9ème édition

Conférence sur

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

ASD 2015

ASD 2015

Actes de la 9ème édition

Conférence sur

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

Edités par

Hassan Badir, Azedine Boulmakoul et Omar Boussaid

10-12 septembre 2015

Tanger, Maroc

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), 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 2015 est accueillie cette année par le Maroc.

ASD 2015 ambitionne de consolider les expériences conduites par les chercheurs, industriels et utilisateurs issus de communautés travaillant sur les systèmes décisionnels. L'objectif de cette neuvième 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 2015 a reçu 34 soumissions d'articles en provenance de six pays (Algérie, France, Lybie, Maroc, Tunisie). Après évaluation par les membres du comité scientifique, composé par 59 chercheurs-experts internationaux du domaine, 15 articles longs et 9 articles courts ont été retenus. Ces papiers couvrent différents thèmes de recherche et d'application sur les systèmes décisionnels.

ASD 2015 est organisé par l'Ecole Nationale des Sciences Appliquées de Tanger, Maroc et a reçu son soutien ainsi que celui de différentes institutions publiques d'enseignement et de recherche que nous tenons à remercier : l'Université Abdelmalek Essâadi, l'Ecole des Nouvelles Sciences et Ingénierie de Tanger (ENSI), Le Centre National pour la Recherche Scientifique et Technique (CNRST), le Laboratoire LabTIC de l'université Abdelmalek Essâadi, le Laboratoire ERIC de l'Université Lyon 2, l'Université HASSAN II Mohammedia-Casablanca, la Faculté des Sciences et Techniques de Mohammedia, la Faculté des Sciences Economiques et de Gestion de Sfax ; le Centre de Recherche en Informatique, Multimédia et Traitement Numérique des Données de Sfax, l'association

AMINTIS, et toutes les autres institutions qui ont aidé de loin ou de près pour la réussite de cette manifestation.

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

Nous sommes très reconnaissants de leur soutien.

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

Les éditeurs
H. BADIR, A. BOULMAKOUL et O. BOUSSAID

Présidents de la conférence

- BADIR Hassan, (Ecole Nationale des Sciences Appliquées de Tanger, Maroc)
- BOULMAKOUL Azedine, (Université Hassan II, Maroc)

Président du comité d'organisation de la conférence

- BADIR Hassan, (Ecole Nationale des Sciences Appliquées de Tanger, Maroc)

Comité de pilotage

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

Comité scientifique

- ABDI Mustapha K., Université d'Oran, Algérie
- ACHI Abdelkader, Université Paris 1, France
- AHMED NACER Mohamed, USTHB Alger, Algérie
- AHMED OUAMER Rachid, Université Tizi Ouzou, Algérie
- AL ACHHAB Mohammed, ENSA of Tetouan, Maroc
- AL-MALAISE AL-GHAMDI Abdullah, FCIT, King Abdulaziz University, KSA
- ASFARI Ounas, Université Lyon2, France
- AYACHI Sonia, ISG, Sousse, Tunisie
- BADACHE Nadjib, CERIST Alger, Algérie
- BADARD Thierry, Université Laval, Canada
- BADIR Hassan, ENSAT, Maroc
- BADRI Abdelmajid, Université Hassan II, Maroc
- BELHAJJAM Khalid, Paris-Dauphine University, LAMSADE, France
- BELLAFKIH Mostafa, INPT Rabat, Maroc
- BELLATRECHE Ladjel, ENSMA Poitiers, France
- BEN ABDALLAH Hanene, Université de Sfax, Tunisie
- BEN BLIDIA Nadjia, Université de Blida Algérie
- BENHARKAT Nabila, INSA de Lyon, France
- BENSLIMANE Djamel, Université de Lyon1, France
- BENTAYEB Fadila, Université Lumière Lyon 2, France
- BIMONTE Sandro, Cemagref, Clermond-Ferrand, France
- BOUCELMA Omar, Université d'Aix-Marseille, France

- BOUFAIDA Mahmoud, Université deConstantine, Algérie
- BOUKHALFA Kamel, USTHB, Alger, Algérie
- BOUKRAA Doulkifli, Univesrité de Jijel, Algérie
- BOULMAKOUL azedine, Université Hassan II, Maroc
- BOURAMAOUL Ramzi Abdelkrim, Université de Constantine, Algérie
- BOUSSAID Omar, Université Lumière Lyon 2, France
- CHKOURI Mohamed Yassin, ENSA of Tetouan, Maroc
- DARMONT Jérôme, Université de Lyon2, France
- EL HEBIL Farid, INPT Rabat Maroc
- EL-MOUADIB Faraj, FIT, Benghazi, Lybie
- FAVRE Cécile, Université Lyon 2, France
- FEKKI Jamel, Université de Sfax, Tunisie
- FISSOUNE Rachida, ENSA of Tangier, Maroc
- GARGOURI Faiez, Université de Sfax, Tunisie
- GHOZZI Faiza, Université de Sfax, Tunisie
- HACHAICHI Yasser, Université de Sfax, Tunisie
- HARBI Nouria, Université Lyon 2, France
- HIDOUCI Walid, ESI Alger, Algérie
- IDRISSI Abdellah, FSR,Université Mohammed V, Rabat, Maroc
- JERBI Houcem, University College Dublin, Ireland
- KABACHI Nadia, Université Lyon1, France
- KAZAR Okba, Université Biskra, Algérie
- KHOLLADI Med-khireddine, Université de Constantine, Algérie
- LEMIRE Daniel, Université du Québec à Montréal, Canada
- MAHDAOUIi Latifa, USTHB, Alger, Algérie
- MARGHOUBI Rabia, Université Hassan II, Maroc
- MELIT Ali, Université de Jijel, Algérie
- MEZIANE Abdelkrim , CERIST, Algérie
- MISSAOUI Rokia, Université d u Québec en Outaouais, Canada
- MOUSSA Rim, University of Carthage, Tunisia
- MOUSSAOUI Abdelouaheb, Université de Sétif, Algérie
- NABLI Ahlem, Université de Sfax, Tunisie
- OMARY Fouzia, FSR,Université Mohammed V, Rabat, Maroc
- OUKID Saliha, Université de Blida, Algérie
- OULAD HAJ THAMI Rachid, ENSIAS Rabat, Maroc
- RAVAT Frank, Université de Toulouse, France
- REGUIEG F Zohra, Université de Blida, Algérie
- SEKHRI Larbi, Université d'Oran
- SIDHOM Sahbi, Université de Nancy, France
- SLIMANI Yahya, FS, Tunis, Tunisie
- TESTE Olivier, Université de Toulouse, France
- ZAROUR Nasreddine, Université de Constantine, Algérie
- ZEGOUR Djamel Eddine, ESI d'Alger, Algérie

Comité d'organisation

- BADIR Hassan , ENSA de Tanger, Maroc
- BELMOKADDEM Houda, ENSA de Tanger, Maroc
- BENSBIH Said, CBI Casablanca, Maroc
- BOULMAKOUL Azedine , Université Hassan II, Maroc
- CHAHBOUNE Asaad, ENSA de Tanger, Maroc
- CHAHBOUNE Nouha, ENSA de Tanger, Maroc
- DERRHI Mustapha, ENSA de Tanger, Maroc
- EL HADDAD Mohamed, ENSA de Tanger, Maroc
- EZZINE Abdelhak, ENSA de Tanger, Maroc
- FISSOUNE Rachida, ENSA de Tanger, Maroc
- FILALI Otman, ENSA de Tanger, Maroc
- LAAZIZ Yassine, ENSA de Tanger, Maroc
- LAHJOMRI Fouad, ENSA de Tanger, Maroc
- MOALLA Mohamed Sahbi, ISET Sfax - Tunisie
- MOUSSA Ahmed, ENSA de Tanger, Maroc
- OUALKADI Ahmed, ENSA de Tanger, Maroc
- SAMADI Hassan, ENSA de Tanger, Maroc
- SBIHI Abderrahmane, ENSA de Tanger, Maroc

