names:tc:xacml:3.0:core:schema:wd-17" PolicyId="Resource-Five-Get-Policy"
RuleCombiningAlgId="urn:oasis:names:tc:xacml:1.0:rule-combining-algorithm:deny-overrides" Version="1.0">
  <Target> <AnyOf> <AllOf>
<Match MatchId="urn:oasis:names:tc:xacml:1.0:function:string-equal">
  <AttributeValue DataType="http://www.w3.org/2001/XMLSchema#string">BDD1/ACHAT/PRODUIT</AttributeValue>
  <AttributeDesignator AttributeId="urn:oasis:names:tc:xacml:1.0:resource:resource-id"
  Category="urn:oasis:names:tc:xacml:3.0:attribute-category:resource" DataType="http://www.w3.org/2001/XMLSchema#string"
  MustBePresent="true"/> </Match> </AllOf> </AnyOf> </Target>
<Rule Effect="Permit" RuleId="Rule-1"> <Target><AnyOf><AllOf>
<Match MatchId="urn:oasis:names:tc:xacml:1.0:function:string-equal">
  <AttributeValue DataType="http://www.w3.org/2001/XMLSchema#string">SELECT</AttributeValue>
  <AttributeDesignator AttributeId="urn:oasis:names:tc:xacml:1.0:action:action-id"
Category="urn:oasis:names:tc:xacml:3.0:attribute-category:action" DataType="http://www.w3.org/2001/XMLSchema#string"
MustBePresent="true"/> </Match> </AllOf> </AnyOf> </Target> <Condition>
<Apply FunctionId="urn:oasis:names:tc:xacml:1.0:function:any-of">
  <Function FunctionId="urn:oasis:names:tc:xacml:1.0:function:string-equal"/>
  <AttributeValue DataType="http://www.w3.org/2001/XMLSchema#string">GESTIONNAIRE_STOCK</AttributeValue>
  <AttributeDesignator AttributeId="http://wso2.org/claims/role" Category="urn:oasis:names:tc:xacml:1.0:subject-
category:access-subject" DataType="http://www.w3.org/2001/XMLSchema#string" MustBePresent="true"/>
</Apply> </Condition>
</Policy>
```

FIG. 3– *exemple de politique de sécurité écrite en XACML 3 qui dit que pour accéder à la colonne produit de la table achat de la base de données 1 avec une action Select il faut que l'utilisateur ait un rôle « Gestionnaire de Stock ».*

Etape 2 : Demande d'accès par un utilisateur

Chaque utilisateur au sein de l'organisation a un ou plusieurs rôles qu'il occupe. Tous ces rôles ainsi que toutes les informations relatives aux utilisateurs sont stockés au niveau de l'organisation dans une base de données au sein du serveur d'autorisation (ii). Lorsqu'un utilisateur U_1 envoie au fournisseur de service une requête de manipulation de données de type (FIG.4) : SELECT, UPDATE ou autre (1) ; le gestionnaire de données retire de cette requête les données concernées, cherche dans le plan des données pour trouver la base concernée (2), ensuite il envoie toutes ces données au PEP (3). Ce dernier crée une requête XACML de type : données hiérarchiques avec décisions multiples et l'envoie au PDP (4). Le PDP retire la base concernée du plan envoyé par le propriétaire, demande au PIP les attributs manquants concernant l'utilisateur (5), le PIP cherche tout le nécessaire concernant les utilisateurs et les données et les renvoie au PDP (6), ensuite le PDP vérifie les métapolitiques de sécurité(7) , si l'utilisateur peut accéder à cette base, le PDP continue de vérifier les politiques concernant le reste des données demandées(8), sinon il envoie un refus comme réponse au PEP (9)qui la transmettra au gestionnaire de données (10), ce dernier la transmet à l'utilisateur (11). Si l'utilisateur a accès à la base, après vérification de l'accès aux autres données, il envoie sous forme d'une réponse multiple la réponse à chaque donnée demandée

(FIG.5). Cette réponse multiple sera envoyée au gestionnaire de données. (Illustré dans la FIG.1)

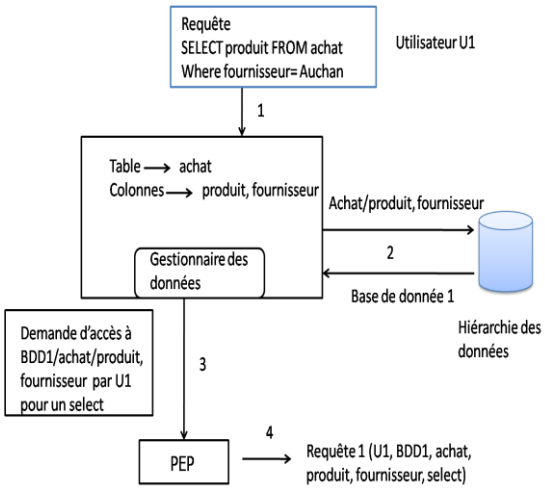


FIG. 4– *Demande d'accès au niveau du fournisseur*

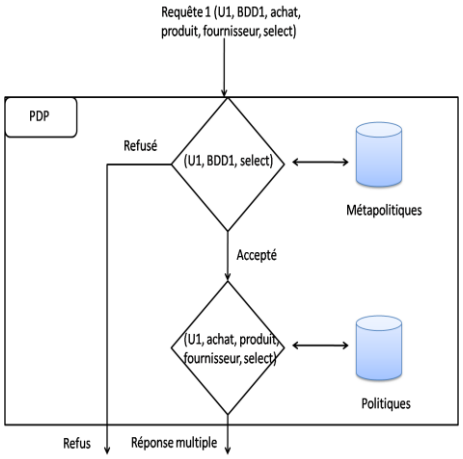


FIG. 5– *La prise de décision par le PDP*

Etape 3 : Gestion d'un résultat avec réponse multiple

Lors de la réception du résultat multiple par le gestionnaire de données, il doit faire des tests pour savoir s'il donne l'accès aux données demandées et à quelle partie exactement l'utilisateur y est autorisé. Si le résultat est refusé pour toutes les données alors le gestionnaire de données envoie un refus à la demande de l'utilisateur. Le gestionnaire suit la règle suivante : Si l'accès à une donnée de niveau supérieur dans la hiérarchie est refusé, alors que l'accès pour l'une de ses descendantes est accepté alors le résultat sera refusé, sinon il sera le même que celui obtenu, si l'utilisateur a accès à une partie de données et pas à une autre, le gestionnaire lui demandera de refaire sa requête car il demande l'accès à des données non autorisées (FIG.6).

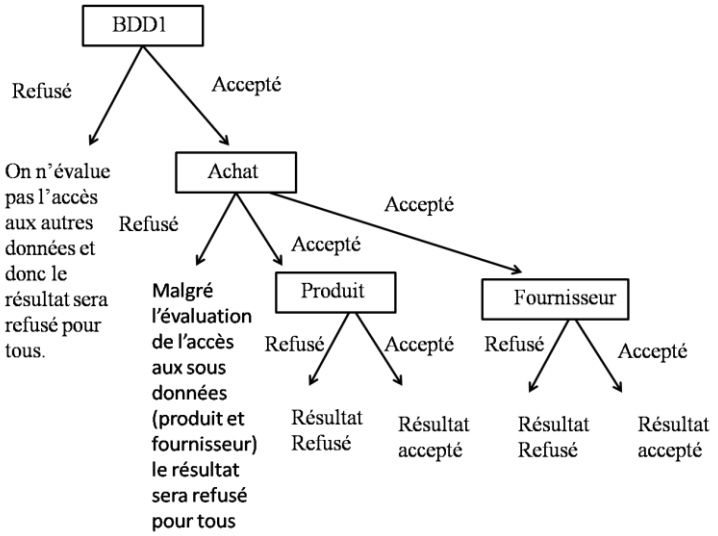


FIG. 6 – *Exemple d'un résultat multiple*

4 Conclusion et perspectives

La confidentialité des données hébergées sur le cloud préoccupe les propriétaires, divers travaux ont assuré la confidentialité en utilisant le cryptage d'attributs, cette technique a permis la protection des données mais elle a aussi rajouté des tâches administratives au propriétaire qui doit gérer la distribution des clés et le cryptage de toutes ses données et ses politiques. Dans cet article, nous avons proposé un modèle de contrôle d'accès qui est une extension du standard XACML basé sur ARBAC. Le modèle proposé prend en considération le niveau de données auquel l'utilisateur peut accéder ce qui facilite la tâche au propriétaire et lui permet de gérer l'accès à ces données selon le niveau souhaité. Le modèle proposé utilise le standard XACML 3 pour la spécification des politiques de sécurité et des métapolitiques, ce standard permet de gérer les demandes d'accès à des ressources hiérarchiques avec des décisions multiples ce qui réduit le parcours de longs fichiers XML de politiques et donne un contrôle d'accès efficace. L'un des points qui reste à établir, est l'implémentation du modèle de contrôle d'accès proposé dans un environnement du cloud computing.

Références

Barkley J., Beznosoz K. Et Uppal J. (1999)., *Supporting Relationships in Access Control Using Role Based Access Control*. Proceeding of the ACM workshop on RBAC, Fairfax, Virginia, USA.

Benantar M. (2006), book: *Access Control Systems Security, Identity Management and Trust Models*, chapter: access control systems, Mandatory-Access-Control Model p 129-146.

Berners-Lee T., Fielding R. et Masinter N. (2005), *Uniform Resource Identifiers (URI): Generic Syntax*, IETF RFC 3986, http://www.ietf.org/rfc/rfc3986.txt, [RFC3986], January.

Chen A., Xing H. , She K. Et Duan G. (2016), *A Dynamic Risk-based Access Control Model for Cloud Computing*, IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom).

Chunlei W., Zhongwei L. and Xuerong C. (2012), *An Access Control Method of Cloud Computing Resources Based on Quantified-Role*, 14th International Conference on Communication Technology, IEEE.

Dos Santos D. , Westphall C. Et Westphall C. (2013), *Risk-based Dynamic Access Control for a Highly Scalable Cloud Federation*, SECURWARE 2013: The Seventh International Conference on Emerging Security Information, Systems and Technologies.

Khaled R., Zhu Y., Hongxin H. Et Ahn G. (2015), *AR-ABAC: A New Attribute Based Access Control Model Supporting Attribute-Rules for Cloud Computing*, IEEE Conference on Collaboration and Internet Computing.

Khan F., Li H., Et Zhang L. (2016), *Owner Specified Excessive Access Control for Attribute Based Encryption*, DOI 10.1109/ACCESS.2016.2632132, IEEE Access.

Mell P., Grance T. (2011) , *The NIST Definition of Cloud Computing*, Computer Security Division Information Technology Laboratory National Institute of Standards and Technology Gaithersburg, MD 20899-8930

Mon E. Et Naing T. (2011)., *The privacy-aware access control system using attribute-and role-based access control in private cloud*, proceedings of IEEE IC-BNMT.

Ninghui L. (2011), *Discretionary Access Control*, pp 353-356, Encyclopedia of Cryptography and Security.

Rissanen E. (2014), *XACML v3.0 Multiple Decision Profile Version 1.0* ,18 May 2014. OASIS Committee Specification 02. http://docs.oasis-open.org/xacml/3.0/multiple/v1.0/cs02/xacml-3.0-multiple-v1.0-cs02.html.Latest version: http://docs.oasis-open.org/xacml/3.0/multiple/v1.0/xacml-3.0-multiple-v1.0.html.

Sun L., Wang R., Yong J. and Wu G. (2012), *Semantic access control for cloud computing based on e-Healthcare*, Proceedings of the 2012 IEEE 16th International Conference on Computer Supported Cooperative Work in Design.

Yuan E. Et Tong J. (2005), *Attributed Based Access Control (ABAC) for Web Service*, The 2005 IEEE International conference on web service (ICWS'05).

Yue-qin F. ET Yong-sheng Z. (2012), *Trusted Access Control Model Based on Role and Task in Cloud Computing*, 7th International Conference on Information Technology in Medicine and Education .

Zhu T. , Liu W. et Song J. (2011), *An efficient Role Based Access Control System for Cloud Computing*, 11th IEEE International Conference on Computer and Information Technology.

Summary

Data storage in cloud computing is a service that has reached unprecedented development. The confidentiality of data remains the major concern of owners. This is why the use of cloud solution leads data owner to reconsider their decision. To face this dilemma several solutions using cryptography were proposed. Nevertheless, when data is encrypted, performing queries becomes more challenging, on one hand and on the other hand; data owners did not have possibility of specifying the accessible part of data (all database, one table from database or some columns from one table). This article presents an access control model which is an extension of XACML standard based on ARBAC. The outsourcing data is considered using the hierarchical level in database structure: database level, table level, and column level. The proposal optimized the access control mechanism by reducing the number of security policies whose be checked and allowed owners to specify the part of data that user can access.

Du réparti vers le cloud et les big data

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

*Université de Tunis El Manar, Faculté des Sciences de Tunis,
Département informatique, URAPOP, El manar 2092, Tunis, Tunisie.
ghorbel.fst@gmail.com,
**Université de Tunis, Ecole Supérieure des Sciences Economiques
et Commerciales de Tunis, Montfleury 1089, Tunis, Tunisie.
karima.tekaya@gmail.com
***Université de Tunis El Manar, Faculté des Sciences de Tunis,
Département informatique, LIPAH, El manar 2092, Tunis, Tunisie.
abdelaziz.abdellatif@fst.rnu.tn

Résumé. La répartition d'un entrepôt de données (ED) devient de nos jours très utile vue la charge d'informations qui s'accroît sans cesse dans n'importe quelle société et l'augmentation des besoins des utilisateurs de cet entrepôt. Outre la répartition, de nouvelles techniques sont apparues pour remédier à cette augmentation cruciale des données comme les Big data et le Cloud computing. Dans cet article, nous proposons une étude comparative des techniques des répartitions des ED pour minimiser le temps d'exécution des requêtes dans un contexte réparti. Nous choisissons la classification comme technique de répartition et de la tester selon deux benchmark APB-1 et TPC-H. Nous ferons aussi un tour d'horizon sur les travaux de répartitions des ED, ensuite nous orientons nos travaux vers les Cloud et les Big data comme solutions.

1 Introduction

Face à la mondialisation et à la concurrence grandissante, la prise de décision est devenue cruciale pour les dirigeants d'entreprises (au sens large du terme, entreprises privées, publiques, institutions, organisations...). L'efficacité de cette prise de décision repose sur la mise à disposition d'informations pertinentes et d'outils d'analyse adaptés. L'objectif des entreprises est de pouvoir exploiter efficacement d'importants volumes d'informations, provenant soit de leurs systèmes opérationnels, soit de leur environnement extérieur, pour l'aide à la décision.
L'informatique décisionnelle a connu et connaît aujourd'hui encore un essor important. Elle permet l'exploitation des données d'une organisation dans le but de faciliter la prise de décision. À l'heure actuelle, la majorité des gros ED souffrent de problèmes de performance causant des problèmes aux usagers, les administrateurs et les développeurs du data warehouse (DW). Il n'y a pas de recettes magiques qui résolvent en un seul coup tous ces problèmes. Il faut quasiment les étudier un par un. Par contre, ils sont pratiquement communs à des ED différents. Ils ne s'appliquent pas à un domaine bien spécifique.

Ces charges de travail volumineuses peuvent dégrader les performances du système de gestion des bases de données (SGBD), et ainsi, ralentir les applications et augmenter ainsi le temps de réponse au client, souvent exigeant dans les délais, en particulier lorsqu'il s'agit d'un décideur. Pour y remédier, diverses techniques d'optimisation ont été proposées, entre autres la fragmentation horizontale. La sélection d'un schéma de fragmentation horizontale (FH) est un problème NP-complet Boukhalfa (2009). Des algorithmes heuristiques ont été proposés afin de définir automatiquement le meilleur schéma de fragmentation.
Nous allons commencer alors par définir la problématique puis nous ferons également un tour d'horizon sur les différentes approches de fragmentation et d'évaluation de l'exploitation des ED.

2 Problématique

La répartition des bases de données classiques se base sur la fragmentation (généralement horizontale) et l'allocation de ces fragments. Cependant, cette technique nécessite une adaptation aux aspects fondamentaux des ED. Faisant face aux contraintes de chargement et aux spécificités de la modélisation multidimensionnelle, leur application demeure inadéquate. Malgré l'insuffisance de ces techniques, elles nécessitent une réévaluation dans le contexte réparti. En effet, très peu de travaux ont considéré le problème de la répartition des ED. Malgré son importance, nous considérons que la répartition des ED n'a pas eu l'attention qu'elle mérite.
D'autre part, le choix d'une solution de répartition doit être, tout d'abord, approuvé par des mesures d'efficacité. Cependant, l'évaluation des solutions de partitionnement et/ou de répartition dans les ED n'a pas été très sollicitée par les chercheurs. Cependant, nous avons aussi remarqué l'absence dans la littérature d'une technique d'évaluation du partitionnement d'un ED au niveau centralisé et réparti et les chercheurs adaptent leurs solutions sur la fonction objective de Chakravarthy et al. (1994).

Malgré l'intérêt accordé aux techniques de fragmentation des données et la diversité des solutions proposées, nous avons constaté que ce problème n'a pas eu l'attention qu'il mérite en dépit de son importance.

3 Démarche préconisée

Ceci nous amène à considérer les points suivants qui constituent les trois grands axes de ce travail :

1. Etude des techniques de partitionnement proposées ;
2. Application du partitionnement dans les ED centralisés et répartis ;
3. Etude des techniques d'évaluation pour le partitionnement des données.

Nous nous sommes fixé comme objectif de faire une étude des travaux de recherche présentés dans l'état de l'art et proposer une répartition et une évaluation de l'exploitation d'un ED. Dans ce contexte, les travaux qui vont être présentés dans cet article, constituent une investigation de l'optimisation par des techniques de calcul évolutif. Nous allons faire une comparaison des travaux présentés dans l'état de l'art pour choisir la meilleure technique de partitionnement et évaluer son efficacité. Ensuite, nous essayerons d'appliquer la meilleure technique de

partitionnement dans un ED réparti. Enfin, nous ferons un tour d'horizon sur les techniques d'évaluation d'un ED et les techniques de fouille de données pour calculer le coût engendré par un certain aménagement du stockage des données.

4 Etat de l'art

Plusieurs solutions ont été proposées pour la FH des ED relationnels. Dans ce qui suit, nous présentons un tour d'horizon sur quelques approches qui ont été développées dans ce contexte.

4.1 Niveau fragmentation

-Nehme et Valduriez (2011) ont défini une technique de Branch and Bound pour l'exécution des requêtes parallèles. C'est une stratégie de partitionnement de données qui minimise le coût coûteux des transferts de données. Cette solution présente comme avantage de trouver le meilleur partitionnement dans des environnements distribués mais reste le délais d'attente pour l'optimiseur qui présente le seul inconvénient.

-Bouchakri et Bellatreche (2012) ont proposé d'effectuer une sélection d'un schéma de fragmentation dite incrémentale basée sur les algorithmes génétiques et permettre l'optimisation de l'exécution de la charge de requêtes décisionnelles et l'adaptation du schéma de fragmentation aux changements de la charge. Cette solution présente l'avantage de prendre en considération les changements au niveau des requêtes mais cela entraîne l'augmentation du temps de mise à jour.

-Barr (2013) a résolu un problème de la sélection de la FH tout en considérant à la fois le nombre de I / O entre la mémoire et le disque. Il a pris les requêtes décisionnelles et le nombre de fragments, comme deux fonctions objectives à minimiser. Cette solution a permis la réduction du temps de réponse lors de la manipulation des requêtes et la reduction des fragments. Cependant ce travail peut être élargi pour prendre en charge d'autres préférences de l'administrateur.

-Elmansouri et al. (2013) ont défini une nouvelle technique (ACP) pour la fragmentation des ED. C'est une technique de description et de réduction des données. Elle comporte une série de décisions critiques portant sur les propriétés des variables soumises à l'analyse, les propriétés de la matrice d'intercorrélation et le nombre de composantes à extraire. Cette solution a comme avantage de mettre en évidence les fragments horizontaux supplémentaires mais cette solution n'est pas encore testé sur ED.

-Ettaoufik et al. (2013) proposent une approche de FH des ED en se basant sur la technologie des services web. C'est un service qui est décrit comme un ensemble de fonctionnalités accomplissant des tâches spécifiques, accessible par un réseau informatique. Le Web Service est un composant logiciel identifié par une URL. Il s'agit donc d'un ensemble de fonctionnalités exposées sur internet ou sur un intranet, par et pour des applications ou machines, sans intervention humaine, et en temps réel. Leur travail présente une amélioration des ED au niveau des fragmentations mais reste à ajouter d'autre préférence pour l'administrateur.

-Boufares et al. (2012), dans leur travail, assurent une meilleure qualité pour les données résultantes. Ils proposent un algorithme séquentiel pour l'élimination des données similaires. Actuellement, un grand nombre d'applications utilisent des données hétérogènes et distribuées de qualité variable. Le besoin d'intégration de données et d'évaluation de la qualité des données

se fait de plus en plus ressentir. De nouvelles méthodes de calculs de distance de similarité pour des types de données complexes sont à développer ainsi que l'automatisation du choix entre elles.

4.2 Niveau évaluation

-Kerkad et al. (2012) donnent une proposition d'étudier conjointement le problème de gestion de tampon (BMP) et le problème d'ordonnancement de requêtes (QSP). Les SGBD manipulant des bases de données volumineuses comme les ED stockent souvent les données sur le disque. En conséquence, l'interrogation nécessite un transfert des données du disque vers la mémoire centrale via le tampon. Un nombre important de travaux sur la gestion de tampon ont été proposés. Malheureusement, ils supposent que les requêtes soient ordonnées. Dans le contexte des ED, les requêtes interagissent du fait qu'elles utilisent la table des faits. Cette interaction pourrait impacter la gestion de cache et offrir un bon ordonnancement de requêtes.

-Perriot et al. (2013) proposent de nouveaux modèles de coût intégrant le paiement à la demande en vigueur dans les nuages. Ils définissent un problème d'optimisation consistant à sélectionner, parmi des vues candidates, celles à matérialiser pour minimiser le coût d'interrogation et de maintenance, ainsi que le temps de réponse pour une charge de requêtes donnée. Leur but d'optimiser les deux critères séparément : le temps est optimisé sous contrainte de coût et vice versa. La performance des ED est classiquement assurée grâce à des structures comme les index ou les vues matérialisées. Dans ce contexte, des modèles de coût permettent de sélectionner un ensemble pertinent de ce type de structures. Toutefois, cette sélection devient plus complexe dans les nuages informatiques.

-Favre (2007), dans sa thèse, présente une solution à la personnalisation des analyses dans les ED. Cette solution se base sur une évolution du schéma de l'entrepôt guidée par les utilisateurs. Il s'agit en effet de recueillir les connaissances de l'utilisateur et de les intégrer dans l'ED afin de créer de nouveaux axes d'analyse. Cette solution présente comme avantage une évaluation de la performance du modèle proposé mais présente aussi des inconvénients lors des mises à jour, maintenance...

-Kerkad et al. (2013) ont proposé une optimisation dans RDW (datawarehouse relationnelle) par HDP (Horizontal Data Partitioning) en considérant l'interaction de la requête. Ils réalisent un codage incrémental pour représenter les schémas de fragmentation et utilisent des prédicats d'élagage et de direction HDP par des requêtes élus. La principale caractéristique des requêtes définies sur un ED relationnel est le fait que leurs jointures passent systématiquement par la table des faits.

-Dehdouh et al. (2013), démontrent que la construction d'un cube OLAP (On-Line Analys Process) est plus performante lorsque l'ED est stocké en colonnes que lorsqu'il est stocké en lignes. Cependant, en absence d'opérateurs d'analyse en ligne, le seul moyen, très coûteux, qui existe pour construire des cubes OLAP consiste à utiliser l'opérateur UNION sur des requêtes de regroupement afin d'obtenir l'ensemble des Group By nécessaires au calcul de cube OLAP.

-Bentayeb et Rakotoarivelo (2007) proposent un nouvel opérateur qui permet d'ajouter de nouveaux niveaux d'analyse intéressants dans la hiérarchie d'une dimension dans le but d'augmenter le périmètre d'analyse avec de nouveaux axes permettant de détecter rapidement des similarités et des tendances dans les données entreposées. Cette intégration montre l'intérêt de combiner la fouille de données et l'analyse multidimensionnelle pour améliorer l'évolutivité des ED. Ainsi, ils envisagent d'exploiter des méthodes d'apprentissage supervisé pour

construire des règles d'analyse sur l'entrepôt ou pour générer des faits prévisionnels à intégrer dans l'entrepôt.

4.3 Niveau cloud computing et big data

-Attasena et al. (2013) ont proposé une analyse des performances qui montre qu'elle peut prévenir les intrusions, garantir la disponibilité des données et leur intégrité, pour un coût réduit (stockage, transfert de données et temps de calcul) dans le paradigme économique de paiement à la demande des nuages informatiques. L'informatique dans le nuage peut contribuer à réduire les coûts et à augmenter la flexibilité en permettant aux entreprises de déployer leurs applications et leurs ED. L'avantage de cette solution est la résolution à la fois des problèmes de sécurité et d'analyse des données. Cependant, le stockage et la gestion des données dans le nuage posent des problèmes de sécurité.

-Ouazzani et al. (2014) ont proposé dans leur perspective un ED liées à l'environnement cloud. Ils ont proposé une solution sur le contrôle d'accès aux ED qui a comme avantage de fonctionner d'une manière indépendante de la plate-forme cible. Mais cette solution présente comme inconvénient l'exigence d'un ED déjà Mis en place avant.
On peut essayer d'appliquer ces deux techniques pour une nouvelle répartition et évaluation. Le cloud computing, abrégé en cloud (le nuage en français) ou l'informatique en nuage désigne un ensemble de processus qui consiste à utiliser la puissance de calcul et/ou de stockage de serveurs informatiques distants à travers un réseau, généralement Internet. Les avantages du cloud computing :
1. Moins cher ;
2. Solution mobile ;
3. Solution flexible ;
4. Des mises à jours automatiques.
Les Big data, ou les grosses données , ou méga données, parfois appelées données massives, désignent des ensembles de données qui deviennent tellement volumineux qu'ils en deviennent difficiles à travailler avec des outils classiques de gestion de base de données ou de gestion de l'information. La différence entre l'informatique décisionnelle et les big data est définie sur les données et leur utilisation :
-Informatique décisionnelle : utilisation de statistique descriptive, sur des données à forte densité en information afin de mesurer des phénomènes, détecter des tendances...
-Big Data : utilisation de statistique inférentielle, sur des données à faible densité en information dont le grand volume permet d'inférer des lois (régressions...) donnant dès lors au big data des capacités prédictives.

5 Etude comparative

Le tableau suivant résume les différents travaux :

Travaux	Approches	Avantages	Limites
Nehme et Valduriez (2011)	Fragmentation	Le meilleur schéma de fragmentation	Le délais d'attente pour l'optimiseur
Bouchakri et Bellatreche (2012)	Fragmentation	Prendre en considération les changements au niveau des requêtes	Augmentation du temps de mise à jour
Barr (2013)	Fragmentation	Réduction du temps d'exécution et des fragments	Manque de préférences d'administrateur
Elmansouri et al. (2013)	Fragmentation	Mettre en évidence les fragments horizontaux supplémentaires	N'est pas encore testée sur un ED
Ettaoufik et al. (2013)	Fragmentation	Amélioration de la fragmentation	Manque de préférences d'administrateur générés
Boufares et al. (2012)	Fragmentation	Meilleure qualité pour les données résultantes	Fonction de calcul et automatisation
Kerkad et al. (2012)	Evaluation	Etudier conjointement le problème de BMP et le problème de QSP	Temps d'attente
Perriot et al. (2013)	Evaluation	Minimiser le coût d'interrogation et de maintenance	Plus complexe dans les nuages informatiques
Favre (2007)	Evaluation	Evaluation de la performance du modèle proposé	Mise à jour, maintenance
Kerkad et al. (2013)	Evaluation	Optimisation au niveau des fragments	Temps d'attente
Dehdouh et al. (2013)	Evaluation	Construction plus performante du cube	Très coûteux
Bentayeb et Rakotoarivelo (2007)	Evaluation	Augmenter le périmètre d'analyse	Manque de prévision
Attasena et al. (2013)	Cloud computing et big data	Résoudre les problèmes d'analyse et de sécurité des données	Problème de sécurité lors du stockage
Ouazzani et al. (2014)	Sécurité	Cible standard	Exigence d'un ED

TAB. 1 – *Synthèse état de l'art*

6 Objectifs et contributions

Nous commençons par mesurer les temps d'exécution pour un ED centralisé puis nous allons partitionner l'entrepôt et recalculer ces temps. Ensuite, nous essayons d'appliquer notre technique de partitionnement sur un ED réparti. Notre travail a pour but d'essayer de produire :
1. Une solution pour le partitionnement d'un ED réparti ;
2. Une solution pour l'évaluation du partitionnement, la comparer avec les techniques précédentes par une évaluation numérique et expérimentale ;
Nous cherchons une méthode et/ou un modèle de coût, les données impliquées et proposer une démarche de calcul. Enfin, nous allons essayer de valider notre solution par le benchmark adéquat. A notre connaissance, il existe très peu de bancs d'essais décisionnels. Cependant, parmi ceux-ci, le plus connu est l'Analytical Processing Benchmark 1 (APB-1)Spoffored (1998). APB-1 est assez simple et s'est révélé limité pour évaluer les spécificités d'activités et de fonctions différentes. Le Transaction Performance Processing Council (TPC) PilHo (2014), qui spécifie des bancs d'essai standards dans le monde relationnel, dispose d'un banc d'essai décisionnel. Les chercheurs utilisent généralement ces benchmark pour valider leurs travaux selon leurs besoins. Tekaya (2012) dans son travail a défini deux nouvelles notions : la corrélation sémantique et la corrélation géographique. La corrélation sémantique permet de fusionner par conjonction deux prédicats inclus dans la même clause WHERE. Une corrélation géographique permet de fusionner deux prédicats n'appartenant forcément pas à la même requête mais sont utilisés par la même localisation géographique. Une localisation géographique désigne le (ou les) site(s) sur lesquels nous envisageons allouer un magasin de données (MD). La solution proposée intègre l'aspect géographique dans le processus de fragmentation. elle englobe quatre phases : (1) détermination d'une liste de prédicats simples, (2) création de la matrice corrélation, (3) application de l'algorithme k-means pour la classification des prédicats et (4) génération des fragments horizontaux. Nous avons ensuite encore réduit le nombre de prédicats dans Ghorbel et al. (2014) par l'algorithme de réduction du nombre de prédicats pour éliminer les prédicats qui se trouvent dans tous les sites de répartition. Nous avons eu les résultats suivantes : Pour la validation de notre solution, nous avons commencé par mesurer les temps d'exécution des requêtes en minute dans le cas d'un ED centralisé, ensuite réparti en appliquant notre solution de réduction des prédicats. Ce travail a été publié dans la revue ISI Ghorbel et al. (2016b). La figure 1 ci-dessous résume nos résultats. Nous pouvons adapter cette solution pour le benchmark TPC-H. Une autre approche que nous allons mettre en valeur c'est d'essayer d'étendre notre travail Ghorbel et al. (2014) par une évaluation au niveau du Benchmark TPC-H avec une évaluation des performances. Nous avons préparé l'environnement expérimental à savoir le benchmark TPC-H pour la réalisation des tests nécessaires pour pouvoir évaluer nos approches. De plus, nous avons commencé a réaliser une charge de requêtes du benchmark TPC-H. Ensuite,nous avons exécuté cette charge au niveau centralisé et enfin nous allons répartir ce benchmark. La figure 2 ci-dessous représente les tables du benchmark. Le tableau 2 résume les caractéristiques de chaque table. Le TPC-H contient huit tables de données distinctes et individuelles. Les relations entre les colonnes de ces tables sont illustrés dans la figure 2. Il contient 8661245 enregistrements alloués sur 1 Gigaoctet de mémoire

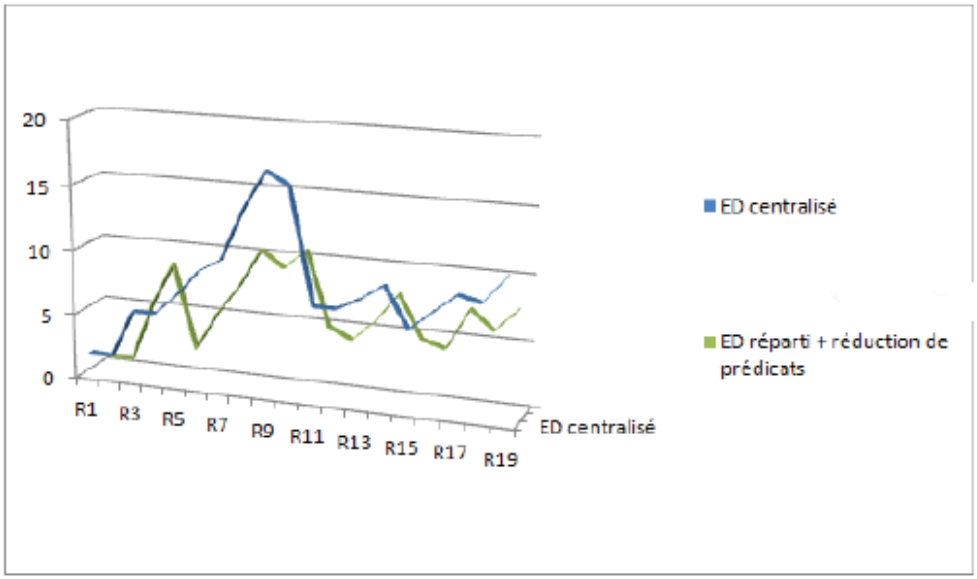


FIG. 1 – *Synthèse répartition APB-1*

PilHo (2014). Nous avons calculer les temps de réponse de notre charge de requêtes au niveau centralisé ensuite nous l'avons réparti suivant la même méthode pour l'APB-1 Spoffored (1998).
Le tableau 3 illustre nos résultats. Ce travail a été présenté à Setit Ghorbel et al. (2016a).Pour les requêtes numéro 1, 2, 3, 8 et 10, les temps d'exécution ont diminué, ce qui constitue pour nous un gain non négligeable. Le temps d'exécution global des requêtes dans un contexte réparti a diminué par rapport au contexte centralisé. De plus, le gain important du TPC-H par rapport l'APB-1 c'est qu'au niveau du TPC-H, les tests se font sur un benchmark réel d'une société, par contre l'APB-1 génère des codes non identifiables et non démonstratifs. N'empêche qu'il est le plus utilisé pour les années 90 et début 2000 avant l'apparition du TPC-H. Par contre, pour les requêtes 4, 5, 6, 7 et 9, les temps d'exécution ont augmenté. Ce sont les requêtes les moins bénéficiaires du partitionnement qui n'ont pas été considérées dans le processus de partitionnement. Ceci explique clairement l'augmentation remarquable des mesures obtenues. De plus, les requêtes contenant plusieurs jointure qui seront ensuite réparties, sont les plus moins bénéficiaire de la répartition. En effet, il faut gagner en terme temps et stockage pour répartir le benchmark. Dans ce contexte, quelques solutions sont envisageables pour la réécriture des requêtes OLAP distantes pour l'optimisation du temps d'exécution des requêtes OLAP distantes notamment les travaux de Liang et al. (2000) et/ou de Kalnis et Papadias (2001) où ils rajoutent une heuristique d'optimisation des requêtes. Vu les avancés dans les réseaux aussi, le temps de réponse est considérablement réduit, ceci n'a pas été mesuré dans notre solution.

Table	Nombre d'enregistrements
Region	5
Nation	25
Supplier	10000
Orders	1500000
Linitem	6000000
Customer	150000
Partsupp	800000
Part	200000

TAB. 2 – *Caractéristiques des tables de l'entrepôt de données*

Numéro de requête	Temps d'exécution centralisé	Temps d'exécution réparti
1	0 :28 :53	0 :10 :12
2	9 :03 :14	0 :57 :94
3	13 :41 :60	2 :23 :44
4	0 :0 :06	4 :31 :24
5	0 :0 :1	3 :42 :85
6	0 :0 :18	7 :20 :11
7	0 :0 :7	9 :49 :62
8	0 :4 :22	0 :0 :62
9	0 :0 :98	9 :49 :62
10	32 :63 :41	10 :05 :72

TAB. 3 – *Temps d'exécution des requêtes utilisées en (mn :s :ms)*

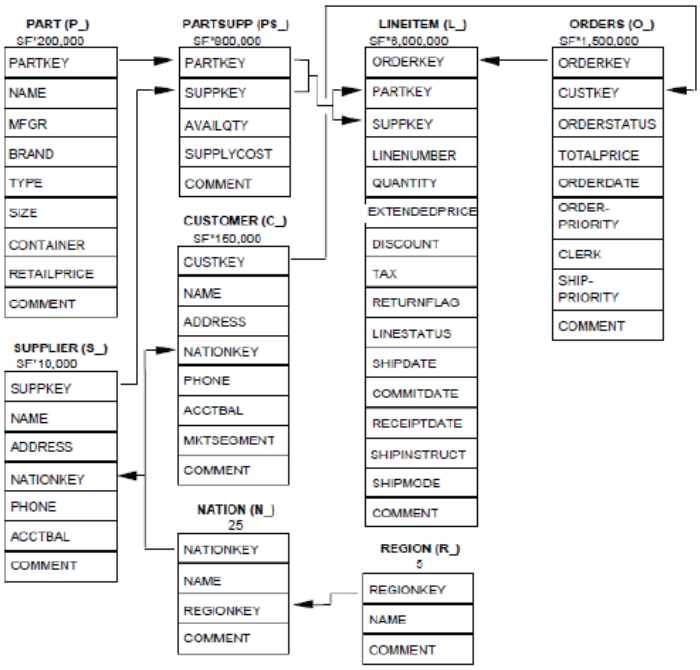


FIG. 2 – *TPC-H*

7 Solutions envisageables

* Nous avons remarqué d'après l'état de l'art qu'on pourra essayer de comparer les différentes approches de partitionnement puis essayer de choisir la meilleur au niveau coût de stockage et temps d'exécution des requêtes.
Nous remarquons aussi qu'au niveau du cloud computing et la sécurité des données, les travaux de recherche sont les nouveaux problèmes à résoudre pour les ED. Donc une solution à faire est une répartition des données au niveau d'un cloud avec une solution d'évaluation de performance. Une deuxième contrainte qu'il faut prendre en considération c'est au niveau de la sécurité des données. En effet, nous avons installé une machine virtuelle et nous avons réalisé nos tests dans un contexte centralisé. Reste à vérifier son gain au niveau réparti.

8 Conclusion

Les utilisateurs des ED sont en augmentation, ils appartiennent généralement à plusieurs domaines, ils effectuent des tâches différentes et ont des requêtes distinctes. De ce fait, le volume des données augmente rapidement posant plusieurs problème de stockage. Pour satisfaire tous leurs besoins, un éclatement de l'ED en plusieurs sous modèles en étoile spécifiques aux domaines est nécessaire. Dans notre travail, nous nous sommes intéressés à l'optimisation des requêtes décisionnelles exécutées sur un ED modélisé en étoile. Pour cela, nous avons proposé

une nouvelle approche pour minimiser le temps d'exécutions des requêtes dans un contexte réparti. Nous avons entamé une recherche approfondie sur l'état de l'art. Ensuite, nous avons adapté notre solution élaboré avec le benchmark APB-1 vers le benchmark TPC-H.

Références

Attasena, V., N. Harbi, et J. Darmont (2013). Sharing-based privacy and availability of cloud data warehouses. *EDA B-9*, 17–32.

Barr, M. (2013). Bi-objective optimization based on compromise method for horizontal fragmentation in relational data warehouses. *International Journal of Machine Learning and Computing 3*, 250–254.

Bentayeb, F. et O. Rakotoarivelo (2007). Evolution de schéma par classification automatique pour les entrepôts de données. *EDA B-3*, 99–112.

Bouchakri, R. et L. Bellatreche (2012). Sélection incrémentale d'un schéma de fragmentation horizontale d'un entrepôt de données relationnel. *In Proceedings of EDA.*, 2–16.

Boufares, F., A. BenSalem, et S. Correia (2012). Un algorithme de déduplication pour les bases et entrepôts de données. *INFORSID*, 497–506.

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

Chakravarthy, S., J. Muthuraj, R. Varadarajan, et S. Navathe (1994). An objective function for vertically partitioning relations in distributed databases and its analysis. *Distributed and parallel databases*.

Dehdouh, K., F. Bentayeb, et N. Kabachi (2013). Performances de requêtes olap dans les bases de données en colonnes. *ASD*, 439–444.

Elmansouri, R., E. Ziati, et O. Elbeqqali (2013). L'analyse en composantes principales normée : Une nouvelle approche pour la fragmentation des entrepôts de données. *ASD*.

Ettaoufik, A., L. Bellatreche, M. Ouzzif, E. Ziyati, et H. Belhadaoui (2013). Service web pour la fragmentation horizontale des entrepôts de données. *ASD*.