ASD'2015

Conférence sur les Avancées des Systèmes Décisionnels

10-12 septembre 2015, Tanger, Maroc

Sommaire

Système d'aide à la décision pour la surveillance d'ouvrage d'art

Derkaoui Aicha
Centre des Techniques Spatiales, Arzew
derkaouia@hotmail.com

Résumé. La sécurité des ouvrages d'art nécessite un contrôle périodique des structures. L'évaluation de l'état de l'ouvrage ainsi que son comportement ne peut se faire sans une surveillance régulière et rapprochée, ceci par la mise en place d'un réseau d'auscultation géométrique. L'apport d'un outil d'analyse dans la gestion de ce réseau est le but du présent travail. L'idée est de proposer un système d'aide à la décision interactif par l'intégration de l'outil SIG et de la technologie OLAP afin de profiter des avantages de chacun pour garantir de meilleures analyses multidimensionnelles et spatiales et faciliter la prise de décision. Il s'agira donc de combiner les capacités cartographiques des systèmes d'information géographique pour le traitement des données géographiques et leur visualisation cartographique avec les outils d'exploration et d'analyse OLAP. Le cadre applicatif de nos travaux se situe dans le domaine de surveillance d'ouvrage d'art. Les données utilisées pour illustrer notre contribution concernent le réseau d'auscultation géométrique du bac de stockage GL4Z.

Mots clés: Système d'Information Géographique (SIG); On Line Analytical Processing (OLAP); Système d'Aide à la Décision (SAD).

1 Introduction

Les structures d'ingénieries telles que les bacs de stockage sont constamment sujettes aux déformations et déplacements sous les contraintes des charges internes et externes qui s'exercent sur leurs structures. Pour s'assurer de leur sécurité, prévenir des détériorations couteuses, vérifier les critères de la construction et suivre leurs comportement en général, une évaluation précise de leurs déplacements dans le temps est nécessaire, pour ce, le principe de base des mesures géodésiques qu'est la détermination à une date donnée, de la position des points d'un réseau établi sur le site d'étude est appliqué. Ainsi, leurs déplacements relatifs permettent de définir les déformations du terrain. Le canevas mis en place pour la surveillance périodique des bacs, se présente comme un outil pour la surveillance et la prévention des risques qui pourraient y surgir, tandis que la détermination de positions des cibles permet de quantifier leurs déplacements par rapport à une position dite initiale, afin de parvenir à une décision.

Le concept de la surveillance a évolué à travers les années ; celle-ci comporte non seulement l'observation des tendances temporelles et spatiales et de ses facteurs déterminants mais aussi la détection des problèmes en émergence dans une perspective d'analyse d'impact des programmes mis en place. L'objectif fondamental de la surveillance reste toujours l'analyse continue de données pertinentes pour la prise de décision. L'horizon de la surveillance s'est donc diversifié, le temps réel est une réalité et les nouveaux défis sont

maintenant liés à la pertinence de données souvent moins spécifiques et à une interprétation juste des résultats obtenus à partir des méthodologies de plus en plus sophistiquées.

L'efficacité de la prise de décision repose sur la mise à disposition d'informations pertinentes et d'outils adaptés. Des solutions connues sous le terme d'OLAP Spatial, qui visent à intégrer la donnée spatiale dans l'OLAP (Bédard, 2009), ont donc été développées. L'OLAP Spatial (SOLAP) a été défini par Yvan Bedard comme "une plateforme visuelle conçue spécialement pour supporter une analyse spatio-temporelle rapide et efficace à travers une approche multidimensionnelle qui comprend des niveaux d'agrégation cartographiques, graphiques et tabulaires" (Bédard, 2002). Le SOLAP enrichit les capacités d'analyse des systèmes OLAP classiques car la visualisation des mesures sur une carte permet de comprendre la distribution géographique d'un phénomène et de mettre en relation les différents phénomènes spatiaux par rapport aux axes d'analyse alphanumériques, et de comparer ces phénomènes à diverses granularités géographiques.

La méthodologie développée a comporté plusieurs phases dont les principales sont :

- Elaboration du modèle décisionnel dédié à la surveillance d'ouvrage d'art en se basant sur la technologie SOLAP. L'approche décisionnelle suggérée est structurée selon les trois étapes de Pictet, à savoir l'étape de structuration, d'exploitation et de concrétisation des résultats.
- Application de l'approche décisionnelle élaborée dans la surveillance du Bac de stockage de GL4Z d'Arzew en détaillant les étapes d'exploitation du modèle décisionnel. Cette conception a été, également, illustrée par la modélisation multidimensionnelle de l'entrepôt de données et des schémas illustratifs du cube de données généré, et par une analyse multidimensionnelle et spatiale à travers l'utilisation de deux puissants outils, l'Olap et le système d'information géographique

2 Etat de l'art sur les travaux connexes

Au cours des dernières années, de nombreux travaux et études ont été réalisés dans le domaine de l'aide à la décision spatiale utilisant la technologie SOLAP dans divers domaines d'application. Ces études sont principalement ceux qui sont dans le centre de recherche en géomatique de l'Université Laval à Québec. Cette nouvelle discipline scientifique définit à la fois le concept et la technologie combinant les capacités cartographiques des systèmes d'information géographique (SIG) aux outils d'aide à la décision OLAP (On-Line Analytical Processing) (Le Rubrus, 2009). On peut citer le travail de (Rosemarie et al., 2007), où les auteurs ont utilisé une approche SOLAP intégré pour développer un système pour explorer l'espace-temps banque de données interactive de l'information d'entreprise en collaboration avec le ministère des Transports du Québec. Sandro présente une introduction de la composante sémantique de l'information géographique et la flexibilité de l'analyse spatiale dans les systèmes Spatial OLAP (Sandro). Une autre application SOLAP est proposée. Les auteurs décrivent comment le SOLAP, peut renforcer la composante technologique des SIG participatifs (PPGIS). Basé sur un cas simulé d'audience publique, la recherche vise à démontrer le potentiel des outils SOLAP pour supporter et améliorer l'accessibilité et l'analyse de l'information géospatiale dans le contexte d'une application PPGIS de gestion environnementale. Dans le travail (Veilleux et al., 2004], une approche intégrée SOLAP a été

développé dans le domaine du sport pour analyser la performance des athlètes par rapport à la météo et le système mécanique. C'est en utilisant un positionnement par satellite (GPS), des outils d'exploration et d'analyse SOLAP fournir une meilleure évaluation et un meilleur suivi des athlètes.

Notre étude vise à fournir un système qui aidera les décideurs à mieux comprendre les déplacements planimétrique et altimétrique de l'ouvrage afin de parvenir à une décision claire sur la stabilité ou non de la structure concernée.

3 Approche décisionnelle adoptée

3.1 Auscultation

L'auscultation regroupe tous les dispositifs permettant de mesurer des grandeurs physiques susceptibles d'évoluer dans le temps, de façon à mettre en évidence son comportement et les phénomènes évolutifs significatifs de son vieillissement. Pour ce faire, on procède à la mesure des déplacements, des déformations, des contraintes, des pressions, des débits… etc. Parmi ces méthodes, on retrouve l'auscultation géométrique.

Les mesures d'auscultation des ouvrages d'art sont d'une très grande importance. Elles ont en effet deux buts principaux :

- Contrôle de la stabilité de l'ouvrage et préviennent des risques majeurs entrainés par sa rupture.
- Permet de suivre le comportement de l'ouvrage dans le temps, c.à.d. quantifier les déplacements et les déformations de certaines parties ou de la totalité de l'ouvrage.

3.2 Méthodologie de l'auscultation géométrique

Ce type de réservoir est bien adapté à l'auscultation par GPS où toute la précision standard de cette technique (EMQ=5mm ±1 ppm) peut être exploitée.

La méthodologie d'auscultation pour ce type de bac se résume comme suit :

1. Implantation du réseau de base: choix des points pas trop éloignés de l'ouvrage sur des sites stables et matérialisation par des bornes en béton. L'altitude des stations de base a été choisie presque identique à celle du bac pour minimiser l'influence de la troposphère.
2. Configuration optimale des cibles
   - Nombre et distribution homogène des points cibles
   - Matérialisation durable des points.
3. Suivi de l'évolution ou de la stabilité de l'ouvrage :
   - Opération 0 : observation en mode statique et détermination des coordonnées des points du réseau qui servira de configuration de référence.
   - Opération i : détermination des positions de ces mêmes points à partir d'une nouvelle campagne d'observations GPS pour quantifier les déplacements de l'ouvrage par rapport à la configuration de référence. La nouvelle configuration ainsi déterminée (opération i) servira de configuration de référence pour la prochaine opération d'auscultation (i+1). Les fréquences d'intervention varient selon l'amplitude des déformations et la vitesse de déplacement du bac et suite à des événements (séismes, glissement de terrain, ...).

3.3 Configuration du réseau d'auscultation du bac de stockage

La configuration générale du canevas d'auscultation du bac de stockage, arrêtée après reconnaissance, est composée de quatre types de réseaux: (réseau de garde, réseau d'auscultation, réseau de cibles et réseau de repères secondaires) (Taibi et al., 2008).

- **Réseau de base:** 15 bornes éloignées de quelques dizaines de mètres du bac.
- **Réseau de cibles:** 42 cibles collées sur la couronne du bac.
- **Réseau de repères secondaires:** 56 piquets bétonnés répartis sur le terrain avoisinant le bac et 19 plaques métalliques scellées sur le trottoir entourant le bac.

3.4 Modèle décisionnel proposé

En s'appuyant sur les éléments de l'auscultation géométrique telle qu'elle est déroulée sur le terrain, nous avons proposé le modèle décisionnel illustré par la Figure 1. Ce modèle suit les étapes de l'approche décisionnelle de Pictet (Hamdadou, 2008). La Figure 1 résume les différentes étapes des phases de structuration, d'exploitation et de concrétisation.

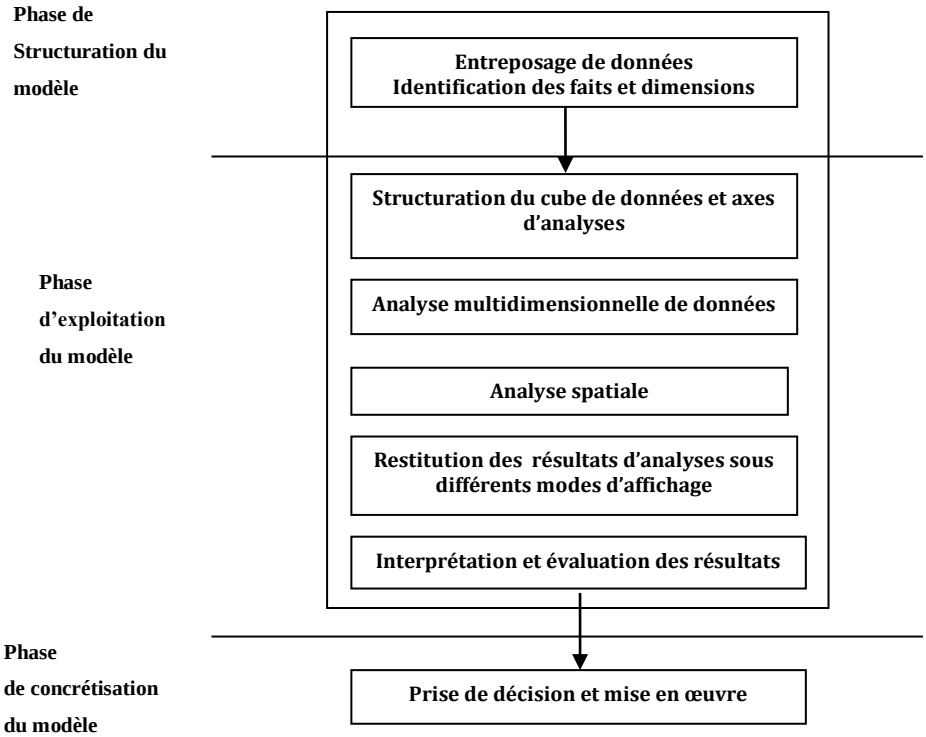


Fig. 1- *Modèle d'Aide à la décision pour l'auscultation géométrique*

3.4.1 Phase de structuration du modèle

Dans cette phase, les données de la surveillance issues des différentes campagnes d'observation subissent une série de traitements dans un processus d'entreposage de données ETL (Extract, Transform, Load) puis intégrées dans un seul entrepôt de données.

La base de données relationnelle est structurée selon un modèle particulier appelé modèle en étoile.

3.4.2 Phase d'exploitation du modèle.

Les données sont présentées sous une vue multidimensionnelle avant de les présenter au module client selon différents modes d'affichage :

- Graphique : sous forme de tableaux, des histogrammes ou camemberts;
- Cartographique par un système d'information géographique.

L'outil OLAP servira pour la navigation dans l'entrepôt de données**.** L'analyse multidimensionnelle permet de connaitre, mesurer et prévoir (prise de décision) au travers de la manipulation des données du magasin.

L'analyse spatiale est une activité essentielle pour la cartographie et les prises de décision. Le SIG servira pour l'analyse spatiale. Il demeure un outil dans le processus décisionnel et il devient un outil intégrateur en fournissant des informations claires et précises aux décideurs, afin de prendre une décision plus juste et éclairée. Les SIG représentent aujourd'hui le meilleur outil pour le traitement numérique de l'information géographique et leur définition les présente notamment comme des outils d'analyses (Hamdadou, 2008).

3.4.3 Phase de concrétisation des résultats.

Cette phase s'intéresse à l'interprétation des résultats d'analyse et la découverte de la connaissance afin de faciliter la prise de la décision.

4 Application de l'approche décisionnelle

4.1 Présentation du bac en sol gelé

Le réservoir en sol gelé du terminal méthanier du complexe SONATRACH GL4/Z d'Arzew, construit en 1965 de capacité d'environ 38.000 m³, le seul mode de stockage souterrain en exploitation dans le monde. Il représente pour le complexe plus de 50% de ses capacités en stockage. Ce réservoir a un diamètre de 37.20 mètres et une profondeur de 36 mètres; il se situe à 100 mètres environ du bord de la mer.

4.2 Intégration des données

L'approche ETL (Extraction, Transformation, Loading) est l'approche traditionnelle pour alimenter un entrepôt de données. Cette phase a pour but d'intégrer les données issues des différentes campagnes d'observations à travers un processus d'entreposage de données ETL, les données de l'auscultation vont être emmagasinées dans un seul entrepôt de données.

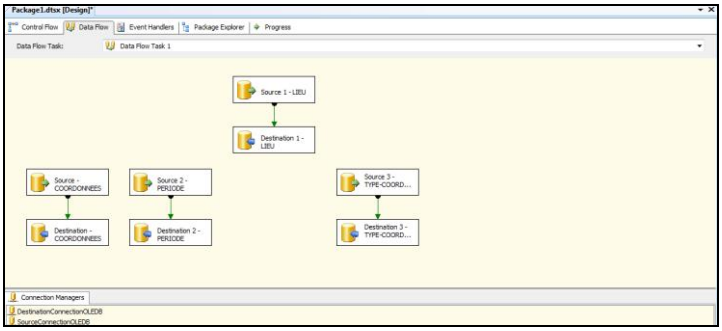


Fig. 2- *Intégration des données*

4.3 Modélisation multidimensionnelle des données

La modélisation constitue une démarche d'investigation permettant de gérer un réseau. C'est une étape fondamentale dans la conception des bases de données (Le Rubrus, 2009).