Favre, C. (2007). *Évolution de schémas dans les entrepôts de données : mise à jour de hiérarchies de dimension pour la personnalisation des analyses*. Thèse de doctorat, Université de Lyon (Lumière Lyon 2).

Ghorbel, M., K. Tekaya, et A. Abdellatif (2014). Réduction du nombre des prédicats pour les approches de répartition des entrepôts de données. *ASD*.

Ghorbel, M., K. Tekaya, et A. Abdellatif (2016a). Fragmentation and evaluation of the operation of data warehouses : state of the art. *SETIT*.

Ghorbel, M., K. Tekaya, et A. Abdellatif (2016b). Réduction du nombre des prédicats pour les approches de répartition des entrepôts de données. *ISI*.

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

Kerkad, A., L. Bellatreche, et D. Geniet (2012). Exploitation de l'interaction des requêtes olap pour la gestion de cache et l'ordonnancement de traitements. *EDA*, 154–163.

Kerkad, A., L. Bellatreche, et D. Geniet (2013). La fragmentation horizontale revisitée : Prise en compte de l'interaction de requêtes. *EDA B-9*, 117–132.

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

Nehme, R. et P. Valduriez (2011). Automated partitioning design in parallel database systems. *SIGMOD : International Conference on Management of data*, 1137–1148.

Ouazzani, A. E., N. Harbi, et H. Badir (2014). Contrôle d'accès aux entrepôts de données fondé sur le profil utilisateur. *ASD*, 95–100.

Perriot, R., J. Pfeifer, L. d'Orazio, B. Bachelet, S. Bimonte, et J. Darmont (2013). Modèles de coût pour la sélection de vues matérialisées dans le nuage, application aux services amazon ec2 et s3. *EDA B-9*, 53–68.

PilHo, K. (2014). Transaction processing performance council (tpc). *Guide d'installation*.

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

Tekaya, K. (2012). *Fragmentation et Allocation Dynamiques des Entrepôts de Données*. Thèse de doctorat, Facultés des sciences de Tunis.

Summary

The distribution of a DW is nowadays very useful considering the load of information that is constantly increasing in any company and the increasing needs of the users of this warehouse. In addition to distribution, new techniques have emerged to remedy this crucial increase in data such as Big Data and Cloud computing.

In this article, we propose a comparative study of DW allocation techniques to minimize query execution time in a distributed context. We select the classification as a distribution technique and test it according to two benchmark APB-1 and TPC-H. We will also give an overview of the DW assignments, then we focus our work on Cloud and Big data as solutions.

Un nouvel algorithme de sauvegarde d'un point de reprise global (*Checkpointing*) pour les bases de données distribuées

Houssem Mansouri*, Mohammed A. El-Dosuky**

*Laboratoire des Réseaux et des Systèmes Distribués (LRSD), Département d'informatique,
Faculté des Sciences, Université Ferhat Abbas Setif1, Algerie.
mansouri_houssem@univ-setif.dz
** Département d'informatique, Faculté d'informatique et de l'information,
Université de El-Mansoura, Égypte.
mouh_sal_010@mans.edu.eg

Résumé. Cet article est consacré à la proposition d'un nouvel algorithme de sauvegarde de point de reprise global (*Checkpointing*) pour la reprise après panne dans les bases de données distribuées. La reprise après panne consiste, comme son nom l'indique, à assurer que les systèmes de gestion de base de données sont capables, après une panne, de récupérer l'état de la base au moment où la panne est survenue. Le terme de panne désigne ici tout événement qui affecte le fonctionnement du processeur ou de la mémoire principale. Il peut s'agir par exemple d'une coupure électrique interrompant le serveur de données, ou d'une défaillance logicielle.
L'algorithme proposé enregistre un état global cohérent de la base de données qui reflète seulement les transactions complétées et non les transactions partiellement exécutées. La comparaison de performances et les résultats de simulations montrent que notre algorithme fonctionne mieux par rapport aux travaux connexes.

1 Introduction

Le développement spectaculaire des systèmes de gestion des bases de données (SGBD) durant cette dernière décennie a entrainé dans son sillage celui des bases de données distribuées. Ces dernières, ayant connu un essor remarquable grâce à leur flexibilité d'utilisation, ont été rapidement adopté par les particuliers que par les entreprises. Bien qu'un internet ubiquitaire, tirant meilleur parti des bases de données distribuées, n'est pas envisageable dans l'immédiat faute d'infrastructures onéreuse et limités en couverture ; ces contraintes s'estomperont à mesure que la technologie du réseau, actuellement en pleine effervescence, se développera d'avantage et assainira les entraves majeurs.

Néanmoins, les préoccupations de l'heure résident dans la maitrise des spécificités des réseaux pour qu'ils puissent offrir une qualité de service satisfaisante. Comme les services attendus dans les bases de données distribuées ne sont, en fait, qu'une extension des services rendus par les systèmes distribués classiques, un grand nombre d'algorithmes ou protocoles usuels accomplissant ces services doivent être modifiés ou repensés pour mieux s'adapter aux contraintes des bases de données distribuées.

Parmi les algorithmes à ajuster pour répondre aux besoins, nous citons en particulier les algorithmes de sauvegarde de point de reprise global ou *Checkpointing* permettant aux systèmes distribués de se doter de capacités de tolérance aux fautes (Sang (1988), Son et Agrawala (1989), Wu et Manivannan (2006, 2009)). Pour s'adapter convenablement aux bases de données distribuées, les algorithmes de *Checkpointing* doivent tenir compte des spécificités de cet environnement par rapport aux systèmes distribués en générale.

2 Modèle du système

Comme montre la figure 1, une base de données distribuée (ou répartie) est une base de données logique dont les données sont distribuées sur plusieurs sites de données (SD) gérés par plusieurs SGBDs et visibles comme un tout (Pilarski et Kameda (1990, 1992), Son (1989), Wu et al. (2008), Sarita et al. (2016)). Les données sont échangées entre les SGBDs par envoi de messages (transactions) à travers un réseau de communication entres les sites de données.

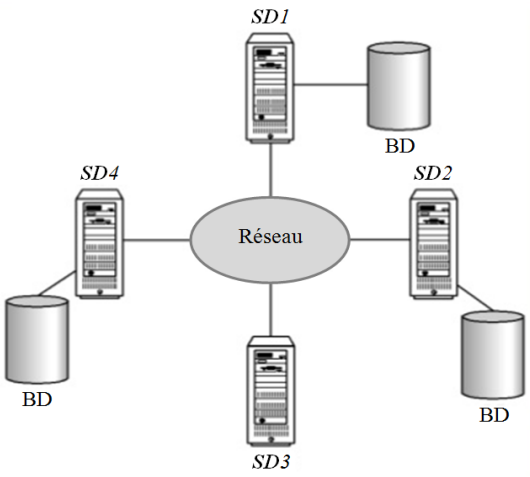


FIG. 1 – Base de données distribuées

3 Elaboration de l'algorithme

Compte tenu des synthèses réalisées dans Lin et Dunham (1997), Roberto et al. (1997), Jiang (2011) et Akanshika (2013), nous en déduisons le résultat suivant :

Un algorithme de *Checkpointing* bien adapté aux bases de données distribuées doit se baser sur les deux axes essentiels suivants :

- Une hybridation entre les techniques de *Checkpointing* classiques pour adapter aux mieux l'algorithme aux nouvelles caractéristiques spécifiques de l'environnement.
- L'introduction de nouveaux mécanismes pour traiter les spécificités des bases de données distribuées.

3.1 Technique de *Checkpointing* utilisées

L'algorithme utilise une hybridation des techniques de *Checkpointing* suivantes :

- Technique de *Checkpointing* coordonnée entre les SGBDs pour sauvegarder un état global cohérent et éviter les messages *Orphelin* (Mansouri (2015a, 2016b)), cette coordination vise la simplicité de mise en œuvre et l'assurance d'obtenir un état global cohérent lors de la sauvegarde des points de reprise. Lorsqu'un point de reprise local est sauvegardé, l'algorithme garantit que tous les points de reprise locaux forment un état global cohérent, les SGBDs coopèrent pour garder un seul point de reprise pour chaque site, d'où minimisation de l'espace de stockage des points de reprise. Le recouvrement est aussi simplifié, puisque chaque site impliqué se réexécutera à son dernier point de reprise permanent sauvegardé.
- La coordination est non bloquante (Mansouri (2015b, 2016a)), absence de blocage momentané des transactions de données entre les sites pendant l'exécution de l'algorithme. Ce qui augmente les performances de manière significative, puisque l'algorithme est exécuté de façon transparente. Cependant, lorsqu'un SGBD*i* envoie un message de données alors il envoie avec le message la valeur courante de son numéro de point de reprise : *Ckpti*. Quand le SGBD*j* destinataire reçoit le message de données alors il complète la transaction si son numéro de point de reprise *Ckptj* est supérieur ou égal au numéro de l'expéditeur *Ckpti*. Sinon, SGBD*j* sauvegarde d'abord un point de reprise de la base de données de son site, avant de compléter la transaction, et mettre à jour le numéro de point de reprise *Ckptj*.
- La coordination est partiellement dynamique ou minimum (Mansouri (2015c, 2017)), l'algorithme n'oblige pas tous les SGBDs à sauvegarder des points de reprise pour chaque exécution, seuls les SGBDs ayant communiqués avec le SGBD initiateur de l'algorithme directement ou indirectement après le dernier point de reprise global doivent sauvegarder un nouvel point de reprise, l'algorithme comprend deux phase : la première consiste à une identification de tous les sites dépendant causalement du site initiateur (depuis son dernier point de reprise) et leur envoi une requête de *Checkpointing*. Sur réception de cette dernière, chaque SGBD identifie à son tour tous les sites de données avec lesquels il a eu à communiquer depuis le dernier point de reprise et leur envoie une requête de *Checkpointing*, et ainsi de suite jusqu'à ce qu'il n'y ait plus de site à identifier. Dans la seconde phase, les SGBDs de tous les sites identifiés pendant la première phase vont sauvegarder des points de reprise permanants, le résultat ainsi obtenu est un état global cohérent avec la participation d'un nombre optimal de sites.

3.2 Déroulement de l'algorithme

L'algorithme se déroule en deux phases : La phase de création des points de reprise provisoires, et la phase de transformation de ces points en points de reprises permanents.

Pour simplifier et éclaircir la représentation, on utilise les abréviations suivantes :

ReqCkpt : requête de *Checkpointing*.
MsgAck : message d'acquittement.
ReqVal : requête de validation.
CkptProv : point de reprise provisoire.
CkptPerm : point de reprise permanent.

3.2.1 Première phase

Lorsqu'un SGBD initie l'algorithme, il sauvegarde un point de reprise de la base de données et détermine l'ensemble des sites dépendants. Il diffuse ensuite une requête de *Checkpointing* ***ReqCkpt*** aux éléments de cet ensemble. Pendant toute l'exécution de l'algorithme, si un SGBD reçoit une requête de *Checkpointing***,** alors il sauvegarde un point de reprise provisoire ***CkptProv*** et détermine également les identificateurs des sites dépendants directement (et qui sont dépendants indirectement du site initiateur), si l'ensemble ainsi déterminée est vide alors le SGBD envoi un message d'acquittement ***MsgAck*** au site initiateur, sinon il diffuse une requête de *Checkpointing* à tous les sites de cette ensemble, et ainsi de suite.

3.2.2 Deuxième phase

Lorsque le SGBD initiateur reçoit tous les messages d'acquittement, alors il diffuse une requête de validation ***ReqVal*** vers tous les sites de données, chaque site ayant reçu cette requête rend le point de reprise permanent ***CkptPerm*** et supprime le point de reprise antérieur.

4 L'algorithme

4.1 Structure de données

Pour chaque site de données le SGBD a la structure de variables suivant :
Ckpt : Entier /* le numéro du dernier point de reprise du site sauvegardé par le SGBD
DepDir : Identificateur /* ensemble des identificateurs des sites dépendants directement
EnsInd : Identificateur /* ensemble des identificateurs des sites dépendants indirectement
Term : Réel /* sert à détecter la terminaison de l'algorithme.

Initialisation :
Ckpt := 1 ;
DepDir := ø ;
DepInd : ø ;
Term : 0 ;

4.2 Le pseudo code

Pendant de l'exécution normale des transactions de la base de données distribuée, si un SGBD*i* reçoit un message de transaction de données d'un autre SGBD*j* alors le SGBD*i* et avant de compléter la transaction et réagit comme suit:

1. ***Receive*** *Transaction*(***SD****i*, ***SD****j*, *Data*)
2. ***DepDir****i* := ***DepDir****i* + ***SD****j*

Pendant l'exécution de l'algorithme :

Algorithme

Rôle du SGBD*i* initiateur:

1. ***Record*** *CkptProv ;*
2. ***Ckpt****i* := ***Ckpt****i* + 1 *;*
3. ***EnsInd*** := ***DepDir****i ;*
4. ***Broadcast*** *CkptReq(****SD****i, **Ckpt**i, **EnsInd**) ;*
5. **A la réception d'un message d'acquittement :**
6. ***Receive*** *MsgAck(****SD****i, **SD**j) ;*
7. ***Term := Term*** + 1 *;*
8. ***If*** (***Term*** = ***n***) ***Then***
9. ***Broadcast*** *ReqVal(****SD****i, **Ckpt**i) ;*
10. ***End if***

Rôle de tout autre SGBD*j*:

1. ***Receive*** *CkptReq(****SD****i, **Ckpt**i, **EnsInd**) ;*
2. ***If*** *(****SD****j* ∈ ***EnsInd***) ***Then***
3. ***If*** *(****Ckpt****j* < ***Ckpt****i*) ***Then***
4. ***Record*** *CkptProv ;*
5. ***Ckpt****j* := ***Ckpt****i ;*
6. ***EnsInd*** := ***EnsInd*** + ***DepDir****j ;*
7. ***Broadcast*** *CkptReq(****SD****i, **Ckpt**n, **EnsInd**) ;*
8. ***Send*** *MsgAck(****SD****i, **SD**j) ;*
9. ***End if***
10. ***End if***
11. **A la réception d'un message de validation:**
12. ***Receive*** *ReqVal(****SD****i, **Ckpt**i) ;*
13. ***If*** *(****Ckpt****j* = ***Ckpt****i*) ***Then***
14. *CkptPerm* := *CkptProv ;*
15. ***Else***
16. ***Ckpt****j* := ***Ckpt****i ;*
17. ***End if***

5 Exemple d'exécution

Nous illustrons l'exécution de l'algorithme à l'aide d'un exemple (figure 2) d'une base de données distribuée composée de cinq site de données ***SD****1* à ***SD****5*. Les SGBDs communiquent uniquement par échange de messages pour exécuter les transactions de donnée.

Si le SGBD3 initie l'algorithme à l'instant *T1*, il sauvegarde un point de reprise pour ***SD***3 et envoi une requête de *Checkpointing* à SGBD2 et SGBD5 pour sauvegarder des points de reprise; Comme ***SD****2* et ***SD****5* sont les seuls sites de données interagissant par des transactions de données avec ***SD****3* entre *T0* et *T1*. Ensuite, et pour la même raison, SGBD2 demandera à SGBD1 de sauvegarder un point de reprise, puisque ***SD****1* est le seul site de données interagissant avec ***SD****2*. On remarque que le SGBD4 dans cet exemple n'est pas concerné par ce processus de *Checkpointing*.

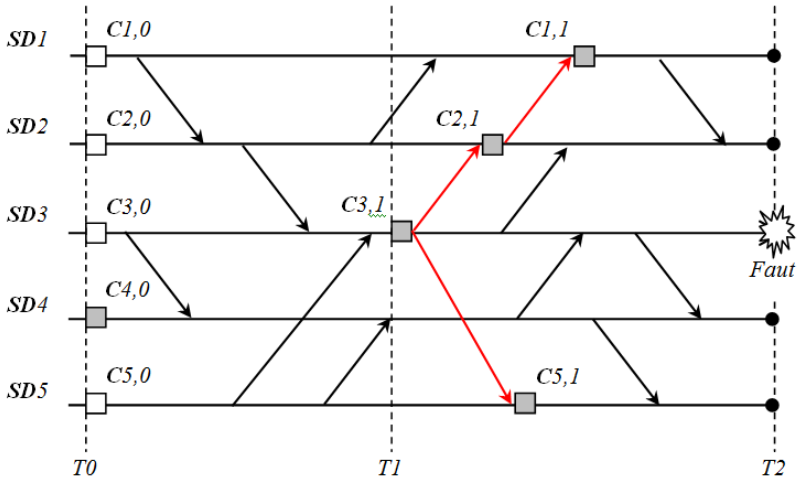


FIG. 2 – Exemple d'exécution

Si une faute ou une défaillance se produite au niveau du site de données ***SD****3* à l'instant *T2* Comme indiqué sur la figure, alors tous le système doit reprendre l'exécution des transactions depuis le dernier état global cohérent sauvegardée {***C****1,1*+***C****2,1*+***C****3,1*+***C****4,0*+***C****5,1*}.

6 Preuve de correction

Lemme 1: si un SGBD*i* initie l'algorithme de *Checkpointing* alors tout autre SGBD*j* qui dépend directement ou indirectement de lui depuis son dernier point de reprise doit sauvegarde un point de reprise.

Preuve : démontrons que si un SGBD*x* sauvegarde un point de reprise alors tout autre SGBD*y* doit sauvegarder un point de reprise.

1. Si le SGBD*x* est l'initiateur de l'algorithme, alors il sauvegarde un point de reprise et envoi une requête de *Checkpointing* à tout les SGBD*y* tel que ***S****y* ∈ ***DepDir****x*, à la réception de cette requête chaque SGBD*y* sauvegarde un point de reprise.
2. Si le SGBD*x* n'est pas l'initiateur de l'algorithme et reçoit une requête de *Checkpointing*, alors :
   a) Si le site ***S****x* appartient à l'ensemble ***EnsInd***, alors il est concerné par la sauvegarde d'un point de reprise.
   b) Et si son numéro de point de reprise est inferieur au numéro de point de reprise de l'initiateur (c.-à-d. le SGDB*x* n'a pas prie précédemment un point de reprise), alors le SGBD*x* sauvegarde un point de reprise.

Donc de 1. et 2. On peut conclure : si un SGBD*x* sauvegarde un point de reprise alors tout autre SGBD*y* doit sauvegarder un point de reprise. Alors par transitivité si un SGBD*i* initie l'algorithme de *Checkpointing* alors tout autre SGBD*j* qui dépend directement ou indirectement de lui depuis son dernier point de reprise doit aussi sauvegarde un point de reprise.

Théorème: L'algorithme assure la sauvegarde d'un état globale cohérent.

Preuve: Afin de prouver ce théorème, nous devons prouver ce qui suit :

Si la réception d'un message de transaction a été sauvegardée dans un point de reprise d'un site de données récepteur, alors l'envoi correspondant de ce message a été sauvegardé dans le point de reprise du site de données expéditeur.

Soit ***SDj*** le site de données récepteur d'un message de transaction *m*, et soit ***SDi*** le site de données expéditeur de *m*.

Si le SGBD*j* sauvegarde la réception de *m* dans le point de reprise de ***SDj*** alors : ***SDi*** ∈ ***DepDirj***; Ainsi et compte tenu du *Lemme* 1, le SGBD*i* doit aussi sauvegarde un point de reprise pour le même processus de *Checkpointing* à cause de la réception d'une requête de *Checkpointing* venant du SGBD*j*, ainsi :

Envoi (*m*) par SGBD*i* → *Réception* (*m*) par SGBD*j*.

Où → est la relation « précédence » (Kuss, 1982 ; Pu, 1985)

Réception (*m*) par SGBD*j* → Sauvegarde de point de reprise du ***SDj*** par SGBD*j*

Sauvegarde de point de reprise du ***SDj*** par SGBD*j* → *Envoi* de *ReqCkpt* par SGBD*j*

Envoi de *ReqCkpt* par SGBD*j* → *Réception* de *ReqCkpt* par SGBD*i*

Réception de *ReqCkpt* par SGBD*i* → Sauvegarde de point de reprise du ***SDi*** par SGBD*i*

Et puisque la relation → est transitive alors :

Envoi (*m*) par SGBD*i* → Sauvegarde de point de reprise du ***SDi*** par SGBD*i*

Donc, l'envoi de *m* a été sauvegardé dans le point de reprise du site expéditeur ***SDi***.

Résultat : Si la réception d'un message de transaction a été sauvegardée dans un point de reprise d'un site de données récepteur, alors l'envoi correspondant de ce message a été sauvegardé dans le point de reprise du site de données expéditeur.

Alors, L'algorithme assure la sauvegarde d'un état globale cohérent.

7 Evaluation des performances

Pour évaluer la performance de l'algorithme nous utilisons les quatre paramètres significatifs suivants :

1. Le nombre de points de reprise crée dans le meilleur cas : d'après le *Lemme 1* l'algorithme force seulement les SGBDs qui sont dépendants directement ou indirectement du SGBD initiateur à sauvegarder des points de reprise.
2. Le temps de blocage dans le plus mauvais cas : selon la conception le temps de blocage de l'algorithme est nul.
3. Le nombre de message de contrôle transférés dans le réseau : pendant la première phase chaque SGBD à besoin de diffuser une requête de *Checkpointing* ***ReqCkpt*** aux SGBDs dépendants de lui, et chaque SGBD qui sauvegarde un point de reprise provisoire à besoin d'envoyer un message d'acquittement ***MsgAck*** au SGBD initiateur. Donc, le nombre de messages de contrôle égale à : $\boldsymbol{n} * \boldsymbol{N}min + \boldsymbol{n} = \boldsymbol{n} * (\boldsymbol{N}min + 1)$.
   Où ***n*** c'est nombre de sites de données dans le système, et ***N****min* c'est le nombre moyen de transactions entre deux sites de données.
   Pendant la deuxième phase, chaque SGBD qui sauvegarde un point de reprise provisoire n'a besoin qu'un seul message de contrôle qui est la requête de validation ***ReqVal***. Donc, le nombre de messages de contrôle égale à : ***n***
   Alors, le nombre totale de messages de contrôle de l'algorithme égale à :
   $\boldsymbol{n} * (\boldsymbol{N}min + 1) + \boldsymbol{n} = \boldsymbol{n} * (\boldsymbol{N}min + 2)$
4. Le degré de distribution de l'algorithme : selon la conception l'algorithme est totalement distribué.

8 Comparaison avec d'autres algorithmes

Dans la littérature, cinq algorithmes de *Checkpointing* pour les bases de données distribuées de type coordonnées ont particulièrement retenu notre attention. Il s'agit de comparer leurs performances avec le notre.

Algorithme	Nombre de point de reprise	Temps de blocage	Nombre de messages de contrôle	Distribution
Cao et Singhal (1998)	***N**min*	2 * ***T**trans*	***n*** * 3 * ***N**min*	Non
Cao et Singhal (2001)	***N**min*	0	***n*** * (4 * ***N**min* + 1)	Oui
Kumar et al. (2003)	***N**min* + ***s***	0	***n*** * (4 * ***N**min* + 1) – ***k***	Non
Weigang et al. (2003)	***N**min* + ***f***	0	***n*** * (4 * ***N**min* + 1) + ***t***	Oui
Kumar et al. (2005)	***N**min* + ***y***	0	***n*** * (4 * ***N**min* + 1) + ***z***	Non
Notre Algorithme	***N**min*	0	***n*** * (***N**min* + 2)	Oui

TAB. 1 – Comparaison

Les valeurs des critères d'évaluation des performances des algorithmes 1 et 2 de la table 1, sont données de façon exacte, alors que ces valeurs pour algorithmes 3, 4 et 5, ont été estimées (par analogie) par les auteurs qui comparent leurs propositions avec celle de 2, on ainsi conclu que le nombre de messages de contrôle pour la proposition 3 est égal au nombre de messages de contrôle pour la proposition 2 moins une valeur réelle positif ***k***. mais le nombre de points de reprise de la proposition 3 a augmenté d'une valeur ***s*** par rapport à la proposition 2.

Idem pour les propositions 4 et 5 mais d'une façon inverse. Les auteures de ces deux algorithmes affirment que leur algorithme minimise le nombre de points de reprise par rapport à la proposition 2, ce que nous représentons dans le tableau par l'ajout des deux variables ***f*** et ***y*** respectivement, mais en même temps ces deux algorithmes augmentent le nombre de messages de contrôle par ***t*** et ***z*** respectivement.

Le temps de blocage est nul pour tous les algorithmes sauf pour la proposition 1, où il est égal à 2 * ***T**trans (*où ***T**trans est* le temps de transfère d'un message entre deux site de données). Le critère de distribution renforce la tolérance aux fautes.

Résultats:

- Notre algorithme réduit le nombre de point de reprises jusqu'à ***N**min*, sachant que : ***N**min* < ***N**min* – *f*
- Le temps de blocage est nul et l'algorithme s'exécute de façon transparente aux opérations de transactions dans la base de données distribuée, ce qui accroît de manière significative les performances.
- L'algorithme réduit le nombre de messages de contrôle jusqu'à ***n*** * (***N**min* + 2), où : ***n*** * (***N**min* + 2) < ***n*** * 3 * ***N**min* < ***n*** * (4 * ***N**min* + 1) – ***k***

9 Paramètres de simulation et résultats

La figure 3 présente le résultat de la simulation du nombre de messages de contrôle pour compléter le processus de *Checkpointing* dans le meilleur cas pour notre algorithme et l'algorithme de Baldoni et al. (1999), par rapport au le nombre de sites de données dans le système, avec deux taux de transactions de données: faible 10 transactions / 20 seconds, et élevé 10 transactions / 5 seconds. L'environnement de simulation comprend un nombre variant de site de données (entre 5 et 55 sites). Nous supposons que chaque site de données a une connexion filaire avec les autres sites d'une bande passante de 2 Mo/s. La taille de chaque message de transaction est supposée être 4 Mo. Chaque message de contrôle est supposé d'une taille de 100 Octets.

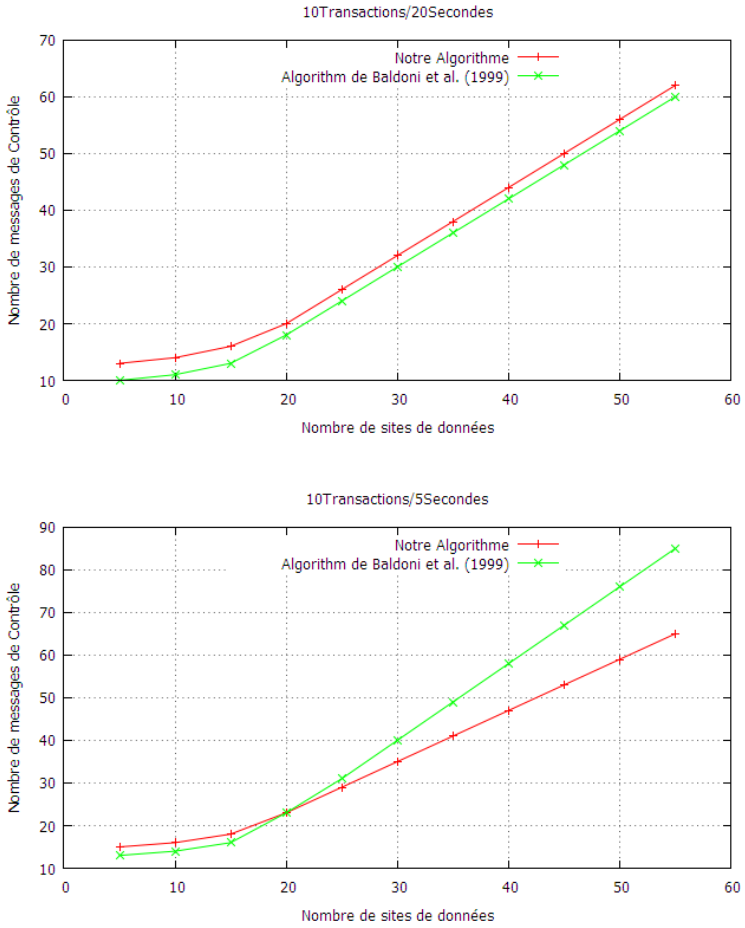


FIG. 3 – Résultats des Simulations

Nous remarquons que l'algorithme de Baldoni et al. (1999) souffre de l'augmentation du nombre de messages de contrôle chaque fois que le taux des transactions dans le système est élevé. En effet, si le taux des transactions est élevé, le nombre de messages de contrôle augmente, ce qui n'a pas été pris en compte par les auteurs de ce document lors de la conception de leur algorithme. Cependant, dans un environnement distribué, les sites de données sont reliés par un réseau (sans fils des fois avec une bande passante faible) - un algorithme de *Checkpointing* appropriée à ce contexte doit tenir compte de cette contrainte. Par conséquent, l'insuffisance de la bande passante peut établir un critère suffisant pour pouvoir invalider l'adoption d'un algorithme de *Checkpointing* pour les bases de données distribuées. Donc, le schéma de Baldoni et al. (1999) ne fonctionne efficacement que si le taux des transactions de données est faible, ou si la bande passante du réseau est élevée.

10 Conclusion

Nous venant de proposer un algorithme de *Checkpointing* qui apport une solution au problème de panne (faute de transaction par exemple) dans les bases de données distribuées, cet algorithme est de type coordonné, minimise le nombre de points de reprise sauvegardés, et évite le blocage des opérations de transaction. Il permet également de minimiser le nombre de messages de contrôle transmis, ce qui diminue avantageusement le trafic sur le réseau au profit des opérations de transactions de données. L'algorithme utilise également les informations de dépendances entre les sites de données pour limiter, au stricte nécessaire, le nombre de SGBD devant sauvegarder des point de reprise. Ainsi, le coût encouru dû aux activités de recouvrement de la base de données après panne est grandement réduit.

Références

Akanshika (2013). Analysis of rollback recovery techniques in distributed database management system. *International Journal of Modern Engineering Research*, 3:1353-1356.

Baldoni, R. F. Quaglia et P. Fornara (1999). An index-based checkpointing algorithm for autonomous distributed systems. *IEEE Transactions on Parallel and Distributed Systems*, 10:181-192.

Cao, G. et M. Singhal (1998). On coordinated checkpointing in distributed systems. *IEEE Transactions on Parallel and Distributed Systems*, 9:1213-1225.

Cao, G. et M. Singhal (2001). Mutable checkpoints: A new checkpointing approach for mobile computing systems. *IEEE Transactions on Parallel and Distributed Systems*, 12:157-172.

Jiang, W. (2011). *Checkpointing and recovery in distributed and database systems*. Thèse de doctorat, Université de Kentucky.

Kumar, L., M. Mishra et R.C. Joshi (2003). Low overhead optimal checkpointing for distributed systems. *Proceedings of the 19th International Conference on Data Engineering*, Bangalore, India.

Kumar, P., L. Kumar, R.K. Chauhan, V.K. Gupta (2005). A non-intrusive minimum process synchronous checkpointing protocol for distributed systems. *Proceedings of the 2005 IEEE International Conference on Personal Wireless Communications*, New Delhi, India.

Kuss, H. (1982). On totally ordering checkpoints in distributed databases. *Proceedings of the ACM International Conference on Management of Data*, Florida, USA.

Lin, J. L. et M. H. Dunham (1997). A survey of distributed database checkpointing. *Journal of Distributed and Parallel Databases*, 5:289-319.

Mansouri, H., N. Badache, M. Aliouat, et A-S.K. Pathan, (2015a). A new efficient checkpointing protocol for distributed mobile computing. *Journal of Control Engineering and Applied Informatics*, 17:43-54.

Mansouri, H., N. Badache, M. Aliouat, et A-S.K. Pathan, (2015b). A non-blocking coordinated checkpointing algorithm for message-passing systems. *Proceedings of the International Conference on Intelligent Information Processing, Security and Advanced Communication*, Batna, Algeria.

Mansouri, H., N. Badache, M. Aliouat, et A-S.K. Pathan, (2015c). Adaptive fault tolerant checkpointing protocol for cluster based mobile ad hoc networks. *Procedia Computer Science*, 73:40-47.

Mansouri, H., N. Badache, M. Aliouat, et A-S.K. Pathan, (2016a). Checkpointing distributed application running on mobile ad hoc Networks. *International Journal of High Performance Computing and Networking, Inderscience Publishers*, Special Issue on: *Wireless Network Technologies and Applications*, to *appear.*

Mansouri, H., N. Badache, M. Aliouat, et A-S.K. Pathan. (2016b). An efficient minimum-process non-intrusive snapshot protocol for vehicular ad hoc networks. *Proceedings of the 13th ACS/IEEE International Conference on Computer Systems and Applications*, Agadir, Morocco.

Mansouri, H., A-S.K. Pathan. et M. Aliouat (2017). A snapshot security protocol for radar network protection. *Proceedings of the 7th Seminar on Detection Systems: Architectures and Technologies*, Algiers, Algeria.

Pilarski, S. et Kameda, T.A. (1990). A novel checkpointing scheme for distributed database systems. *Proceedings of the 9th ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, Tennessee, USA

Pilarski, S. et T. Kameda, (1992). Checkpointing for distributed databases: starting from the basics, *IEEE Transactions on Parallel and Distributed Systems*, 3:602-610.

Pu, C. (1985). On-the-fly, incremental, consistent reading of entire databases. *Proceedings of the 11th Conference on Very Large Database*, Stockholm, Sweden.

Roberto B., Q. Francesco, et M. Raynal (1997). *Consistent data checkpoints in distributed database systems: a formal approach*. Rapport de recherche, Université de Rome.

Sang, H.S. (1988). Efficient decentralized checkpointing in distributed database systems. *Proceedings of the 21st Annual Hawaii International Conference on Computer Systems, Vol.II. Software Track,* Hawaii, USA.

Sarita, S., A. Priyanka, G. Rakesh, M. Shivlal et A. Pradeep (2016). Analysis of recovery techniques in data base management system. *Research Journal of Computer and Information Technology Sciences*, 4:4-8.

Son, S.H. et A.K. Agrawala, (1989). Distributed checkpointing for globally consistent states of databases, *IEEE Transactions on Software Engineering*, 15:1157-1167.

Son, S.H. (1989). An algorithm for non-interfering checkpoints and its practicality in distributed database systems. *Information Systems*, 14:421-429.

Weigang, N., S.V. Vrbsky et S. Ray (2003). Low-cost coordinated nonblocking checkpointing in mobile computing systems. *Proceedings of the 8th IEEE International Symposium on Computers and Communication,* Kiris-Kemer, Turkey.

Wu, J., et D. Manivannan, (2006). An efficient non-intrusive checkpointing algorithm for distributed database systems. *Springer Lecture Notes in Computer Science Series*, 4308, 82-87.

Wu, J., et D. Manivannan et B. Thuraisingham, (2008). Transaction-consistent global checkpoints in a distributed database system. *Proceedings of the 2008 International Conference on Data Mining and Knowledge Engineering*,

Wu, J., D. Manivannan, et B. Thuraisingham, (2009). Necessary and sufficient conditions for transaction-consistent global checkpoints in a distributed database system, *Information Sciences*, 179:3659-3672.

A Genetic Algorithm with Heterogeneous Population for Data Clustering

Amina Bedboudi*, Cherif Bouras** , Mohamed T. Kimour***
University of Badji Mokhtar-Annaba, Po. Box. 12, Annaba, Algeria
*bedboudi.amina@hotmail.fr
**bourascdz@yahoo.fr
***kimour@yahoo.com

Abstract. Clustering has been recognized as a primary data mining method for knowledge discovery. It has been widely used in several domains, such as biology, system engineering and social sciences, in order to identify natural groups in large amounts of data. The most popular methods for data clustering such as K-means suffer from the drawbacks of requiring the number of clusters and their initial centroids, which should be provided by the user. In this paper we present an approach to automatically generate such parameters and achieve optimal clusters using a modified genetic algorithm. The latter operates on heterogeneous populations. Furthermore, we have introduced new crossover and mutation operators. Experimental results show that our modified genetic algorithm is better efficient alternative to the existing approaches.

1 Introduction

Clustering is one of the most important tasks of spatial data mining. It is a typical unsupervised learning technique for grouping similar data points. A clustering algorithm assigns a large number of data points to a smaller number of groups (or clusters) such that data points in the same group share the same properties (similar) while, in different groups, they are dissimilar (Scott, 1992). There are many applications of clustering such as image segmentation, market segmentation, part family formation for group technology, web pages grouping, statistical prediction, etc. (Hartano, 2015) (Vo-Van , 2017).

Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis (Junji, 2012) (Sawant, 2015). Data mining adds to clustering the complications of very large datasets with very many attributes of different types. This imposes unique computational requirements on relevant clustering algorithms. A variety of algorithms have recently emerged that meet these requirements and were successfully applied to real-life data mining problems. In the literature, many clustering methods have been proposed where K-means and genetic algorithms are the most notable ones (Hartano, 2015), (Han, 2001), (Scott, 1992), (Abdul Nazeer, 2009), (Ettaouil; 2013), (Romany,2012) (Delavar, 2014). However, most of them suffer from the drawbacks of requiring the number of clusters and their initial centers to be provided by the user, but also from the poor clustering quality (Sivanandam, 2008).

In this paper, we present a genetic algorithm to not only generate the above-mentioned parameters, but also to improve the speed and accuracy of the clustering process. It is based on an appropriate data structure and process that handles heterogeneous populations in a modified genetic algorithm. Generally, using parameters such as the crossover and mutation probabilities that adapt to the evolution of the algorithm is a good choice, since higher diversity in the population can be achieved, preventing the algorithm to fall in local minima.

Moreover, to accelerate the genetic algorithm process and increase the individual diversity of the initial population, we generate that initial population in two manners: the first subpopulation is obtained with a deterministic way and the second one with a random way. Increasing the population diversity will allow to achieve better quality.

The rest of the article is structured as follows: section 2 gives an analysis of the two main data clustering techniques; k-means and genetic algorithms. Section 3 details our genetic-based data clustering approach by describing the proposed structure of chromosomes and genetic operators. Section 4 gives the results of applying our approach and discusses their advantages. Finally, we draw a conclusion and define some future works in section 5.

2 Clustering

Clustering can be thought of as data partitioning or segmenting into homogeneous groups. The clustering is usually accomplished by determining the similarity among the data on predefined attributes. The most similar data are grouped into clusters (Al Malki, 2016), (Vo-Van , 2017) (García, 2014).

There is a close relationship between clustering techniques and many other disciplines. Clustering has always been used in statistics and science (Rui, 2009) (Allaire, 2000), (Zaki, 2014). Typical applications include speech and character recognition. Viewed as a density estimation problem, machine learning clustering algorithms were applied to image segmentation and computer vision (Fukunaga, 1990) In the literature, various methods have been used to handle the clustering problem. The K-means [1] is considered one of the major algorithms widely used in clustering (Chittu, 2011), (Sawant, 2015), (Al Malki, 2016), (Junji, 2012), (Ettaouil, 2013). Genetic algorithms (Kumar, 2009), (Romany, 2012), (Valarmathi, 2009), (Najmah, 2016) are also used in clustering, either in a separated way or combined with-the k-means algorithm.

K-means first randomly select the k objects; each object initially represents a cluster center (Al Malki, 2016). For each remaining object according to its distance from the center of each cluster, assign it to the nearest cluster. Afterward, each cluster center is replaced by the average value on the respective cluster. This process is repeated, until k centers do not change. K-means algorithm attempts to provide a partition of the given data such that the centroids of provided classes are the minima of the following objective function: The algorithm is as follows :

Input: a set of objects, K: number of clusters,
Output: objects partitioned into k clusters
1. Select k objects as initial centers;
2. Assign each data object to the closest center;
3. Recalculate the centers of each cluster;
4. Repeat steps 2 and 3 until centers do not change;
5 End

Due to its practice properties, K-means algorithm is considered as the most popular technique to cluster information. Its main advantages are: i) easy to implement, ii) the comparison is conducted only between the observations and the center of classes, and iii) it detects and isolates the outliers.

However, the algorithm exhibits some drawbacks, especially on the fact that it can be really slow since in each step the distance between each point to each cluster has to be calculated, which can be really expensive in the presence of a large dataset. Moreover, it is sensitive to the initial cluster centers that are provided by the user as well as the k number of clusters.

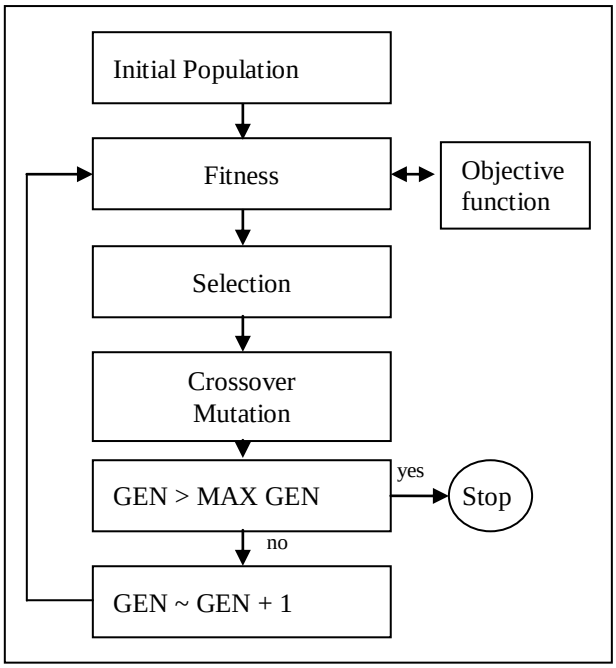


FIG. 1– Outline of the genetic algorithm.