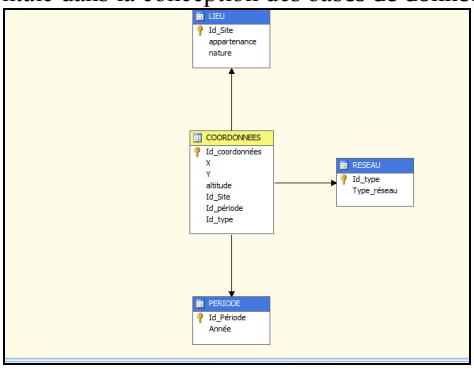


Fig. 3- *Modélisation de l'entrepôt à l'aide d'un schéma en étoile*

4.4 Analyse multidimensionnelle

4.4.1 Exploration du cube

La navigation est un terme utilisé pour décrire le processus employé pour explorer un cube de façon interactive, habituellement en utilisant un client OLAP graphique connecté à un serveur OLAP. Le processus utilisateur interactif pour une requête multidimensionnelle est appelé "Slicing" et "Dicing". Le résultat d'une requête multidimensionnelle est soit une cellule, un "slice" bidimensionnel, ou un sous-cube multidimensionnel.

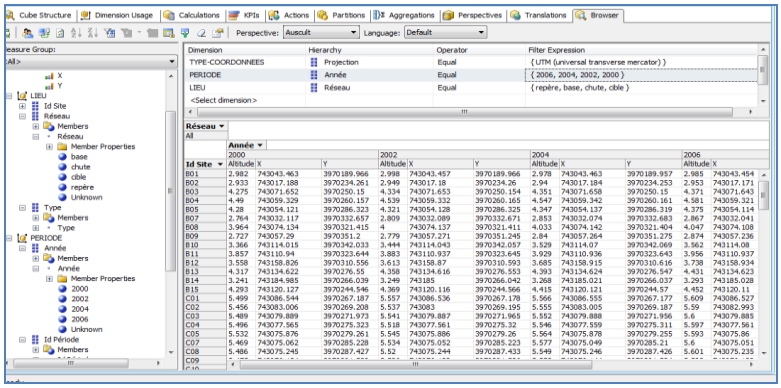


Fig. 4 - *Exploitation du cube de données*

Exemples d'exploration de données par filtre :

Une coupe sur 2006, les données affichées concernent uniquement cette année ou par année et type de réseau, le tableau concerne uniquement les données du réseau de repères :

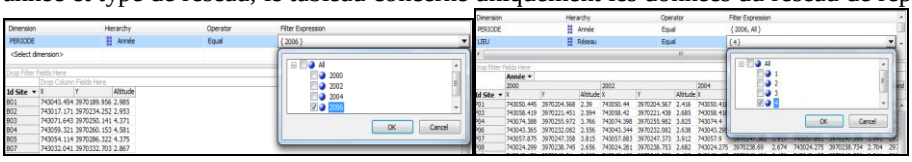


Fig. 5- *Opération slice et dice*

4.4.2 Détermination des déplacements du réseau.

Le langage MDX, est un langage d'interrogation des bases multidimensionnelles plus adapté que le classique SQL pour le traitement des requêtes de type OLAP. MDX signifie "Multi-Dimensional Expressions"

Afin de calculer les différents déplacements durant ces périodes d'observations, on exécute quelques requêtes MDX :

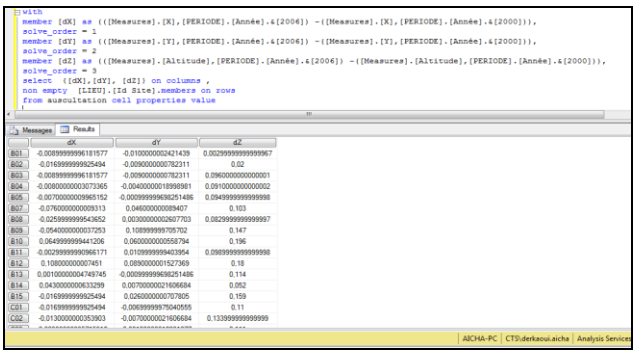


Fig. 6 - *Résultat de la requête de calcul du déplacement*

4.4.3 Restitution des résultats d'analyse

Un entrepôt de données doit permettre de trouver et d'extraire l'information, de la stocker et de l'interroger, de l'analyser et de l'enrichir, et d'offrir des outils de visualisation et de reporting. Cet outil basé sur une base multidimensionnelle permet une interaction rapide et intuitive pour naviguer à travers l'information.

Les figures suivantes illustrent les variations du réseau de base en X, Y et en altitude.

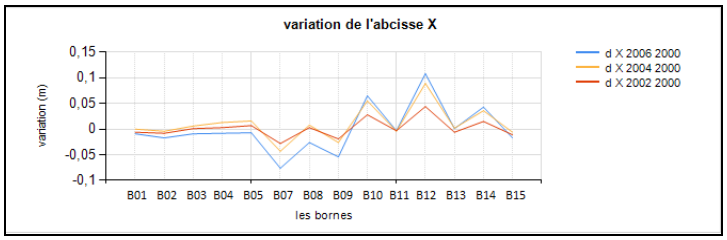


Fig. 7- *Variation de l'abscisse X du réseau de base*

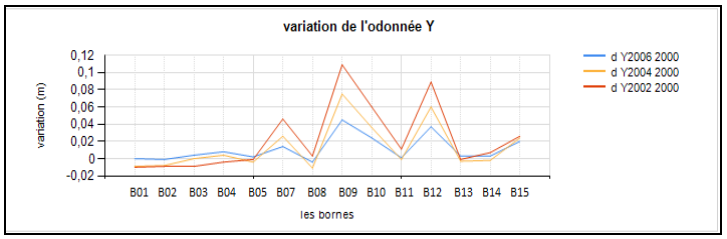


Fig. 8 - *Variation de l'ordonnée Y du réseau de base*

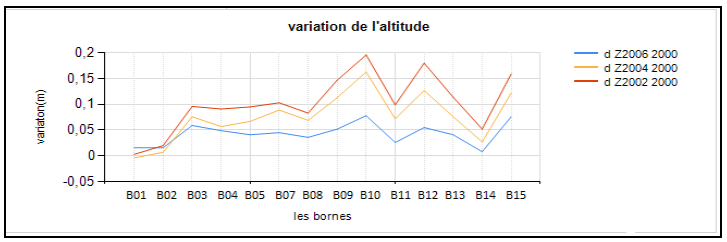


Fig. 9 - *Variation de l'altitude du réseau de base*

4.5 Analyse spatiale

L'analyse spatiale est une activité essentielle pour la cartographie et les prises de décision. Elle permet de déterminer la distribution spatiale d'une variable, les relations entre la distribution spatiale des variables et l'association de celles-ci à une zone géographique (Taibi et al., 2008).

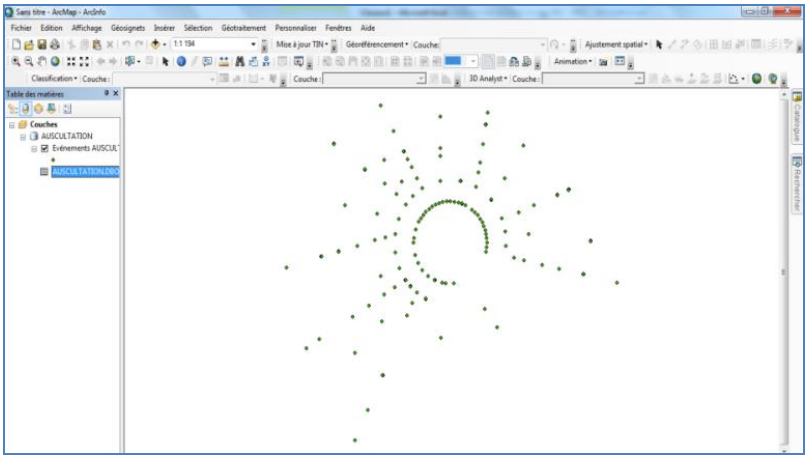


Fig. 10- *Réseau d'auscultation*

Le traitement des données multi-époques a permis la détermination des altitudes de tous les points et la restitution des déplacements verticaux au cours du temps.