2.1 Genetic algorithms

Genetic algorithms (Kumar, 2009), are robust, stochastic optimization algorithms, used to solve a wide variety of problems. They were developed with the goal of better understanding natural processes such as adaptation, and it belongs to a type of search techniques that mimic the principles of natural selection to develop solutions of large optimization problems.

A genetic algorithm finds the optimal value for a particular objective function depending on the problem to be solved. The standard approach to an optimization problem begins by designing an objective function that can model the problem's objectives while incorporating any constraints (Allaire, 2000). A genetic algorithm consists of: chromosomal representation, initial population, fitness evaluation, selection and reproduction (i.e., crossover and mutation). GA operates by maintaining and manipulating a population of potential solutions

called chromosomes (Valarmathi, 2009). Each chromosome has an associated fitness value which is a qualitative measure of the goodness of the solution encoded in it.

In a genetic algorithm process, we determine firstly an initial solution, composed of a set of chromosomes (initial population), and iteratively apply reproduction operators(selection, crossover and mutation) until achieving a certain quality parameter or a predefined number of iterations. A use of a fitness function guides the stochastic selection of chromosomes which are then used to generate new candidate solutions through crossover and mutation.

Therefore, basic operations are: i) Crossover, which generates new chromosomes by combining sections of two or more selected parents, ii) mutation, which acts by randomly selecting genes which are then altered; thereby preventing suboptimal solutions from persisting and increases diversity in the population.

There are three main types of selection methods: fitness proportionate selection, ranking method and tournament selection. In tournament selection, individuals are selected randomly from the population, based on the fitness function. Tournaments are often held between pairs of individuals, although larger tournaments can be used. Simple outline of a genetic algorithm is shown in Fig. 1.

Genetic algorithms are characterized by attributes such as objective function, encoding of the input data, crossover, mutation, and population size (Romany, 2012),.

1. Objective function. It is used to assign each individual in the population a fitness value; an individual with a higher fitness represents a better solution to the problem than an individual with a lower fitness value;
2. Encoding. Genetic algorithms operate on an en-coding of the problem's input data (which represent in-dependent variables for the objective function);
3. Elitism. This is a way to ensure that the highly fit-ting chromosomes are not lost and copied to the new population. Elitism has been found to be very important to the performance of genetic algorithms;
4. Crossover. It is a procedure in which a highly fit-ting chromosome is given an opportunity to reproduce by exchanging pieces of its genetic information with other highly fitting chromosomes;
5. Mutation. This is often applied after crossover by randomly altering some genes to individual parents;
6. Population size. It is the number of individuals in a population. The larger the population size, the better the chance that an optimal solution will be found.

Genetic algorithms iterate a fixed number of times. Since the function's upper bound (the maximum fitness value possible for an individual) may not be known or cannot be reached, we must limit the number of generations in order to guarantee the termination of the search process. This may result in a suboptimal solution. Moreover, combined approaches of genetic algorithms with other clustering techniques are still not satisfying on some user requirements especially for the accuracy and execution time.

3 The proposed approach

The aim of our approach is to enhance clustering results in minimizing errors and execution time. It is a modified form of the classical genetic algorithm. Indeed, the input of such modified genetic algorithm process is a dataset, which is stored in a vector, and automatical-

ly generates the number of clusters with their initial centroids, while defining appropriate crossover and mutation operators. Moreover, it is based on a structure of chromosomes that is different in number of constituent genes.

3.1 Chromosome representation

Chromosome representation is a problem encoding performed at the most important steps in using genetic algorithm to solve a problem (Najmah, 2016). To encode both the number of clusters and their centers, we propose a chromosome structure that applies a real coded genetic algorithm to the clustering problem, where crossover and mutation operators are applied directly to real parameter values. The use of real parameter values in the GA representation has a number of advantages over binary coding. The efficiency of the GA is increased as there is no need to convert the solution variables to the binary type, less memory is required, there is no loss in precision by discretization to binary or other values, and there is greater freedom to use different genetic operators.

In this work, the existing population is processed in the form of several subpopulations with different size. In other words, the number of genes in every cluster differs. A chromosome in a subpopulations may contain a number K of genes varying from 2 until k_{max}. K_{max} is the maximum number of clusters. These genes hold corresponding information about the cluster centers. Here, each gene in the chromosome represents cluster center, and number K of genes in a chromosome stands for number of clusters. The value of K is assumed to lie in the interval [2; K_{max}]. Thus, each solution i is a fixed-length string represented by cluster centers cij ; j = 2,…,K_{max}. Then, solution i of the population is represented via a vector as follows:

c_{i1}	…	c_{ij}	…	c_{iKmax}

FIG. 2– An example of chromosomes as defined by our approach.

3.2 Fitness function

The fitness function is a measure of profit we need during optimization. It is an objective function that is used to summarize how close a given solution is to achieving the aims and has an important effect on success of a genetic algorithm.

Fitness is proportional to the utility or ability of individual which the chromosome represents. Measure of fitness helps in evolving good solutions and implementing natural selection. In this work, the fitness of a chromosome is computed using Mean Square Error (MSE) (Allaire, 2000). The MSE calculate the error between the cluster

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 \quad \text{.....(1)}$$

$$Fitness = \frac{1}{MSE}$$

After the definitions of the fitness function the different parameters of the genetic algorithm operators are fixed.

3.3 Initial population generation

To accelerate the genetic algorithm process and increase the individual diversity of the initial population, we generate that initial population in two phases. The first phase consists in a deterministic producing manner, and the second one consists in a randomly generation. We proceed such generation so to obtain 20% of the initial population from the first phase and 80% from the second phase.

TAB. I –. List of initial parameters for the genetic algorithms.

Parameter	Value
Maximum number of clusters, K_{max}	15
Population size, P	80
Crossover probability	0.6
Mutation probability	0.3
Maximum no. of iterations,	100

In this phase, we deterministically generate the first sub-population that constitutes 20% the entire population. From the input dataset, we produce a sorted dataset. Then, the sorted dataset is divided into k equal segments and the means of each segment I considered as a center. This will be done according to three manners: 1) Kmax-1 populations are created using the segment modal value, 2) K_{max-1} populations are created using the segment mean value, 3) K_{max-1} populations are created using the segment min-max value. For instance, the first manner is detail as follows:

```
Sort the data set into a sorted vector in increasing order;
For (k=2, k<=Kmax-1, k++){
Sorted vector is divided into k equally segments
We select from each segment the modal value, which will be
considered as a center;
Create the kth population
}
```

3.4 Reproduction

Genetic Algorithms aids to look for the best solution among a number of possible solutions throw reproduction. Reproduction can be implemented in an algorithmic form, based on reproduction operators, suited to the encoding scheme. The objective of the reproduction operators is to ensure diversity in the population, such that the fittest solutions can be derived through the evolutionary process. Such evolution process is composed of crossover and mutation operations.

During such process, invalid chromosomes may be produced. In the proposed chromosome representation, repetition of genes produces invalid chromosomes because a data point cannot be center point of more than one cluster. So to be able to detect production of invalid chromosomes, we sort the chromosome genes in ascending order. In doing so, we aim to achieve small length, fast detection of invalid chromosomes, and faster crossover and mutation operations. Fig.3. illustrates our crossover process using a mask.

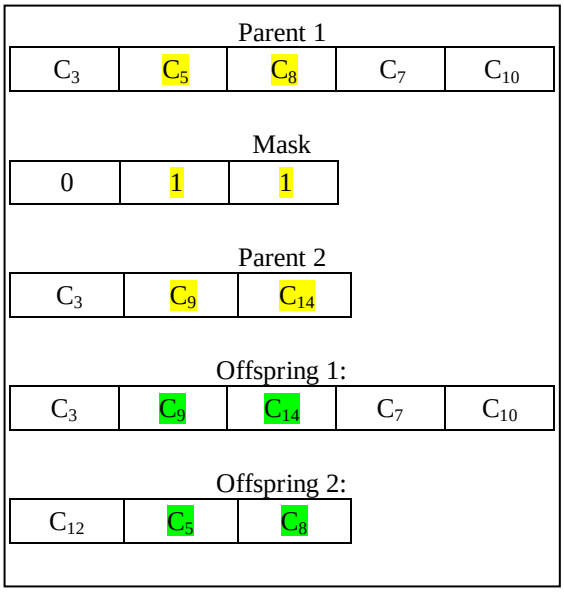


FIG. 3– The crossover process.

1) Crossover

It is a reproduction process that the children chromosomes are generated according to the fitness function values of their parents. We define a crossover operator in such a way that it could accept parent chromosomes with different number of genes. It occurs with probability *crossp*.

For each two selected parents, crossover is applied, producing two offsprings. To this end, a binary mask is created. Its length is equal to that of the shorter parent. In this mask (Fig.3.), digit "1" occurs with probability *pone*.

$$C^{(1)}_j = C^{(1)}_j + \alpha\, C^{(2)}_j \qquad (1)$$

$$C^{(2)}_j = C^{(2)}_j + (1-\alpha)\, C^{(1)}_j \qquad (2)$$

FIG. 4– Formulas of the crossover between parent P_1 and parent P_2, producing gene j of the offspring $C^{(1)}$ and gene j of the offspring $C^{(2)}$.

Offspring O_1 is created by applying the formula (1) and offspring O_2 is created by applying the formula (2). These formulas (Fig.4) are applied on genes in the parents corresponding to the positions in the mask with a digit "1".

Input: data points, k_{max}
Output: data points partitioned into K clusters
Procedure:

1. Generate the initial populations of P heterogeneous chromosomes;
2. Evaluate the fitness values of all chromosomes in the initial population;
3. Select the three fittest parent chromosomes from the population, denoting them as (P_1, P_2, P_3);
4. Apply the crossover and mutation operators on the two fittest chromosomes in (P_1, P_2, P_3) to generate two offsprings (O_1, O_2);
5. Evaluate the fitness values of the offsprings
6. Select the three fittest chromosomes from (P_1, P_2, P_3, O_1, O_2) and insert them in the population;
7. In the case of an offspring O_1 or O_2 hasn't been selected at the step 6, replace the worst chromosome in the population with such offspring;
8. Update crossover and mutation probabilities according to an adjustment function;
9. Repeat from step (3), n iteration times;
10. Extract the fittest chromosome,
11. Extract clusters and their centers from the fittest chromosome and apply the KMO operator to obtain the clusters.

FIG. 5– The main steps of the proposed genetic algorithm to clustering.

Our overall genetic process organizes operations as depicted by Fig. 5. To find the closest cluster to every point, we apply a k-means operator (KMO) (Najmah, 2016), which is a one step of the classical k-means algorithm. In other words, we assign data points to their clusters for each new chromosome using the KMO operator, in order to compute its fitness. It is worth noting that we should check the validity of the resulting chromosomes to avoid redundant centers. In the case of a redundancy, we readopt the previous chromosome value. Fig.4 depicts the crossover formulas with α as a uniformly distributed random number such that $\alpha \in [-1, 1]$.

1) Mutation

Mutation is intended to prevent falling of all solutions in the population into a local optimum of the solved problem or enhancing the obtained chromosome. Mutation takes place with a lower probability than that of the crossover. We define two types of mutation to be taken place after the crossover. Topological mutation is intended to add or delete genes from an offspring. Gene mutation is intended to modify a selected gene from an offspring. It is

applied by randomly selecting a gene from an offspring and replacing it with a data point randomly extract from a dataset. Topological mutation procedure is as follows:

1. For each offspring, randomly generate α as a uniformly distributed random number such that $\alpha \in [0, 1]$.
2. If $\alpha \in [0, 0.5]$ then randomly select a center from the longest offspring and delete it.
3. If $\alpha \in]0.5, 1]$ then randomly select a data from the dataset and append it to the shortest offspring.

Afterward, a validity check is performed to avoid redundant centers and to keep the cluster number in $[2, K_{max}]$.

4 Experimental results

To evaluate our approach and test its efficiency, we compare it with K-means, and the closest work to ours of (Sivanandam, 2008) using two selected datasets from UCI repository of machine learning database, which are Iris and Lymphoma datasets. It is worth noting that the first difference of our approach and the two above mentioned methods is that we automatically determine the k number of clusters and their initial centroids. In the following, we show that our approach outperforms at least such two methods when considering two important parameters, that is, the average error and the average execution time.

In the Iris dataset, each data point has four feature values, which represents the length and the width of sepal and petal, in centimeters. It has three classes with 50 samples per class. The value of k is therefore chosen to be three (clusters).

In the Lymphoma dataset, we find 62 samples consisting of 4026 genes spanning three classes, which include 42 Diffuse Large B-Cell Lymphoma samples, nine Follicular Lymphoma samples, and 11 B-cell Chronic Lymphocytic Leukemia samples. This dataset is to be partitioned into three clusters. Each dataset was use for each method for 10 times and then we determined the average time and error as mentioned in TAB II.

TAB. II – Two used UCI data sets.

Dataset	paramters	k-means	[9]	Proposed approach
Iris	Avg Error	36.53	29.59	19.59
	Avg Time	18.34	10.56	9.34
Lymphoma	Avg Error	44.12	38.49	22.15
	Avg Time	16.15	10.25	8.28

TAB II shows the efficiently experiments of the proposed method over K-means and the work of (Sivanandam, 2008), which closest to ours. The experiments have been conducted to measure the average error and average execution time parameters for the three methods. As shown in TAB II, for each dataset, average error and average time are listed to show the

trade-off between them. From the table it is observed that, our method outperforms the other two ones, as it enhanced the average error and the average execution time.

5 Conclusion and future Works

In this paper we have introduced a modified genetic algorithm, which is based on different structures of chromosomes in a population, in order to be applied on the data clustering problem. It allows the used in automatically determining the initial values of the cluster centroids, providing better results than using random numbers. The approach allowed acceleration of the search process by reducing the average execution time and obtain the best data partitions. As a future works, we plan to improve the fitness function by investigating other reproduction operators to gain better performance.

References

Abdul Nazeer K. A., M. P. Sebastian Improving the Accuracy and Efficiency of the k-means Clustering Algorithm. WCE 2009, July 1 - 3, London, U.K. 2009.

Allaire G. And M. Bena, Conception Optimal Des Structures, Berlin Heidelberg New York: Springer, 2000.

Al-Malki A., Mohamed M. Rizk, M. A. El-Shorbagy, A. A. Mousa1Hybrid Genetic Algorithm with K-Means for Clustering Problems, Open Journal of Optimization, 5, 71-83, 2016.

Chittu.V,N.Sumathi, A Modified Genetic Algorithm Initializing K-Means Clustering, Global Journal Of Computer Science And Technology,Vol. 11 Issue 2 February 2011

Delavar A.G., G.H.Mohebpour, A New Genetic Center Based Data Clustering Algorithm Based On K-Means, International Journal Of Mechatronics, Electrical And Computer Technology Vol. 4(13), , Pp. 1820-1839, October, 2014.

Ettaouil M.; E. Abdelatif, F. Harchli, Improving The Performance Of K-Means Algorithm Using An Automatic Choice Of Suitable Code Vectors And Optimal Number Of Clusters, Journal Of Theoretical And Applied Information Technology. Vol. 56 No.2. 20th October 2013.

Fukunaga K, Introduction To Statistical Pattern Recognition. Academic Press, San Diego, Ca. ,1990.

García S.; J.Luengo; F.Herrera (2014), Data Pre-processing In Data Mining, Cham : Springer, P 327

Han J.; M.Kamber, Data Mining: Concepts And Techniques. Morgan Kaufmann, 2001.

Hartono, Erianto Ongko, and Dahlan Abdullah,, Determining a Cluster Centroid of K-Means Clustering Using Genetic Algorithm, International Journal of Computer Science and Software Engineering (IJCSSE), Volume 4, Issue 6, June 2015.

Junji W., Advances In K-Means Clustering: A Data Mining Thinking, Berlin ; New York : Springer, P186, 2012.

Kumar S.K., P.Renuga, Reactive Power Planning Using Real GaComparison With Evolutionary Programming, International Journal Of Recent Trends In Engineering, Vol 1, No. 3. 2009.

Massart D. and L.Kaufman, (1983). The Interpretation Of Analytical Chemical Data By The Use Of Cluster Analysis. John Wiley & Sons, New York, Ny.

Mcgregor A.; M.Hall; P. Lorier, And J. Brunskill, Flow Clustering Using Machine Learning Techniques, The National Laboratory Of Applied Network Research (Nlanr), San Diego Supercomputer Center, University Of California San Diego, 10100 Hopkins Drive, Ca

MLD, Index of /Ml/Machine-Learning-Databases/Iris, last access: 2017.

Najmah A. et al, A Variant Of Genetic algorithms For Non-Homogeneous Population, Intl Conf On Applied Mathematics, Rome. 2016.

Romany F.M., Using Genetic Algorithm For Identification Of Diabetic Retinal Exudates In Digital Color Images, Journal of Intelligent Learning Systems And Applications, 4, 188-198, 2012.

Rui X, Wunsch DC II, Clustering. IEEE Press series on computational intelligence, John Wiley & Sons, 2009.

Sawant K. B.. Efficient Determination of Clusters in K-Mean Algorithm Using Neighborhood Distance. The International Journal of Emerging Engineering Research and Technology 3(1): 22-27, 2015.

Sivanandam , S.N, And Deepa S.N. Introduction To Genetic Algorithms. Berlin Heidelberg New York: Springer, 2008.

Valarmathi K.; D. Devaraj And T.K.Radhakrishnan Real-Coded Genetic Algorithm For System Identification And Controller Tuning, Applied Mathematical Modelling 33 3392–3401, 2009.

Vo-Van T., Trung Nguyen-Thoi, Trung Vo-Duy, Vinh Ho-Huu & Thao Nguyen-Trang, Modified genetic algorithm-based clustering for probability density functions, Journal of Statistical Computation and Simulation, Pages 1-16 , Published online: 12 Mar 2017.

Zaki, Mohammed J., Meira Jr, Wagner. Data Mining and Analysis. Cambridge University Press: Cambridge, 2014.

Résumé. Le clustering a été reconnu comme une principale méthode d'exploration de données pour la découverte de connaissances. Il a été largement utilisé dans plusieurs domaines, comme la biologie, l'ingénierie des systèmes et les sciences sociales, afin d'identifier les groupes naturels dans de données massives. K-means est la méthode la plus répandue pour le regroupement de données. Cependant, elle souffre des inconvénients d'exiger le nombre de clusters et leurs centres initiaux, qui doivent être fournis par l'utilisateur. Dans cet article, nous présentons notre approche pour générer automatiquement de tels paramètres afin d'obtenir des clusters optimaux en utilisant un algorithme génétique modifié, opérant sur des populations hétérogènes. En outre, nous avons introduit de nouveaux opérateurs de croisement et de mutation. Les résultats expérimentaux montrent que notre algorithme génétique modifié est une alternative plus efficace aux approches existantes.

Outil d'aide à la prédiction de défauts logiciels

Ahmed Taha Haouari, Labiba Souici-Meslati, Fadila Atil

Laboratoire LISCO, Université Badji Mokhtar-Annaba, BP 12, Annaba 23000, Algérie
ahmed-taha.haouari@univ-annaba.org, labiba.souici@univ-annaba.org,
fadila.atil@univ-annaba.dz

Résumé. La fouille de données en ingénierie de logiciels est un domaine en pleine expansion, notamment la prédiction de défauts logiciels qui consiste, essentiellement, à appliquer des méthodes d'apprentissage automatique sur des métriques logicielles pour classer les entités d'un système comme étant sujettes, ou non, aux défauts.
Malgré les nombreux modèles proposés pour la prédiction de défauts logiciels, aucun n'a atteint une applicabilité étendue en raison du manque d'outils d'automatisation du processus de prédiction. Ces outils permettraient aux gestionnaires de projets de cibler les entités sujettes aux défauts pour mieux gérer les ressources disponibles et améliorer la phase de test.
Dans ce contexte, notre objectif est de proposer un outil d'aide à la prédiction de défauts logiciels, suite à la construction de classifieurs basés sur six méthodes d'apprentissage automatique pour la prédiction de défauts, en utilisant cinq bases de données issues du référentiel PROMISE.

1 Introduction

Depuis les années 90, la prédiction des défauts logiciels (Software Fault Prediction) est devenue un des principaux axes de recherche pour l'amélioration de la qualité des logiciels. Dans ce domaine, les chercheurs utilisent l'historique des programmes précédents, principalement sous forme de métriques logicielles, pour construire des modèles de prédiction. Des techniques statistiques ou des méthodes d'apprentissage automatique sont ainsi utilisées afin de pouvoir classer les entités logicielles (classes ou méthodes, par exemple) des nouvelles applications comme étant sujettes aux défauts (fault prone) ou non sujettes aux défauts (not fault prone) (Catal, 2011).

L'objectif principal de cette classification est de permettre aux gestionnaires de projets de mieux gérer les ressources disponibles. Par exemple, si un module est classé comme étant "*fault prone*", il faudra allouer les ressources nécessaires afin de se concentrer sur son test (Catal et Diri, 2007 ; Jiang, 2008 ; Kaur et Kaur, 2015 ; Malhotra, 2015 ; Radjenović et al, 2013). Les avantages d'un tel procédé sont multiples (Erturk et Sezer, 2015 ; Catal, 2011) :

- Raffinement du processus de test, permettant ainsi d'augmenter la qualité du système réalisé, tout en réduisant le coût lié à cette activité qui pourrait atteindre 50% du coût total de développement.
- Détection des modules qui nécessitent un refactoring lors de la phase de maintenance.

- Possibilité de trouver de meilleures conceptions alternatives si on applique le procédé de prédiction sur les métriques au niveau des classes, pendant la phase de conception.
- Réduction du temps et des efforts consacrés au processus de révision du code.

Malgré un nombre important de travaux de recherche dans le domaine de la prédiction de défauts logiciels, ce procédé n'est pas encore couramment appliqué dans l'industrie logicielle. Ceci est principalement dû au fait que les travaux publiés ne se concentrent que sur l'efficacité prédictive des méthodes construites, sans prendre en considération les besoins réels de l'industrie logicielle (Rana et al, 2014). Par conséquent, on constate un manque accru d'outils permettant de faciliter le travail d'un chef de projet afin de prédire les défauts logiciels avant leur apparition (Lessmann et al, 2008). Ceci est particulièrement le cas des nouveaux projets pour lesquels on ne dispose pas de données antérieures permettant de simplifier le processus de prédiction, qui devient, de ce fait, inter-projets.

Dans cette optique, l'objectif de notre travail est de proposer un outil d'aide à la prédiction de défauts logiciels inter-projets. Afin d'avoir une idée précise des caractéristiques que notre outil devrait avoir pour être utile et efficace, nous avons mené une étude expérimentale de construction et d'évaluation de classifieurs pour la prédiction de défauts logiciels. Cette étude inclut six algorithmes d'apprentissage automatique appliqués sur cinq bases différentes, issues de cinq logiciels orientés objet se trouvant dans le référentiel PROMISE (Menzies et al, 2016) spécialisé dans les données liées au domaine de génie logiciel.

La suite de cet article est organisée comme suit : la section 2 présente la prédiction de défauts logiciels inter-projets. La section 3 détaille l'étude expérimentale que nous avons menée en décrivant les algorithmes, les bases données et les métriques utilisées. Les résultats des expérimentations effectuées sont résumés et discutés dans la section 4. Enfin, la section 5 est dédiée à la présentation de l'outil proposé. Une conclusion et des perspectives futures clôtureront cet article.

2 Prédiction de défauts logiciels inter-projets

Dans le cas commun de la prédiction de défauts logiciels, les chercheurs utilisent les données historiques d'un projet réalisé pour construire un modèle de prédiction pour sa nouvelle version, on parle alors de prédiction intra-projet. Cependant, il arrive souvent que ces données ne soient pas disponibles ou qu'il s'agisse d'une nouvelle application (pas de version antérieure). On parle, dans ce cas, de prédiction inter-projets où nous avons particulièrement constaté la rareté des travaux de recherche (Zimmermann et al, 2009 ; He et al, 2012, 2015).

Zimmermann et al. (2009) ont mené des expérimentations où ils ont utilisé les données historiques de deux navigateurs internet qui sont Firefox et Internet Explorer. En employant la méthode de régression logistique, ils ont trouvé que les données relatives au premier permettaient de prédire les défauts du second avec de bonnes performances (rappel ou recall égal à 81.25 %) alors que l'opposé a donné de mauvais résultats (rappel égal à 4.12%).

He et al (2012, 2015) ont mené des expérimentations sur des données du référentiel PROMISE avec plusieurs algorithmes d'apprentissage automatique (Arbres de décision, Machines à vecteurs supports, Régression logistique, Naïve Bayes…). Leur étude a permis de conclure que c'est complètement faisable d'effectuer la prédiction inter-projets à condition

de bien sélectionner les données d'apprentissage d'après des mesures de performances telles que le rappel et la f-mesure (voir section 3.4).

3 Étude expérimentale effectuée

Cette section décrit les algorithmes d'apprentissage automatique, les bases données et les métriques que nous avons utilisés pendant notre étude expérimentale, ainsi que la méthode d'évaluation des performances des algorithmes sélectionnés pour la prédiction de défauts logiciels.

3.1 Méthodes d'apprentissage automatique

Les méthodes utilisées pour la construction des modèles de prédiction de défauts sont soit des méthodes issues des statistiques comme la régression logistique, qui est l'une des technique les plus fréquemment implémentées, soit des méthodes d'apprentissage automatique (Machine Learning) pouvant être supervisées ou non supervisées, comme les réseaux de neurones, les arbres de décision ou le Clustering par les k-means (Kaur et Kaur, 2015 ; Malhotra, 2015 ; Radjenović et al, 2013).

Plusieurs algorithmes d'apprentissage ont été adoptés pour la prédiction de défauts logiciels. Selon plusieurs études (Catal, 2011 ; Beecham, et al, 2010 ; Kaur et Kaur, 2015 ; Malhotra, 2015 ; Radjenović et al, 2013), ces méthodes donnent de meilleurs résultats que les méthodes statistiques, considérées comme des méthodes à boîte noire où les relations entre les entrées et les sorties ne sont pas faciles à détecter ou analyser, de plus, elles sont très dépendantes des données.

Pour la construction de nos classifieurs, nous avons choisi l'outil WEKA (http://www.cs.waikato.ac.nz/ml/weka/). C'est un logiciel open source qui fournit un ensemble de classes et d'algorithmes en Java implémentant les principales méthodes de data mining.

Les méthodes d'apprentissage que nous avons sélectionnées pour la construction de nos classifieurs de prédiction de défauts logiciels sont les suivantes :

Système immunitaire artificiel pour la reconnaissance (AIRS): Artificial Immune recognition System (AIRS) a été introduit par Watkins (2001). C'est un algorithme d'apprentissage automatique supervisé, nécessitant l'utilisation d'un ensemble d'antigènes comme données d'apprentissage dont le système doit produire un ensemble d'anticorps utiles pour la phase de classification. En plus de la première version AIRS1, une deuxième, nommée AIRS2, a été proposée comme résultat de la collaboration de Watkins A., Timmis J. et Boggess L. (Watkins et al, 2004). AIRS2 présente quelques différences mineures par rapport à AIRS1, mais est moins complexe que son prédécesseur. A la fin de la phase d'apprentissage des deux versions de l'algorithme AIRS, nous obtenons un ensemble de cellules mémoires permettant d'effectuer une classification et cela grâce à la méthode des k-plus proche voisin (kppv). AIRS a été utilisé avec succès dans plusieurs problèmes de classification (Brownlee, 2005) et a été aussi appliqué dans le domaine de la prédiction de défauts logiciels (Catal et Diri 2007, 2009 ; Abaei et Selamat, 2013).

Les réseaux de neurones artificiels (MLP): Le réseau neuronal artificiel le plus connu est le perceptron multicouches (Multi Layer perceptron ou MLP). C'est l'un des algorithmes les plus utilisées dans la prédiction. Une étude récente montre que le MLP est l'un des algorithmes les plus stables pour la prédiction de défauts logiciels (Kaur et Kaur, 2015).

Les arbres de décision (J48): Plusieurs études ont montré que les arbres de décision s'adaptent très bien au problème de prédiction de défauts logiciels (Catal et Diri, 2007, 2009 ; Jiang et al, 2008 ; Hall et al, 2011 ; Abaei et Selamat, 2013). Nous avons choisi d'utiliser J48 qui est une implémentation en java de l'algorithme C4.5.

Les forêts aléatoires d'arbres décisionnels (Random Forest : RF) : ce sont des classifieurs basés sur l'algorithme des arbres. Les arbres sont construits sans élagage donc ils sont de très grande taille. Lorsque tous les arbres de la forêt sont construits, une nouvelle instance (classe de l'arbre) est fixée sur tous les arbres pour lancer un processus de vote pour la classification. Dans RF, la classe qui reçoit le plus de votes sera sélectionnée pour classer la nouvelle instance (Jiang, 2008). Selon la communauté des chercheurs dans le domaine de prédiction de défauts, RF est le meilleur algorithme d'apprentissage automatique recensé pour la prédiction de défauts (Hall et al, 2011 ; Malhotra, 2015 ; Radjenović et al, 2013)

Le Naïve Bayes (NB) : Selon Jiang et al (2008) et He et al (2015), c'est l'un des algorithmes les plus simples mais les plus efficaces pour la prédiction de défauts.

3.2 Métriques choisies

Pour la construction de nos modèles de prédiction, nous avons choisi les métriques orientées objet car ce sont les métriques les plus efficaces lors de la phase de pré-livraison des logiciels (Malhotra, 2015 ; Radjenović et al, 2013). De plus, le paradigme orienté objet (OO) est le plus dominant à l'heure actuelle. En plus de la métrique LOC (nombre de ligne de code), le tableau 1 montre toutes les métriques utilisées dans notre travail. Pour plus d'information, se référer à Jureczko et Spinellis (2010).

Nom	Signification	Référence source
DIT : Depth of Inheritance Tree	Profondeur de la classe dans l'arbre d'héritage	Chidamber et Kemerer (1994)
WMC : Weighted Methods per Class	Nombre de méthodes d'une classe, chaque méthode a une complexité égale à 1.	
CBO : Coupling Between Objects	Nombre de classes auxquelles une classe est couplée	
NOC : Number Of Children	Nombre de classes qui héritent directement de cette classe	
RFC : Response For a Class	Ensemble des méthodes qui peuvent être directement appelées lors de l'exécution de n'importe quelle méthode de cette classe.	
LCOM : Lack of Cohesion in Methods	Différence entre le nombre de paires de méthodes qui n'accèdent pas aux mêmes attributs et le nombre de paires de méthodes qui accèdent aux mêmes attributs d'une classe. Elle est égale à 0 si le résultat est négatif.	

LCOM3: Lack of Cohesion in Methods	$LCOM3 = \frac{(\frac{1}{a}\sum_{j=1}^{a}\mu(Aj))-m}{1-m}$ avec m : nombre de méthodes par classe. a : nombre d'attributs par classe. μ(A) : nombre de méthodes qui accèdent à l'attribut A.	Henderson-Sellers (1996).
Ca : Afferent couplings Ce Efferent couplings	Nombre de classes qui dépendent de la classe mesurée. Nombre de classes dont la classe mesurée est dépendante.	Martin (1996)
NPM : Number of Public Methods	Nombre de méthodes d'une classe qui sont déclarées comme publiques.	Bansiya et Davis (2002)
DAM : Data Access Metric	Ratio entre le nombre d'attributs privés (ou protégés) et le nombre total d'attributs de classe.	
MFA : Measure of Functional Abstraction	Ratio entre nombre de méthodes héritées par une classe et le nombre total de méthodes accessibles par la méthode de la classe	
MOA : Measure of Aggregation	Nombre de champs de classe dont les types sont des classes définies par le développeur.	
CAM: Cohesion Among Methods of Class	Connexité entre les méthodes d'une classe, elle est calculée sur la base de leurs paramètres..	
IC : Inheritance Coupling	Nombre de classes parentes auxquelles une classe donnée est couplée. Une classe est couplée à sa classe parente, si l'une de ses méthodes héritées est fonctionnellement dépendante de la méthode redéfinie ou nouvelle de la classe.	Tang et al (1999)
CBM : Coupling Between Methods	Nombre total de méthodes nouvelles ou redéfinies auxquelles toutes les méthodes héritées sont couplées	
AMC : Average Method Complexity	Taille moyenne des méthodes pour chaque classe.	
CC : Cyclomatic complexity	Métrique de niveau méthode adaptée en OO et divisée en deux parties MAX_CC= CC maximal d'une méthode qui se trouve dans la classe AVG_CC= moyenne CC de toutes les méthodes de la classe	McCabe (1974)
Bug	**True** pour dire la classe est fault prone (sujette aux défauts) **False** pour dire la classe est not fault prone (non sujette aux défauts)	

TAB. 1- Métriques OO utilisées dans notre étude expérimentale

3.3 Bases de données

Les valeurs des métriques choisies sont extraites des bases de données que nous avons utilisées. Nous avons sélectionné 5 bases différentes issues de 5 programmes orientés objet se trouvant dans le référentiel de données PROMISE. Ces données sont de plus en plus utilisées dans le domaine, elles sont issues des travaux de Jureczko et Madeyski (2010) ainsi que Jureczko et Spinellis (2010) qui donnent plus d'information sur la manière avec laquelle les

données sont rassemblées. Les logiciels open source d'où sont extraites ces bases de données, appartiennent à *The Apache Software Foundation*, ils sont tous écrits en langage Java. Le tableau 2 présente un aperçu sur les bases de donnés utilisées. Les instances représentent les classes qui composent le programme.

Base de données	Nombre d'instances (classes)	Nombre d'instances fault prone	Taux % d'instance fault prone
Camel 1.6	965	188	19.50
Lucene 2.4	340	203	59.70
Poi 3.0	442	281	63.60
Xalan 2.5	803	387	48.20
Xerces 1.4	588	437	74.30

TAB. 2- Bases de données utilisées

Avant d'utiliser ces données, nous avons effectué un prétraitement afin de les adapter à notre problème de classification binaire.

À l'origine, la caractéristique "bug" était représenté par des chiffres indiquant le nombre d'erreurs relevées dans la classe inspectée. Pour remédier à cela, nous avons utilisé la même démarche effectué par certains chercheurs (Erturk et Sezer, 2015 ; He et al, 2012, 2015), en modifiant toutes les métriques bugs de toutes les instances des bases par ce qui suit :

Si bug = 0, ceci indique qu'il n'y a pas de défauts, donc il est remplacé par "false" pour indiquer que la classe est not fault prone

Si bug > 1, Alors elle remplacé par "true" pour dire que la classe est fault prone.

3.4 Méthode d'évaluation des performances

Pour l'évaluation des classifieurs dans le domaine de l'apprentissage automatique et la fouille de données, on se base souvent sur la matrice de confusion. C'est une matrice N x N où N est le nombre de classes, elle reporte comment le modèle a classé l'échantillon de test par rapport à sa vraie classe. Dans notre cas de prédiction de défauts logiciels, il s'agit d'un problème à deux classes (True : fault prone/false : not fault prone).

	True	False
True	TP	FN
False	FP	TN

TAB. 3 Matrice de confusion

Dans cette matrice de confusion, **TP** indique le nombre de vrais positifs qui représente le nombre d'exemples *True* qui sont classés *True*. **FN** : le nombre de faux négatifs qui représente le nombre d'exemples étiquetés *True* classés *False* par le modèle. **FP**: le nombre de

faux positifs (faux rejet), exemples *False* classés *True*. **TN** : le nombre de vrais négatifs, qui représente le nombre d'exemples *False* classés comme tels par le modèle.

Plusieurs mesures d'évaluation des performances de l'apprentissage peuvent êtres calculées à partir de la matrice de confusion.

Taux d'erreur =$\frac{FP+FN}{TP+FP+TN+FN}$.

Rappel (recall) =$\frac{TP}{TP+FN}$. (Sensibilité ou TVP : taux de vrais positifs)

Précision=$\frac{TP}{TP+FP}$.

Anti-Spécifité= 1-Spécificité =$1-\frac{TN}{TN+FP}=\frac{FP}{TN+FP}$ (taux de faux positifs)

Taux d'exemples bien classés (puissance ou accuracy) = $\frac{TP+TN}{TP+FP+TN+FN}$.

F-mesure (mesure F) = $\frac{2*(\text{Recall}*\text{Precision})}{(\text{Recall}+\text{Precision})}$

Selon plusieurs travaux, la puissance (accuracy), le rappel (recall) et la précision ne sont pas de bons indicateurs quand il s'agit de comparer des classifieurs pour la prédiction car les deux dernières mesures ne prennent pas en compte le taux des faux positifs, ce qui conduit à une vue partielle de la classification (Catal et Diri, 2009 ; Lessmann et al, 2008 ; Erturk et al, 2015).

Suite à notre étude bibliographique, la méthode d'évaluation des performances que nous avons choisie est l'aire sous la courbe ROC (AUC-ROC : Area Under Curve-Receiver Operating Characteristic). Dans le cas d'un classifieur binaire, il est possible de visualiser les performances du classifieur sur cette courbe. La courbe ROC est une représentation du taux de vrais positifs (recall) en fonction du taux de faux positifs (anti-spécificité). Son intérêt est de s'affranchir de la taille des données de test quand les données sont déséquilibrées, et c'est notre cas, comme le montre le Tableau 2 de nos bases d'apprentissage. Cette représentation met en avant un nouvel indicateur qui est l'aire sous la courbe (AUC). Plus elle se rapproche de 1, plus le classifieur est performant.

Lorsque la valeur de AUC est entre 0.5 et 0.7, elle dite faible, et si AUC est entre 0.7 et 0.9 la méthode est dite bonne pour certains types d'applications et si c'est plus, on dit généralement, que la méthode a un très bon taux de prédiction (Zhou et al, 2010) . De plus, si AUC=0.5, la classification effectuée est dite aléatoire.

Pour évaluer les performances de nos classifieurs, le déroulement de nos expérimentations est effectué comme suit :

- Phase 1 : les données sont divisées en deux parties : une des bases (par exemple Camel 1.6) va être la base d'apprentissage, alors que les 4 autres, une à une, vont servir comme base de test.
- Phase 2 : nous sélectionnons une des autres bases qui n'a pas servi comme base d'apprentissage puis répétons la première phase.
- L'expérience est conduite en utilisant les algorithmes d'apprentissages automatiques cités dans la section 3.2.
- L'expérience est répétée 5 fois afin d'avoir des résultats plus fiables.

4 Résultats obtenus

Les résultats de notre étude expérimentale, relative à la construction et l'évaluation des classifieurs, vont nous servir de base pour configurer notre outil d'aide à la prédiction de défauts logiciels. Le Tableau 5 détaille les performances obtenues lors des expérimentations décrites dans la section précédente sous forme (AUC-ROC/puissance %). Ces performances sont classées par base d'apprentissage, et indiquent les valeurs de AUC-ROC et de la puissance (%) obtenues avec chaque classifieur, en utilisant les autres bases comme échantillon de test des modèles construits.

Base d'apprentissage	**Camel 1.6**			
Bases de test	**Lucene 2.4**	**Poi 3.0**	**Xalan 2.5**	**Xerces 1.4**
AIRS 1	0.528/47.94	0.545/49.09	0.513/52.30	0.594/43.53
AIRS 2	0.489/43.23	0.537/45.70	0.504/51.55	0.493/34.01
J48	0.558/54.11	0.575/59.95	0.573/55.66	0.553/38.77
NB	0.621/52.05	0.548/47.05	0.543/54.91	0.531/37.75
MLP	0.535/43.23	0.664/57.23	0.515/55.79	0.506/39.62
RF	0.66/45.29	0.708/42.30	0.581/54.67	0.718/32.65
Base d'apprentissage	**Lucene 2.4**			
Bases de test	**Camel 1.6**	**Poi 3.0**	**Xalan 2.5**	**Xerces 1.4**
AIRS 1	0.534/53.78	0.528/49.32	0.536/53.3	0.694/59.69
AIRS 2	0.522/49.32	0.529/48.86	0.545/54.17	0.652/58.84
J48	0.593/54.19	0.824/78.95	0.538/54.54	0.939/95.40
NB	0.626/70.88	0.806/61.99	0.595/57.40	0.854/66.15
MLP	0.616/53.47	0.682/62.66	0.547/54.04	0.924/94.89
RF	0.64/47.25	0.788/77.14	0.59/53.05	0.999/99.65
Base d'apprentissage	**Poi 3.0**			
Bases de test	**Camel 1.6**	**Lucene 2.4**	**Xalan 2.5**	**Xerces 1.4**
AIRS 1	0.54/60.41	0.559/55.88	0.527/53.05	0.595/46.59
AIRS 2	0.57/63.21	0.626/61.47	0.526/52.92	0.615/46.93
J48	0.53/48.29	0.687/67.64	0.578/57.53	0.485/53.06
NB	0.612/74.09	0.672/55.29	0.561/54.91	0.666/40.13
MLP	0.547/63.93	0.687/60	0.581/56.03	0.705/50.17
RF	0.621/55.85	0.705/65	0.586/54.54	0.663/51.53

Base d'apprentissage	**Xalan 2.5**			
Bases de test	**Camel 1.6**	**Lucene 2.4**	**Poi 3.0**	**Xerces 1.4**
AIRS 1	0.577/60.10	0.622/59.70	0.555/50.67	0.603/44.89
AIRS 2	0.537/67.66	0.582/55.58	0.519/45.24	0.625/51.70
J48	0.535/51.60	0.552/54.41	0.578/61.08	0.709/60.20
NB	0.635/78.03	0.705/48.52	0.72/44.11	0.713/36.05
MLP	0.532/63.52	0.497/48.23	0.478/44.79	0.494/43.36
RF	0.574//65.59	0.572/55.29	0.681/54.52	0.668/48.80
Base d'apprentissage	**Xerces 1.4**			
Bases de test	**Camel 1.6**	**Lucene 2.4**	**Poi 3.0**	**Xalan 2.5**
AIRS 1	0.485/32.64	0.567/62.05	0.586/66.06	0.492/48.19
AIRS 2	0.564/35.33	0.572/63.82	0.578/68.09	0.526/51.30
J48	0.561/26.21	0.536/59.11	0.558/64.93	0.507/48.06
NB	0.579/46.52	0.698/70	0.796/76.24	0.569/53.54
MLP	0.629/26.63	0.684/60.29	0.705/65.38	0.532/46.94
RF	0.568/23.73	0.634/59.41	0.782/64.93	0.544/48.69

*TAB. 4- Résultats détaillés des expérimentations sous forme (**AUC-ROC/ puissance %**)*

Le Tableau 5 synthétise les résultats que nous avons obtenus lors des expérimentations décrites dans la section précédente, ils sont classés par base d'apprentissage, et indiquent la moyenne AUC-ROC obtenue avec chaque classifieur, en utilisant les autres bases comme échantillon de test des modèles construits.

	Camel 1.6	Lucene 2.4	Poi 3.0	Xalan 2.5	Xerces 1.4
AIRS 1	0.545	0.573	0.55525	0.58925	0.5325
AIRS 2	0.50575	0.562	0.58425	0.56575	0.56
J48	0.56475	0.7235	0.57	0.5935	0.5405
NB	0.56075	0.72025	0.62775	0.69325	0.6605
MLP	0.555	0.69225	0.63	0.50025	0.6375
RF	**0.66675**	**0.75425**	**0.64375**	0.62375	0.632

TAB. 5 - Moyenne AUC-ROC par base apprentissage/classifieur

La figure 1 donne un meilleur aperçu des résultats obtenus. D'après nos expérimentations, comme le montrent le Tableau 5 et la Figure 1, le meilleur algorithme pour la prédiction inter-projet est RF (Random Forest), ce qui concorde avec la littérature du domaine où RF est considéré comme le meilleur algorithme d'apprentissage automatique pour la prédiction intra-projet. De plus, dans notre cas la base Lucene 2.4 est la base qui a permis d'avoir les

meilleures performances par tous les algorithmes que nous avons utilisés dans notre étude. Par exemple, si nous prenons Lucene 2.4 comme base d'apprentissage pour l'algorithme RF, cela nous a permis d'avoir 99.65 % de taux de classification des instances de la base Xerces 1.4, ce qui nous encouragé à utilisé RF comme la méthodes de fouille de données et que Lucene 2.4 sera la base d'apprentissage par défaut pour notre outil.

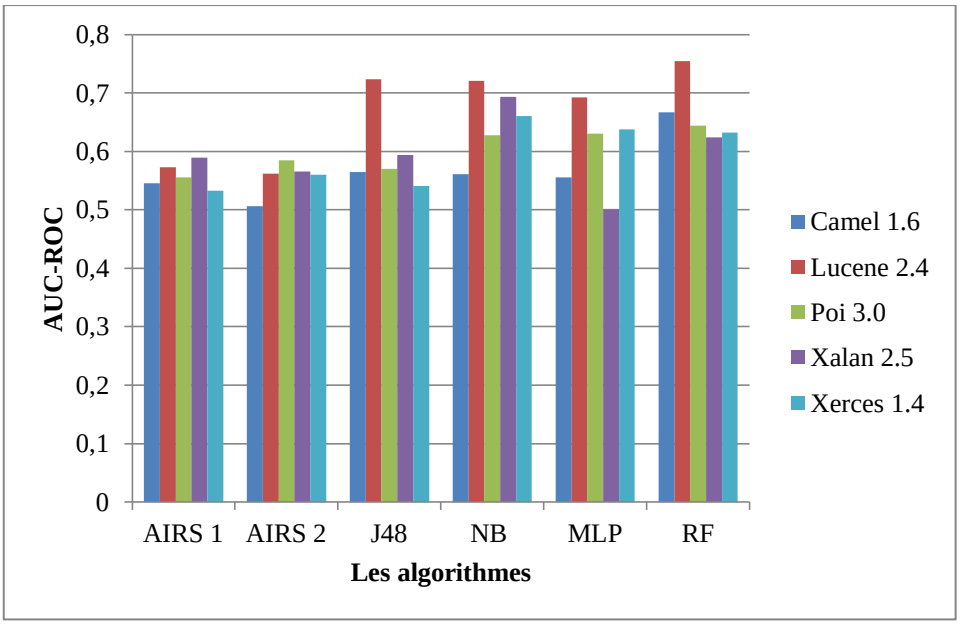


FIG. 1- Synthèse des résultats obtenus

5 Présentation de l'outil réalisé

Malgré un nombre assez élevé de travaux sur la prédiction de défauts logiciels, il y a un manque notable d'outils dédiés précisément à cet effet. Il y a certes beaucoup de logiciels capables de calculer les métriques, qu'elles soient orientées objet ou autres, mais ces logiciels sont majoritairement propriétaires et payants. Par contre, plusieurs outils proposent en open source des algorithmes d'apprentissage automatique facilement accessibles, tels que WEKA.

Les quelques rares propositions d'outils automatiques de prédiction de défauts logiciels se focalisent sur la prédiction intra-projet (voir section 2), comme l'outil de Ostrand et Weyuker (2010) qui utilise l'algorithme de régression binomiale négative ainsi que celui de Catal et al (2011) qui utilise l'algorithme Naïve Bayes (NB).

Dans la suite de cette section, nous allons présenter notre propre outil d'aide à la prédiction de défauts logiciels. Cet outil doit prendre en compte tous les résultats trouvés dans l'étude expérimentale effectuée afin de mieux cibler les modules sujets aux défauts lors de la phase de test. Il doit être capable de calculer les métriques orientés objets, et doit utiliser un des algorithmes de prédiction.

Une des raisons pour lesquelles il y a un manque d'outils capables de prédire les défauts logiciels est lié au coût de développement de ce genre d'applications. Pour cette raison, nous avons utilisé dans notre outil des composants open source et gratuits, et qui sont :

CKJM (Chidamber and Kemerer Java Metrics) : Ce logiciel nous a permis de construire les bases de données que nous avons utilisées lors de nos expérimentations. Cependant, nous avons adapté cette API afin de pouvoir calculer les métriques des programmes écrits en langage java. Ce logiciel, combiné avec **WEKA**, représentent le cœur de notre outil. Pour des raisons de portabilité, nous avons choisi d'utiliser le **langage XML** pour stocker à la fois les métriques des logiciels étudiés, ainsi que les résultats produits par notre outil.

Concernant les méthodes d'apprentissage des classifieurs intégrés dans l'outil proposé, nous avons retenu les trois meilleurs algorithmes qui nous ont permis d'obtenir les meilleurs résultats pendant notre étude expérimentale, et qui sont respectivement RF, NB et MLP. Nous avons choisi d'appliquer le principe de vote majoritaire entre ces trois classifieurs pour classer une nouvelle entité par notre outil.

En ce qui concerne l'interface graphique de notre outil, elle est simple et conviviale et intègre toutes les fonctions dont un développeur a besoin pour intégrer la prédiction de défauts dans le cycle de vie d'un logiciel développé ; plus précisément, avant la phase de test, et lors de la factorisation dans la phase de maintenance. Notre outil fournit deux modes d'utilisation : *normale* et *avancé*. Le premier mode est assez simple. Il faut juste importer le dossier qui contient le byte code (.class) de l'application que l'utilisateur veut analyser, et la base d'apprentissage sera directement intégrée (aucune connaissance supplémentaire n'est requise). Notons que la base d'apprentissage par défaut est Lucene 2.4.

Le deuxième mode d'utilisation (Figure 2) est aussi simple que le premier, mais donne l'avantage à l'utilisateur d'utiliser sa propre base d'apprentissage. Le fichier doit être de type arff. Bien entendu, une certaine connaissance est requise pour manipuler ce genre de fichiers; ce qui facilite l'utilisation du plug-in Weka lors de l'apprentissage. En appuyant sur le bouton point d'exclamation, une aide est fournie pour permettre de bien écrire l'en-tête de fichier arff manipulé pour éviter les problèmes de classification.

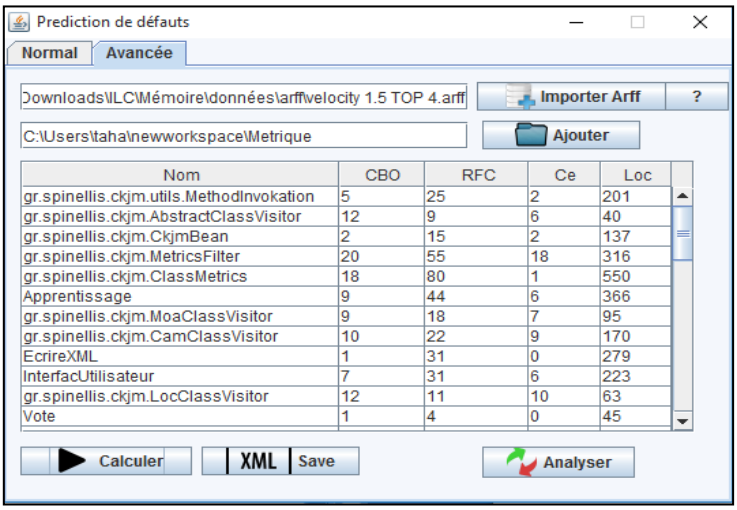


FIG. 2- Mode avancé de l'outil d'aide proposé

Pour obtenir les résultats de l'analyse, c'est à dire la classification (Figure 3), il suffit juste d'appuyer sur le bouton *Analyser* dans n'importe quel mode d'utilisation. Rappelons que le terme *fault prone* désigne le fait que cette classe contienne des défauts ou pas (*true* pour oui, *false* pour non). Cette interface permet de sauvegarder les résultats soit en format XML, soit en format de fichier arff pour une future utilisation.

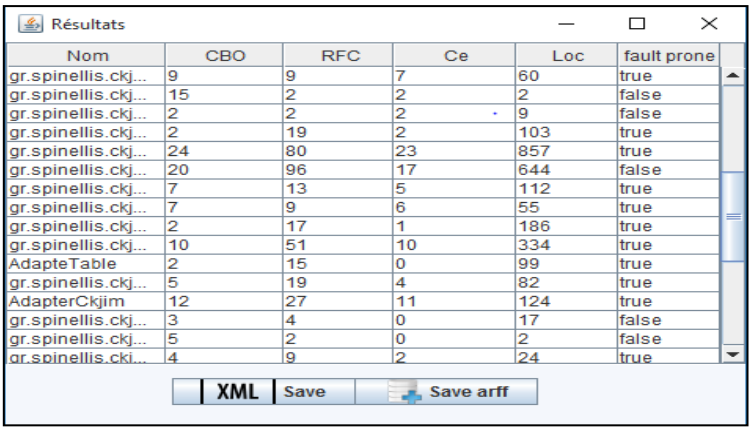


FIG. 3- Résultats d'analyse fournis par l'outil proposé

6 Conclusion et perspectives

Notre étude se place dans le cadre de la prédiction de défauts logiciels qui fait généralement intervenir de méthodes d'apprentissage automatique afin de classer les entités logicielles (méthodes, classes) d'un système comme étant sujettes ou pas à contenir des défauts (fault prone ou not fault prone). Pour y parvenir, il faut utiliser des métriques logicielles calculées à partir d'une version antérieure du système (prédiction intra-projet) ou d'une autre application (prédiction inter-projets).

Dans cet article, nous proposons principalement un outil d'aide à la prédiction de défauts logiciels inter-projets. Notre proposition se base sur l'étude expérimentale que nous avons menée afin de construire et d'évaluer des classifieurs pour la prédiction de défauts logiciels inter-projets. Nos classifieurs utilisent six algorithmes d'apprentissage automatique : les systèmes immunitaires AIRS1 et AIRS2, le perceptron multicouches, les arbres de décision, random forest et naïve bayes. L'évaluation a été effectuée sur cinq bases de données différentes issues du référentiel de données PROMISE (Menzies et al, 2016) spécialisé dans les données liées au domaine de génie logiciel.

Les résultats que nous avons obtenus, suite à notre étude, nous ont permis de confirmer la faisabilité de la prédiction inter-projets mais aussi sa difficulté relative aux choix des métriques et des méthodes d'apprentissage pour la construction des modèles de prédiction.

Concernant notre outil, il permet d'aider à mieux exploiter les ressources disponibles, surtout pendant la phase de test, pour améliorer la qualité des logiciels développés et d'en réduire les coûts.

Actuellement, cet outil ne fonctionne que sur des programmes écrits en langage java, et présente encore quelques problèmes de performances. Nous travaillons sur une nouvelle version pour prendre en compte plusieurs améliorations:

- Collecter des données dans des bases qui contiennent des données assez équilibrées entre les deux classes *fault prone* et *not fault prone*, comme la base Lucene 2.4, par exemple.
- Choisir un algorithme d'apprentissage assez stable comme RF et MLP (Kaur et Kaur, 2015)
- Etendre l'approche utilisée afin de prendre en compte des programmes écrits en des langages autres que Java.
- Augmenter la prédictibilité des classifieurs intégrés dans l'outil en remplaçant, au fur est à mesure, les données de la base d'apprentissage par des données issues d'autres réalisations.

Références

Abaei, G., et Selamat, A. (2013) A survey on software fault detection based on different prediction approaches. *Vietnam Journal of Computer Science*, 1: 79–95.

Bansiya, J., Davis, C.G. (2002) A hierarchical model for object-oriented design quality assessment. *IEEE Transactions on Software Engineering*, 28: 4–17.

Beecham, S., Hall, T., Bowes, D., Gray, D., Counsell, S., Black, S., (2010) A Systematic Review of Fault Prediction approaches used in Software Engineering. Technical Report, University of Limerick, Ireland.

Brownlee, J. (2005) Artificial immune recognition system (AIRS) a review and analysis, Technical report No. 1-02, Swinburne University, Australia.

Catal, C., Diri, B. (2007) Software Fault Prediction with Object-Oriented Metrics Based Artificial Immune Recognition System, LNCS 4589, pp. 300–314.

Catal, C., Diri, B. (2009). Investigating the effect of dataset size, metrics sets, and feature selection techniques on software fault prediction problem. *Information Sciences*, 179: 1040–1058.

Catal, C. (2011) Software fault prediction: A literature review and current trends, *Expert Systems with Applications*, 38: 4626–4636.

Catal, C., Sevim, U., Diri, B. (2011) Practical development of an Eclipse-based software fault prediction tool using Naive Bayes algorithm. *Expert Systems with Applications*, 38: 2347–2353.

Chidamber, S.R, Kemerer, C.F. (1994) A metrics suite for object-oriented design, *IEEE Transactions on Software Engineering*, 20: 476–493.

Erturk , E., Sezer, E. A. (2015). A comparison of some soft computing methods for software fault prediction, *Expert Systems with Applications*, 42: 1872–1879.

Hall, T., Beecham, S., Bowes, D., Gray, D., Counsell, S. (2011) A Systematic Review of Fault Prediction Performance in Software Engineering , *IEEE Transactions on Software Engineering*, 38: 1276-1304.

He, Z., Shu, F., Yang, Y., Li, M., & Wang, Q. (2012). An investigation on the feasibility of cross-project defect prediction. *Automated Software Engineering*, 19: 167-199.

He, P., Li, B., Liu, X., Chen, J., & Ma, Y. (2015). An empirical study on software defect prediction with a simplified metric set. *Information and Software Technology*, 59: 170–190.

Henderson-Sellers, B. (1996). *Object-Oriented Metrics, measures of Complexity*. Prentice Hall.

Jiang, Y., Cukic, B., Ma, Y. (2008) Techniques for evaluating fault prediction models, *Empirical Software Engineering*, 13: 561–595.

Jureczko, M., Madeyski, L. (2010) Towards identifying software project clusters with regard to defect prediction. Proceedings of the 6th International Conference on Predictive Models in Software Engineering - PROMISE '10, 1.

Jureczko, M., Spinellis, D. (2010) Using Object-Oriented Design Metrics to Predict Software Defects. Models and Methods of System Dependability. 69–81.

Kaur A., Kaur, K. (2015) An Empirical Study of Robustness and Stability of Machine Learning Classifiers in Software Defect Prediction. In: El-Alfy ES., Thampi S., Takagi H., Piramuthu S., Hanne T. (eds) Advances in Intelligent Informatics. Advances in Intelligent Systems and Computing, Vol. 320, pp. 621–631.

Lessmann, S., Member, S., Baesens, B., Mues, C., Pietsch, S. (2008). Benchmarking Classification Models for Software Defect Prediction: A Proposed Framework and Novel Findings, *IEEE Transactions on Software Engineering*. 34: 485–496.

Martin, R. (1996). OO Design Quality Metrics. *Quality Engineering*, 8: 537–542.

Malhotra, R. (2015). A systematic review of machine learning techniques for software fault prediction. *Applied Soft Computing Journal*, 27: 504–518.

Menzies, T., Krishna, R., Pryor, D. (2016). The Promise Repository of Empirical Software Engineering Data; North Carolina State University, Department of Computer Science, http://openscience.us/repo.

McCabe, J. (1976). A complexity measure, *IEEE Transactions on Software Engineering*, SE-2: 308–320.

Ostrand, T. J., Weyuker, E. J. (2010). Software fault prediction tool. Proceedings of the 19th International Symposium on Software Testing and Analysis - ISSTA '10.

Radjenović, D., Heričko, M., Torkar, R., & Živkovič, A. (2013). Software fault prediction metrics: A systematic literature review. *Information and Software Technology*, 55: 1397–1418.

Rana, R., Staron, M., Hansson, J., Nilsson, M., Meding, W. (2014). A Framework for Adoption of Machine Learning in Industry for Software Defect Prediction, 9th International Conference on Software Engineering and Applications (ICSOFT-EA)

Tang, M. H., Kao, M. H., Chen, M. H. (1999).An empirical study on object-oriented metrics, Proc. 6th Int. Softw. Metrics Symp, 8: 242–249.

Watkins, A. (2001) *AIRS: A Resource Limited Artificial Immune Classifier*, M.S Thesis, Mississippi State University, 2001.

Watkins, A., Timmis, J., Boggess,L.(2004).Artificial immune recognition system (AIRS): an immune inspired supervised learning algorithm, *Journal of Genetic Programming and Evolvable Machines*, 5: 291-317.

Zhou, Y., B. Xu, et H. Leung. (2010). On the ability of complexity metrics to predict fault-prone classes in object-oriented systems. *Journal of Systems and Software*, 83: 660-674.

Zimmermann, T., Nagappan, N., Gall, H., Giger, E., Murphy, B.(2009) Cross-project defect prediction , Proceedings of the 7th joint meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering (ESEC/FSE 2009), Amsterdam, The Netherlands.

Summary

Data mining in software engineering is a fast growing field. It includes the prediction of software defaults, which mostly consists of applying machine-learning algorithms on software metrics to classify system entities as fault prone, or not.

Due to the lack of automatic tools for the software default prediction, numerous existing models have not reached an extended applicability. Automatic tools would guild project managers to target entities prone to defects, manage available resources efficiently and improve the testing phase.

In this context, our objective is to propose a tool to aid in the prediction of software defaults, after the construction of classifiers using six machine-learning methods with five databases from the PROMISE repository.

Introducing Big Data into Digital Control Systems

Djilali DAHMANI*, Sidi Ahmed RAHAL**, Ghalem BELALEM***

*Departement of mathematics and computer science, USTO-MB University, Oran, Algeria.
ddahmani@yahoo.fr , djilali.dahmani@univ-usto.dz
**Departement of mathematics and computer science, USTO-MB University, Oran, Algeria.
rahalsa2001@yahoo.fr
***Departement of computer science, Oran University 1, Ahmed Ben Bella, Oran, Algeria.
ghalem1dz@yahoo.fr

Abstract. Many industrial companies are provided with digital control systems (DCS). These systems collect automatically huge real-time data from different electronic equipments and sensing detectors. While these data are being collected, specific applications explore and manage them in order to provide services to users like critical information, graph evolutions, real-time alarms, etc. Later, these great amounts of data are just stored as historical archives without any dealings. Many users hope to take advantage of these data and use them in other areas beyond their specific applications such as Data Mining. However, these huge archived data cannot be handled in a simple manner since the most of the DCS systems use relational databases. So, the huge data need adjustment before any processing. With the emergence of Big Data, many concepts come into sight such as NoSQL. To deal with DCS data issue, we propose an approach to migrate historical DCS data from relational to an appropriate Big Data NoSQL system, and use a distributed environment containing many nodes. The processing and the storage of data are automatically performed on nodes by system sharding. As experimentation, we have used industrial data generated by an oil and gas DCS, and some data mining queries are done and analyzed to validate the performance.

1. Introduction

Industrial companies and manufacturers are increasingly equipped with digital control systems (DCS) that generate very large amounts of real time data. These data are used by specific applications to provide real time critical information, in-time graphs of evolution, real-time alarms, etc. Later, these data are stored for historical archives, and in many cases they are often deleted later. However, lots of users are increasingly interested in these historical data in order to use them in many business processes, and especially in data mining area, like extracting useful knowledge, providing early feedback means, improving future requests, etc. Nevertheless, this huge accumulated data needs progressively capacity and power to support processing and storage. Every machine or server arrives at its limits to support the storage and processing of huge data, whatever its physical capacity CPU, memory,

or disk. Also, machines' upgrade or extension cannot be considered as a permanent solution. So, distributed platforms contained a lot of servers (nodes), such as clusters, grids can be efficiently used to deal with this issue. Unfortunately for DCS databases, since most of them are relational, they cannot be directly deployed and take benefit from distributed platforms. This limit comes from the relational model which it is not scalable in such environments.

Nowadays, with the emergence of Big Data, many new systems like NoSQL and New SQL come into sight and occupy good places in different areas. These systems propose to deal with structured and unstructured huge data in distributed platforms as they give guarantee for automatic distribution of big data in such platforms. So, NoSQL systems can be used as alternatives to the relational model (Leavitt, 2010). The objective is to take benefit from the resources of distributed platforms by using scalability, and improve data management. These Big Data systems are in favor of large data volume with various structured and unstructured types, and they are very efficient in distribution storage and processing.

In this article, we propose an approach to migrate relational data of DCS to a NoSQL system, and deployed result data in a distributed multi-nodes platform. The gain obtained is the elasticity and flexibility given from this environment, since the distribution of storage and processing between nodes is provided automatically by the native NoSQL Sharding. As a result, the resources will be allocated and adapted as needs for the new migrated data, and we give best performance and improvement for both the storage and data processing time. So, users and administrators of DCS can profit of the migrated NoSQL data and carry out their queries without worry.

This article has five sections. The first section is an introduction, and in the second, we present relational data issue on distributed environment and some related work. We give in the third section an approach to use distributed NoSQL system and data migration. In the fourth section, we prepare DCS data for a company (SH, 2015) on an experimental platform. In the fifth section, we carry out some interesting queries to test and compare performance. A conclusion closes this paper at the last section.

2. Distributed data issue and related work

In this section, we present the constraints of the relational model in distributed world, the NoSQL technology, and some related works given to deal with this issue.

A distributed environment is a set of physical machines (nodes) which participate jointly to accomplish parallel processing and storage. All resources of the nodes (CPU, memory, disks) can be shared or not. The number of nodes can differ from a few to thousands of nodes. So, we can find simple clusters with few nodes which typically share disks (like SAN or NAS), or more complex like grids that contain hundreds or thousands of nodes where resources are often not shared.

The main part of the DCS databases are based on the relational model which is build on the concept of table (relation between data) and operations of set-algebra. This model is suitable for transactional needs due to the ACID properties (Atomicity, Consistency, Isolation, and Durability) (Sharma et Dave, 2012), and it works very well in a single-node environment. Conversely, relational data cannot be well distributed and deployed in a distributed multi-nodes environment. The ACID properties become constraints for the model and then prevent data from being distributed effectively between nodes (Degroodt, 2011). Note also that ACID constraints, although they ensure consistency, may become in some cases a block-

ing factor (Moniruzzaman et Hossain, 2013) (Salminen, 2012). For instance, an investigator in internet is often interested in having immediate response even if that response is not up-to-date. Consequently, new systems need to be created in order to dynamically distribute and manage data between nodes with more efficiency and usefulness. New systems for Big Data have been emerged like NoSQL and NewSQL (Grolinger et al. 2013).

NoSQL (Not Only SQL) is a set of concepts that allows quick and efficient data processing with emphasis on performance, reliability and agility (Creary et Kelly, 2014). NoSQL offer new architectures in order to align with current technological developments relating to Big Data and distributed environments. NoSQL is distinguished from relational by the absence of pattern and structure imposed, and the lack of ACID properties in favor of performance. The main characteristics of a NoSQL system are (Creary et Kelly, 2014): more than rows in tables, free of joins, schema-free, working on multiple processors, use shared-nothing commodity, support linear scalability, and innovative. NoSQL have begun to emerge since 2009 and continues to develop now (Salminen, 2012). Their reason is not to substitute the relational model, but to give an alternative to the new needs related to Big Data (Grolinger et al. 2013).

The CAP theorem invented by Brewer (2000), states that a database of a distributed system can guarantee at the same time only two of the following 3 constraints:
Consistency, Availability , and Tolerance to Partitioning [(Creary et Kelly, 2014). So, we have 3 types: CA (consistency/ availability), AP (availability/ partitioning) and CP (consistency/ partitioning). Note that this theorem was criticized by some authors Wade et al., 2013) (Abadi, 2012).

There are four main NoSQL classes (Moniruzzaman et Hossain, 2013):

- Key-value store: Redis, Riak, Memcached, Oracle NoSQL, Voldemort, etc.
- Graph stores: Neo4j, Orient DB, HyperGraphDB, FlockDB, etc.
- Column family databases: Cassandra, HBase, Hypertable, Accumulo, etc.
- Document store: MongoDB, CouchDB, Couchbase, RavenDB, etc.

Many works have explained the data issue of the relational model on distributed platform. Some authors refer to this issue as database elasticity issue (Degroodt, 2011). As solution, we can distinguish two proposed classes:

Some authors suggest the possibility to expand the relational model to support the elasticity of databases on the distributed environment. It is in this sense that some providers' relational databases (Oracle Clusterware, 2016) (MySql Cluster, 2016) have expanded their systems capabilities by integrating new features to support at a certain degree the management of their data in multi-nodes environments. We cite as an example Oracle Real Application Cluster RAC (Degroodt, 2011) and MySQL Cluster (2016). These opportunities can provide load balancing and failover, but with a reduced number of nodes because they need shared space between nodes. More and more nodes are added, data management becomes more complex and inefficient than a given limit. However, in a distributed environment where the number of nodes is very large, these proprietary solutions are unusable. Briefly, these solutions are suitable for clusters with few nodes.

In contrast, other authors propose to abandon the relational model in favor of NoSQL for distributed environments (Creary et Kelly, 2014) (N. Leavitt, 2010). But, it should be noted that NoSQL is still recent and there are no standards that define a typical architecture for a particular case of data. Only a detailed study of each case allows picking a specific NoSQL type.

For our case, to deal with DCS data issue, we have big data that need large storage and high-speed processing which bypass any node capacity. Thus, do we call for a distributed environment or it is not beneficial? What type of model expanded relational or NoSQL is favorable for DCS? In case, which NoSQL is best fitting? How to migrate data to this new system? And how much is the gain on performance like time, storage and distribute processing compared to single node, etc. We have to perform experiments and evaluate performance to bear out any suggestion.

3. Our proposal

As proposal, we suggest migrating DCS data to a suitable NoSQL system. We choose a NoSQL system according to specific criteria which are predominantly based on the fitting of NoSQL class to DCS data. Also, since DCS data must be consistent, the type of the chosen NoSQL system must be CP (Consistency/ Partitioning). Migrated NoSQL DCS data will be distributed on a platform consisted of many nodes. The Sharding is a NoSQL property which distributes and allocates dynamically resources between nodes and adapt them according to data needs. Fig. 1 represents a schema for our proposed migration. This can greatly improve the performance on either the data storage or the processing time for ad hoc queries.

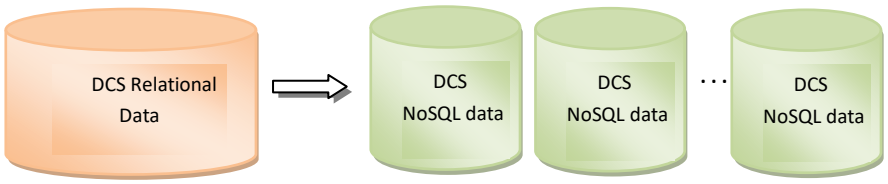


FIG. 1. *Proposed approach*

As experimentation, we use the historical DCS database of an oil and gas company (SH,2015). These data include all information captured from sensing electronic detectors, like real-time quantities of different equipments (temperature, pressure, quantities flux of fluid, flow gas, electrical charges, vapor of turbines, electricity of turbo-generators, etc). We use this database to perform useful data mining queries.

The original DCS database is Oracle version 10g release 2. So, we have to choose a suitable type of NoSQL system. There are currently several NoSQL systems like MongoDB, Cassandra, Redis, HBase, Memcached, Neo4j, CouchDB, Riak, etc. In the following, we select a NoSQL system depending on the following various useful.

- Volume of data: The target NoSQL system must support and manage very large data, like Cassandra, Hbase, MongoDB, Neo4j, CouchDB, etc.
- Best ranking: The target NoSQL system must have a best rank among popular existing systems. Many international agencies publish yearly classification for different systems. For example, the top 10 popular database systems published by Solid IT (2014), which shows Oracle as the most popular relational system, and MongoDB and Cassandra for NoSQL, as 6th and 10th position respectively.
- Consistent system (CP): Since DCS database is relational and then consistency is always ensured, so CP systems (HBase, MongoDB, Redis, etc) are more suitable.

- Suitability to DCS nature: DCS data are organized around well-defined tables such as: equipments, products, utilities, etc. These data are hierarchically structured between items and their elements. Thus, document store systems like MongoDB, CouchDB, Couchbase are adequate for this case.

As result, we intuitively select MongoDB as a suitable and efficient NoSQL system for DCS data. MongoDB implements a centralized distributed architecture of multiple nodes, and supports data replication via a master-slave model. It uses BSON (Binary JSON, 2012) objects, an optimized derivative of JSON. MongoDB installation and configuration are explained in the MongoDB official document MongoDB Release 2.6.4 (2014).

Finally, we need an approach to migrate our Oracle DCS data to MongoDB. There are many migration's approachs in literature, some of them propose to go through XML files, others use Text files, etc. For instance, we can find *Pelica Data Migraton* (Pelica, 2016) and *Spviewer Software* (Spviewer, 2016). However, these tools are not free or open source, since they are destined for commercial using.

For our migration context, we propose an approach using JSON files since JSON types are directly manipulated in MongoDB (Dahmani et al., 2016). Our approach consists of two steps as shown in Fig. 2.

1. Generate JSON data from Oracle database using any tool like OraMongo.
2. Load the JSON data into MongoDB database using the integrated tool mongoimp presented in mongodb-enterprise-tools package (MongoDB, 2014). Note that while loading JSON objects into MongoDB, they are automatically converted to BSON.

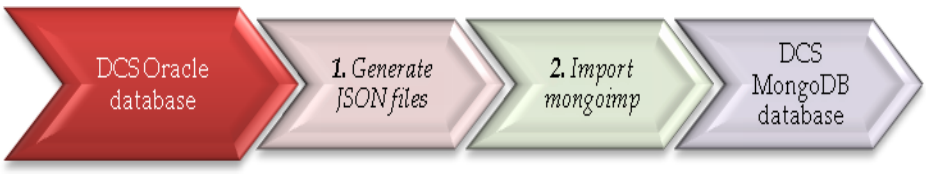


FIG. 2. *APPROACH FOR MIGRATING DCS DATA FROM ORACLE TO MONGODB*

As result, Oracle tables become MongoDB collections, and their rows become documents.

4. Preparation of target DCS data

To prepare DCS data, we start implementing a distributed platform, and we perform migration data from Oracle 10gR2 to MongoDB version 2.6[1]. Some experiments and their results will be done in the next section. Firstly, we configure a distributed platform managed directly by MongoDB. This MongoDB platform is composed of nodes called Shards, and it does not require many resources like other distributed environments (Cloud computing IaaS, Hadoop, etc). MongDB has three processes: (Heinrich et Stettler, 2012)

- Configuration server process: stores metadata of each shard.
- Mongos process: redirects users' requests to the appropriate shards and groups the results before sending them back to these users.

[1] MongoDB version 2.6 appeared in Sep. 2014

- Mongod process: hosts and manages data in the Shard.

As shown Fig.3, we use a MongoDB platform composed of 11 nodes:

- One *Mongo server* that run *Mongos* process which manages and distributes data.
- One *configuration* server used by *Mongos* to manage shards and replications.
- 9 shards, each one runs *mongod* and hold a data partition distributed by Mongos.

After configuration the distributed platform, we install a MongoDB distributed database according to the steps listed in the MongDB Sharding procedure (MongoDB, 2014).

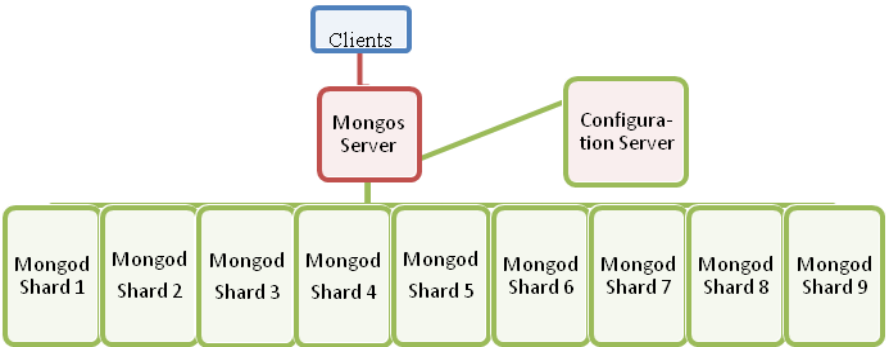


FIG. 3. *MONGODB DISTRIBUTED PLATEFORM USED.*

As the platform is ready, we migrate data from Oracle to MongoDB by using the two steps of our approach (refer to Fig. 2).

4.1. Step 1: Generate JSON data from the Oracle database.

We use a scripting tool OraToJSON to generate JSON files from Oracle relational tables. Once connected to DCS database, we can select tables and generate their JSON data. An extract of generated JSON file is shown in Fig. 4.

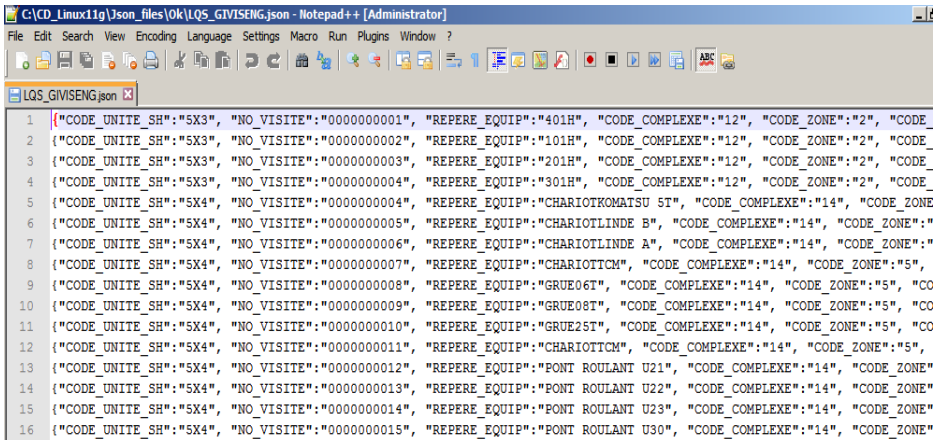


FIG. 4. AN *EXTRACT FROM A JSON FILE GENERATED BY SCRIPTING TOOL*

4.2. Step 2: Load the JSON data in MongoDB database.

We use the mongoimp tool to load one JSON file or all JSON files of a directory into the MongoDB database. Fig.5 shows two scripts load and loadall that can be used to call mongoimp. Note that loading can be done from any shard. On MongoDB database, Oracle tables become collections, and just in case a manual work are done on some collections to adapt hierarchical relations between linked data.

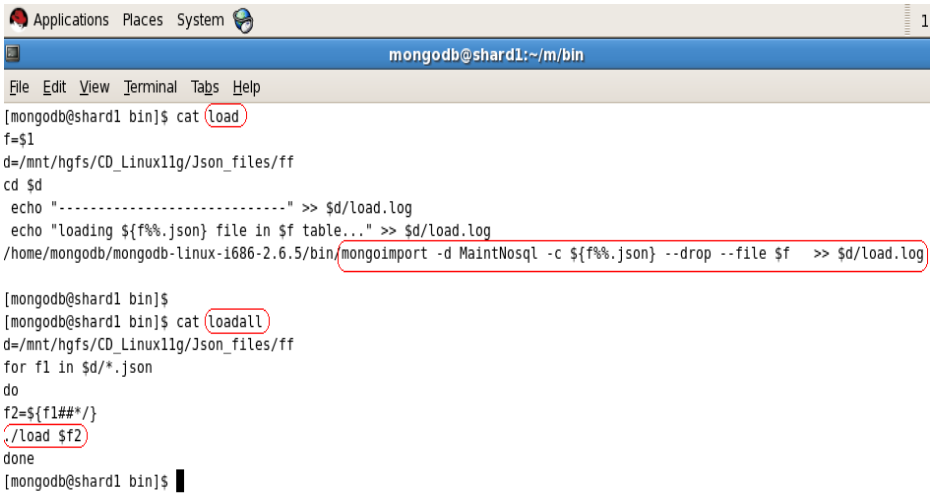


FIG. 5. *SCRIPTS USED TO LOAD JSON FILES IN MONGODB DATABASE.*

Once the migration is complete, NoSQL DCS data are ready for experimentation. In practice, these data are frequently used in data mining process based on association rules (Agrawal, 1994). This technique allows detecting links between data by searching frequent itemsets and rules. It is based on two steps:

- Search frequent patterns (itemsets).
- Generate association rules from these itemsets.

The first step is very costly in time and space, because it navigates several times the entire context (database) to find all potential frequent itemsets. The number of these itemsets is exponential, i.e., for an itemset I with size m, the number is 2^{m-1} (Agrawal, 1994).

The basic well-known algorithm for association rules is Apriori (Agrawal et Srikant, 1995). This algorithm has been implemented to search all association rules in DCS database.

Searching queries concerns only itemsets that surpass a predetermined threshold called the minimum support (Agrawal et Srikant, 1995). A rule X→Y verifies a factor support S if and only if at least S% of transactions (records) in the database that contain X and Y.

In our experiments, test queries are used to look for frequent itemsets with variation of minimum support for each test.

5. Results and Discussion

The new DCS MongoDB database is now ready; our goal is to prove performance acquired by using elasticity given by distributed NoSQL data. This elasticity signifies the ability to distribute data and queries processing on multiple nodes, contrary to the rigidity of relational data. So theoretically, good performance is expected.

In the following, we use the same data mining queries for both the distributed MongoDB data and the Oracle mono-node data. We will compare and analyze the results of the storage, run time, and shared query processing by varying each time the number of shares from 2 to 9. As experiments, we use collections which containing real-time data used frequently in looking for frequent itemsets. For data storage, we can see in Fig.6 a graph showing that MongoDB has shared automatically DCS data on different shards, but for Oracle only one node supports data storage, the other nodes remain idle.