Pour avoir une vue d'ensemble de la déformation verticale, on a tracé sous forme de courbes de niveaux en Figures (A) et (B) les **déplacements verticaux** interpolés sur la zone d'étude. La valeur maximale du déplacement vertical est de l'ordre de 252 mm. Les profils altimétriques restitués montrent que le terrain a subi un gonflement plus important côté terre (Sud) que côté mer (Nord).

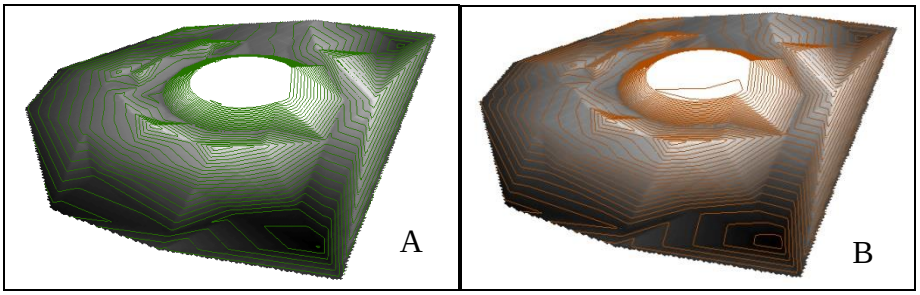


FIG. 11 – *Représentation du terrain avoisinant le Bac de stockage en 2000 (A) et 2006 (B)*

Afin d'arriver à une comparaison plus conviviale, une superposition des contours des iso lignes des deux époques (Figure 12, A), et une carte de surface soustraite à partir des différences entre les deux surfaces numériques (Figure 12, B), ont été réalisé.

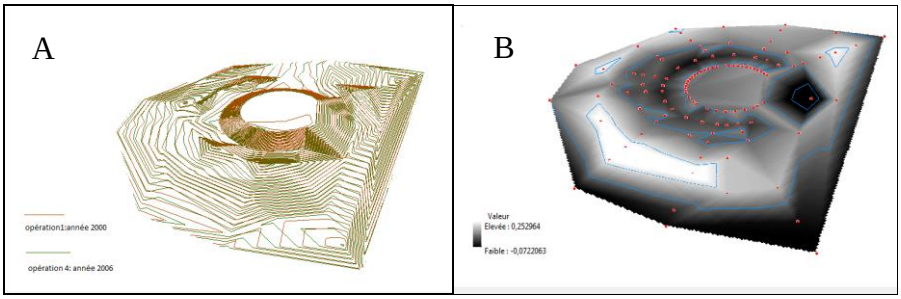


FIG. 12 - *Les contours iso lignes représentant le terrain de la 1ère et la 4ème opération (**A**) et la surface soustraite entre ces deux époques (**B**)*

La Figure (12, B) représente la variation de l'altitude et donc la significativité des déformations de la zone d'étude en termes de tassement et gonflement, ce dernier est important au niveau de la surface blanche, et atteint 252 mm. Ces modèles numériques réalisés à partir de nos données, sont une représentation tridimensionnelle proche de la réalité du terrain permettant d'identifier les caractéristiques les plus significatives du relief terrestre de la zone avoisinante du bac.

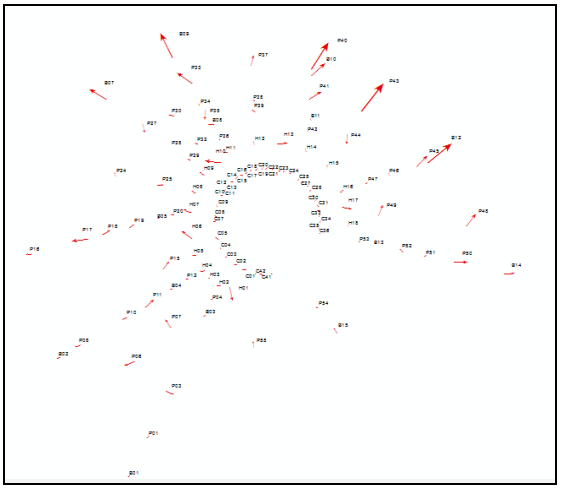


Fig. 13 - *Déplacements planimétriques des points du réseau entre les époques 2000 et 2006*

La Figure 13 montre que la majorité des points du réseau d'auscultation du bac en sol gelé sont soumis à un phénomène de déplacement horizontal dans la direction Nord-est. La valeur maximale du déplacement est de l'ordre de 163 mm.

La comparaison des résultats obtenus lors des quatre campagnes d'observations GPS a permis de mettre en évidence des déplacements de l'ordre de 163 mm en planimétrie et de 252 mm en altimétrie.

5 Conclusion

Ce travail constitue une contribution pour la surveillance d'ouvrage d'art sur les réponses technologiques possibles permettant de mieux maitriser et comprendre l'exposition aux risques. Cette étude se situe dans le cadre des systèmes d'aide à la décision. Ces systèmes se basent, généralement, sur une approche OLAP afin de faciliter l'interrogation et l'analyse des données. Cette approche adopte la modélisation dimensionnelle organisant les données d'une manière adaptée aux analyses permettant une étude analytique poussée des données, procurant une efficace aide à la décision. Une autre catégorie d'outils, dérivée des premiers, est également en plein essor. Il s'agit des outils SOLAP, qui allient la puissance analytique des outils OLAP, avec les capacités de visualisation des cartes géographiques

Dans un premier temps, il a fallu élaborer un modèle décisionnel dédié à la surveillance d'ouvrage d'art par auscultation géométrique. L'approche décisionnelle suggérée est structurée selon les trois étapes de Pictet. La méthode de modélisation de données est multidimensionnelle c'est-à-dire reposant sur les concepts de dimension, de fait, de mesure et de hiérarchie. L'entrepôt de données a été schématisé par un modèle en étoile. L'exploration de cet entrepôt a été effectuée par des opérateurs d'analyse OLAP. Notre application a été enrichie par la conception de cet entrepôt de données avec des outils offrant des possibilités d'analyse et d'exploration du cube de données avancées et non supportée par les systèmes transactionnels classiques. Elle a porté sur le réseau d'auscultation du bac de stockage GL4Z, établi pour la surveillance de cet ouvrage très particulier. Les informations issues des 04 époques d'observation émergent comme un élément décisionnel de premier ordre, et la capacité de les gérer et de les utiliser à bon escient constituent un apport appréciable des nouvelles technologies d'information pour l'atteinte des objectifs d'aide à la décision. Cette application nous a permis une bonne gestion de la dimension temporelle, l'ajout de mesures calculées, le filtrage sur les membres des dimensions, des opérations d'exploration et une gestion flexible des affichages cartographique, tabulaire et graphique.

Cette étude de cas a montré l'intérêt de l'intégration des SIG et de l'OLAP dans un système d'aide à la décision pour la surveillance d'ouvrages d'art. Ce genre d'application justifierait également l'intérêt de solutions logicielles plus ouvertes qui associeraient étroitement, au sein d'une même plate-forme, les composantes du SIG et de l'OLAP : "solution SOLAP Intégrée".

L'élaboration de notre travail ne constitue nullement une fin en soi, mais représente sûrement une base pour d'autres études du même aspect. A travers cette étude, un certain nombre de limites et de freins ont été identifiés positionnant le travail dans une logique plus prospective qu'opérationnelle à court terme. Les limites portent aussi bien sur les données que sur la technologie elle-même.

Références

Bédard, Y. (2009)**.** *OLAP et SOLAP. Notions avancées de bases de données SIG. Entrepôts de données spatiales*. Centre de recherche en géomatique. Université Laval, Québec.

Bédard, Y. (2002)**.** *Introduction aux systèmes SOLAP*. Cours sujet spécial. Centre de recherche en géomatique. Université de Laval, Québec..

Derkaoui, A., Ghezali, B. (2008)**.** *Elaboration d'un SIG pour la gestion des réseaux géodésique*. Bulletin des sciences géographiques, n° 21.