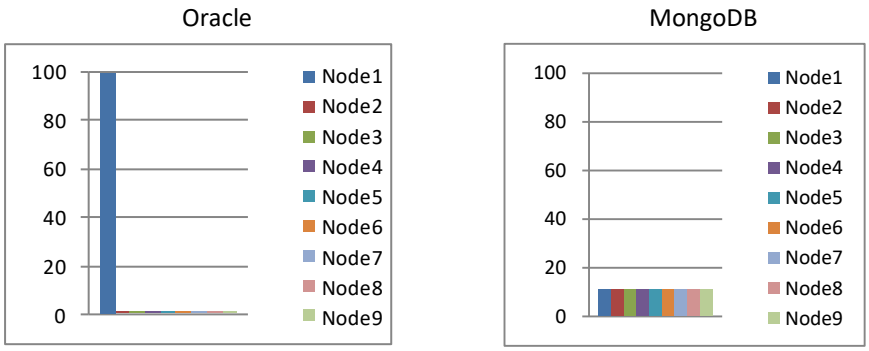


FIG. 6. *COMPARE DCS DATA STORAGE REPARTITION BETWEEN MONGODB AND ORACLE*

For query processing, frequent itemsets queries are executed by just a single node for Oracle, but for MongoDB, thanks to sharding, queries are shared and all nodes are involved in the query processing. Fig. 7 shows a comparison of the query processing repartition between Oracle and MongoDB, for 9 nodes.

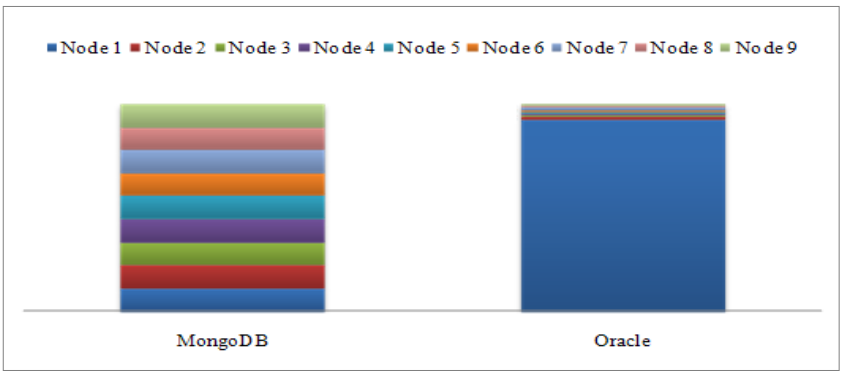


FIG. 7. *COMPARISON OF QUERIES REPARTITION BETWEEN MONGODB AND ORACLE.*

Finally, for the query run time, we compare the time of the frequent itemssets search algorithm between MongoDB and Oracle. Table 1 shows the run time results using 0.2 and 0.75 as minimum support. Note that these values are calculated as medians after four tries for each test, and the results are with incertitude of ± 1second.

Node	Min Support	Run Time (seconds)	
		Oracle	MongoDB
1	0.2	150	203
	0.75	42	78
2	0.2	150	213
	0.75	42	84.5
3	0.2	150	161
	0.75	42	53
4	0.2	150	150
	0.75	42	39.5
5	0.2	150	137
	0.75	42	31.5
6	0.2	150	123
	0.75	42	29
7	0.2	150	113
	0.75	42	25.5
8	0.2	150	95
	0.75	42	22.5
9	0.2	150	89
	0.75	42	19

TAB 1-*RESULTS OF DATA MINING ALGORITHM RUN TIME ON DCS DATA BETWEEN MONGODB & ORACLE*

The graphs of figures Fig. 8 shows the run time evolution by adding nodes for minimum support values 0.2 and 0.75 respectively. We note that at the beginning and with a single node, the run time is favorable for Oracle, a more deterioration with 2 nodes for MongoDB, however with adding more nodes, the run time becomes more favorable for MongoDB.

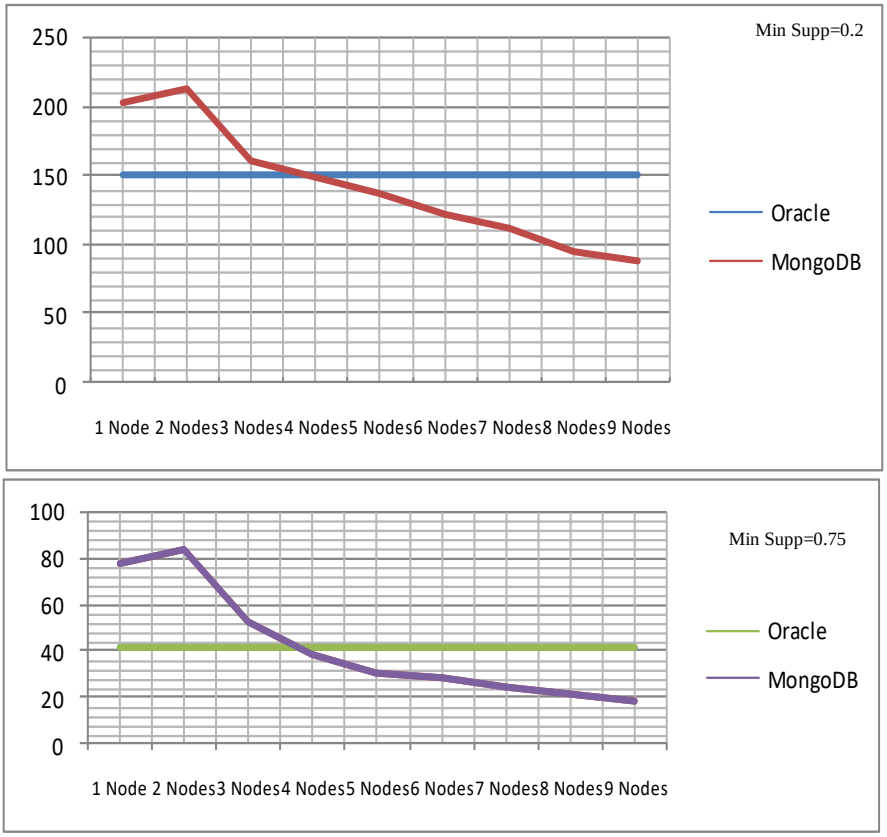


FIG. 8. *COMPARISON THE RUN TIME BETWEEN DCS DATA IN MONGODB AND ORACLE*.

We can consolidate the gains obtained in execution time of MongoDB compared to Oracle for the three minimum support values. The result is displayed in the Fig. 9.

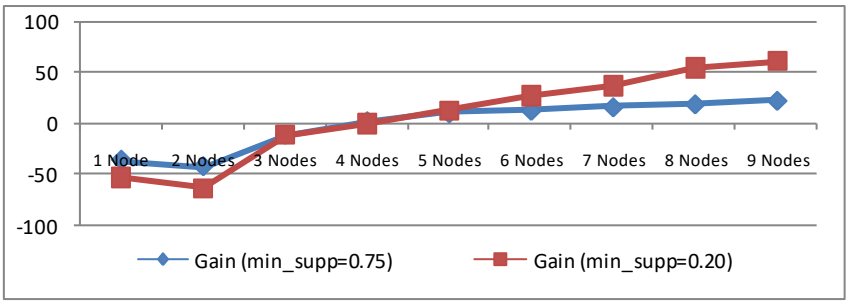


FIG. 9. GAIN *IN RUN TIME GIVE BY MOVING DCS DATA FROM ORACLE TO MONGODB*

6. Conclusion

In this article, we have presented a proposal to improve performance of processing and storage for the digital control systems (DCS) that generate a lot of real-time data. After being used, historical DCS data are often archived and put out of action. Data Mining operations become very hard in time and space on these huge data since they are relational and mono-node. So, by introducing big data, our approach suggests migrating relational DCS database to an adequate NoSQL and use a distributed platform contained many nodes. Processing and storage are shared between nodes to deal with all Data Mining operations on DCS data. Also, we have select MongoDB as NoSQL system for DCS data, and we have proposed a method to migrate data from relational to MongoDB.

As experiment, after configuring a distributed environment with a number of nodes, we have installed MongoDB and migrated DCS data. The NoSQL sharding property allows repartition of storage and processing between all nodes. Some experiments Data Ming queries have been done in order to compare performance of results between multi-nodes NoSQL data and mono-node relational data.

Overall, the final results demonstrate an interesting improvement in run time, in addition of the data storage gained by joining all disk nodes.

Finally, in perspective we look forward to enlarge this experimentation in widespread platforms environments with many nodes like Hadoop platform. Also, since NoSQL is new-fashioned and still in development, we are hearing of potential new NoSQL systems to test them and find the most appropriate for DCS data.

References

Abadi D. J. (2012). *Consistency Tradeoffs in Modern Distributed Database System Design*, IEEE Computer Society, vol. 45, no. 2, pp. 37-42.

Agrawal R., Srikant R. (1995). *Mining Generalized Association Rule*, in Proceedings of the 21st International Conference on VLDB, San Francisco, CA, pp. 407–419.

Agrawal R., Srikant R., Methta M., Shafer J. (1994). *Fast Algorithms for Mining Association Rules in large databases*. In Proceedings of the 20th International Conference on Very Large Databases , Santiago de Chile, Chile, pp.478-499.

Brewer E. A. (2000). *Towards robust distributed systems*, (Invited Talk). Principles of Distributed Computing, Portland, Oregon, July.

BSON(2012). *Binary JSON*, available at http://bsonspec.org.

Creary D. M. and Kelly A. (2014). *Making Sense of NoSQL*. Edition: Manning Publications Co.

Degroodt N. (2011). *L'élasticité des bases de données sur le Cloud Computing*. Master thesis in Sciences computer, Free University of Bruxelles, p12-20.

Dahmani D, Rahal S., Belalem G. (2016). *Improving the Performance of Data Mining by Using Big Data in Cloud Environment"*. Journal of Information & Knowledge Management, Vol. 15, No. 4, World Scientific Publishing Co. ISSN: 0219-6492.

Grolinger K. et al. (2013). *Data management in cloud environments: NoSQL and NewSQL data stores*, Journal of Cloud Computing: Advances, Systems and Applications, a Springer Open Journal, pp. 1-22.

Heinrich L., Stettler C. (2012). *Bachelor Thesis*, HES, High school Management of Geneve (HEG-GE), IT management, Genève, pp. 21-24.

Leavitt N. (2010). *Will NoSQL Databases Live Up to Their Promise?* ISSN: 0018-9162.

MongoDB Documentation Project Release 2.6.4 (2014). Available at: http://docs.mongodb.org/manual/tutorial/

Moniruzzaman A.B.M and Hossain S. A. (2013). *NoSQL Database: New Era of Databases for Big data Analytics -Classification, Characteristics and Comparison,* International Journal of Database Theory and Application, Vol. 6, No. 4.

MySql Cluster (2016). Site https://www.mysql.com/products/cluster/.

Oracle Clusterware (2016). Web site
http://docs.oracle.com/cd/B28359_01/rac.111/b28255/intro.htmSalminen (2012). *Introduction to NoSQL*, NoSQL Seminar @ TUT.

Pelica (2016). Pelica Data Migration, http://www.techgene.com/it-solutions/data-migration.

Sharma V., Dave M. (2012). *SQL and NoSQL Databases*. International Journal of Advanced Research in Computer Science and Software Engineering, Vol.2, Issue 8, ISSN: 2277 128X. Research paper available: www.ijarcsse.com, page 2-8.

Solid IT (2014). *Classement NoSQL*, site: http://developpez.com.

Sonatrach, integrated oil and gas Company, http://www.sonatrach.com/en/

Spviewer (2016). Spviewer Software, http://www.spviewer.com.

Wade Diackl B., Ndiaye1 S. and Y. Slimani (2013). *CAP Theorem between Claims and Misunderstandings: What is to be sacrificed?*.International Journal of Advanced Science and Technology, Vol. 56, pp3-10.

Approche de sélection du processus métier à externaliser vers le cloud

Mouna Rekik[1], Khouloud Boukadi[1], Hanêne Ben Abdallah[2]
1 Université de Sfax, Tunisie
2 Université King Abdulaziz, Jeddah, KSA

Résumé.
Les entreprises désirant s'inscrire dans une démarche d'externalisation de leurs processus métier vers le cloud doivent identifier judicieusement leurs attentes de ce choix et s'assurer de ses risques potentiels. En effet, elles doivent orienter les efforts de l'externalisation vers les processus les plus prioritaires qui méritent entre autres, une amélioration urgente de leurs performances et une réduction éminente de leurs coûts. Ceci leur assurera un gain en termes de temps et de coût. Par ailleurs, le risque et les conséquences de l'échec d'une externalisation peuvent être fatals pour l'entreprise.
Ainsi, nous proposons dans ce papier une approche d'aide à la décision qui consiste à la sélection des processus à externaliser vers le cloud. Elle part de l'extension des modèles de processus métier par les facteurs d'externalisation jusqu'à l'application de la méthode AHP (Analytic Hierarchy Process) comme méthode d'aide à la décision multicritère.

1 Introduction

Le paradigme d'externalisation des processus métier (souvent appelé en anglais Business Process Outsourcing (BPO)) n'est pas récent. Néanmoins, l'émergence de nouveaux modèles de prestation, comme l'environnement cloud, ainsi que l'exigence accrue des entreprises ont engendré des transformations radicales dans le domaine du BPO. Les entreprises désirant s'inscrire dans une démarche d'externalisation de leurs processus métier doivent identifier judicieusement leurs attentes de ce choix et s'assurer de ses risques potentiels. La réduction des coûts et la concentration sur le cœur du métier sont des exemples de facteurs qui influeront sur les choix des entreprises en matière d'externalisation de processus métier. Cependant, une question primordiale se pose : quels sont les facteurs qui guident le choix des processus pour une externalisation vers le cloud ? Certes, plusieurs travaux ont largement abordé les facteurs d'externalisation des processus métier, mais quelques-uns seulement ont considéré les facteurs liés à l'environnement cloud comme cible d'externalisation des processus.

L'environnement cloud reste un domaine relativement récent, où les clients souhaitent s'assurer que les données manipulées par les activités du processus sont sécurisées. La confiance des clients est naturellement plus importante lorsque les données sont traitées, stockées et contrôlées au sein de l'entreprise puisque l'externalisation du traitement ou encore du stockage de ces données dans le cloud peuvent s'accompagner de risques de sécurité. L'examen des risques liés aux données des activités permet d'évaluer la pertinence de l'externalisation.

Malgré l'intérêt accordé à l'externalisation des processus métier et l'expansion continue du cloud, la majorité des travaux de recherche comme Afshari et al. (2010) se sont intéressés à l'externalisation des applications monolithiques.
La sélection des applications à externaliser se fait manuellement en se référant aux avis des experts. Nous considérons la sélection des processus à externaliser comme un problème

de décision dont les critères sont les facteurs d'externalisation et le résultat concerne le processus qui mérite le plus d'être externalisé.

Ainsi, nous proposons dans ce papier une approche de sélection des processus métier qui méritent d'être externalisés et ce, conformément à un ensemble de facteurs d'externalisation liés à la fois au processus et au cloud.
L'approche proposée part de l'extension des modèles de processus métier par les facteurs d'externalisation jusqu'à l'application de la méthode de décision multicritère AHP (Analytic Hierarchy Process) Saaty (1987).

Nous exposerons dans la Section 2 les détails relatifs aux facteurs d'externalisation que nous retenons et les techniques et fonctions nécessaires pour les quantifier. Par la suite, nous proposerons dans la Section 3 l'extension du modèle du processus métier par ces facteurs. La Section 4 détaillera l'élaboration de la méthode AHP pour la sélection du processus à externaliser. La concrétisation de l'approche à travers une étude de cas industriel sera présentée dans la Section 5. L'évaluation est exposée dans la Section 6.

2 Facteurs d'externalisation retenus

La prise de décision d'externalisation des processus métier en général, se base essentiellement sur différents facteurs que nous adoptons en considérant l'environnement cible qui est le cloud. Les différents facteurs adoptés sont :

2.1 Réduction des coûts

La réduction des coûts est parmi les facteurs majeurs de la décision d'externalisation des processus métier. En effet, l'entreprise est motivée par le fait que l'environnement d'externalisation cible est capable de réaliser l'exécution de ses processus avec le moindre coût. Ainsi, plus les dépenses relatives à un processus métier sont élevées, plus ce dernier mérite le plus d'être externalisé.

Le coût de réalisation des processus métier dans l'entreprise dépend des coûts de la maintenance logicielle et matérielle et des dépenses mensuelles associées. En externalisant un processus vers le cloud et plus précisément vers le modèle IaaS, nous estimons réduire les coûts relatifs aux dépenses matérielles nécessaires pour l'exécution du processus. Les experts IT de l'entreprise sont censés fournir les dépenses monétaires relatives à chaque activité du processus métier.

2.2 Concentration sur l'importance stratégique

L'importance stratégique du processus métier influence la décision de son externalisation. Généralement, les entreprises externalisent les processus métier qui n'ont pas d'importance stratégique pour elles, afin de se focaliser sur ceux considérés comme cœur métier (ou d'importance stratégique élevée). Dans ce contexte, l'expert métier de l'entreprise doit préciser l'importance stratégique de chaque processus métier impliqué dans la décision d'externalisation : cœur métier, critique non cœur, et non critique non cœur.

Un processus qui est non critique non cœur mérite le plus d'être externalisé vu qu'il ne présente aucune importance stratégique pour l'entreprise.

2.3 Diminution des risques

Avant de présenter les risques associés à l'externalisation des processus métier vers le cloud, une définition du terme risque est essentielle. Les guides ISO/IEC ISO (2008) le définissent comme étant : "*une combinaison de la probabilité d'un événement et de ses conséquences* ".

Généralement, pour pouvoir mesurer, évaluer et éviter les risques, une entreprise doit élaborer un ensemble d'étapes qui constituent le processus de gestion des risques. Ce processus inclut :

- L'élaboration d'une étude préalable relative à l'entreprise y compris la définition des objectifs et du contexte du processus de la gestion des risques et la collecte des informations utiles ;
- L'évaluation des risques doit être réalisée pour : (i) identifier les sources des risques et les domaines d'impact, (ii) analyser les risques via l'estimation de leurs conséquences et de la probabilité que ces conséquences peuvent réellement se produire et (iii) évaluer quels sont les risques qui nécessitent d'être traités et leur niveau de priorité ;
- Le traitement du risque via la sélection des options du traitement convenable (par exemple, éviter le risque en décidant de ne pas continuer ou de ne pas commencer une certaine activité considérée comme source de risques) et la définition des plans de traitement ;
- L'acceptation du plan de traitement du risque par les experts et dirigeants de l'organisation.

Pour éviter les risques de l'externalisation des processus métier vers le cloud, nous suivons le processus standard auparavant discuté pour la gestion des risques. En effet, nous commençons d'abord par une étude préalable qui permet d'identifier les processus métier impliqués dans la décision de l'externalisation et qui sont inclus dans le processus de gestion de risques. Par la suite, nous procédons à l'évaluation et au calcul de l'importance d'un risque par rapport à un processus métier. Le calcul de l'importance d'un risque nécessite (i) l'identification des besoins de sécurité liés à un processus métier et à ses activités, (ii) l'évaluation des influences d'un risque sur ces besoins et (iii) le calcul du risque total de l'externalisation d'un processus.

Le traitement de risques est réalisé par l'externalisation des processus qui posent moins de risques. En effet, l'environnement cloud expose différents risques dont chacun influence un besoin de sécurité spécifique. Par conséquent, moins les besoins de sécurité d'un processus sont influencés par les risques d'externalisation vers le cloud, plus le processus mérite d'être externalisé.

Nous nous intéressons dans le cadre de ce travail de recherche aux cinq besoins de sécurité les plus communément connus, à savoir la confidentialité, l'intégrité, la disponibilité, la non-répudiation et l'authenticité.

Dans le cadre de notre travail de recherche, ces différents besoins de sécurité sont liés aux objets de données des processus métier. Un objet de données représente une structure d'information généralement traitée dans les activités comme les documents, les courriers, etc. Les experts de l'entreprise (métier et IT) doivent collaborer pour préciser les besoins de sécurité requis pour chaque objet de donnée. Pour évaluer le besoins de sécurité requis pour chaque activité, il suffit d'évaluer les besoins de sécurité relatifs à chaque objet de données

que cette activité produit ou consomme. Il est assigné à chaque besoin de sécurité, requis pour chaque activité, une valeur égale à 1. Si un besoin de sécurité n'est pas exigé pour une activité, la valeur assignée est égale à 0. Ensuite, pour calculer et évaluer le risque associé à l'externalisation d'un processus vers le cloud en considérant les besoins
de sécurité requis, il est indispensable de présenter d'abord les risques liés à l'adoption du cloud et qui sont identifiées par le CSA (2011) :

- La perte des données : plusieurs raisons peuvent engendrer la perte de donné comme la suppression ou la modification d'un document sans recourir à la sauvegarde du document original. Les risques de perte de données augmentent en adoptant le cloud à cause des caractéristiques architecturales ou opérationnelles de cet environnement (Armbrust et al. 2010).
- Le piratage des comptes : il est réalisé par le recours à plusieurs méthodes comme la fraude et l'exploitation des vulnérabilités logicielles.
- Les interfaces non-sécurisées : les fournisseurs de cloud présentent un ensemble d'interfaces logicielles ou des APIs que les clients utilisent pour gérer et interagir avec les services du cloud. La sécurité et la disponibilité des services cloud dépendent essentiellement de ces interfaces et de ces API. Ainsi, ces derniers doivent être conçus pour protéger les utilisateurs du cloud contre toute tentative malveillante ou accidentelle.
- La violation des données : elle se produit lors d'un vol ou lors d'une manipulation des données sensibles.
- Le déni de service : il a lieu lorsque les attaquants tentent explicitement d'empêcher les utilisateurs légitimes d'utiliser leur service.

Chacun de ces risques influence négativement un ou plusieurs besoin(s) de sécurité requis par les activités du processus métier :

- La perte de données (PD) influence négativement la disponibilité (d) et la non-répudiation (nr) ;
- Le piratage des comptes (PC) influence négativement la confidentialité (c), l'intégrité (i), la disponibilité, la non-répudiation et l'authenticité (a),
- Les interfaces non sécurisées (INS) influence négativement la confidentialité, l'intégrité et l'authenticité ;
- La violation des données(VD) influence négativement la confidentialité ;
- Le déni de service (DS) influence négativement la disponibilité.

Une fois l'influence des risques du cloud sur les besoins de sécurité est identifiée, le calcul de la valeur de risques associée à l'externalisation d'un processus s'avère possible. Prenons l'exemple d'un objet de données appelé OD1 relatif à une "Facture" du processus métier "Gestion des formations du personnel ". Les besoins de sécurité requis par cet objet de données sont les suivants : {confidentialité, intégrité et disponibilité}. Le risque associé à l'externalisation de l'objet de données OD1 vers le cloud relatif à la perte de données est :

- Risque (PD;OD1) = (Confidentialité×0)+(Intégrité×0)+(Disponibilité×1)+ (Non-répudiation×0)+ (Authenticité×0)=1.
- La valeur du risque total relatif à un processus métier (PM) pour chaque influence i se calcule en utilisant la fonction 1.

$$RiT\ (PM)\ = \sum_i^n \sum_j^m Risque(i, OD_j) \qquad (1)$$

Où n correspond au nombre total des influences qui est dans notre cas cinq et m est le nombre des objets de données d'un processus métier.

2.4 Amélioration de la performance

Les entreprises qui externalisent leurs processus métier s'appuient sur l'hypothèse que les prestataires de services cloud sont capables d'exécuter les processus externalisés plus efficacement. En effet, les services cloud sont caractérisés par leur élasticité et le réajustement de leur capacité selon le besoin des clients et à tout moment.

Dans le cadre de nos travaux de recherche, les valeurs des indicateurs clé de performance (en anglais Key Performance Indicator (KPI)) (Wetzstein et al., 2011).sont observées pour évaluer la performance des processus métier. Les valeurs des KPIs sont comparées par leur valeur cible. Si la valeur cible (ou seuil) est dépassée, la performance des processus est considérée comme violée/dégradée. Notamment, la performance étudiée correspond à la durée du processus métier.

Durant une période d'exécution, la performance des instances du processus métier varient entre violation et non violation. Le processus qui a un pourcentage d'instances de performance violée élevé, est plus susceptible d'être externalisé. En effet, les ressources offertes par les services de type IaaS sont capables d'exécuter le processus externalisé avec une durée minimale par rapport à celle de l'entreprise.

Afin d'évaluer la performance des processus métier de l'entreprise, nous nous basons sur les logs d'événements d'un processus qui permettent d'extraire les informations pertinentes liées aux exécutions passées des processus métier. A partir de ces logs, la performance des processus métier est identifiée en calculant le pourcentage des instances du processus qui sont violées (le temps de réponse dépasse la valeur cible) en utilisant la formule 2.

$$Perf(PM) = \frac{NbInstancesV\ iolees}{NbreT\ otalInstances}$$

3 Extension du BPMN par les facteurs d'externalisation

Actuellement, les langages de modélisation des processus métier et particulièrement le standard BPMN 2.0 ne fournissent pas la plupart des facteurs d'externalisation. Cette lacune est expliquée par le fait que la préoccupation essentielle de ces langages est la modélisation et non pas l'externalisation. Pour combler cette lacune, nous proposons une extension légère du BPMN 2.0 qu'on intitule OutyBPMN pour la spécification des caractéristiques d'externalisation des processus métier. La spécification du BPMN2.0 introduit un mécanisme d'extension, permettant d'ajouter des éléments graphiques et de nouveaux concepts pour étendre les éléments standards du BPMN par des attributs. L'extension de BPMN2.0 se base essentiellement sur quatre éléments qui sont :"*Extension*", "*ExtensionDefinition*", "*ExtensionAttributeDefinition*" et "*ExtensionAttributeValue*". La classe " *ExtensionDefinition* " définit les

attributs supplémentaires, cependant la classe " *ExtensionAttributeDefinition* " présente la liste des attributs qui peuvent être attachés à tout élément du BPMN. L'élément "*ExtensionAttributeValue*" contient les valeurs de ces attributs. Finalement l'élément *"Extension*" importe l'élément " *ExtensionDefinition* " et ses éventuels attributs à la définition du modèle BPMN.

Dans ce contexte, de multiples travaux ont proposé l'extension du BPMN pour différents objectifs. Cette proposition s'explique par le fait que l'extension permet d'ajouter une meilleure compréhension de la modélisation des processus métier. En outre, l'ajout de nouveaux concepts à la modélisation des processus métier permet de modifier la façon d'utiliser et de voir le BPMN. Il s'agit de passer d'une utilisation contemplative de la modélisation à une utilisation productive qui assure l'automatisation de l'analyse et de la mise en œuvre des processus métier (Bocciarelli et al., 2011). Dans ce contexte, Rodríguez et al., (2007) présentent une nouvelle extension du BPMN pour incorporer les besoins de sécurité aux modèles de processus métier. De plus, Saeedi et al., (2010) proposent d'ajouter la qualité de service dans la modélisation des processus métier. L'ajout des contraintes temporelles à la modélisation du BPMN est aussi largement étudié (Cheikhrouhou et al., 2015).

Malgré la multitude des travaux de recherche qui ont abordé l'extension du standard BPMN, aucun travail n'a considéré l'extension du BPMN par des facteurs liés à l'externalisation. Ainsi, nous sommes les premiers à proposer une extension du BPMN pour l'externalisation des processus métier. En respectant le mécanisme d'extension offert par BPMN 2.0, nous définissons notre extension pour enrichir la modélisation des processus métier par les facteurs d'externalisation comme l'illustre la Figure 1.

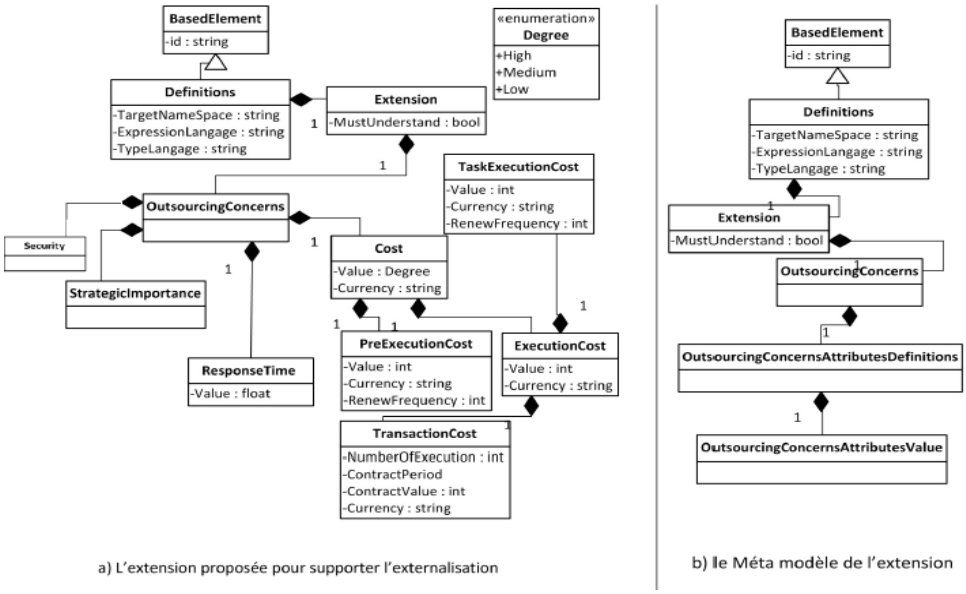


FIG. 1 - *Méta-modèle OutyBPMN proposé*

La Figure 1 b) présente le diagramme de classes correspondant au méta modèle utilisé pour étendre le BPMN 2.0. La classe "*OutsourcingConcerns*" présente la classe "*ExtensionDefinition*" qui contient les classes des attributs d'extension à savoir les classes : *"Cost* ", "*Security*", *" ResponseTime* " et "*StrategicImportance*". Ces classes présentent la classe " *ExtensionAttributeDefinition* " détaillée dans la Figure 1 a). L'extension proposée, permet d'ajouter des nouveaux concepts à l'élément *activity* et *pool* du BPMN 2.0.

4 Modèle de décision à base d'AHP pour la sélection des processus à externaliser vers le cloud

La variété des facteurs d'externalisation retenus complique davantage la décision d'externalisation et exige le recours à des méthodes d'aide à la décision multicritère. La méthode multicritère Analytic Hierarchy Process (AHP) est adoptée pour assister les experts de l'entreprise dans la sélection du processus le plus approprié pour l'externalisation ainsi que le classement des processus impliqués dans la décision d'externalisation. Cette méthode, proposée par Saaty (1987), sert à organiser et structurer les informations et les préférences des experts nécessaires pour la prise de la décision. En outre, l'utilisation de l'AHP permet d'intégrer les opinions subjectives des experts métier de l'entreprise (décideurs). Par la suite, une synthèse de différentes opinions et préférences est élaborée pour donner une recommandation relative à l'alternative (processus métier) la plus convenable aux décideurs.

Afin de classer les processus métier à des fins d'externalisation vers le cloud, nous réalisons les quatre principales étapes nécessaires recommandées par la méthode AHP.

1. La décomposition hiérarchique : l'utilisation de la hiérarchie permet de s'abstraire la structure du problème étudié afin d'analyser et évaluer les interactions entre les éléments du problème et l'effet de ces derniers sur la solution finale. La Figure 2 présente la structure hiérarchique appelée aussi modèle de décision de notre problème d'externalisation réalisée suivant l'AHP.

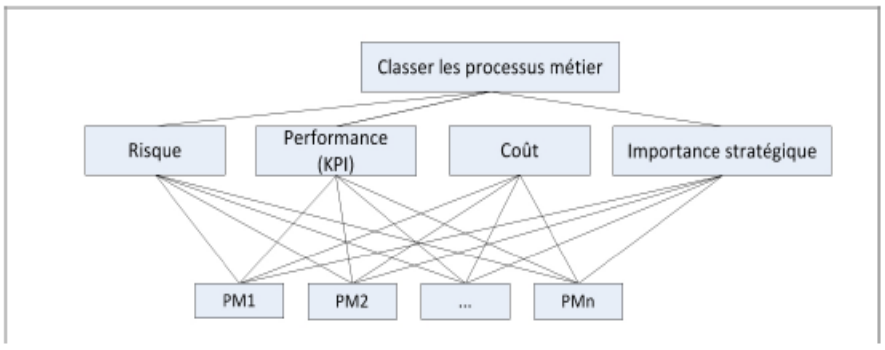


FIG. 2 - *Structure hiérarchique du problème d'externalisation basée sur l'AHP*

Comme le présente la Figure 2, le problème est décomposé en une hiérarchie d'éléments qui sont inter-reliés. Essentiellement, la hiérarchie est décomposée en trois niveaux. Le premier niveau (le sommet de la hiérarchie) correspond à l'objectif de la décision à savoir "classer les processus métier " pour identifier ceux considérés comme les plus convenables pour l'externalisation. Le deuxième niveau correspond aux critères relatifs aux facteurs d'externalisation adoptés. Le dernier niveau présente les différentes alternatives de la décision qui sont dans notre cas les processus métier impliqués dans la décision d'externalisation.

2. La comparaison par paires : afin de pouvoir réaliser une décision judicieuse, il faut quantifier le problème. La quantification est réalisée essentiellement par le recours à des comparaisons par paires des éléments relatifs à chaque niveau hiérarchique du modèle de décision avec ceux correspondant au niveau supérieur. La comparaison permet de fournir un ensemble de matrices de comparaison obtenues principalement suite à l'attribution des poids W = (w1,w2, w3, w4) relatifs aux facteurs de décision adoptés. Les poids attribués sont accordés selon l'échelle proposée par Saaty.
3. Calcul des valeurs des priorités : les jugements élaborés permettent d'évaluer l'importance de chaque élément de la hiérarchie. En effet, une matrice de comparaison

binaire relative à chaque critère doit être remplie par le décideur. Cette étape permet de générer un vecteur de priorité V = (V1, …, Vn) où n est le nombre de critères de la décision. L'obtention du vecteur final qui correspond aux poids finaux de chaque choix est élaborée en calculant le produit V ×W.

4. Le calcul de la cohérence de jugements : cette dernière étape requise par la méthode AHP permet d'évaluer le ratio de compatibilité CR relatif aux jugements des décideurs. La valeur tolérable d'incompatibilité relative au ratio de compatibilité CR est de 0.10. Si cette valeur est dépassée, une révision des jugements élaborés doit être réalisée.

5 Etude de cas

Nous considérons le cas d'une entreprise pétrolière "Abid group » gérant quatre processus métier : (PM1) "Gestion des formations du personnel ", (PM2) "Création des tiges de forage en 3D", (PM3) "Présenter les tiges de forage sur le site Web de l'entreprise " et (PM4)"Gérer la vente des tiges de forage". Le Tableau 1 présente les valeurs des facteurs d'externalisation relatifs à la performance, le coût, l'importance stratégique et les besoins de sécurité pour chaque processus métier.

Processus	Performance (% des instances violées) (P)	Coût (C)	Importance Stratégique (IS)	Risque (R)
PM1	55%	400	Cœur métier	35
PM2	89%	341	Critique Non Cœur	47
PM3	90%	196	Critique Non Cœur	200
PM4	93%	150	Non Critique Non Cœur	45

TAB. 1 – *Détails relatifs aux processus métier étudiés*

Afin de pouvoir sélectionner le processus métier à externaliser, la première étape consiste à réaliser une matrice de comparaison (4 × 4) des quatre facteurs d'externalisation pour montrer leur importance vis-à-vis de l'objectif global du problème. Le Tableau 2 présente les valeurs de comparaison élaborées. L'attribution des jugements subjectifs est élaborée par les experts métier de l'entreprise étudiée.
Le ratio de cohérence (CR) est égal à 0.05 qui est inférieur à 0,1. Le jugement est donc considéré comme compatible. Le vecteur de priorité obtenu (dernière colonne du Tableau 2) indique que le facteur coût est le facteur le plus influant sur la décision d'externalisation.

	P	C	IS	R	Vecteur de priorités
P	1	1/5	1/7	1/3	0.054
C	5	1	1/3	5	0.248
IS	7	3	1	7	0.05
R	3	1/5	1/7	1	0.104

TAB. 2 – *Matrice de comparaison deux à deux des critères de décision*

L'étape suivante consiste à relever la priorité des quatre alternatives relatives aux différents processus métier par rapport aux quatre critères.
Le Tableau 3 illustre le classement relatif à chaque alternative. Les résultats de ce classement montrent que le processus métier PM4 "Gérer la vente des tiges de forage" est le plus approprié pour l'externalisation.

Alternatives	classement	poids
PM1	3	0.234
PM2	2	0.195
PM3	4	0.145
PM4	1	0.426

TAB. 3 – *Résultat final*

6 Analyse de sensibilité

Pour assurer la consistance de la décision finale, nous appliquons l'analyse de sensibilité (Bouh et al., 2016). Il s'agit, d'évaluer l'impact de la variation des valeurs du poids des critères sur la sélection du processus métier à externaliser. En utilisant l'analyse de sensibilité, il est possible aussi de visualiser différents scénarios de "what-if" qui aident à observer l'impact du changement des critères sur l'ordre final des alternatives. En effet, puisque les poids des critères sont généralement basés sur un jugement purement subjectif, la stabilité de l'ordre des alternatives par rapport à la variabilité des poids des critères doit être testée. L'analyse de sensibilité dans notre cas a été réalisée en augmentant ou en diminuant les valeurs des poids des critères. Les changements des priorités des alternatives sont par la suite observés pour évaluer la stabilité de l'ordre ou du classement fourni. Si l'ordre des alternatives est très sensible aux petits changements, une révision des valeurs des poids doit être faite.

La Figure 3 présente les valeurs initiales relatives aux poids des critères. Ces valeurs sont sélectionnées par les experts de l'entreprise "Abid group".

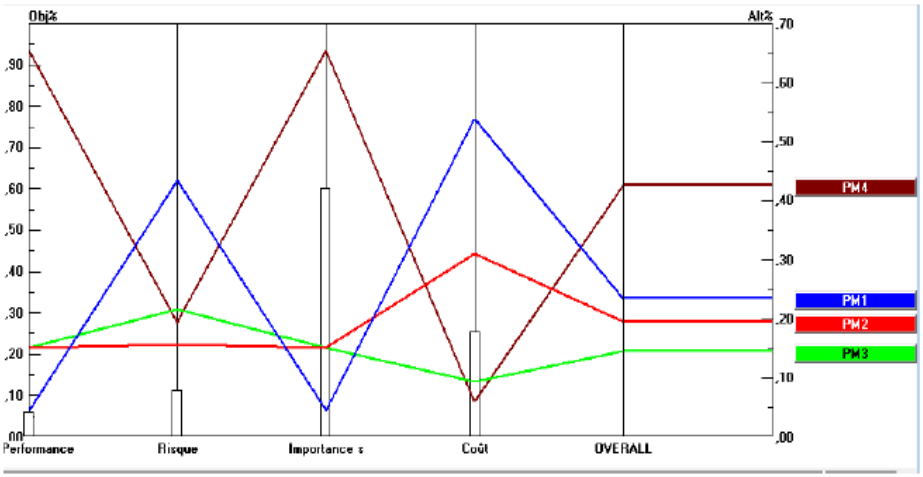


FIG. 3 - *Valeurs des critères initialement sélectionnées par les experts*

Comme nous avons auparavant mentionné, le processus métier PM4 a la valeur maximale du poids par rapport aux autres alternatives qui correspond à 0.426.
Afin d'évaluer l'impact du changement de la priorité des critères sur l'ordre des alternatives, nous augmentons la valeur du poids relatif au critère coût pour passer de 0.248 à 0.45. Le processus métier PM4 reste toujours la meilleure alternative comme l'illustre la Figure 4.

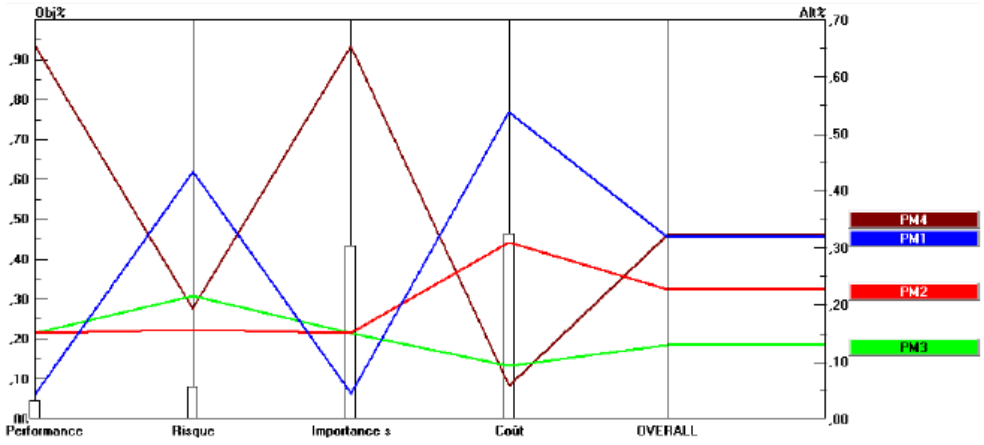


FIG. 4 - *Analyse de sensibilité par la variation du critère coût*

De même, nous varions la valeur du poids relatif à la performance pour passer de 0.054 à une valeur au-delà de 0.65. Le processus métier PM4 reste toujours la meilleure alternative à externaliser. L'analyse de sensibilité élaborée montre que la variation des valeurs des poids des critères n'entraine pas un grand changement sur l'ordre des alternatives et que le processus métier PM4 reste le meilleur choix pour l'externalisation vers le cloud. Ainsi, nous pouvons considérer le jugement élaboré comme étant fiable et efficace.