Derkaoui, A., Ghezali, B. (2009). *Intégration de l'analyse multidimensionnelle dans la gestion d'un réseau d'auscultation*. Revue XYZ n°118.

Hamdadou, D. (2008). *Un Modèle de Prise de Décision pour l'Aménagement de territoire, Une Approche Multicritère et une Approche de Négociation*. Thèse de Doctorat, Laboratoire d'Informatique d'Oran LIO, Université d'Oran.

Le Rubrus , B. (2009). *Capacité de rendu cartographique autour des technologies SOLAP*. UE ENG111 - Epreuve TEST Travail d'Etude et de Synthèse Technique en informatique: Conservatoire National des Arts et Métiers. Centre d'Enseignement de Grenoble. France.

Taibi, H., Kahlouche, S., Zeggai, A., Ghezali, B., Ayouaz O., Belhadj, L. (2008). *Auscultation du bac de stockage en excavation de Gaz Naturel Liquéfié GNL par GPS*. Revue XYZ No. 117

McHugh, R., Roche, S., Bédard, Y. (2007). *Vers une solution SOLAP comme outil participatif*. Centre de Recherche en Géomatique, Chaire de recherche industrielle CRSNG en bases de données géospatiales décisionnelles, Université Laval.

Rivest, S., Gignac, P., Charron, J., Bédard, Y. (2004). *Développement d'un système d'exploration spatio-temporelle interactive des données de la Banque d'information corporative du ministère des Transports du Quebec*. Colloque Géomatique 2004 - Un choix stratégique! Montréal, 27-28 octobre. Montréal, Canada

Sandro, B. *L'information géographique et les entrepôts de données*. INSA de Lyon, Laboratoire d'Informatique en Image et Systèmes d'information *UMR CNRS 5205, INSA, France*

Veilleux, J-P, Lambert, M., Santerre, R ., Bédard, Y. (2004). *Utilisation du système de positionnement par satellites (GPS) et des outils d'exploration et d'analyse SOLAP pour l'évaluation et le suivi de sportifs de haut niveau*. Colloque Géomatique 2004 - Un choix stratégique! Montréal, 27-28 octobre. Montréal, Canada

Summary

The safety of engineering structures requires periodic inspection. Assessing the condition of the structure and its behavior cannot be done without a regular and frequent monitoring by establishing a network of geometric auscultation. The contribution of an analytical tool in the management of this network is the purpose of this work. The idea is to propose an interactive decision support system by integrating GIS and OLAP technology to take advantage of each to ensure better multidimensional and space analysis to facilitate decision making. It will therefore be to combine the technology of geographic information systems for the treatment of geographic data and map viewing with exploration and analysis tools OLAP, this achievement involves many concepts and techniques. The practical application of our work lies in the field monitoring of engineering structures, the data used to illustrate our contributions concern the geometric network auscultation of the storage tank GL4Z.

Keywords: Geographic Information System (GIS), On Line Analytical Processing (OLAP) and Decision Support System (DSS).

Géo-sémantique analytique des trajectoires

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

*, ** Faculté des Sciences et Techniques de Mohammedia (FSTM),
Université Hassan II – Casablanca
*lkarim.lkarim@gmail.com
**azedine.boulmakoul@gmail.com
*** Université Joseph FourierGrenoble, France
ahmed.Lbath@ujf-grenoble.fr

RÉSUMÉ
Pour aider les services basés sur la localisation à répondre aux demandes du marché, nous proposons l'architecture générale d'un système scalable et performant pour gérer les trajectoires des objets mobiles. Nous détaillons également les prototypes des composants logiciels développés. En effet, nous évaluons la scalabilité et la performance de chaque composant proposé pour collecter, traiter, stocker, entreposer et analyser les trajectoires des objets mobiles. Le module de collecte permet la collecte des différents types de données géographiques à partir des appareils de positionnement et ce en utilisant les sockets en mode asynchrone avec pool d'objet. Ensuite, nous présenterons le module de stockage et d'entreposage dans une base de données NoSQL, adaptée au volume des données des trajectoires, de type MongoDB. Enfin, pour avoir un système d'analyse scalable et accessible dans le cloud, nous offrons un module d'analyse des trajectoires dans le système hadoop et nous évaluons le système proposé à travers l'étude de cas «Système de suivi des chemins pris par les clients dans les espaces commerciaux».

1. Introduction

Les différents domaines d'applications et de recherches ont besoin de collecter, de représenter et d'explorer les connaissances des trajectoires, tels que le domaine de transport et de la logistique, de la gestion, du marketing et des affaires, de la criminologie, de la surveillance des territoires, et de la gestion de flotte, etc. Satisfaire les attentes des utilisateurs des services de localisation en termes de vitesse de temps de réponse, performance et évolutivité est la clé du succès des entreprises. En effet, la production croissante des données spatio-temporelles engendre de très gros volumes de données disponibles et intéressantes à analyser. Ces volumes de données à très grandes échelles nécessitent des moyens de collecte, de traitement, de stockage (Agrawal et al., 2011), d'analyse et de visualisation appropriés. Les bases de données traditionnelles, support des entrepôts de données, ne sont plus adaptées pour gérer et traiter ces grandes masses de données.

Nous proposons dans le présent article une architecture globale, les composants logiciels et les technologies utilisées dans le système scalable conçu pour la collecte, le

traitement, le stockage, l'entreposage, l'analyse et la visualisation des différents types des trajectoires.

2. Concepts de base des trajectoires

Une trajectoire est une description de l'évolution au fil du temps, du mouvement physique des objets en mouvement. On cite dans ce qui suit les présentations de base des trajectoires :
(a) Trajectoire Raw ou brute est l'enregistrement des positions d'un objet dans un domaine spécifique de l'espace et du temps. Pour un objet en mouvement et un intervalle de temps donné, elle est présentée comme une séquence de lieu géométrique dans le système spatial 2D (xi, yi, ti). (b) Trajectoire structurée (Spaccapietra et al., 2008) est définie comme une trajectoire brute structurée en segments correspondant à des étapes significatives dans la trace de la trajectoire (voyages). (c)Trajectoire sémantique (Spaccapietra et al., 2008) possède une sémantique liée au domaine des applications, elle utilise les quatre composants (arrêt, déplacer, début et fin). Arrêter (S), déplacer (M), début (B) et fin (E) ne sont plus des positions spatio-temporelles, mais plutôt des objets sémantiques liés à la connaissance géographique générale et aux données géographiques de l'application. (d) Autre approche décrit les schémas de déplacement dans des contextes à la fois spatiales et temporelles basés sur la notion de région d'intérêt (Giannotti et al., 2007) en définissant la notion de voisinage spatial et de la tolérance temporelle. (e) Yu et al. 2007 ont étendu le concept du chemin espace-temps pour représenter à la fois les activités physiques (marche, conduite, etc.) et virtuelles (envoi de courrier électronique, appel téléphonique). Comme chaque activité possède un emplacement géographique et un intervalle de temps, le chemin spatio-temporel a été profilé en tant que conteneur de toutes les activités se produisant par un objet en mouvement.

3. Système proposé

Dans la littérature, il existe différentes présentations des trajectoires, chacune d'elles modélise la trajectoire d'une facette. Dans (Boulmakoul et al., 2012), nous avons présenté un méta modèle unifié pour modéliser tous les types de trajectoires des objets en mouvement, tenant compte de l'aspect structural et géo-sémantique. Le modèle « activité » a été intégré dans cette modélisation, et permet ainsi de capturer l'activité au sens sociogéographique.
Le système spatial mobile, présenté dans ce travail, permet de stocker, tracer les objets mobiles et de répondre aux requêtes spatio-temporelles destinées aux trajectoires.
Les recherches dans l'entreposage des trajectoires sont encore à un stade précoce. Des travaux préliminaires ont été trouvés pour modéliser et maintenir un entrepôt de données des trajectoires, définir un cube de données simple consistant en des dimensions spatiale et temporelle, et des mesures concernant les trajectoires numériques (Marketos et al., 2008). Cependant, les travaux existants des entrepôts des trajectoires ne prennent pas en compte, de façon explicite, les contraintes du mouvement, comme les mouvements dans un réseau routier, ni les différentes facettes des trajectoires (brutes, structurées, sémantiques, basées sur les régions d'intérêt, de plus, ces modèles rendent le système non convenable pour entreposer les données volumineuses des trajectoires.
L'analyse des trajectoires facilite le processus de prise de décision dans différents domaines d'applications. En outre, les données collectées en temps presque réel doivent être stockées dans un entrepôt de données avec une faible sinon aucune latence dans certains domaines

d'applications. Pour ce faire, nous entreposons les différents types de trajectoires pour faciliter la prise de décision instantanée en répondant à une série de questions complexes dans différents domaines (ex. Trouver instantanément, les régions où se trouvent des goulots d'étranglement). Pour des contraintes de scalabilité, de tolérance de pannes, de distribution du stockage et de traitement des données des trajectoires, l'entreposage est effectué dans des bases de données NoSQL de type cloud (Boulmakoul et al., 2014; Boulmakoul et al., 2012). Prendre des décisions, à partir des données brutes, collectées des appareils de positionnement sous forme d'une liste de points spatio-temporels en vrac s'avère coûteux et aussi pauvre en terme d'informations pour différents domaines d'application. D'où, il est nécessaire de procéder à la reconstruction de ces trajectoires(Boulmakoul et al., 2014).
Une fois les différents types de trajectoires sont entreposés une autre problématique apparaît. Cette problématique consiste à analyser et extraire des informations spatio-temporelles utiles. Nous appliquons la méthode d'analyse simpliciale et issue de la topologie algébrique, sur l'entrepôt des trajectoires, pour extraire les informations implicites et potentiellement utiles.
L'analyse simpliciale a été développé à l'origine par R. Atkin (Atkin, 1974), comme une approche issue de la topologie algébrique pour étudier les caractéristiques structurales des systèmes sociaux dans lesquels deux ensembles d'indicateurs, fonctionnalités ou caractéristiques sont liées les unes aux autres. L'utilisation des caractéristiques topologiques de la méthode d'analyse simpliciale comme connectivenness, et l'excentricité permettra de répondre à une variété de requêtes comme par exemple, trouver l'ensemble optimal des régions accroches et aussi l'excentrique. Pour extraire davantage d'informations sur l'entrepôt des trajectoires. Les données sont prétraitées et puis entreposées dans un entrepôt NoSQL dans un environnement cloud.

4. Architecture générique et composants réutilisables

L'architecture du système proposé pour notre méta-model unifié des trajectoires des objets mobiles s'appuie sur l'Architecture Orientée Service (SOA) afin d'améliorer les performances et l'interopérabilité des applications (Boulmakoul et al., 2012; Boulmakoul et al., 2013a). En architecturant les composants du système par les services web basés sur les standards publics de l'OGC en les plaçant dans un substrat de messagerie SOA, nous pouvons intégrer les différents services des trajectoires (suivi, visualisation et interrogation de l'entrepôt des trajectoires des objets mobiles) avec d'autres applications et services basés sur la localisation.
L'architecture du système, présenté sur la Figure 1, est constituée de cinq couches principales : la couche de collecte de données, couche de prétraitement, couche de génération des différents types de trajectoires, couche de base de données et entreposage des trajectoires, couche d'analyse et des applications des trajectoires.

1. **Couche de collecte de données**

Cette couche permet de collecter les données à partir des appareils mobiles, guichets automatiques et les caméras IP à l'aide des protocoles HTTP, FTP et SOAP. Tout dépend de la criticité des données à gérer, les services de collecte de données sont utilisés pour collecter les données en ligne ou hors ligne. L'utilisation des interfaces des sockets visent à rendre possible des applications de collecte en temps presque réel entre les sources de données et le serveur de collecte de données. Au niveau du serveur de collecte de données, nous avons utilisé l'API des sockets .Net en mode asynchrone avec un pool d'objets Sockets

(Boulmakoul et al., 2013; Boulmakoul et al. 2012) pour pouvoir les réutiliser dans plusieurs collectes des données spatio-temporelles et recevoir des notifications en cas d'erreur ou d'opération réussie. Les résultats des tests de performances de notre système de collecte des données spatio-temporelles montrent que la performance et la scalabilité de notre cadre proposé n'a pas changé et le serveur de collecte peut recueillir 600 000 messages de 600 objets en mouvement simultanés. En ce qui concerne l'évolutivité du serveur et de la performance, les essais en chiffres montrent que la variation des ressources système, tels que la mémoire, le disque physique et le processeur, est légère et contrôlée. Tandis que la durée totale consommée pour la collecte et l'enregistrement des messages dans MongoDB varie entre 9 ms lors de la collecte d'un message d'un objet en mouvement, et 67s lors de la collecte et l'enregistrement des messages de 10000 objets mobiles, et environ 4s pour collecter et enregistre 50 000 messages d'un objet en mouvement.

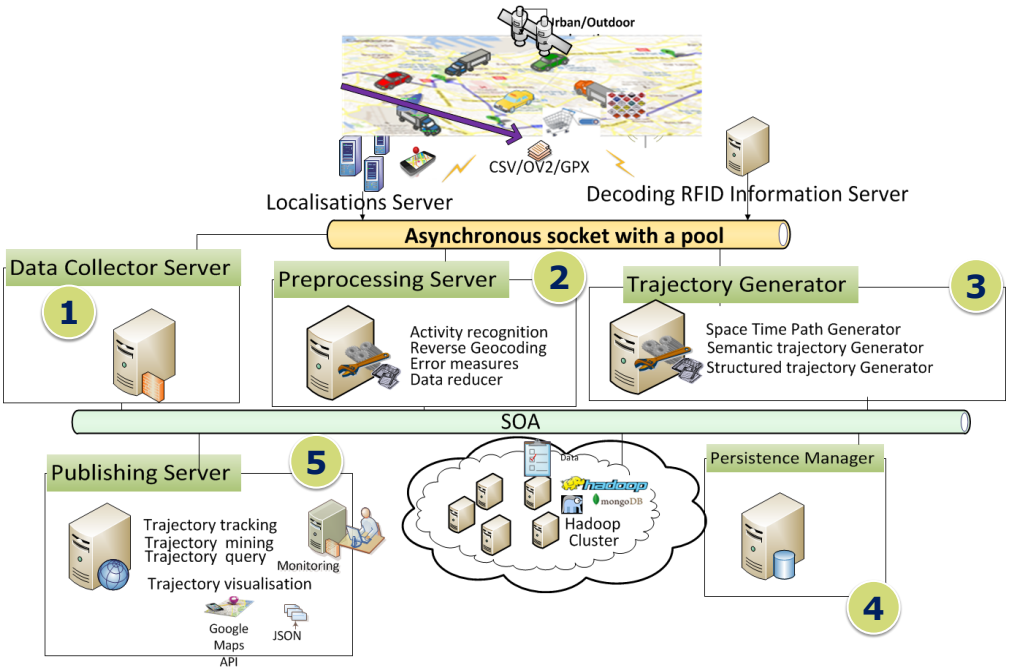


Figure 1: Architecture générale du système.

Par conséquent, notre système de collecte est évolutif et efficace pour différents types d'applications tels que: guides touristiques de la communauté; gestion des flottes, suivi des marchandises, des camions et des taxis; la publicité; protection des biens, des véhicules et antivol; et les études comportementales des êtres humains.

2. **Couche de prétraitement des trajectoires**

Elle emploie les services de réduction de données, de reconnaissance d'activité et du géocodage inverse.

3. **Couche de générateur de trajectoire**

Grâce à cette couche, un ensemble de points spatio-temporels est transformé de trajectoire brute nettoyée en structurée, sémantique, avec région d'intérêt ou chemin espace-temps. La construction des trajectoires structurées nécessite la segmentation de la trajectoire brute en trajectoire structurée, et ce en identifiant les stops (Figure 2).