7 Conclusion

La sélection des processus métier à externaliser n'est pas une tâche triviale vu la variété des critères considérés ainsi que la diversité des alternatives du problème de décision étudié. Ainsi, assister les décideurs/experts de l'entreprise pour la réalisation de cette décision s'avère primordial. Nous avons proposé dans ce papier notre approche de sélection des processus à externaliser qui consiste à classer les processus métier pour pouvoir en identifier celui qui mérite le plus d'être externalisé vers le cloud. Elle commence par l'extension des modèles de processus métier par les facteurs d'externalisation comme : la réduction des coûts, l'amélioration de la performance, l'importance stratégique et la minimisation des risques jusqu'à l'application de la méthode d'aide à la décision multicritère AHP.

Références

Afshari, A., Mojahed, M., et Yusu_, R. M. (2010). Simple additive weighting approach to personnel selection problem. *International Journal of Innovation, Management and Technology*, 1(5) :511.

Armbrust, M., Fox, A., Gri_th, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., et al. (2010). *A view of cloud computing. Communications of the ACM*, 53(4) :50_58.

Bocciarelli, P. and D'Ambrogio, A. (2011). *A bpmn extension for modeling non functional properties of business processes*. In Proceedings of the 2011 Symposium on Theory of Modeling & Simulation. 160-168, San Diego, CA, USA. Society for Computer Simulation International.

Bouh, M. A. and Riopel, D. (2016). *Système hybride flou multicritère á base de connaissance pour la sélection des systèmes d'entreposage des charges palettisées*.

Cheikhrouhou, S., Kallel, S., Guermouche, N., and Jmaiel, M. (2015). The temporal perspective in business process modeling : a survey and research challenges. *Service Oriented Computing and Applications*, 9(1) :75-85.

CSA (2011). *https ://cloudsecurityalliance.org/star/. cloudsecurityalliance, https ://cloudsecurityalliance.org/star/.*

ISO (2008). Iso/iec 27005 information technology - security techniques -information security risk management. *Technical report*.

Rodríguez, A., Fernández-Medina, E., and Piattini, M. (2007). A bpmn extension for the modeling of security requirements in business processes. *IEICE Trans. Inf. Syst*., 745-752.

Saaty, R. (1987). *The analytic hierarchy process what it is and how it is used*. Mathematical Modelling, 9(3) :161- 176

Saeedi, K., Zhao, L., and Sampaio, P. (2010*). Extending bpmn for supporting customer-facing service quality requirements*. In IEEE International Conference on Web Services (ICWS), 616-623.

Wetzstein, B., Leitner, P., Rosenberg, F., Dustdar, S., and Leymann, F. (2011). Identifying influential factors of business process performance using dependency analysis. *Enterprise Information Systems*, 5(1) :79-98.

Summary

Enterprises aiming to outsource their business processes to the cloud, must properly identify their expectations of this choice and ensure its potential risks. Indeed, they must direct the efforts of outsourcing to the highest priority processes that deserve an urgent improvement in their performance and a significant reduction in their costs. This outsourcing ensures a gain in terms of time and cost. Moreover, the risk and the consequences of the failure of an outsourcing can be fatal for the enterprise.

In this paper, we propose a decision support approach which consists in selecting the processes to be outsourced to the cloud. It starts from the extension of business process models by the outsourcing factors until the application of the Analytic Hierarchy Process (AHP) as a multi-criteria decision support method.

Un SaaS Composite Auto-Adaptatif (Self AdaptiVe-CSaaS)

Rima Grati[1], Khouloud Boukadi[1], Hanene Ben-Abdallah[2]

[1] Faculté des Sciences Économiques et de Gestion de Sfax
[1]Route Aeoport Km4, P14, Sfax, Tunisia
Rima.grati@fsegs.rnu.tn
Khouloud.boukadi@fsegs.rnu.tn
[2]King Abdulaziz University
[2]Umm Al Muminin, Al-Sharafeyah, Jeddah 23218, Saudi Arabia
[2]hbenabdallah@kau.edu.sa

Résumé. Le terme SaaS, pour Software as a Service, est utilisé pour désigner une application dans le cloud computing. Afin d'offrir un SaaS avec des fonctions flexibles à faible coût, un nouveau concept a été introduit connu sous le nom SaaS Composite. Un SaaS Composite est né d'une combinaison de Cloud et de services Web. De nos jours, le composite SaaS fonctionne dans un environnement dynamique et volatile comme le Cloud Computing et a donc des paramètres de QoS (Qualité de service) très changeants. Par conséquent, la surveillance et l'adaptation des SaaS composites sont d'une grande importance pour garantir les paramètres QoS définis dans le contrat de service (SLA Service Level Agreement). Pour résoudre ces problèmes, nous proposons dans ce papier deux Systèmes Fuzzy pour SaaS auto-adaptatifs qui visent à prévenir les violations SLA pour SaaS composite. Nous présentons les résultats des expériences que nous avons menées et qui montrent l'efficacité des systèmes flous.

1 Introduction

L'adoption et la réussite des services composite (appelée aussi SaaS composite) dans un environnement Cloud est conditionnée par une exécution correcte de ces services et un respect du SLA agréé entre le client et le fournisseur. Cependant, dans le Cloud, les services sont soumis à des variations de charge qui risquent de modifier les conditions initialement posées par les fournisseurs pour formuler leurs offres. Ainsi, il est indispensable de doter ces services par des mécanismes d'adaptation dynamique qui leur permettent de s'adapter eux même au changement de leurs contextes d'exécution. En agissant ainsi, ces services évitent aux fournisseurs de payer des remboursements monétaires dus à la violation des SLA et permettent au client de s'assurer que les services exécutés ont respecté le contrat agréé.

L'adaptation dynamique des services dans le Cloud vise à modifier le service en fonction de l'environnement dans lequel il s'exécute. Ainsi, un service auto-adaptatif évolue et modifie son propre comportement quand il n'a atteint pas ses objectifs, en se concentrant généralement sur le niveau applicatif.

Etant donné qu'un SaaS composite né d'une combinaison entre le Cloud et le service Web, il hérite les concepts proposés par ces derniers et rajoute ses propres spécificités. De la technologie service Web, il hérite les concepts liés aux relations client-fournisseur, aux services abstraits/concrets, à la composition des services, etc. Du paradigme Cloud, il hérite les concepts de couche, l'élasticité des ressources et le paiement à l'usage, etc.

Nous nous intéressons particulièrement dans ce travail de recherche au SaaS composite implémentant des processus de longue durée d'exécution. Aujourd'hui, les SaaS composites exécutées dans un environnement dynamique et volatil tel que le Cloud sont sujet à des modifications fréquentes de leurs paramètres de qualité de service. Par conséquent, la surveillance et l'adaptation des SaaS composite est d'une grande importance pour garantir les paramètres QoS définis dans le SLA. C'est dans ce contexte que nous essayons dans ce travail de répondre aux questions de recherche suivantes :

- **Choix de la ou des stratégie (s) d'adaptation à appliquer**

En fonction des variations du contexte (dégradation des paramètres de qualité du niveau métier), il existe un ensemble de stratégies d'adaptation possible. Néanmoins, le choix de la stratégie d'adaptation la plus appropriée est une tâche complexe en raison du nombre de critères qui interviennent dans le processus de prise de décision (Avila 2014). En outre, ces stratégies ont différents coût, complexités et temps d'exécution. D'une part, l'adaptation au niveau métier traite de la substitution ou du remplacement d'un service élémentaire par un autre avec une fonctionnalité équivalente et dynamiquement lie le service à la composition.

Dans la littérature, les travaux considèrent uniquement la substitution du service élémentaire lorsque la qualité du service se dégrade. Notre approche examine également le fait que l'amélioration d'un paramètre de qualité dans un certain point de l'exécution peut être considérée pour améliorer d'autres paramètres en ajustant les poids pendant la phase de sélection du service. En effet, pour réaliser une adaptation niveau métier, il est indispensable de bien étudier et analyser les variations du contexte. Par conséquent, la conception d'un système qui peut quantifier objectivement les informations du contexte sur le service avec une précision raisonnable et en temps réel est devenue un besoin primordial qui constitue un challenge intéressant à résoudre. La logique floue apporte des solutions intéressantes aux problèmes déjà soulignés. Le besoin est fort à un ou éventuellement à plusieurs systèmes d'aide à la décision basés sur la logique flou et capables à un moment donné de décider de la stratégie d'adaptation ainsi que des actions d'adaptation à mettre en place.

- **Coût et impact de l'adaptation**

Le coût des actions d'adaptation du niveau métier ainsi que l'impact du changement doivent être considérés avant de décider de la meilleure action d'adaptation à appliquer. En effet, une action d'adaptation particulière possède un coût d'exécution qui peut être parfois plus élevé que le bénéfice attendu. Par exemple, le coût d'une substitution de service est également un facteur important pour déterminer la pertinence de la substitution. En outre, le déclenchement des actions d'adaptation à chaque variation du contexte (par exemple une variation de la performance d'un service élémentaire) n'est pas toujours la meilleure des solutions. Ces différentes constations nous poussent à poser un certain nombre de questions :

- Est-ce qu'une action d'adaptation est indispensable dans le niveau métier?
- Quelle est l'utilité de ces actions d'adaptation par rapport à leurs coûts d'exécution?

Principales contributions

Afin de répondre aux différentes problématiques déjà soulignées, ce travail de recherche propose :

- Une catégorisation de contexte qui facilite les mécanismes d'auto-adaptation des services. Cette catégorisation comporte : le contexte des services élémentaires et le contexte des SaaS composites.

- Un modèle de service composite sensible à son contexte, et en réponse aux changements pertinents, évalue l'utilité de l'adaptation et exécute la ou les actions d'adaptation afin d'éviter les violations de SLA. Ce modèle de service repose en réalité sur deux systèmes de logique flou.

- L'implémentation et l'évaluation du modèle de service (SAV-CSaaS) dans un environnement de Cloud.

Le reste de ce document est organisé comme suit. La section 2 décrit les principaux travaux connexes. La section 3 décrit la définition du contexte pour les SaaS composites auto-adaptatifs. La section 4 présente l'architecture globale du système. La section 5 détaille les systèmes flous pour les SaaS composites auto-adaptatifs. Les résultats des expériences concernant l'application des systèmes flous sont présentés à la section 6. La section 7 conclut le document et discute nos futures perspectives.

2 Travaux connexes

Dans (Pernici and Siadat 2011), les auteurs proposent un système d'inférence floue (FIS) pour capturer une QS globale et choisir des stratégies d'adaptation en fonction des changements de QS. Ce système repose sur deux moteurs d'inférence Fuzzy : le moteur d'évaluation QoS et le moteur de prise de décision. Le premier est utilisé pour déduire le degré global de QoS. Le deuxième moteur est utilisé pour la prise de décision des stratégies d'adaptation. Trois stratégies d'adaptation sont prises en considération par les auteurs : ne rien faire, renégociation et substitution. L'approche proposée permet de choisir des stratégies d'adaptation en fonction des changements de QoS.

Dans (Aschoff and Zisman 2012), les auteurs décrivent un framework ProAdapt (Proactive Adaptation) qui prend en charge l'adaptation de compositions de services déclenchées par différents types de problèmes. Ce framework permet de remplacer une opération de service ou un groupe d'opérations par une autre opération de service ou un groupe d'opérations dynamiquement composées. L'adaptation proactive de la composition du service se compose de trois étapes : la première étape est l'identification et la prédiction des problèmes tandis que la deuxième étape est l'analyse des problèmes déclenchés par la prédiction suivie par la troisième étape qui est la décision des actions à prendre face aux problèmes pour exécuter les actions d'adaptation.

Dans (Avila 2014), l'auteur présente la conception et la mise en œuvre du framework pour l'adaptation des services composés basée sur une logique floue. Ce cadre vise à prévenir la dégradation de la QoS et à améliorer les niveaux de QoS d'un service composite. Il utilise deux systèmes d'inférence Fuzzy qui évaluent les valeurs QoS des services composites, basées sur des données historiques afin d'identifier le besoin d'adaptation. Le cadre se compose de six composantes : composition engine, adaptation manager, service binder, service selector, predictor and sensors.

Dans (La and Kim 2010), les auteurs présentent un framework pour l'activation de services mobiles sensibles au contexte. Le framework permet aux tâches de capturer le contexte, de déterminer quelles adaptations spécifiques au contexte sont nécessaires, adapter les services candidats pour le contexte et exécuter le service adapté. Le résultat de la sensibilité des

services au contexte est pour les consommateurs qu'ils reçoivent de meilleurs services adaptés au contexte actuel du consommateur.

Dans (Qu, Wang et al. 2015), les auteurs proposent CCCloud qui est un modèle de sélection du service de cloud computing et basée sur la comparaison et l'agrégation des évaluations subjectives extraites des consommateurs de cloud et des évaluations objectives de de tests quantitatives de performance des parties. Les auteurs proposent une nouvelle approche pour évaluer la crédibilité des utilisateurs de cloud. Dans leur modèle, les évaluations objectives sont utilisées comme repères pour filtrer les évaluations subjectives potentiellement biaisées, puis les évaluations objectives et les évaluations subjectives sont agrégées pour évaluer les performances globales d'un service cloud. Leur modèle prend en compte les contextes des évaluations objectives et des évaluations subjectives. En calculant la similitude entre les différents contextes, le niveau de référence des évaluations objectives est ajusté dynamiquement en fonction de la similitude du contexte et les scores définitifs agrégés des services cloud alternatifs sont pondérés par la similitude entre les contextes d'un éventuel consommateur en nuage et de chaque groupe de test. Cela rend le modèle de sélection de service plus efficacement pour les besoins personnalisés des consommateurs de cloud.

L'analyse de la revue de la littérature montre que les approches présentées ne sont conscients de la QoS. Nous avons essayé de couvrir le contexte de l'environnement de travail pour proposer des actions d'adaptation adéquates. Un autre point fort de notre approche est l'inclusion d'une évaluation de la décision d'adaptation. Ce critère n'a pas été abordé par les travaux présentés. Au cours du processus d'adaptation, nous considérons également l'optimisation de la QoS qui n'a pas été envisagée par aucune autre approche.

3 Définition et catégorisation du contexte pour un SaaS auto-adaptatif

L'étude de la littérature montre que la sensibilité au contexte est devenue un élément central pour la conception et la mise en place des services auto-adaptatifs. En ayant cette sensibilité au contexte, les services traditionnellement très peu informés de leur entourage deviennent plus conscients de l'environnement dans lequel ils évoluent et aussi de l'environnement du demandeur de services. Par ce biais, ils permettent une meilleure flexibilité et renforcent leur capacité à se reconfigurer de manière automatique afin de respecter les exigences du client. Nous considérons que le contexte est « l'ensemble des paramètres qui peuvent appartenir à l'environnement du SaaS (SaaS composite, service élémentaire) et qui influencent son comportement en définissant de nouvelles vues sur les fonctionnalités proposées par les services participants. Ces paramètres peuvent être statiques ou dynamiques » (Boukadi, Ghedira et al. 2008)

Étant donnée la diversité des informations composant le contexte, il est utile d'essayer de les classer par catégories. Nous proposons une catégorisation qui détaille les paramètres contextuels pouvant influencer le comportement du SaaS composite et qui sont : le contexte du service élémentaire et le contexte du service composite.

- **Contexte du service élémentaire** (*elementary service related context*) décrit les services participants au SaaS composite. Il comporte deux sous-catégories de base à savoir : le contexte fonctionnel et le contexte non-fonctionnel.

- Le contexte fonctionnel : décrit la fonctionnalité proposée par un service élémentaire à travers un ensemble d'opérations. Un service élémentaire dans un SaaS composite est un service à fine granularité ayant un fichier WSDL (Web Services Description Language) et publié dans le registre de service. Afin d'accomplir efficacement sa fonctionnalité de base, le contexte du service élémentaire regroupe également un ensemble de besoins en termes de ressources comme la mémoire requise, le stockage, etc.
- Le contexte non fonctionnel: comprend les paramètres de qualité (QoS) que soit agrées au mesurées, la charge de travail du service ainsi que son coût de substitution. Les paramètres de qualités agrées ainsi que le coût de substitution constituent le contexte non fonctionnel statique acquis à partir du document WSLA (web service level agreement) (Avila 2014). Tandis que les paramètres de qualités mesurées et la charge du travail forment le contexte non fonctionnel dynamique. En effet, le service élémentaire possède une charge de travail qui fait partie de son contexte non fonctionnel et qui influence largement ses paramètres de qualité. En général, les paramètres de qualité différencient les services élémentaires ayant la même fonctionnalité. Ainsi, les clients peuvent évaluer ces paramètres et sélectionner le service élémentaire à inclure dans leurs schémas de composition. Les paramètres QoS peuvent être quantitatifs (par exemple, temps de réponse, disponibilité, coût, etc.) ou qualitative (par exemple, réputation et sécurité). Dans le cadre de ce travail, nous nous intéressons aux paramètres de QoS quantitatifs suivants :
  - ✓ Temps de réponse (Response Time) : caractérise le temps écoulé depuis la soumission de la requête jusqu'à la réception de la réponse.
  - ✓ Prix (Cost) : représente le montant monétaire échangé contre la consommation du service. D'une manière générale, il illustre le prix d'invocation des opérations du service.
  - ✓ Disponibilité (Availability) : représente le pourcentage de temps pendant lequel un service est opérationnel. Les valeurs les plus grandes montrent une disponibilité élevée tandis que les petites valeurs impliquent une disponibilité basse.

- **Contexte du SaaS Composite** (*Composite SaaS related Context*) : tout comme le contexte du service élémentaire, le contexte du service composite distingue deux sous-catégories de contexte : le contexte fonctionnel et le contexte non fonctionnel. Le contexte fonctionnel décrit la fonctionnalité offerte par le SaaS à ses clients à travers un document WSDL. Rappelons que le SaaS composite fourni par une entreprise représente un processus ou un sous-processus métier englobant les compétences et le savoir-faire de cette entreprise. Ce service offre des opportunités d'affaire aux entreprises désireuses d'initier des coopérations interentreprises à la demande. Quant au contexte non fonctionnel, il s'intéresse aux paramètres de qualité qui caractérisent un SaaS composite (agrées dans le SLA-SaaS et mesurés pendant l'exécution) ainsi qu'un ensemble de points de conformité. Ces derniers sont des points dans l'exécution de la composition où une décision d'adaptation peut être prévue. En effet, ils visent à vérifier si l'exécution des services se comporte conformément au document SLA déjà agrée. Le concept de points de conformité est similaire à celui proposé dans (Aschoff and Zisman 2012) et surtout applicable pour les processus de longue durée d'exécution. Nous avons suivi les consignes décrites dans (Aschoff and Zisman 2012) pour définir les

points de conformités. Ces dernières stipulent que les points de conformité très tôt dans l'exécution d'une composition, ont accès à peu d'informations sur l'exécution des services élémentaires, par contre le temps de réaction est élevé. Inversement, les points de conformité, définis très tard dans le schéma de composition, augmentent la probabilité de la violation du SLA. En général, les points de conformité doivent être conçus de telle sorte que suffisamment d'informations contextuelles sur l'exécution des services est disponible pour générer des prévisions utiles.

4 Architecture d'un Composite SaaS auto-adaptatif (SAV-CSaaS)

Les architectures présentées dans les travaux existants pour assurer la sensibilité au contexte accordent une très grande importance à la gestion du contexte sans toutefois présenter comment modifier le comportement d'une application pour qu'elle s'adapte au contexte. Dans notre approche d'adaptation, nous accordons une importance capitale à la fois à la gestion du contexte et à l'adaptation. Pour ce faire, nous définissons une architecture qui aide le SaaS composite à être conscient de son contexte, et en réponse aux changements pertinents, évalue l'utilité de l'adaptation et s'adapte en cas de besoin pour éviter les violations de SLA.

Une vue d'ensemble de l'architecture du SAV-CSaaS est illustrée dans la Figure 1. Cette architecture respecte le principe de modularité (elle se compose de plusieurs modules) qui est un principe phare dans le domaine du développement des logiciels. Les modules sont : *Adaptation Manager*, *Context Manager*, *QoS Calculator*, *SLA violation predictor* et *Discovery Manager*. Ces différents modules sont décrits dans ce qui suit.

- Le module « Adaptation Manager » : l'objectif ultime de ce module est de maintenir la qualité de service agréé au niveau du SLA. Pour ce faire, il se base sur les informations capturées par le Context Manager et invoque les systèmes de logique flou afin de décider, à chaque point de conformité, si une adaptation est nécessaire au niveau métier. En plus, il réalise l'action d'adaptation suggérée (substitution ou ne rien faire).
- Le module « Context Manager » : le gestionnaire de contexte que nous proposons est un module qui se base sur d'autres modules à savoir : le Conformity points Collector, Instance Context Collector et le Smart Listener, afin de collecter les paramètres de contexte. Le Conformity points Collector et l'Instance Context Collector capturent les paramètres du contexte statiques relatifs aux différents services élémentaires et au service composite lui-même. Quant aux paramètres dynamiques, ils sont surveillés et détectés par le Smart Listener. En effet, le Smart Listener détecte les changements dynamiques des paramètres de contexte qui surviennent au moment de l'exécution du service et qui sont susceptibles de déclencher des adaptations. Outre les fonctions de capture du contexte, le Context Manager assure la communication avec les autres modules d'adaptation. Il les informe des informations contextuelles et des changements éventuels du contexte.

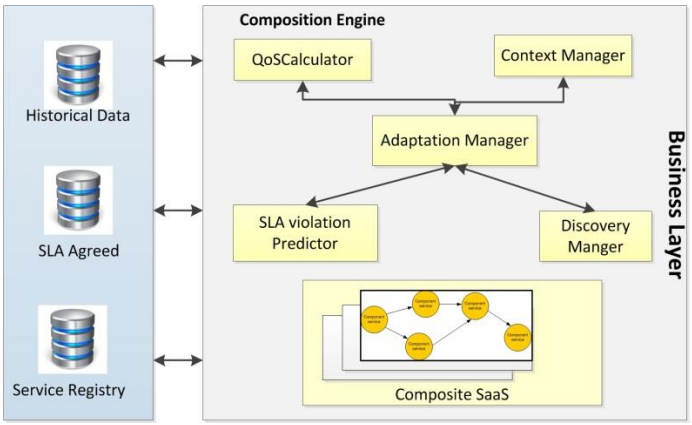


FIG 1 – *ARCHITECTURE D'UN SAV-CSAAS*

- Le module « QoS Calculator » : permet le calcul des paramètres QoS. Pour ce faire, il exploite les détails du QoS représentant le SLA agréé et invoque les plugins nécessaires afin de calculer les valeurs des paramètres de QoS du service composite. Pour réaliser ce calcul, il considère les branchements utilisés pour connecter les services élémentaires ainsi que les valeurs de paramètres de QoS des services élémentaires (exprimés ou pas par des mesures de bas niveau fournis et convertis par le Moniteur).
- Le module « SLA violation predictor » : utilise les réseaux Bayésiens afin de prédire les violations du SLA. À cette fin, il utilise les valeurs des paramètres de QoS du chemin exécuté précédant le point de conformité. Ces valeurs sont obtenues au moment de l'exécution à partir du module QoSCalculator et sont utilisées ensuite pour construire le modèle probabiliste qui peut prédire la violation du SLA agréé. La prédiction se base fondamentalement sur les valeurs de QoS de l'exécution en cours et l'exécution passé de chemins similaires stockées dans le référentiel de données historiques (Historical Data). Le SLA violation predictor est implémenté comme celui présenté dans (Leitner 2011).
- Le module « Discovery Manager » : si l'utilité de l'adaptation du niveau métier est élevé et la stratégie d'adaptation de type substitution est maintenue, ce module est invoqué. Il considère le contexte fonctionnel du service à substituer et cherche dans le référentiel des services (Service Registry), pour constituer une liste des services candidats. Ensuite, en se basant sur le poids suggéré par la stratégie de substitution (substitution avec poids élevé pour le coût ou temps de réponse, etc.), sélectionne à partir de la liste des services candidats le service adéquat.

5 Systèmes de logique Flou pour SAV-CSaaS

Comme nous l'avons déjà mentionné auparavant, SAV-CSaaS repose sur des systèmes de logique flou, comme des systèmes d'aide à la décision multicritères, pour les adaptations métier. Chacun de ces systèmes propose ses propres variables linguistiques et ses propres règles. Ces différents systèmes sont décrits dans ce qui suit. Les deux systèmes de logique flou du niveau métier sont illustrés dans la Figure 2.

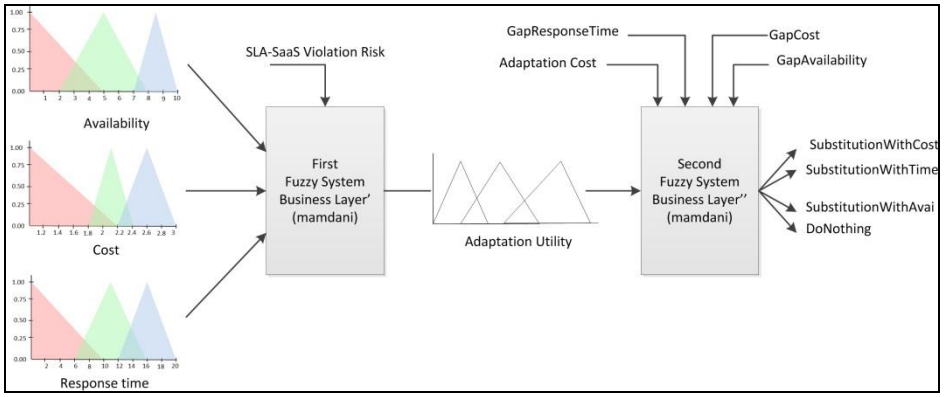


FIG *2– Les deux systèmes de logique flou du niveau métier*

Le premier système flou est invoqué par l'Adaptation Manager à chaque point de conformité. Son objectif est d'évaluer les valeurs des paramètres de QoS du service composite afin de décider si des adaptations doivent être prises en compte au niveau métier.

Deux variables d'entrée sont prévues pour ce système à savoir : (1) les paramètres QoS du chemin exécuté du SaaS composite tels que le temps d'exécution, le coût, etc. et (2) le risque de violation du SLA récupéré à partir du SLA Violation Detector. Quant à la seule variable de sortie, elle illustre l'utilité de l'adaptation (Adaptation Utility). Ces variables sont exprimées selon trois termes linguistiques : faible, moyen et élevé.

Afin de regrouper les paramètres de QoS en des ensembles flous, nous utilisons l'algorithme Fuzzy C-Means (FCM). Nous estimons que ce choix est plus judicieux et moins couteux que celui de se baser sur les exécutions antérieures des services ou sur l'avis des experts. En effet, dans le contexte de service Web et Cloud, les experts ne sont pas en mesure de juger facilement l'appartenance d'un paramètre de QoS à un ensemble flou.

L'algorithme Fuzzy C-Means (FCM) que nous utilisons est un algorithme de classification non-supervisée flou. Issu de l'algorithme des C-moyennes (C-means), il introduit la notion d'ensemble flou dans la définition des classes : chaque point dans l'ensemble des données appartient à chaque cluster avec un certain degré, et tous les clusters sont caractérisés par leur centre de gravité.

Comme les autres algorithmes de classification non supervisée, il utilise un critère de minimisation des distances intra-classe et de maximisation des distances inter-classes, mais en donnant un certain degré d'appartenance à chaque classe. Cet algorithme nécessite la connaissance préalable du nombre de clusters et génère les classes par un processus itératif en minimisant une fonction objective. Ainsi, il permet d'obtenir une partition floue de chaque qualité de service grâce à un degré d'appartenance (compris entre 0 et 1) à une classe donnée. Le cluster auquel est associé un paramètre de qualité de service est celui dont le degré d'appartenance sera le plus élevé.

À cette fin, nous définissons trois classes de paramètres de QoS à savoir : faible, moyen et élevé. Pour chaque contexte fonctionnel de service élémentaire et à chaque paramètre de qualité de service, nous appliquons le Fuzzy C-means pour organiser les services autour de trois clusters de qualité. Après l'application de l'algorithme, les centroïdes des classes $C = C = \{ c_1 , ... , c_c \}$ (Alexander and Heiko 2003) et une matrice de partition $W = w_{ij} \in [0,1]$ sont définis . Chaque élément w_{ij} indique le degré auquel le paramètre de qualité de service du service élémentaire j remplit le terme linguistique i. Les degrés d'appartenance du service

aux clusters correspondants sont obtenus en minimisant itérativement la fonction objective suivante :

$$J_m(W, C, S) = \sum_{j=1}^{n} \sum_{i=1}^{c} (W_{ij})^m \|S_j - C_i\|2 \text{ avec} \qquad \text{(Equation 1)}$$

$$W_{ij} = \frac{1}{\sum_{k=1}^{c} (\frac{\|S_j - C_i\|}{\|S_j - C_k\|})^{\frac{2}{m-1}}}$$

L'avantage d'utiliser l'algorithme Fuzzy C-means est que chaque valeur d'un paramètre de QoS d'un service élémentaire est classée dans au moins une classe avec un degré d'appartenance. Une fois classés, ces paramètres de qualité peuvent être évalués. Par exemple, un service élémentaire appelé Plan assembling pourrait être classé dans la classe "élevé" pour le coût avec un degré d'appartenance de 0,2; il peut être également classé dans la classe "moyen" avec un degré d'appartenance de 0,8.

Afin d'aider à la prise de décision, le système de logique floue se base sur un ensemble de règles de type if-then (voir Figure 3). Nous avons défini 32 règles composées, identifiées en combinant les différents variables d'entrée du système flou et les relations possibles avec la variable de sortie à savoir l'utilité d'adaptation. Ces règles représentent le scénario qui peut avoir lieu pour la première prise de décisions à savoir si une adaptation au niveau métier est nécessaire ou une stratégie de rien faire sera maintenue.

Rule 1
IF (Response Time IS high AND Cost IS low AND Availability IS high AND SLA-SaaS Violation Risk IS high) OR (Response Time IS low AND Cost IS high AND Availability IS low AND SLA-SaaS Violation Risk IS high) OR (Response Time IS low AND Cost IS low AND Availability IS high AND SLA-SaaS Violation Risk IS high) OR (Response Time IS low AND Cost IS low AND Availability IS low AND SLA-SaaS Violation Risk IS high) THEN Adaptation Utility IS high

FIG 3– *Extrait des règles du premier système de logique floue du niveau métier*

Quant au deuxième système de logique floue (FC2), il est invoqué quand l'utilité de l'adaptation est élevée ou moyenne et permet ainsi de retrouver la meilleure stratégie d'adaptation. Les variables d'entrée de ce système sont : l'utilité de l'adaptation, le coût de l'adaptation et l'écart entre les paramètres de QoS agréés et ceux mesurés[1]. La variable de sortie de ce système est la stratégie d'adaptation qui peut être : ne rien faire ou substitution avec Temps/Coût/ Disponibilité. Les différents variables de ce système sont exprimées en utilisant les même termes linguistiques que le FC1 à savoir : élevé, moyen et faible.

- **Les deux systèmes de logique flou en action**

[1] Calculé en utilisant la formule suivante : $GapX(S_i) = \frac{agreedX(S_i) - mesuredX(S_i)}{mesuredX(S_i)}$ Avec X= coût, temps d'exécution ou disponibilité et S_i est un service élémentaire. Ensuite une fonction gaussienne est utilisée pour définir le degré d'appartenance de l'écart aux ensembles flous : élevé, moyen et faible.

Les deux systèmes de logique flou déjà décrits sont invoqués par le gestionnaire d'adaptation (Adaptation Manager) conformément à l'algorithme présenté dans la Figure 4. Les notations utilisées par l'algorithme sont les suivantes :

- Au moment de l'exécution, tout SaaS composite, comme mentionné auparavant possède un contexte nommé SaaS-Ctxt
- Toute exécution d'un SaaS composite trace un chemin d'exécution particulier P = $\{S_1,. . . , S_k\}$qui décrit les services invoqués et les branchements utilisés par ces services.
- Chaque service élémentaire appartenant au chemin d'exécution possède un contexte nommé e-Ctxt.
- GapTime, GapCost et GapAvailability désignent respectivement l'écart observé entre le temps d'execution, coût ou disponibilité agréé(e) et celui mesuré(e) des services élémentaires présents dans le chemin P.

La logique de fonctionnement des deux systèmes de logique flou débute par l'invocation du module SLA Violation Predictor (ligne 6). Si le risque de violation est faible, l'algorithme est arrêté. Sinon, la liste des services participants au chemin d'exécution P est organisée en utilisant un tri descendant selon le GapTime (ligne 9) après avoir récupéré à partir du gestionnaire de contexte les temps d'exécution agréés et mesurés. Le service élémentaire qui a le plus d'écart entre le temps d'exécution agréé et celui mesuré est sélectionné par la suite (ligne 11). Le temps d'exécution de ce service est extrait. La même procédure est appliquée pour les autres paramètres de qualité de service (le coût et la disponibilité) (ligne 12 à 29) pour invoquer le premier système de logique floue (FS1) qui va décider à son tour si une stratégie d'adaptation est nécessaire (ligne 31). Si oui (c'est à dire l'utilité de l'adaptation est élevée ou moyenne), le deuxième système de logique floue (FS2) est invoqué pour obtenir la stratégie d'adaptation adéquate (ne rien faire ou substitution avec Temps/Coût/ Disponibilité). Cette stratégie est envoyée par la suite au gestionnaire d'adaptation qui va sélectionner le service substituant en tenant en considération les recommandations du FS2 ou ne rien faire le cas échéant (ligne 43).

6 Evaluation expérimentale et validation

Afin de tester l'approche SAV-CSaaS nous avons utilisé l'environnement de Cloud appelé Amazon EC2. L'ensemble des expérimentations a été réalisé en utilisant l'étude de cas d'un processus de longue durée d'exécution adapté de (Leitner 2011). Cette étude de cas considère un revendeur de robots sophistiqués (ACME BOT).

Nous évaluons l'efficacité de l'approche d'adaptation proposée au niveau métier. Tout d'abord, nous évaluons la décision d'adaptation suggérée par le système de logique floue (FS1) en la comparant avec une adaptation classique (substitution à chaque variation du contexte). Ensuite, nous évaluons la stratégie d'adaptation proposée par le système de logique floue FS2 en la comparant avec une approche de substitution basée sur une pondération fixe.

5.1 Evaluation du nombre de substitution des services

Nous avons exécuté 10 fois le CSaaS en variant à chaque exécution les paramètres de contexte et en adoptant deux approches d'adaptation différentes : notre approche basée sur les systèmes de logique flou et l'approche d'adaptation classique. Une approche classique effectue une substitution de services élémentaires quand une déviation de QoS se produit. Une comparaison entre ces deux approches est illustrée dans la Figure 5.

L'analyse des résultats obtenus de la Figure 5 nous permet de constater que notre approche d'adaptation réduit le nombre de substitutions de service par rapport à un adaptation classique. En effet, le nombre de substitutions augmente dans le cas d'une adaptation classique de façon exponentielle avec le nombre d'exécutions simultanées du CSaaS. Ceci peut être expliqué par le fait que notre approche vise à garantir un compromis entre la qualité globale du service composite, la pertinence de l'adaptation et le coût de substitution. En effet, dans certains cas de variations des paramètres de contexte (temps d'exécution élevé d'un service élémentaire) ; celle-ci peut être ignorée en raison du coût élevé de la substitution.

```
1  Input SaaS_id, CP_id;
2  Output Adapted_SaaS;

3  Path Vector <Service>=null;
4  Context SaaS_ctxt=ContextManager.getContext(SaaS_id)
5  SaaS_ctxt.getmesuredQoS();
6  SLAViolRisk=SLASaaSViolationPredictor(Path,mesuredQoS[]);
7  Path=ExtractPath(SaaS_id, CP_id);
8  If (SLAViolRisk is high or meduim)
9  {        Sort Path by GapTime Desendent
10   {
11   S_id=Path[0];
12   GapTime=getGapTime(S);
13   Context e_ctxt=ContextManager.getContext(S);
14   RT=ctxt.getMesuredRT();
15   }
16        Sort Path by GapCost desendent;
17  {
18  S_id=Path[0];
19  GapCost=getGapCost(S);
20  Context.e_ctxt=ContextManager.getContext(S);
21  Cost=e_ctxt.getmesuredCost();
22  }
23        Sort Path by GapAvail desendent
24  {
25  S_id=Path[0];
26  GapAvail=getGapAvail(S);
27  Context.e_ctxt=ContextManager.getContext(S);
28  Avail=ctxt.getmesuredAvail();
29  }
30  }
31   AdaptationUtility=FS1(SLAViolRisk, RT, Cost, Availability)
32  If AdaptationUtility is high or meduim
33 {
34  Service S_id=getSucesseurCP(CP_id);
35  Contexte e_ctxt=ContextManager.getContext(S_id);
36  SubstitutionCost=e_ctxt=getAdaptationCost();
37  Service_funct=e_ctxt.getFunctionalContext();
38  SubstitutionStrategy=FS2(GapTime, GapCost, GapAvail, AdaptationCost);
39  If (SubstitutionStrategy !=DoNothing)
40  {
41 CandidateService[]=DiscoveryManager(SubstitutionStrategy, Service_funct);
42 }
43  AdaptationManager(CandidateService[0], SaaS_id);
44 }}
```

FIG 4– *Logique de fonctionnement des deux systèmes de logique floue*

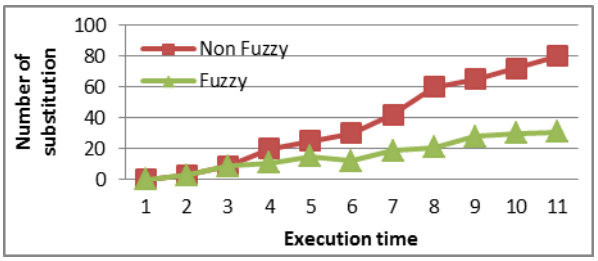


FIG 5– *Evaluation du nombre de substitutions selon les deux approches d'adaptation*

5.2 Adaptation des services basée sur les poids fixes

Dans cette série d'expérimentations nous évaluons notre approche d'adaptation en la comparant avec une approche d'adaptation basée sur une substitution à pondération fixe des paramètres de QoS (W1= W2=W3=0.33 avec W1 est le poids associé au temps de réponse, W2 est le poids pour le coût et W3 est le poids de la disponibilité).

Le SaaS composite de l'étude de cas a été exécuté 50 fois et ceci en adoptant les deux approches d'adaptation. Les figures 6 et 7 illustrent une comparaison entre l'approche proposée et l'approche de substitution à pondération fixe pour chacun des paramètres de qualité de service (temps de réponse, coût et disponibilité). L'analyse de ces figures prouve que l'approche d'adaptation proposée améliore les valeurs de qualité de service globales de la composition. En effet, notre approche d'adaptation permet un temps de réponse plus réduit que celui de l'approche de substitution à pondération fixe (voir Figure 6). Il s'agit en effet d'une réduction moyenne de 2,5 % et un écart-type moyen de 4,7 %. Quant à la comparaison selon le coût, elle montre que l'utilisation du système de logique floue pour l'adaptation permet une réduction moyenne de 3,7 % avec un écart type moyen de 5,2 %, par rapport à une approche de substitution à pondération fixe (voir Figure 7).

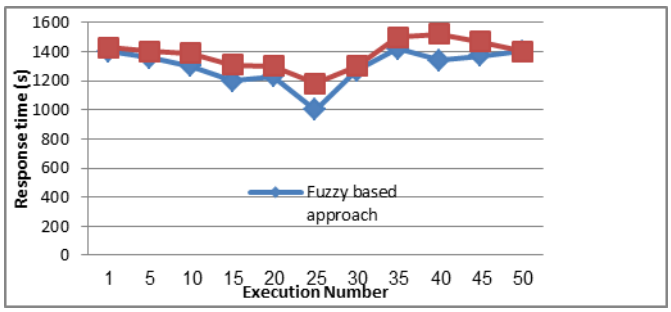


FIG 6– *Comparaison selon le temps d'exécution*

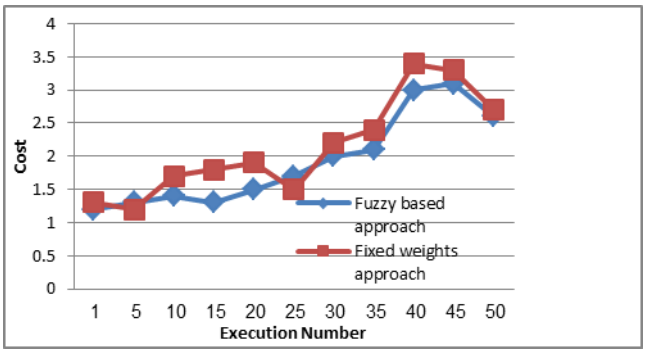


FIG 7– *Comparaison selon le coût*

7 Conclusion

De nos jours, la surveillance et l'adaptation des SaaS composites sont d'une grande importance pour garantir les paramètres QoS définis dans le contrat de niveau de service (SLA). Le travail décrit dans ce document propose : (i) Une catégorisation de contexte qui facilite les mécanismes d'auto-adaptation du service.. (ii) Un système qui aide le SaaS composite à être conscient de son contexte, et en réponse aux changements pertinents dans son contexte, évalue l'utilité de l'adaptation et s'adapte si nécessaire pour éviter la violation des SLA. (iii) Deux systèmes flous pour le SaaS auto-adaptatif : Le premier évalue les valeurs QoS du chemin exécuté dans le service composite afin de décider si des mesures d'adaptation appropriées doivent être prises dans la couche métier. La seconde est invoquée lorsque l'utilitaire d'adaptation est haut ou moyen et vise à sélectionner la meilleure stratégie d'adaptation. (iv) La mise en œuvre et l'évaluation de l'adaptation des entreprises en utilisant un environnement Cloud et un processus long.

Références

Alexander, K. and L. Heiko (2003). "The WSLA Framework: Specifying and Monitoring Service Level Agreements for Web Services." Journal of Network and Systems Management 11(1): pp. 57-81.

Aschoff, R. R. and A. Zisman (2012). Proactive adaptation of service composition. 2012 ICSE Workshop on Software Engineering for Adaptive and Self-Managing Systems (SEAMS).

Avila, S. D. G. (2014). Proactive Adaptation in Service Composition using a Fuzzy Logic Based Optimization Mechanism. 4th International Conference on Cloud Computing and Service Science (CLOSER'14),, Barcelona, Spain.

Boukadi, K., C. Ghedira, S. Chaari, L. Vincent and E. Bataineh (2008). CWSC4EC: how to employ context, web service, and community in enterprise collaboration. Proceedings of the 8th international conference on New technologies in distributed systems., Lyon-France, ACM.

La, H. J. and S. D. Kim (2010). A Conceptual Framework for Provisioning Context-aware Mobile Cloud Services. 2010 IEEE 3rd International Conference on Cloud Computing.

Leitner, P. (2011). On Preventing Violations of Service Level Agreements in Composed Services Using Self-Adaptation, Fakultät für Informatik der Technischen Universität Wien.

Pernici, B. and S. H. Siadat (2011). Selection of Service Adaptation Strategies Based on Fuzzy Logic. 2011 IEEE World Congress on Services (SERVICES).

Qu, L., Y. Wang, M. A. Orgun, L. Liu, H. Liu and A. Bouguettaya (2015). "CCCloud: Context-Aware and Credible Cloud Service Selection Based on Subjective Assessment and Objective Assessment." IEEE Transactions on Services Computing 8(3): 369-383.

Summary

Nowadays, composite SaaS runs in a dynamic and volatile environment such as Cloud Computing and has hence very changing QoS parameters compared to those running in a classical service environment. Therefore, the Composite SaaS monitoring and adaptation are of a great importance to guarantee the QoS parameters defined in the Service Level Agreement (SLA). To address these issues, in this paper we propose some Fuzzy Systems for self-adaptive SaaS that aim to prevent the SLA violations for composite SaaS. The proposed approach objective is to satisfy three important issues such as the selection of service adaptation strategies.. And the cost as well as the impact of changes. We present results of experiments that we have conducted to evaluate the work. These experimental results show the effectiveness of the fuzzy systems.

Vers une logique non monotone distribuée pour l'analyse de l'interaction *conducteur-piéton*

Azedine Boulmakoul[1], Lamia Karim[1,2], Meriem Mandar[1,3], Zineb Besri[1,4], Mohamed Tabaa[5]

[1]LIM/IOS., FSTM, Hassan II University of Casablanca, B.P. 146 Mohammedia, Morocco,
[2]Higher School of Technology EST Berrechid, Hassan 1st University, Morocco
[3]National School of Applied Sciences Bd Béni Amir, BP 77 Khouribga, Morocco
[4]National School of Applied Sciences of Tetouan, BP 2222 M'hannech , Tetouan, Morocco
[5]LPRI Lab. Moroccan School of Engineering Sciences, Casablanca, Morocco

Résumé. Dans ce papier nous définissons les fondations d'une théorie logique de défauts non monotone et distribuée pour modéliser le système «conducteur-piéton». Cette voie, se distingue de l'approche floue et intuitionniste que nous avons proposée dans nos travaux antérieurs. Les deux approches sont de nature sémantique. L'approche fondée sur la théorie du flou est symbolique, la présente approche est de nature logique. Des processus inférentiels dynamiques entre acteurs intelligents distribués sont exploités. Nous prouvons la validité formelle de notre proposition et nous soulignons la contribution du formalisme du raisonnement révisable et de la représentation des connaissances des systèmes intelligents distribués pour la représentation de l'interaction conducteur-piéton. Le raisonnement non monotone acquiert les réflexes décisionnels nécessaires aux acteurs dont l'objectif est de déployer leurs comportements, ceci assure une réactivité et un meilleur usage des connaissances locales et avoisinantes. Ces dispositifs logiques permettent de comprendre la dynamique et les échanges hypothétiques entre acteurs, où les faits et les observations de base sont souvent assujettis à l'interprétation relative.

1 Introduction

Les piétons sont des usagers vulnérables du système routier. La capacité de répondre à la sécurité des piétons est une composante importante des efforts visant à prévenir les accidents de la circulation routière. Les accidents des piétons, comme d'autres accidents de la circulation routière, ne devrait pas être acceptés comme étant inévitables parce qu'ils sont, le produit d'un système associant comportement individuel, outils de transport et environnement, dans des conditions qui rendent le dysfonctionnement à la fois prévisible et évitable. Outre les attitudes à l'égard des risques et leur évaluation, tout usage de la route doit également prendre en compte un ensemble de règles gouvernant les interactions de l'environnement routier. La transgression de ces règles est à l'origine des comportements dangereux pouvant induire des accidents. Les accidents de piétons se produisent majoritairement en milieu urbain, du fait d'une exposition aux risques supérieure en ville. Le taux de gravité exprimé par le ratio du nombre de personnes tuées et le nombre de celles accidentées varie en fonction du lieu de survenance des accidents. Ce taux est plus important en dehors des agglomérations sur les autoroutes et les routes nationales. Et ce à cause d'une part de l'importance des vitesses et densités de circulation et que les conducteurs s'attendent moins à trouver des piétons. Par ailleurs, beaucoup d'accidents de moindre gravité se produisent aux passages des piétons où

les vitesses de circulation sont plus faibles et les conducteurs sont avertis de la présence des piétons. Les conducteurs sont plus susceptibles à s'arrêter lorsque les piétons semblent déterminés à traverser, contrairement à ceux qui attendent passivement leur tour pour traverser. Cependant des différences existent selon les catégories de piétons : les enfants sont percutés au début de la traversée puisqu'ils s'élancent sans regarder, tandis que les piétons âgés sont percutés en fin de traversée vu qu'ils n'ont pas le temps de traverser alors que la circulation a déjà repris. Le changement de mode de transport (monter ou descendre de voiture, des transports en commun, etc.) entraine également des accidents. Les piétons enfants sont particulièrement concernés vu qu'ils ne se rendent pas compte qu'ils deviennent très vulnérables en sortant du premier moyen de transport. Les chocs piétons-véhicules n'impliquent généralement qu'un seul véhicule. Les poids lourds engendrent un taux de gravité plus important que les autres types. Et les chauffeurs reportent les accidents aux problèmes de visibilité qu'ils rencontrent vis-à-vis des piétons. Dans ce papier, nous projetons de modéliser le modèle de risque par une approche logique qui capture la sémantique des piétons et des conducteurs pour le partage de l'espace routier. Notre idée est fondée sur le postulat suivant : *l'entendement portant sur le partage de l'espace-temps et les connaissances spécifiques et relatives de l'ontologie des objets routiers créent une coupure dans les référentiels de connaissances et provoquent des singularités sémantiques que nous mesurons par des accidents*. *Nous suggérons de modéliser avec la logique dynamique les principaux facteurs psychologiques impliqués dans la prise de décision de traversée de rue*. Certes, le besoin de développer des théories logiques sur le diagnostic et le contrôle des systèmes constitue une question essentielle dans les recherches en intelligence artificielle et de la modélisation des systèmes complexes. La dimension logique dans la problématique de la représentation de la connaissance et du raisonnement est multiple. Son rôle dans la modélisation de l'incertitude et dans la gestion dynamique des connaissances est d'un grand secours pour modéliser le raisonnement révisable. Inévitablement, les logiques non monotones sont qualifiées pour la mise au point de systèmes d'inférence permettant la modélisation des raisonnements révisables. Plusieurs issues ont été développées dans la littérature de l'intelligence artificielle, en particulier nous allons citer les grandes classes de logiques non monotones utilisées : les logiques des défauts, les logiques non monotones de McDermott, les logiques autoépistémiques et les logiques basées sur le principe du monde clos et le principe de circonscription (Davis, 1980), (Kleer et al., 1992). Dans la modélisation logique du diagnostic avec les logiques non monotones, les travaux de Reiter (1987) sont principalement les plus célèbres. Dans la continuité des travaux de Reiter (1980, 1987), nous modélisons la théorie de l'action définie comme complément de la théorie du diagnostic modélisée par la théorie logique avec des défauts.

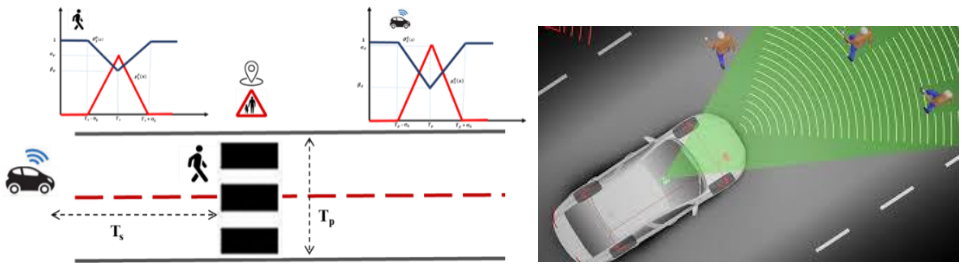


FIG. 1 – *Modèle intuitionniste flou de l'interaction conducteur-piéton.*

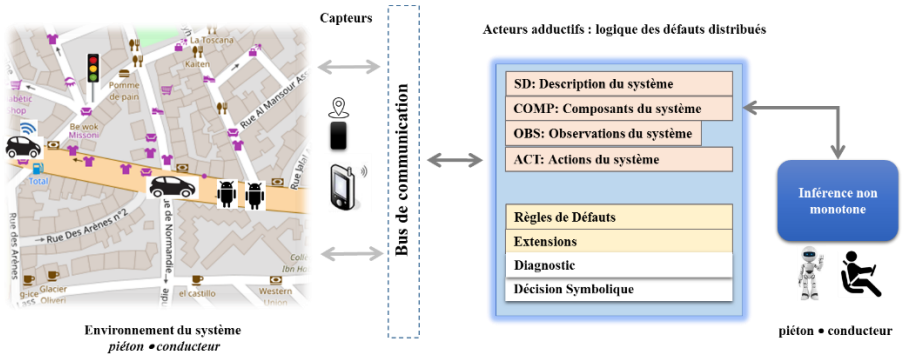


FIG. 2 – *Interaction conducteur-piéton : Agents à logique des défauts distribués*

La figure 1, illustre l'approche intuitionniste floue que nous avons initiée dans (Mandar, Boulmakoul, 2016 a). La figure 2 décrit le contexte et les composants du piéton virtuel logique.

La suite du papier s'organise de la manière suivante: la section 2 propose des rappels brefs sur la théorie logique avec défauts et la théorie du diagnostic à partir des premiers principes, elle présente aussi une nouvelle formalisation du contrôle symbolique. La section 3 donne les prémices d'une théorie de défauts pour modéliser le système *conducteur-piéton*. Enfin, la section 4 conclut et oriente les développements futurs.

2 Logique des défauts, théorie du diagnostic et du contrôle symbolique

Dans cette section nous rappelons les éléments préliminaires portant sur les théories logiques avec défauts. Par la suite nous définissons une nouvelle approche pour formaliser le contrôle symbolique associé à un diagnostic, basée sur la théorie logique avec défauts. Enfin nous décrivons sommairement une société d'agents intelligents à base de théorie avec défauts distribués.

2.1 Théorie logique avec des défauts

La logique des défauts a été introduite par Reiter (1980), pour formaliser le raisonnement simplement consistant. Ci-après nous donnons succinctement le formalisme sous-jacent à cette logique .Désignons par $\mathcal{L}$ un langage prédicatif de premier ordre. Une règle de défaut D est une expression de la forme $\frac{\alpha(x):M\beta_1(x),M\beta_2(x),\ldots,M\,\beta_m(x)}{\gamma(x)}$, où $\alpha(x), \beta_1(x), \beta_2(x), \ldots, \beta_m(x), \gamma(x)$ sont des formules de $\mathcal{L}$.

- $\alpha(x)$ est appelé le prérequis du défaut D,
- $\beta_1(x), \beta_2(x), \ldots, \beta_m(x)$ est appelé justification du défaut D,
- $\gamma(x)$ est appelé conséquent du défaut D,
- M est un symbole du métalangage.

Une théorie de défauts est un couple (W,Δ), où W est un ensemble consistant de formules fermées de $\mathcal{L}$ et Δ un ensemble de règles de défaut.

2.1.1 Extension de théories avec défauts

Une théorie avec défauts (W,Δ) sous-entend un certain nombre d'ensembles de croyances que l'on peut inférer de façon consistante à partir de l'ensemble W. Ces ensembles de croyances sont appelés extensions de la théorie avec défauts.

Soit $Th_{\mathcal{L}}(X) = \{w | w \in \mathcal{L}, w\ est\ fermé\ et\ X \vdash w\}$, c'est l'ensemble des formules fermées qui peuvent être inférées de X de façon valide par les lois d'inférence classiques de $\mathcal{L}$.

Soit E un ensemble de formules fermés de $\mathcal{L}$ et soit (W,Δ) une théorie avec défauts fermés.

Soit la suite de formules E_i construite comme suit :

$$\begin{cases} E_0 = W \\ E_{i+1} = Th_{\mathcal{L}}(E_i) \cup \left\{ \gamma \left| \frac{\alpha : M\beta_1, M\beta_2, \dots, M\,\beta_m}{\gamma} \in \Delta \right. \right\} \end{cases}$$

E est une extension de (W,Δ) si set seulement si $E = \bigcup_0^\infty E_i$

L'exemple donné ci-dessous illustre le concept d'extension :

Soit la théorie avec défaut (W,Δ), où $W = \emptyset$ et $\Delta = \left\{ \frac{:MA}{A}, \frac{:MB}{\neg A} \right\}$. Cette théorie possède deux extensions : $E_1 = Th_{\mathcal{L}}(\{A\})$ et $E_2 = Th_{\mathcal{L}}(\{B, \neg A\})$

2.2 Théorie du diagnostic et du contrôle revisitée

2.2.1 Théorie du diagnostic selon les principes de base

Dans la théorie du diagnostic proposée par Reiter (1987), un système est décrit par le triplet (DS, COMP, OBS), désignant respectivement la description du système (DS) par un ensemble de formules logiques de premier ordre, COMP, les composantes du système, un ensemble de constantes propositionnelles ; OBS, les observations capturées du système, un ensemble de formules logiques de premier ordre. Supposons que nous avons pu déterminer qu'un système (SD, COMP) est défectueux, nous entendons par là de manière informelle, que nous avons fait des observations sur le système représentées par l'ensemble OBS qui est en contradiction avec la description du système prédit, et qui postule que tous ses composants se comportent correctement. D'où le fait que l'observation OBS génère des conflits avec l'hypothèse affirmant que toutes les composantes du système se comportent correctement peut être formalisé par l'ensemble des formules logiques suivant : $SD \cup OBS \cup \{c \in \text{COMP} | \neg AB(c)\}$ est inconsistant. Le principe de parcimonie formule que le diagnostic d'un système est un ensemble minimal de composants en état anormal ($AB(c)$ signifie *abnormal*) Ceci nous amène à ce qui suit:

Définition du diagnostic : Un diagnostic pour le système (SD, COMP, OBS) est l'ensemble minimal $\Delta \subseteq$ COMP tel que :

: $SD \cup OBS \cup \{c \in \Delta | AB(c)\} \cup \{c \in COMP - \Delta | \neg AB(c)\}$ est consistant.

Inférence à partir d'un diagnostic : Un diagnostic Δ pour le système (SD, COMP, OBS) prédit Π (une formule logique de premier ordre) si et seulement si

$$SD \cup OBS \cup \{c \in \Delta | AB(c)\} \cup \{c \in COMP - \Delta | \neg AB(c)\} \models \Pi$$

Le théorème suivant définit la relation entre la théorie des défauts d'un système et ses diagnostics.

Théorème de Reiter (1987) Soit le system (SD, COMP, OBS) alors E est une extension de la théorie des défauts $DT = \left(SD \cup OBS, \left\{c \in COMP \left| \frac{:\neg AB(c)}{\neg AB(c)} \right.\right\}\right)$ Si et seulement si pour un ensemble diagnostic Δ de (SD, COMP, OBS) $E = \{\Pi | \Delta \text{ prédit } \Pi\}$

2.2.2 Théorie du contrôle selon les principes de base

Dans le paragraphe §2.1.2., un système a été décrit par le triplet (DS, COMP, OBS) auquel nous ajoutons un ensemble ACT de formules logiques de premier ordre correspondant aux actions à entreprendre pour un état du système désigné par ses diagnostics. L'intérêt de lier le diagnostic à la stratégie d'instrumentation des actions en concordance avec l'état du système est de gérer les situations anormales en termes de comportement du système. Dans notre cas, il est question des gérer les congestions qui sont des situations anormales du trafic. Le processus d'activation des actions pour l'asservissement d'état d'un système peut être paraphrasé par le texte suivant : « *Si le processus est dans l'état X et si l'on applique l'action U, alors on observe le processus dans un état Y* ». Le contrôle d'un système est intrinsèquement lié à son diagnostic. Pour chaque diagnostic doit correspondre des actions à mener pour réguler le système vers les états désirés *et ce selon une stratégie donnée. A savoir la stratégie est un « ensemble d'actions coordonnées, d'opérations habiles, de manœuvres en vue d'atteindre un but précis».*

Définition Un contrôle symbolique {ensemble d'actions} associé à un diagnostic **Δ** pour le système (SD, COMP, OBS, ACT) est l'ensemble A ⊆ ACT tel que :
$SD \cup OBS \cup \{a \in \Delta | AB(a)\} \cup \{a \in COMP - \Delta | \neg AB(a)\} \vDash$ A
$SD \cup OBS \cup \{a \in \Delta | AB(a)\} \cup \{a \in COMP - \Delta | \neg AB(a)\} \cup \neg$A $\vDash \varnothing$

Corolaire : Soit le système (SD, COMP, OBS, ACT), soit E est une extension de la théorie des défauts $DT = \left(SD \cup OBS, \left\{a \in ACT \left| \frac{:\neg AB(a)}{\neg AB(a)} \right.\right\}\right)$, alors $E \cap ACT$ est un contrôle symbolique du système (SD, COMP, OBS, ACT).

L'ensemble $\mho = \{A_i = E_i \cap ACT | E_i \; extension \; de \; DT\}$ est l'ensemble des actions symboliques possibles pour le système (SD, COMP, OBS, ACT). L'ensemble des actions symboliques déduit peut faire usage de métarègles pour ordonnancer les actions et simuler leur exécution.

3 Logique des défauts distribuée et application aux piétons virtuels

Dans ce travail, nous développons un cadre formel pour modéliser le contrôle symbolique fondé sur les théories avec défauts. Le formalisme que nous proposons réutilise les travaux de Reiter (1987) portant sur la modélisation logique du diagnostic des systèmes. Notre formalisme considère le contrôle du système de la même manière que le diagnostic, i.e. « le contrôle est un ensemble minimal d'actions qui sont en état anormal », où une action est anormale si son activation éventuelle n'est pas conforme avec les observations effectuées sur les composantes du système, ainsi qu'avec ses comportements prévus. Dans ce travail le contrôle symbolique sera associé à un diagnostic du système. Intuitivement, c'est l'état global du système caractérisé par son diagnostic au sens de Reiter, qui peut inférer les actions à

mener pour atteindre des objectifs. Nous définissons une société d'agents (Bajo et al. 2015) selon la construction ci-dessous :

Définition. Un système multi-agents (SMA) est une collection d'agents SMA = $\{(A)_i , i=1..n\}$, qui opèrent dans le même environnement. Les agents peuvent interagir avec l'environnement et les uns avec les autres en utilisant des interfaces, selon des protocoles et des langages de communication prédéfinis.

Le système multi-agents est asynchrone qui signifie que les agents peuvent effectuer leurs actions à tout instant tant qu'ils sont dans un état valide. A chaque agent de SMA est associée une théorie logique avec défaut $(W_i,\Delta_i) \mid i = 1..n$. Une théorie de défaut distribuée associée à SMA est l'ensemble des théories : DDT = $\{(W_i,\Delta_i) \mid i = 1..n \}$. Les propriétés et les problèmes d'inférence de cette structure ne seront pas traités dans ce le présent papier.

3.1 Les facteurs de risque

Les principaux risques pour les piétons sont relatifs à un large éventail de facteurs: Les véhicules en termes de conception surtout aux niveaux des fronts durs. La vitesse de déplacement des véhicules influence à la fois le risque d'accident et les conséquences de ce dernier. L'effet sur le risque d'accident provient principalement de la relation entre la vitesse et la distance d'arrêt. Par ailleurs, l'absence de visibilité claire à la fois des piétons et des conducteurs permet d'expliquer le taux élevé des accidents de piétons durant les mois d'hiver, du fait de l'obscurité et de mauvaises conditions climatiques. Le risque d'accidents des piétons augmente lorsque la conception des routes et de l'aménagement du territoire ne parviennent pas à planifier et fournir des installations comme les trottoirs, les passages, les points de refuges ou les médianes soulevées, ou un examen adéquat des accès des piétons aux intersections. Ces infrastructures permettent aux piétons de traverser la route en toute sécurité. D'autres facteurs contribuent également aux accidents des piétons. Ces facteurs concernent d'une part les conducteurs en termes de pratiques de conduite dangereuses, du taux d'alcoolémie élevé, la fatigue, la distraction y compris l'utilisation du téléphone mobile, les attitudes agressives, l'échec des conducteurs à respecter le droit de passage pour les piétons aux points de passage de ces derniers. Et d'autres part les piétons en termes des attitudes de prise de risque, du taux d'alcoolémie élevé, la distraction y compris l'utilisation du téléphone mobile, la réduction du temps de réaction et la réduction de la vitesse de marche pour les personnes âgées; l'incapacité des enfants à jauger la vitesse des véhicules et d'autres informations pertinentes pour une traversée sécurisée. Le paragraphe suivant traite la virtualisation logique des piétons et des conducteurs. Elle admet que la perception de l'espace-temps est fondamentale dans la modélisation de cette interaction. Les capteurs intelligents seront d'une grande utilité aussi pour capturer les comportements et les observations du système.

3.2 Acteurs logiques adductifs « piéton-conducteur »

Les systèmes complexes sont caractérisés par des dynamiques non linéaires, stochastiques et chaotiques. La systémique constitue le paradigme d'excellence pour modéliser cette classe de système. L'un des outils de l'approche systémique, très certainement aujourd'hui l'une des techniques de modélisation les plus utilisés est la modélisation par le système multi-agent (SMA). Un agent peut être défini comme un objet réactif communicant autonome

capable d'agir dans son environnement, ses actions étant dirigées vers un but, de percevoir et éventuellement de se représenter partiellement son environnement et d'interagir avec d'autres agents. Dans le cadre de ce travail, le système multi-agents construit est développé autour des théories avec défauts. Chaque agent A_i est une machine logique abstraite dénotée par le système :
(SD_i, $COMP_i$, OBS_i,ACT_i). La société d'agents est formulée par la théorie des défauts distribués$DDT = \{(W_i, \Delta_i) | i = 1..n\}$, où $W_i = SD_i \cup OBS_i$ et $\Delta_i = \left\{a \in ACT_i \left| \frac{:\neg AB(a)}{\neg AB(a)} \right.\right\}$

Dans l'objectif d'associer un ensemble d'actions à un diagnostic donné du système ; cette théorie pourra en particulier être dotée d'un ensemble de défauts relatifs au diagnostic. Dans la modélisation des interactions conducteur-piéton plusieurs facteurs sont à considérer. Dans ce papier de positionnement, pour des raisons de simplicité nous développons les théories des défauts seulement pour les piétons et conducteurs. Il est tout à fait intuitif de considérer les autres entités de l'environnement comme des agents influençant cette interaction.
La structure d'acteurs est schématisée dans la table donnée ci-dessous :

Acteur	Attribution
	Agent Piéton Virtuel (APV)
	Agent Conducteur Virtuel (ACV)

TAB. 1 – *Acteurs intelligents distribués associés au système conducteur-piéton.*

3.3 Théorie des défauts distribuée des agents intelligents

Un agent est défini par un système holistique approprié à la perception de sa finalité. Il est décrit par un tuple (DS, COMP, OBS, ACT), désignant respectivement la description du système par un ensemble de formules logiques de premier ordre, COMP, les composantes du système, un ensemble de constantes propositionnelles, OBS, les observations capturées du système, un ensemble de formules logiques de premier ordre, enfin ACT un ensemble de formules logiques de premier ordre correspondant aux actions à entreprendre par un agent intelligent.

3.3.1 Agent piéton virtuel (APV)

L'agent APV est destiné à s'approprier la perception de la réalité locale de l'environnement de son parcours (les obstacles, les point d'intérêts, la signalisation, la chaussé, etc.). Les éléments de sa théorie logique sont résumés dans la table 1.

Elément	Contenu
SD	Relations structurales
COMP	Les segments routiers adjacents à la zone de passage, les feux de signalisation, les obstacles, les panneaux publicitaires, les points d'intérêt, l'espace d'influence, les autres piétons, les voitures, etc.
OBS	Etat du trafic des segments, la dynamique des véhicules (temps de sécurité, temps de traversée, etc.)
ACT	Traverser avec une vitesse donnée, s'arrêter, courir.

TAB. 2 – *Théorie locale d'un agent de type piéton virtuel*

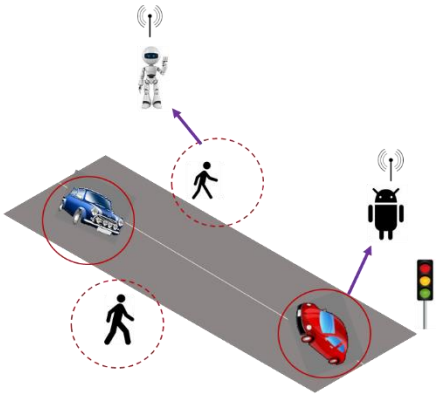


FIG. 3 – *Environnement conducteur-piéton.*

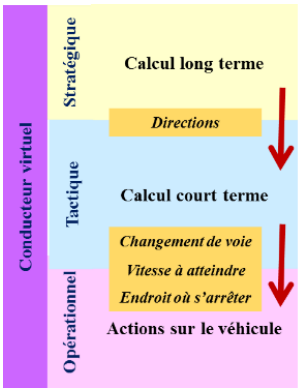


FIG. 4 – *Actions du conducteur virtuel*

La figure 3 illustre brièvement un contexte d'interaction conducteur-piéton. La figure 4 schématise les niveaux d'action du conducteur virtuel. Le système logique associé à un acteur de type piéton virtuel (APV) est donné ci-après :
Les composants correspondent aux segments du tronçon concerné par l'interaction. Cet ensemble est donné par $COMP = \{(F_i)_{i=1..f}, (S_i)_{i=1..n}, (SV_i)_{i=1..k}, (P_i)_{i=1..m}\}$. Cet ensemble de composants dénote les segments de la route, les points d'intérêts, la signalisation verticale et horizontale, les feux de signalisation, panneaux publicitaires, les obstacles, etc.
La description de ce système est formulée comme suit :
SD = {relations structurales, relations topologiques, relations de conflits, prédicats d'usage de la signalisation, etc.}

Pour décrire le comportement normal de l'action de traversée d'un piéton, la règle suivante permet de le spécifier avec le prédicat AB(.), qui sera ajouté à SD :
$APV(p,t,X,act) \wedge \mathbb{C}(act, [\![t, t+\varepsilon]\!], danger) \rightarrow AB(act)$
Le prédicat $APV(p,t,X,act)$ exprime la fait suivant : *le piéton p à l'instant t, en localisation X, a l'intention de déployer l'action act*

Le prédicat $\mathbb{C}(act, [\![t, t+\varepsilon]\!], danger)$ signifie : *Si l'action act est dangereuse sur l'intervalle de temps* $[\![t, t+\varepsilon]\!]$ *alors act est en état anormal.* Cette validité sera élaborée par des indicateurs de risque que nous avons produits dans nos travaux (Mandar, Boulmakoul, 2015, 2016).
Les observations du système concernent l'état de la signalisation, l'état de trafic de chacun des segments, appréciation de la vitesse des véhicules, le temps de sécurité des véhicules, le temps de traversé moyen des piétons, etc.

OBS = {*état de la signalisation, temps de sécurité d'arrêt, etc.}*
L'ensemble ACT de l'action considérée pour l'exemple ci-dessus prend la forme suivante :
$ACT = \{traverser, stop, courir, etc.\}$

3.3.2 Agent conducteur virtuel (ACV)

La conduite d'un véhicule est un processus complexe qui nécessite un traitement de l'information dynamique et temps réel de l'automobiliste pour s'adapter aux conditions de conduite rencontrées. La perception à elle seule est une source d'erreurs plus importante que l'action, la décision et les défaillances générales réunies. Dans le code de la route, le conducteur doit rester constamment maître de sa vitesse, et réguler cette dernière en fonction de l'état de la chaussée, des difficultés de la circulation, et des obstacles prévisibles. Le facteur principal sur lequel le conducteur peut agir est le temps de perception et de réaction face au danger. La théorie de l'agent conducteur virtuel ACV, dispose des éléments perçus au niveau de son environnement routier. Le tronçon de passage, les piétons, les autres voitures, le carrefour, et les éléments cognitifs de décision pour la conduite. La décomposition en niveau des activités du conducteur virtuel donnée dans (Michon, 1985) est structurée comme suit (voir Figure 4) : le niveau stratégique assure au conducteur l'établissement des plans à long terme, le niveau tactique lui confère le choix en fonction de ses buts stratégiques des buts à court et moyen terme, comme la sélection de voies, le choix d'une vitesse, etc. ; le niveau opérationnel attribue au conducteur l'exécution des commandes sur son véhicule en fonction des décisions imposées par le niveau tactique.

Elément	Contenu
SD	Relations structurales
COMP	Segments routier, signalisation, piétons, etc.
OBS	Etat du carrefour, Variables de trafic, vitesse de conduite, etc.
ACT	Actions de conduite: accélérer, ralentir, freiner, s'arrêter, etc.

TAB. 2 – *Théorie d'un agent conducteur virtuel.*

La description du système SD contient la règle suivante :
$ACV(c, t, X, act) \wedge \mathbb{C}(act, [\![t, t+\varepsilon]\!], danger) \rightarrow AB(act)$
Le prédicat $ACV(c, t, X, act)$ formule la fait suivant : *le conducteur c à l'instant t, en localisation X, a l'intention de déployer l'action act.* Le prédicat $\mathbb{C}(act, [\![t, t+\varepsilon]\!], danger)$ signifie : *Si l'action act est dangereuse sur l'intervalle de temps* $[\![t, t+\varepsilon]\!]$ *alors act est en état anormal.*

3.3.3 L'interaction APV ~ACV

Disposant des théories des défauts (W_i^c, Δ_i^c) et (W_j^p, Δ_j^p) associées respectivement aux agents ACV et APV (figure 5). Chacune de ces théories génère chaque période de temps des extensions. Pour un agent, une extension exprime un monde possible où les actions/tâches à entreprendre respectent son utilité et garantissent sa survie en assurant gestion de risque optimale. L'idée que nous soutenons ici, est de considérer la fusion des extensions où la consistance logique est respectée et par conséquent dans cette fusion il faudra chercher les conflits accidentogènes. L'autre possibilité concerne l'émergence des fusions contradictoires et ce sont ces structures logiques qu'elle faudrait aussi suivre pour mettre en exergue les

dangers à haut risque et dont les conséquences sont des accidents fortement probables. Ces éléments seront développés dans nos futurs travaux.

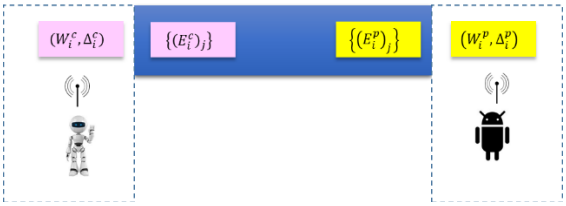


FIG. 5 – *Agents et théorie de défauts distribués : interaction de deux théories.*

Certaines considérations techniques ont été volontairement omises, et ce pour des considérations de protection industrielle. Une grande partie de ce travail est en cours d'évaluation dans le cadre de dépôt de brevet international déposé le mois de février 2016 et dont la phase d'évaluation n'est pas encore achevée.

4 Conclusion

Dans ce travail, nous proposons une théorie avec défauts distribuée pour modéliser le système *conducteur-piéton*. L'intelligence distribuée supportée par la logique non monotone répartie est mise à disposition de la modélisation des interactions conducteur-piéton, pour investir les scénarios accidentogènes. La prochaine étape vise le déploiement de cette approche et la comparaison avec l'approche basée sur théorie des ensembles flous. L'intégration de la psychologie comportementale est aussi nécessaire pour étudier en détail les profils des piétons et des conducteurs en simulation et en situation réelle.

Références

Bajo J., et al. (2015) Trends in Practical Applications of Agents, Multi-Agent Systems and Sustainability, ISBN 978-3-319-19628-2, Advances in Intelligent Systems and Computing, Volume 372, Springer International Publishing Switzerland 2015.

Davis M. (1980) The mathematics of non-monotonic reasoning. *Artificial Intelligence, Volume 13, Numbers 1,2 pp. 73-80, April 1980 (Special Issue)*.

Kleer J. Mackworth A., Reiter R. (1992) Characterizing diagnoses and systems, Artificial Intelligence 56 (1992) 197-222 Elsevier.

Mandar, M., Boulmakoul, A. (2016 a) Pedestrians' fuzzy intuitionistic risk exposure model: foundation and first development, onference: 5ème Edition du Workshop International sur l'Innovation et Nouvelles Tendances dans les Systèmes d'Information INTIS 2016, 25-26 Novembre 2016 - Fès, Maroc. ISBN 978-9954-34-378-4, ISSN 2351-9215, At Fes, Maroc.

Mandar, M., Boulmakoul, A. (2016 b) Pedestrian risk exposure using fuzzy ant model simulation, in 3rd IEEE International Conference on Logistics Operations Management - GOL'2016, Fes - Morocco.

Mandar, M., Boulmakoul, A. (2014) Virtual pedestrians' risk modelling, International Journal of Civil Engineering and Technology. 10/2014; 5(10):32-42.

Michon. J. (1985) A critical view of driver behavior models: what do we know, what should we do? In Human behavior and tra-c safety. L. Evans, R. Schwing, 1985.

Reiter R. (1980) A logic for default reasoning. *Artificial Intelligence, Volume 13, Numbers 1,2 pp. 81-132, April 1980 (Special Issue)*.

Reiter R.(1987) Theory of diagnosis from first principles, *Artificial Intelligence 32, pp. 57-95, 1987.*

Distributed non-monotonic logic for *driver-pedestrian* interactions analysis

Summary

In this work we define foundations of a logical theory of non-monotonic and distributed defaults to model the "driver-pedestrian" system. This process differs from the fuzzy and intuitionistic approach that we proposed in our previous work. Both approaches are naturally semantic. The fuzzy theory approach is symbolic, the present approach is logical in its ontology. Dynamic inferential processes between distributed intelligent actors are exploited. We prove the formal validity of our proposition and underline the contribution of the formalism of revisable reasoning and distributed intelligent systems knowledge representation for representation of the conductor-pedestrian interaction. The non-monotonic reasoning acquires the decision-making skills necessary for the actors whose aim is to deploy their behaviors, this ensures a reactivity and a better use of the local and neighboring knowledge. These logical components make it possible to understand the dynamics and the hypothetical exchanges between actors, where the basic facts and observations are often subjected to relative interpretation.

Vers l'utilisation des évidences syntaxiques, sémantiques et temporales dans le PRF pour améliorer la recherche d'information dans les tweets

Zouhel Boucetta*, Abdelkrim Bouramoul*
Mourad Bouznada***
* Département Informatique, Université de Saad Dahleb Blida,
Route Soumaa, Blida 9000-Algérie
boucetta.zouhel@gmail.com
** Département Informatique Fondamentale et ses Applications, Laboratoire MISC
Université Abdelhamid Mehri Constantine2
B.P. 325, Constantine 25017 - Algérie
abdelkrim.bouramoul@univ-constantine2.dz
*** Département Informatique Fondamentale et ses Applications, Laboratoire MISC
Université Abdelhamid Mehri Constantine2
B.P. 325, Constantine 25017 - Algérie
mourad.bouznada@univ-constantine2.dz

Résumé. La recherche d'informations dans le corpus des tweets est une tâche difficile considérant le volume accru de ce dernier, la taille courte des tweets et la qualité du langage utilisé pour écrire ces documents. L'expansion de la requête via le Pseudo Relevance Feedback (PRF) est une technique qui a réussi à apporter des solutions à ces problématiques. Néanmoins, les termes d'expansion sélectionnés par le PRF peuvent être non pertinents vu que le modèle de recherche utilise simplement la similarité syntaxique pour la sélection des tweets pertinents. Pour remédier à ce problème, nous proposons dans cet article une approche qui combine trois types d'évidences : sémantiques, temporales et syntaxiques pour le reclassement de la première liste des résultats du PRF et par la suite sélectionner les plus pertinents pour le choix des termes d'expansion qui vont servir à l'élargissement de la requête lors de la prochaine session de recherche. Dans notre proposition, nous avons pris également en considération l'avis de l'utilisateur concernant la pertinence temporale des résultats.

1 Introduction

De nos jours, les plates-formes de microblogging sont les réseaux sociaux les plus récents et les plus utilisés du Web 2.0. Elles présentent une masse volumineuse d'informations. Twitter est le service de microblogging le plus populaire avec 320 millions d'utilisateurs actifs par mois et plus de 500 millions de tweets envoyés par jour [1]. Ce volume de publications complique l'opération d'accès à l'information par les Microblogeurs.

Le tweet est un document court dont la longueur ne dépasse pas 140 caractères. Ecrit

[1] http://www.blogdumoderateur.com/chiffres-reseaux-sociaux/

souvent avec un langage mal orthographie, contenant des abréviations et des argots afin de transcrire l'information avec un nombre minimum de caractères.

La recherche d'informations dans le corpus des tweets présente un véritable défi selon Choi et al. (2012) pour les modèles actuels de recherche d'informations, cela est dû au volume du corpus d'une part et aux caractéristiques des tweets d'autre part. En effet, quand l'utilisateur soumet une requête, le modèle de recherche sera confronté à deux problèmes : d'abord l'absence des termes de la requête dans le tweet, et le fait que chaque terme apparait au plus une seule fois dans le texte.

La sélection des tweets via un Pseudo Relevance Feedback (Rocchio, J. J. (1971)) se base sur un appariement syntaxique entre la requête et les tweets. De ce fait, il ya une grande probabilité que dans le Top de la liste figurent des tweets non pertinents, ce qui va affecter la pertinence des termes d'expansion. Pour améliorer le classement des tweets pertinents et obtenir ainsi de meilleurs termes pour l'expansion de la requête, beaucoup de travaux ont introduit les évidences temporales (Miyanishi et al. (2013a), Willis et al. (2012)) dans leurs propositions en les combinant avec les évidences syntaxiques pour le reclassement des tweets résultats de la première recherche.

C'est dans ce contexte que s'inscrit notre travail qui vise à améliorer l'efficacité de la recherche dans les tweets, mais en combinant trois sources d'évidence au lieu de deux (syntaxique, temporale et sémantique).

Cet article s'articule autour de deux sections, la première sera consacrée à la présentation des travaux de l'état de l'art, la deuxième présente notre système de sélection des termes d'expansion pour la reformulation de la requête. Elle décrit l'architecture de l'approche proposée. Nous terminons cet article par une conclusion et des perspectives.

2 Etat de l'art

L'opération de recherche des microblogs pertinents présente un handicap pour les outils de recherche, vue l'évolution exponentielle du corpus et les propriétés des tweets. Pour surmonter ce défi, plusieurs travaux ont traité cette thématique en démontrant l'importance de l'aspect temporal pour améliorer l'efficacité de la recherche. En effet, Efron et Golovchinsky (2011) ont proposé une méthode basée sur l'estimation bayésienne pour l'incorporation du temps dans le modèle de recherche afin de montrer l'efficacité de l'introduction de la dimension fraicheur, considérée comme importante pour de nombreux types de requêtes. Massoudi et al. (2011) ont présenté un modèle pour la récupération des microblogs en se basant sur l'expansion de la requête avec des termes récents. Ils ont utilisé aussi dans leur Framework des indicateurs de qualité spécifiques pour les microblogs. Willis et al. (2012) ont proposé trois méthodes pour la recherche temporale des tweets, la première favorise les termes récents ayant une cooccurrence élevée avec tous les termes de la requête, la deuxième favorise les tweets pertinents qui appartiennent aux périodes de grande concentration des tweets, la troisième favorise les termes qui appartiennent à des tweets pertinents qui figurent dans les grandes concentrations des tweets et qui ont une occurrence élevée avec tous les termes de la requête. Dakka et al. (2012) ont construit une Framework, en introduisant le temps dans leur modèle de recherche. Ils ont proposé des techniques automatiques pour identifier les périodes du temps important pour une requête. Miyanishi et al. (2013a) ont proposé trois méthodes pour l'expansion temporale de la requête. Deux méthodes individuelles basées sur la variation temporelle et la fraicheur (TVQE et TRQE) et leur combinaison (TVRQE) pour surmon-

ter les limites des méthodes individuelles. Miyanishi et al. (2013b) ont proposé aussi une approche en deux étapes pour l'expansion de la requête: la première étape consiste à lancer une première recherche, puis donner à l'utilisateur la possibilité de choisir un tweet de la liste des résultats, les termes de ce dernier sont utilisés comme des termes d'expansion dans le deuxième tour de recherche, la deuxième consiste à utiliser les évidences lexicales et les profils temporaux de la requête et des tweets dans un modèle de pertinence pour le choix des meilleurs termes d'expansion. A cet effet, Efron et al. (2014) ont proposé une framework, qui se base sur le feedback et l'hypothèse du cluster temporal, pour le choix des tweets pertinents qui se regroupent ensemble dans le temps. A notre tour, nous visons à améliorer la pertinence de la recherche des tweets par la contribution suivante:

1. Donner à l'utilisateur la possibilité de choisir la pertinence temporale qui convient à sa requête ;
2. La prise en compte de deux types d'évidences temporales: fraicheur, concentration des tweets et leur combinaison, pour estimer l'importance temporale d'un tweet pour une requête.
3. La considération de l'aspect sémantique de la requête.
4. Combiner les trois sources d'évidence syntaxique, temporale et sémantique pour le reclassement des tweets.

3 Architecture générale de notre solution

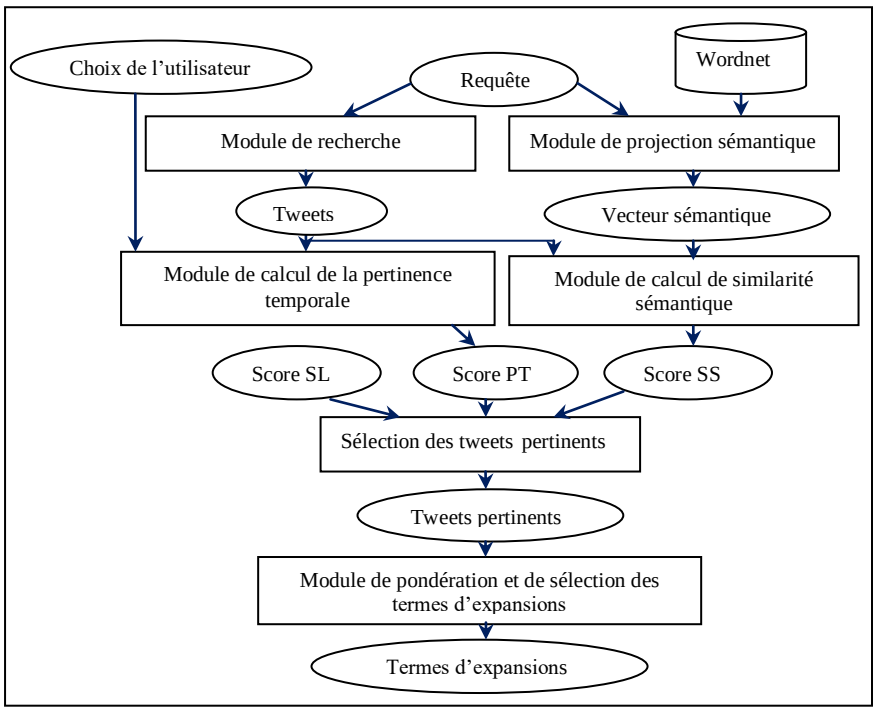


FIG. 1 – *Architecture général de notre système.*

Dans cette section, nous détaillons l'architecture de notre système qui est la concaténation d'un ensemble de modules qui se complètent pour améliorer la recherche d'informations dans les tweets. Notre système offre à l'utilisateur la possibilité de choisir la pertinence temporale qui convient sa requête. Ce dernier peut choisir de restituer les tweets récents du PRF, ou bien ceux qui figurent dans les grandes regroupements des tweets, ou seulement les plus récents dans ces grandes concentrations. Le score temporal de chaque tweet est calculé selon le type de pertinence sélectionné par l'utilisateur. La similarité sémantique entre la requête et chaque tweet du PRF est calculé pour favoriser les tweets proches sémantiquement de la requête. Notre système combine les évidences temporales, sémantiques et syntaxiques (score de Lucene) pour reclasser les tweets résultats du PRF. Ceci permettra d'avoir les documents performants dans le top de la liste. Ces derniers sont utilisés par la suite pour extraire les termes utiles pour l'expansion de la requête. Les deux modules (PS) et (MC) utilisés dans notre architecture, ont été proposés dans notre travail précédent Bouramoul et al. (2011) dans le cadre de la recherche d'informations dans le web. La vue d'ensemble de notre architecture est représentée dans la figure1.

Dans ce qui suit, nous présentons les différents modules de notre architecture.

3.1 Module de recherche(RM)

Dans le cadre de ce travail, nous avons opté pour le moteur de recherche Lucene open source, vu qu'il est le moteur le plus utilisé dans la tache de recherche d'informations dans les microblogs dans TREC. La recherche du texte intégral par Lucene se fait sur deux étapes (Smart (2006)):

- Créer un index pour tous les tweets du corpus ;
- Analyser les requêtes soumises par l'utilisateur, puis fournir une liste des résultats de recherche, dont l'index a été construit à l'étape précédente.

Le module de recherche transmet la requête de l'utilisateur au moteur de recherche, indexe la requête puis lance sa recherche par index dans le corpus des tweets, à la fin il récupère les documents pertinents syntaxiquement à la requête. Le module de recherche récupère par la suite la liste des documents triés selon leur score de pertinences syntaxiques calculé par Lucene.

3.2 Le module de calcul du score temporal (STM)

La recherche des tweets par la prise en compte du temps, consiste à utiliser leurs dates de soumission pour estimer leurs importances visant à vie une requête. Le score de chaque tweet récupéré du module de recherche est calculé selon une des trois techniques inespérées des travaux de l'état de l'art (Dakka et al. (2012), Miyanishi et al. (2013a), Efron et al. (2014), Choi, J., W. B.Croft (2012)), nous détaillons chacune par la suite.

Le score fraicheur. Dans ce cas, les tweets pertinents sont ceux du PRF, dont leurs dates de publication est proche de celle de la soumission de la requête. L'estimation de la fraicheur d'un document revient a calculer sa proximité temporale de la requête. Le score de la fraîcheur du tweet t pour une requête Q, est calculé comme suit:

$$SPT1(Q,t) = \frac{(Dt-DQ)}{\sigma} \quad (1)$$

Dt représente la date de publication du tweet t et Dq la date de soumission de la requête Q. σ est la différence temporale entre Dq et la date de soumission du plus ancien tweet dans PRF.

Le score concentration temporale. Le but ici est de déterminer la pertinence temporale des tweets selon leur distribution avec les retweets dans le temps. Sachant que nous avons utilisé comme granularité temporale le jour (Dakka et al. (2012)). Nous avons classé les tweets et les retweets du PRF dans un histogramme par date de publication des tweets. Les documents qui appartiennent à la même barre ont le même score de pertinence. La barre qui contient le plus de tweets et retweets, est la plus importante selon l'aspect concentration temporale. Un tweet est pertinent s'il appartient au grand regroupement. Nous avons défini le score de concentration d'un tweet par la formule suivante :

$$SPT2(Q,t) = \frac{card(s_p)}{card(s)} \quad (2)$$

Avec card(s) : le nombre total des tweets et retweets publiés avant la soumission de la requête et $card(s_p)$ et le nombre des tweets et retweet publiés dans le jour « p ».

Le score combinaison. Nous visons ici à sélectionner les tweets récents dans les grandes concentrations des tweets. Cela est réalisable par la combinaison des scores SPT1 et SPT2. Alors la pertinence temporale d'un tweet t publié au jour Dt par rapport à une requête Q publiée au jour DQ est mesurée comme suit :

$$SPT3\ (Q,t)\ =\ SPT1(Q,t)\ *\ SPT2(Q,t)\ \ (3)$$

3.3 Module de projection sémantique (PSM)

Afin de prendre en compte la sémantique de la requête pour favoriser les tweets du PRF similaire sémantiquement avec cette dernière, nous associons à chaque terme de la requête l'ensemble des mots qui lui sont sémantiquement liés. L'idée est de projeter les termes de la requête sur les concepts de l'ontologie WordNet en utilisant les deux relations sémantiques : 'synonymies' et 'hyperonymies' pour extraire les différents sens de la requête. Par la suite l'ensemble des concepts récupérés pour chaque terme sont utilisés en conjonction avec le terme lui-même lors de la pondération par le module de calcul. L'objectif est de favoriser un tweet qui contient des mots sémantiquement proches à ceux que l'utilisateur cherche, même si ces mots n'existent pas comme termes dans la requête. Nous utilisons, à cet effet, l'ontologie WordNet selon le principe suivant : au départ nous accédons à la partie de l'ontologie contenant les concepts et les relations sémantiques, ces derniers sont utilisés pour récupérer tous les synsets et hyperonymies relatifs à chacun des termes de la requête. Ces derniers sont enfin utilisés pour la construction du vecteur sémantique qui contient pour

chaque terme de la requête, les synonymes et les hyperonymes appropriés. La figure 1 illustre le fonctionnement du module de projection sémantique.

3.4 Module de calcul (CM)

Une fois que le vecteur sémantique est construit, par le module de projection sémantique, le module de calcul procède à la construction des vecteurs tweets et du vecteur requête à base des coefficients calculés à l'aide de la fonction de pondération appropriée (formule 4). Le module de calcul mesure par la suite la similarité sémantique entre ces deux vecteurs en utilisant la fonction de calcul de similarité entre deux vecteurs (formule 5).

Le fonctionnement de ce module est réalisé donc en deux étapes, la formule utilisée dans chaque étape est proposée dans Salton (1971):

Pondération des termes. Cette étape calcule le poids de chacun des termes du tweet. Elle se déroule comme suit : Un coefficient t_{ji} du vecteur tweet Tj mesure le poids du terme i dans le tweet j. la formule de pondération des termes des tweets est la suivante :

$$t_{ji} = \frac{occ(w)}{card(t_j)} \quad (4)$$

occ(w) c'est le nombre d'occurrences du terme w dans le tweet , card(t_j) est le nombre de termes du tweet t_j.
Un coefficient q_{ki} du vecteur requête Q_k mesure le poids du terme w dans tous les tweets.

Appariement « tweets/requête». La comparaison entre le vecteur tweet et celui de la requête revient à calculer un score qui représente la pertinence sémantique du tweet vis-à-vis la requête. Alors la similarité entre la requête et un tweet n'est que le cosinus entre leurs vecteurs. La formule est donnée par la suite :

$$PS(Q_k, T_j) = \frac{\sum_{i=1}^{M} q_{ki} - t_{ji}}{\sqrt{\sum_{i=1}^{M} q_{ki}^2 \sum_{i=1}^{M} t_{ji}^2}} \quad (5)$$

M est le nombre des termes d'un document, q_{ki}-t_{ji} est la différence entre le poids d'un terme q_{ki} de la requête et le poids du terme de même rang du tweet t_{ji}.

3.5 Le module du calcul de score final de pertinence (PFM)

Pour prendre en considération les différentes sources d'évidence : syntaxique, temporale et sémantique, pour le calcul de la pertinence finale de chaque tweet t pour une requete Q, nous avons proposé une formule linéaire qui cumule le score de similarité sémantique SS et le score de pertinence temporale SPT (SPT1ou SPT2 ou SPT3) et le score du Lucene SL. Chaque score est pondéré par son poids d'importance (ω ou λ ou (1- ω - λ)). La formule est détaillée par la suite :

$$PF(Q,t) = \omega * SS(Q,t) + \lambda * SPT(Q,t) + (1 - \omega - \lambda) * SL(Q,t)$$
$$\omega \epsilon\ [0,1]\ et\ \lambda\ \epsilon\ [0,1] \quad (6)$$

3.6 Sélection des tweets pertinents (SM)

Le fonctionnement de ce module s'effectue selon les étapes suivantes**:**
- Sélectionner la première liste des résultats de recherche;
- Récupérer le score final de pertinence de chaque tweet «PF»;
- Classer la liste suivant l'ordre décroissant du «PF»;
- Sélectionner les α Top tweets « αT» pour extraire les termes d'expansion;

3.7 La pondération des termes d'expansion(PTM)

Ce module se charge de calculer la fréquence de chaque terme dans « αT» avec la formule (4), puis choisir les β termes les plus fréquents pour l'expansion de la requête. Comme il pondère les termes Q_0 de la requête originale avec un poids Ω, et les termes d'expansion Q_E avec un poids (1-Ω). Pour favoriser en termes d'importance les termes de Q_0 par rapport aux termes Q_E.

4 Conclusion et perspective

Dans cet Article nous avons présenté une approche pour la recherche des Microblogs pertinents, en combinant les deux méthodes suivantes: le PRF et l'expansion de la requête dont l'objectif principal est d'améliorer les performances de recherche dans le Corpus des tweets. Notre contribution est divisée en trois grandes parties: la première consiste à offrir à l'utilisateur la possibilité de proposer manuellement le type de pertinence temporale qui convient à sa requête, ce qui permettra de diriger le système pour la sélection des tweets pertinents. La seconde consiste à proposé trois techniques pour la recherche des tweets pertinents temporellement à la requête, à savoir, la technique basée sur la fraicheur, celle basée sur la concentration des tweets et une technique basée sur leur combinaison. La troisième prend en considération l'aspect sémantique de la requête. La dernière consiste à combiner trois évidences, temporelles, sémantiques et syntaxiques dont le but est d'améliorer le classement des tweets pertinents de la liste des résultats du PRF.

L'expérimentation de notre travail est en cours de réalisation. Nous avons opté pour le Corpus de test de la tache microblogs de TREC 2011, afin de tester la performance de notre système.

Références

Bouramoul, A., M.K. Kholladi, B.L. Doan (2011). *How Ontology Can be Used to Improve Semantic Information Retrieval: The AnimSe Finder Tool*. In International Journal of Computer Applications (IJCA) – ISSN : 0975 - 8887, Vol.21, No.9 : 48-54.

Efron, M. (2011).*The university of illinois graduate school of library and information science at TREC 2011*. In TREC.

Efron, M., J. J. Lin, J.He, P. Arjen (2014). *Temporal feedback for tweet search with non-parametric density estimation*.In SIGIR, 33-42.

Dakka, W., L.Gravano, P.G. Ipeirotis (2012). *Answering general time-sensitive queries*. In TKDE 24(2):220–235.

Choi, J., W. B.Croft (2012). *Temporal Models for Microblog Retrieval.* In CIKM'12.

Massoudi, K., M.Tsagkias, M .Rijke, W.Weerkamp (2011). *Incorporating query expansion and quality indicators in searching microblog posts*. In ECIR, 362–367.

Miyanishi, T., K. Seki, K.Uehara (2013a).*Combining Recency and Topic-Dependent Temporal Variation for Microblog Search*. In Proceedings of the 35th European Conference on Information Retrieval, 331– 343.

Miyanishi, T., K. Seki, K. Uehara (2013b). *Improving pseudo-relevance feedback via tweet selection.* In CIKM, 439–448.

Rocchio, J. J. (1971). *Relevance feedback in information retrieval The SMART retrieval system - experiments in automatic document processing*. (G. Salton ed), Chapter 14, pp 313-323.

Smart, J. F. (2006). Web Mining Lab in UCLA.

Salton, G. (1971). *A comparison between manual and automatic indexing methods*. Journal of American Documentation, 20(1):61–71.

Willis, C., R. Medlin, J. Arguello (2012). *Incorporating Temporal Informationin Microblog Retrieval* . In Proceedings of the Twenty-First Text REtrieval Conference.

Summary

The information retrieval in the corpus of tweets becomes a difficult task considering its increased volume, also the short size of the tweets and the quality of the language used to write these documents. The expansion of the query via the pseudo relevance feedback (PRF) and among the Techniques that have managed to overcome this challenge. But the expansion terms selected by the PRF may be irrelevant view that the search model simply uses syntactic similarity for selecting relevant tweets. To fix this problem we have presented an approach to extend the query with terms of expansion, it is based on three types of evidences: temporal, syntactic, and semantic. We have also considered the opinion of the user concern the temporal interval of the results.

Contrôle de la conformité organisationnelle basée sur la mesure des distances de partitions de l'ensemble des complexes simpliciaux

Zineb BESRI*,
Azedine BOULMAKOUL**

*Ecole Nationale des Sciences Appliquées de Tetouan BP : 2222 M'hannech
z.besri@gmail
**Faculté des Sciences et Techniques de Mohammedia BP 146 Mohammedia 20650 Maroc
azedine.boulmakoul@gmail.com

Résumé. Les organisations multinationales de tous les secteurs d'activité doivent se conformer aux lois, règlements et politiques en matière de protection de la vie privée et des données, conçus pour protéger les renseignements confidentiels et sensibles. La conformité exige que les organisations adoptent et mettent en œuvre diverses activités qui consistent à veiller à ce qu'elles disposent d'un personnel professionnel dédié à la conformité ainsi que des technologies permettant de réduire les risques. Dans ce contexte s'inscrit notre papier. Ce travail de recherche propose une nouvelle approche de mesure de conformité par la mesure des distances de partitions. Ces partitions représentent les complexes simpliciaux résultant de l'analyse simplicial de la structure organisationnelle de l'entreprise. Le contrôle de son degré de conformité par rapport à la stratégie du top management de l'entreprise. On propose des mesures alternatives nouvelles de la conformité par le calcul de distance entre deux partitions quelconques d'un ensemble P et d'une distance principale existante, à savoir la distance de partition $\Theta(.,.)$. La comparaison permet de vérifier leurs restrictions à des éléments modulaires de la structure, ainsi qu'en termes de réorganisation appropriés.

Mots clé : Conformité, distance partition, topologie algébrique, analyse structurale, refonte organisationnelle.

1 Introduction

Au cours de la dernière décennie, un intérêt considérable a été porté à la mesure de la distance entre les partitions (ainsi qu'entre et / ou dans les collections de partitions). La question se pose, en général, lorsque l'on fait des comparaisons de similitude entre les graphes. (S. Ben-Davidet al, 2006), (A. D'yachkov et al, 2006), (D. Gusfield, 2002), (D. A. Konovalov et al, 2005), (C. Yu, B. C. Ooi et al, 2001)

Le problème de la quantification de la distance entre les partitions d'un ensemble fini est ici approché avec une cible combinatoire spécifique, en ce que la mesure proposée vise à tenir compte des relations, de rencontre et d'union de l'ensemble de partition exactement de la même manière que les distances de Hamming (J. Pinto da Costa, 2004) entre les sous-

ensembles fait avec l'inclusion, l'intersection et l'union. En d'autres termes, l'objectif est de reproduire la différence symétrique entre les sous-ensembles lors de la mesure des distances entre les partitions.

Les organisations multinationales de tous les secteurs d'activité doivent se conformer aux lois, règlements et politiques en matière de protection de la vie privée et des données, conçus pour protéger les renseignements confidentiels et sensibles. La conformité exige que les organisations adoptent et mettent en œuvre diverses activités coûteuses liées aux processus, aux personnes et aux technologies. Ces activités consistent à veiller à ce qu'elles disposent d'un personnel professionnel dédié à la conformité ainsi que des technologies permettant de réduire les risques. Dans ce contexte s'inscrit notre papier. Ce travail de recherche propose une nouvelle approche de mesure de conformité par la mesure des distances de partitions. Ces partitions représentent les complexes simpliciaux résultant de l'analyse simplicial de la structure organisationnelle de l'entreprise à contrôler son degré de conformité par rapport à la stratégie du top management de l'entreprise. On propose des mesures alternatives nouvelles de la conformité par le calcul de distance entre deux partitions quelconques d'un ensemble P, ainsi qu'une distance principale existante, à savoir la distance de partition $\Theta(.,.)$. La comparaison permet de vérifier leurs restrictions à des éléments modulaires de la structure, ainsi qu'en termes de réorganisation appropriés.

La vérification de conformité, également appelée analyse de conformité, vise à détecter les incohérences entre un modèle de processus et son journal d'exécution correspondant et leur quantification par la formation de métriques.

Cet article est structuré comme suit : section 2 nous rappelons le principe de l'analyse structurale et la méthode canonique Q-analysis. La section 3 propose la nouvelle approche de mesure de conformité par distance partition. Puis un exemple illustrant la contribution de cette mesure dans le contrôle de conformité et aidant à la prise de décisions pour une éventuelle refonte organisationnelle. Enfin une conclusion pour synthétisé notre travail de recherche.

2 L'analyse Simplicial

L'analyse structurale, dans les sciences humaines, est une approche interdisciplinaire qui se fonde sur le postulat que les acteurs sociaux se caractérisent par leurs relations plutôt que par leurs attributs (le sexe, l'âge, la classe sociale, etc.). Ces relations ont une plus ou moins grande densité, la distance qui sépare deux acteurs est plus ou moins grande, et certains acteurs occupent des positions plus centrales que d'autres. Des théories permettent d'expliquer ces phénomènes, dont celle des liens forts et des liens faibles, et celle sur les trous structuraux où se trouvent les acteurs qui ne peuvent communiquer entre eux que par l'intermédiaire d'un tiers. L'analyse structurale est la modélisation de ces relations et l'analyse de leur impact sur la performance d'une entreprise. Nous modélisons ces relations au moyen d'une simple association 1/0 (oui / non). Par exemple, si un employé exécute une seule activité spécifique dans l'entreprise, l'employé sera "associé à cette activité » et non pas « associée » à toute autre activité. L'employé peut aussi être un membre d'une unité organisationnelle si une association serait modélisée entre l'employé et l'unité d'organisation.

Soit I un ensemble d'éléments et D une base de données de transactions, où chaque opération a un identifiant unique (tid) et contient un ensemble d'éléments. L'ensemble de tous les identifiants tids est désigné par T. Les entrées de la base de données sont des relations

$\lambda \subseteq$ IxT. Quand un article i survient dans une transaction t, nous notons (i,t)$\in \lambda$. Un ensemble X $\subseteq$ I est aussi appelé un jeu d'éléments et un ensemble Y $\subseteq$ T est appelé tidset. Un itemset avec des articles de k items est appelé un k-itemset. Pour un itemset X, on note son tidset correspondant comme $t(X) = \bigcap_{x \in X} t(x)$, l'ensemble de tous les tids des transactions qui contiennent X comme un sous-ensemble. Pour un tidset Y, on note son itemset correspondant comme $i(Y) = \bigcap_{y \in Y} i(y)$, c'est à dire, l'ensemble des éléments communs à toutes les transactions avec les tids dans Y. Le support d'un itemset X, notée σ(X) est le nombre de transactions dans lesquelles elle se produit comme un sous-ensemble, c'est-à-dire, σ(X) = t(X). Un itemset est fréquent si son support est supérieure ou égale à un support minimum (min_sup) de valeur spécifié par l'utilisateur σ(X) $\geq$ min_sup . Un itemset fréquent est appelée maximal s'il n'est pas un sous-ensemble d'un autre motif fréquent. Soit c : P(I)$\rightarrow$P(I) l'opérateur de fermeture, définie comme c(X) = i(t(X)) où X $\subseteq$ I un itemset fréquent. X est fermé si et seulement si c(X) = X est alternativement, un itemset fréquent. X est fermé s'il n'existe pas d'un sous-ensemble propre Y, X $\subset$ Y avec σ(X) =σ(Y).

2.1 Chaines de q-connection dans K

Ayant deux simplexes σ_p, σ_r dans K, on peut dire qu'ils sont joint par une chaine de (JH Johnson. (1982) s'il existe une séquence finie de simplexes $\sigma_{\alpha 1}$, $\sigma_{\alpha 2}$, … , $\sigma_{\alpha h}$ telle que :

1. $\sigma_{\alpha 1} \leq \sigma_p$,
2. $\sigma_{\alpha h} \leq \sigma_r$,
3. $\sigma_{\alpha i}$, $\sigma_{\alpha i+1}$ ont une face partagée $\sigma_{\beta i}$ (i = 1, …, h-1).

Cette séquence est dite chaine de q-connexion (q-connectivité) si q est le dernier des entiers $\sigma_{\alpha 1}$, $\sigma_{\alpha 2}$, … , $\sigma_{\alpha h}$. La longueur de la chaîne sera prise comme (h-1) et, en cas de besoin de la chaîne peut être désignée par $[\sigma_p, \sigma_r]_q$.

2.2 Q-analysis

Cette méthode permet d'analyser les systèmes de structures à la fois au niveau global (système dans son ensemble) et au niveau local (niveau d'éléments qui sont connectés les uns aux autres pour former une structure), ainsi que pour l'estimation de la complexité structurale de systèmes sur la base des résultats d'une telle analyse (L Duckstein, et al,1997). Cette approche connue sous le nom de Q-analyse (ou, dynamique polyèdre), utilise les idées de (C.Dowker, 1952) (A.C Gatrell et al 1982), (CL Freeman, 1980)comme fond mathématique. Le Q-analyse est basé sur les relations de q-proximité et de q-connectivité entre les simplexes d'un complexe donné (ou complexe simplicial). Le Q-analyse d'un complexe K détermine le nombre de classes d'équivalence distinctes, ou les composants q-connectés, pour chaque niveau de dimension q allant de 0 à q-1. Les classes d'équivalence sont déterminées par une règle de la manière suivante : Si deux simplexes sont q-connectés (soit q-proche ou q-relié), alors ils sont dans la même classe. Pour mieux voir cela, nous introduisons, pour un q fixé, une relation γ_q sur les simplexe de K, définie par:

- $(\sigma_p, \sigma_r) \in \gamma_q$ si et seulement si σ_p est q-connecté à σ_r . La relation γ_q est une relation réflexive, symétrique et transitive donc une relation d'équivalence.
- Les classes d'équivalences de la relation γ_q sont les éléments membre de l'ensemble quotient K/γ_q, et constitue une partition de tous les simplexes du complexe K qui sont de

l'ordre ≥ q. Nous notons la cardinalité de K/γ_q par Q_q. Cela est égale au nombre des composantes distinctes q-connectées de K. Quand nous analysons K par la recherche de toutes les valeurs de Q_0, Q_1,…, Q_N où N= dim K, nous disons que nous avons effectué une Q-analyse sur K. Pour chercher toutes les faces partagées q-valeur entre toutes les paires de Y dans KY(X;λ), les étapes suivantes peuvent être effectuées :

1. A partir de ΛxΛ T ,
2. Evaluer Λx Λ T -Ω =(ω_{ij}), avec ω_{ij} = 1

3 Distance partition

Mesurer le degré de conformité entre les deux structures organisationnelles, est une question de comparaison entre les deux structures (la structure Le réelle fournie par l'audit organisationnel (Boulmakoul et al 2014) traditionnel et la structure formelle proposée par le top management) et de calculer la distance entre elles. Par la suite, évaluer si elle est conforme ou non par rapport à la structure référentielle. La vérification de conformité, également appelée analyse de conformité, vise à détecter les incohérences entre un modèle de structure organisationnelle et son équivalent sur la réalité de l'entreprise et leur quantification par la formation de métriques.

On propose des mesures alternatives nouvelles de la conformité par le calcul de distance entre deux partitions quelconques d'un ensemble P, ainsi qu'une distance principale existante, à savoir la distance de partition Θ(. , .). La comparaison permet de vérifier leurs restrictions à des éléments modulaires de la structure, ainsi qu'en termes de réorganisation appropriés.

3.1 Définitions

Soient P et Q deux partitions dans l'ensemble X de n éléments avec respectivement p et m classes. Nous admettons que p ≤ m. P^q={$C_1^q, C_2^q \dots C_p^q$}et Q^q={$C_1^q, C_2^q \dots C_m^q$}, définit par la relation d'équivalence γ_q sur Ω. q est le maximum de la dimension de la partition.

Soit Θ: PxP → $|R^+$, la distance entre deux partitions, Θ(P,Q).

Le nombre minimum de transferts pour changer P en Q, noté Θ(P^q,Q^q),est obtenue par l'établissement d'une bijection entre les classes de P^q et ceux de Q^q en gardant un nombre maximum d'éléments dans des classes correspondant à ceux qui ne nécessitent pas d'être déplacé. Par conséquent, nous commençons à ajouter (m – p) classes vides à Pq, de sorte que Pq est considéré comme une partition avec p classes.

Soit γ la correspondance de P × Q → N qui associe à une paire de classes le cardinal de leur intersection. Classiquement, $n_{i,j}$ = | C_i ∩ C'_j | Et n_i = | C_i | Et n'_j = | C'_j | Dénotent les cardinaux des classes.

Soit Δ la correspondance qui associe à chaque paire de classes (C_i, C'_i) le cardinal de leur différence symétrique, notée $\delta_{i,j}$. Nous avons δ(i, j) = n_i + n'_j − 2 × n_i. Nous considérons donc le graphe bipartite complet $K_{q,q}$ dont les sommets sont les classes de P et Q, avec des arcs pondérés soit par γ soit par Δ.

La bijection minimisant le nombre de transferts entre deux partitions avec q classes P et Q correspond à un assemblage de poids maximum ω_1 en K_q, q pondéré par γ ou de manière équivalente, à un assemblage du poids minimum ω_2 dans K_q, q pondéré par Δ; de plus,

$\Theta(P^q,Q^q) = n - \omega_1 = \omega_2 /2$
si $\dim(P^q) \neq \dim(Q^q)$ alors $\Theta(P^q,Q^q) = \infty$,
si $\cup_q\{C_i^P\} \neq \cup_q\{C_i^Q\}$ alors $\Theta(P^q,Q^q) = \infty$,

Et la distance entre les deux partitions P et Q est définie par:

$$\Theta(P^q,Q^q) = \sum_{q=0}^{\min(\dim(P),\dim(Q))} \Theta^q(P^q, Q^q)$$

Par exemple, nous avons deux partitions P^q et Q^q (figure 1)

λ	F1	F2	F3	F4	F5	F6	F7
L1	1	0	0	0	0	0	0
L2	1	1	0	0	0	0	0
L3	1	0	1	0	0	0	0
L4	1	1	1	1	0	0	0
L5	0	1	0	1	1	0	0
L6	0	0	0	0	1	1	1
L7	0	0	0	1	1	0	1
L8	0	1	0	0	0	0	0
L9	0	0	1	1	0	0	1
L10	0	0	0	0	0	1	1

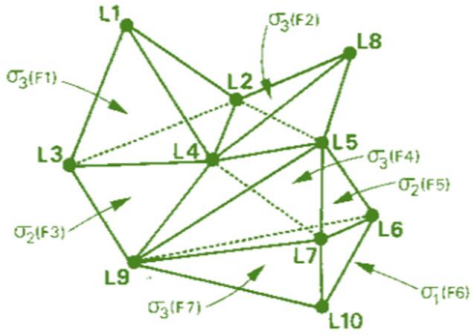


Figure 1 le complexe assemblé $KF(L,\lambda^{-1})$

Table 1 Deux exemples de partitions

P	Q
q=3 $\{L_4\}$	
q=2 $\{L_4\},\{L_5\},\{L_6\}, \{L_7\}, \{L_9\}$	q=2 $\{L_4\}$
q=1$\{L_{2,} \dots L_{7,} L_{9,} L_{10}\}$	q=1 $\{L_4\},\{L_5\},\{L_6\}$
q=0 $\{L_{1,} \dots L_{10}\}$	q=0$\{L_1\},\{L_{2,} \dots L_{6,} L_8\}, \{L_{7,} L_{10}\}$

Pour q= 3 nous avons $\Theta(P^3,Q^3) = \infty$ pour le reste des niveaux de connectivité nous trouvons les valeurs suivantes :

Table 2 Calcul de distance entre les partitions P et Q

Distance partition
q=3 $\Theta(P^3,Q^3) = \infty$
q=2 $\Theta(P^2,Q^2) = 4$
q=1 $\Theta(P^1,Q^1) = \infty$
q=0 $\Theta(P^0,Q^0) = \infty$

3.2 Control de conformité

Les exigences de conformité découlent de plus en plus des normes industrielles émergentes, des lignes directrices interministérielles ou éthiques.

L'idée principale est de mesurer le degré de non-conformité entre la véritable organisation d'entreprise et celle formelle proposée par le top management. Nous utilisons les deux méthodes pour diagnostiquer et extraire la structure organisationnelle réelle d'une entreprise.

(Boulmakoul et al 2014) Maintenant, comment pouvons-nous comparer et contrôler la conformité entre les deux structures organisationnelles ?

En utilisant des techniques de distance de partition, nous pouvons calculer la distance entre les partitions générées à partir de la topo-scopie pour chaque niveau dimensionnel des composants q-connectés.

Nous proposant des niveaux de conformité par intervalle de mesure de distance

- Si $\Theta(P^i,Q^i) = 0$ alors élément P^i conforme à Q^i
- Si $\Theta(P^i,Q^i) \neq 0$ et ne tends pas vers ∞ alors P^i est partiellement conforme à Q^i
- Si $\Theta(P^i,Q^i) = \infty$ alors P^i n'est pas conforme à Q^i

Le tableau suivant interprète les distances calculées par leur degré de conformité

Table 3 Degré de conformité par niveau de classe d'équivalence

Distance partition	Degré de conformité
q=3 $\Theta(P^3,Q^3) = \infty$	Non-Conforme
q=2 $\Theta(P^2,Q^2) = 4$	Conformité partielle
q=1 $\Theta(P^1,Q^1) = \infty$	Non-Conforme
q=0 $\Theta(P^0,Q^0) = \infty$	Non-Conforme

A partir de ces résultats et principalement par le moyen correspondance par degré de conformité. Le top-management va avoir une sorte de carte qui met en relief les zone qui nécessitent une refonte organisationnelle pour être plus aligné avec leur stratégie.

Le système de conformité ciblé et le résultat de l'auto-évaluation sont utilisés pour identifier les lacunes et de définir les mesures nécessaires pour les fermer. Par exemple, une banque mondiale universelle qui a appliqué la méthode de l'autoévaluation a révélé que son alerte précoce pour les expositions individuelles des entreprises était inadéquate, voire en baisse de certaines exigences réglementaires. Les lacunes identifiées et les mesures définies fournissent à la haute direction une vue d'ensemble des lacunes les plus importantes afin que les dirigeants puissent formuler des recommandations finales sur les mesures d'atténuation.

Il est essentiel à cette étape d'avoir une discussion ouverte avec la direction et les experts à travers les divisions sur les causes profondes des carences qui ont été découverts et les mesures définies pour affiner et créer à travers l'organisation. Par exemple, la haute direction d'une banque mondiale universelle a découvert que la gestion des limites était considérablement compromise parce qu'il était impossible d'agréger les expositions entre les unités en temps opportun en raison des problèmes de qualité des données provoqués par des interfaces insuffisantes de différents systèmes informatiques

4 Conclusion

Dans ce papier, nous rappelons l'approche d'analyse simplicial et la représentation de la structure organisationnelle de l'entreprise à travers l'ensemble des complexes simpliciaux. C'est un prés-requis pour pouvoir générer les partitions à évaluer leurs conformités. Nous proposons une nouvelle mesure qui calcul exactement le degré de conformité entre la structure réelle et celle de faite. En calculant le nombre d'itérations et changement à faire pour atteindre la structure référence.

En perspective appliquer cette nouvelle approche dans une étude de cas réelle pour l'évaluer et voir ses limites afin de fournir une mesure applicable dans le processus d'audit et de contrôle organisationnelle

Références

A. D'yachkov, V. Rykov, D. Torney, and S. Yekhanin. (2006). On application of the partition distance concept to a comparative analysis of psychological or sociological tests. Stochastic Analysis and Applications, 24:61–78.

A.C Gatrell and JR Beaumont. (1982).An introduction to Q-Analysis. Geo Abstracts.: Nor-Wich,

Azedine Boulmakoul and Zineb Besri. 2014. "Processus analytique structurale d'évaluation et de refonte organisationnelle fondée sur le réfentiel CMMI-DEv," 37042,

C. H Dowker, (1952). "Homology groups of relations," JStor, vol. 56, no. 1, pp. 84-95.

C.L Freeman, (1980)."Q-Analysis and the structure of friendship networks," Int J Man Mach Stud, vol. 12, no. 1, pp. 367-78.

C. Yu, B. C. Ooi, K.-L. Tan, and H. V. Jagadish. (2001). Indexing the distance: an efficient method to KNN processing. In Proceedings of the 27th International Conference on Very Large Data Bases, VLDB '01, pages 421–430,

2001.

D. A. Konovalov, B. Litow, and N. Bajema. (2005) Partition-distance via the assignment problem. Bioinformatics, 21(20):3912–3917.

D. Gusfield.(2002) Partition-distance: A problem and a class of perfect graphs arising in clustering. Information Processing Letters, 82:159–164.

JH Johnson. (1982) "Q-Transmission in simplicial complexes," Int J Man-Mach Stud , vol. 16, no. 4, pp. 351-77.

J. Pinto da Costa and P. Rao. (2004). Central partition for a partition-distance and strong pattern graph. REVSTAT – Statistical Journal, 2(2):127–143.

L Duckstein, A Steven, and Nobe, (1997). "Q-analysis for modeling and decision making.," European journal of operational Research, vol. 103, pp. 411-425.

M. Meil˘a. (2008). Local equivalences of distances between clusterings. Technical report, University of Washington, Department of Statistics.

S. Ben-David, U. von Luxburg, and D. P´al. (2006). A sober look at clustering stability. In Learning Theory - Lecture Notes in Computer Science, volume 4005/2006, pages 5–19.

Summary

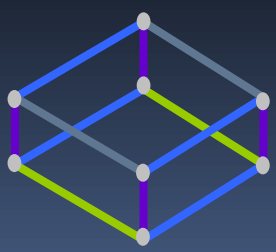
Multinational organizations in all areas of business comply with privacy and data laws, regulations and policies for sensitive and sensitive information. Compliance requires organi-

zations to adopt and implement a variety of activities to ensure that they have professional compliance staff and technologies to reduce risk. In this context fits our paper. This research proposes a new approach to measuring conformance to the measurement of partition distances. These partitions represent the simplicial complexes resulting from the simplicial analysis of the organizational structure of the company. Controlling the level of compliance with the company's senior management strategy. On proposal of the new alternative measures of conformity by the computation of distance between two partitions of the set a set of the main distance existing, namely the distance of the partition Θ (.,.). The comparison makes it possible to check their restrictions on the modular elements of the structure, as well as the appropriate reorganization terms.

10ième édition de la Conférence Maghrébine sur les Avancées des Systèmes Décisionnels

27-29 avril 2017, Tabarka — Tunisie

L'importance accordée par la communauté scientifique et par les industriels à l'informatique décisionnelle (ou *Business Intelligence*) ne cesse d'augmenter, comme en témoigne le nombre de travaux théoriques et d'outils mis sur le marché. En effet, les systèmes décisionnels permettent à l'entreprise de prendre des décisions à différents niveaux hiérarchiques en analysant son existant et son passé pour mieux prédire le futur. Sur le plan de la recherche, la conférence maghrébine sur les Avancées des Systèmes Décisionnels (ASD) est dédiée aux systèmes décisionnels permettant de consolider les efforts des chercheurs d'une part, et de répondre aux aspirations des professionnels d'autre part. Les contributions que cette édition accueille portent en particulier sur les thèmes suivants : architecture, organisation et conception des entrepôts de données, modélisation multidimensionnelle, sémantique et ontologies décisionnelles, big data et entrepôts de données, entrepôts de données complexes, business intelligence et cloud computing, systèmes d'information décisionnels et applications industrielles …



Edition 2017