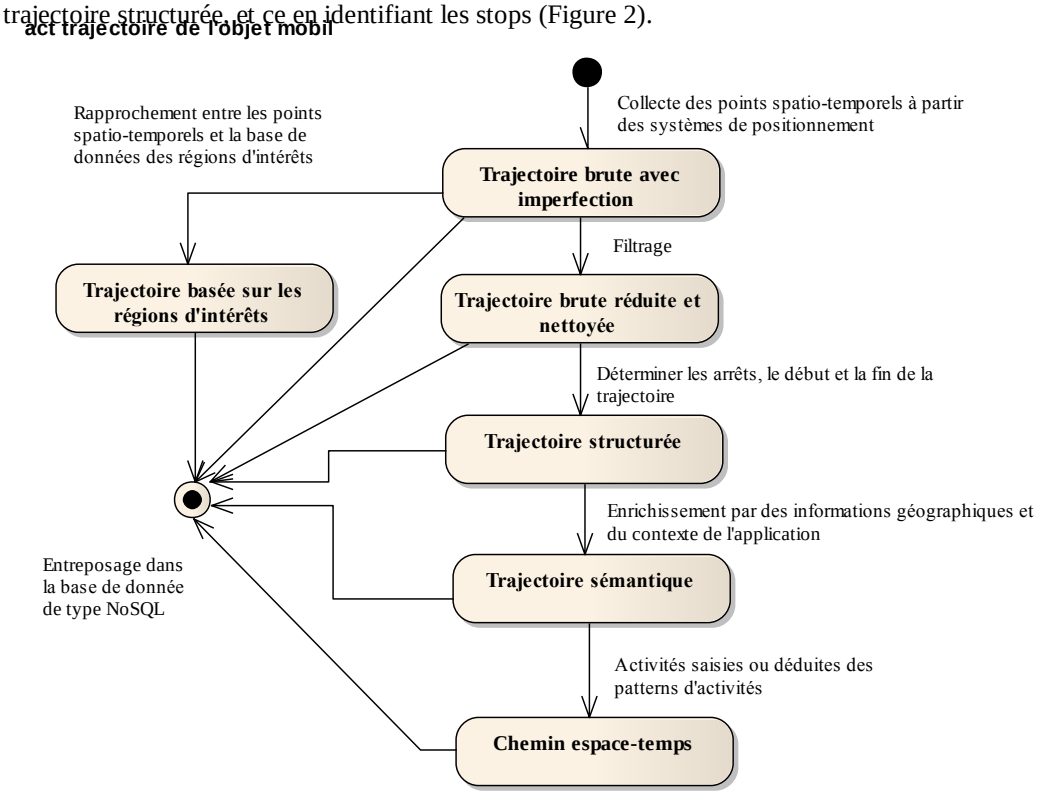


Figure 2: Diagramme d'état de transition de l'objet trajectoire.

Un stop est déterminé par un mouvement avec une vitesse qui tend vers 0 ou inférieure à un seuil paramétré, si l'intervalle temporel du stop dépasse le seuil paramétré, il est considéré comme la fin de la trajectoire, les points spatio-temporels correspondants aux changements de position entre deux stops, entre le début de la trajectoire et le premier stop, ou entre le dernier stop et la fin de trajectoire sont les déplacements (M). La construction des trajectoires sémantiques est faite en enrichissant les quatre composants début, fin et la séquence consécutive et alternée des déplacements et arrêts de la trajectoire structurée par les données de contexte de l'application pour avoir plus de sens lors de l'analyse de l'entrepôt des trajectoires. La construction des trajectoires basées sur les régions d'intérêts est faite en utilisant un rapprochement de la trajectoire de type brute ou structurée avec une table de correspondance contenant les zones géométriques et les noms des régions dans une période donnée. La construction des chemins espace-temps est faite en enrichissant les quatre composants de la trajectoire sémantique par les activités de l'objet mobile.

4. **Couche de stockage et d'analyse dans un environnement hadoop cloud**

Cette couche contient les données géographiques, et les entrepôts des trajectoires. Les trajectoires entreposées sont automatiquement réparties dans des espaces de stockage et des lots de traitement sur un ensemble de cluster grâce à la solution évolutive Hadoop qui embarque une capacité de tolérance aux pannes. En effet, le stockage s'effectue dans les bases de données MongoDB de la catégorie NoSQL, et l'analyse avec Hadoop dans un environnement Cloud.

a. **Stockage des trajectoires**

La scalabilité des bases des données relationnelles est obtenu en utilisant du matériel plus performant et en ajoutant plus de mémoires, par contre une base de données de type NoSQL profite de la montée en puissance en répartissant la charge sur plusieurs systèmes de base. Par conséquent, NoSQL est une base de données peu coûteuse pour s'approprier une base de données de trajectoires. Les deux bases de données sont évaluées sur la même machine 64 bits, en utilisant les dernières versions de PostgreSQL avec des extensions de PostGIS (version 9.2.2) et bases de données MongoDB (version 2.2.3). Le but de cette étude est de tester les performances des bases de données pour l'entreposage des données géo-spatiales lors de l'enregistrement des données brutes générées par notre simulateur de suivi des véhicules. Chaque message envoyé à partir de l'appareil GPS est un objet composé d'une position et plusieurs balises. Dans notre test, nous avons utilisé les fichiers générés par notre simulateur de suivi des véhicules ; la structure d'un message de suivi collecté est : « identifiant de l'objet mobile ; latitude ; longitude ; date ; heure ; observation »

D'un point de vue géo-spatiale, MongoDB permet nativement des points d'indexation géo-spatiales (sans extension), contrairement à POSTGIS qui permet d'utiliser non seulement des points géométriques mais aussi d'autres objets géographiques plus avancés (par exemple des lignes ou polygones). Ce benchmark, montre que MongoDB est beaucoup plus rapide dans l'insertion de données géographiques. Nous expliquons cette performance par le fait que MongoDB supporte les indexes géo-spatiales avec aucun type de données géo-spatiales dédié. Alors que PostGIS utilise deux tables en arrière-plan (Ramsey, 2012), spatial_ref_sys et geometry_columns, pour récupérer et en apprendre davantage sur les types de géométrie disponibles dans une base de données spécifique. En outre, l'architecture de MongoDB supporte l'évolutivité horizontale et la haute disponibilité grâce à l'ensemble des réplications (replication sets). L'auto-fragmentation (Auto-sharding) permet de distribuer la charge sur plusieurs serveurs et garder les données équilibrées sur ces serveurs.

b. **Analyse des trajectoires**

- **Approche d'analyse des trajectoires**

L'analyse de l'entrepôt des trajectoires permet de compléter la compréhension des dimensions spatio-temporelles du réseau social des trajectoires. Notre approche pour analyser l'entrepôt des trajectoires des objets mobiles est basée sur les démarches suivantes (Boulmakoul et al., 2013b) :

- Démarche fondée sur l'analyse structurale exploitant des outils issus de la topologie algébrique.
- Démarche faisant référence à l'analyse des réseaux sociaux.

Les valeurs des mesures obtenues en appliquant les méthodes d'analyse simpliciale et l'analyse des réseaux sociaux sur l'entrepôt des trajectoires permettent de décrire comment un acteur, une région ou une activité est intégré dans le réseau. Par exemple, le degré de centralité entrant (indegree) mesuré pour un acteur, qui pourra être un objet mobile une région ou une activité, quantifie la façon dont cet individu se rapporte avec les autres dans le réseau. A titre d'exemple, en appliquant l'approche proposée sur l'entrepôt des trajectoires dans un intervalle de temps, nous pouvons déduire l'importance d'une région R, d'un objet mobile MO ou d'une activité A dans le réseau, et par la suite aider à résoudre les problèmes de différentes applications tels que les congestions. La centralité des objets mobiles, des régions et des activités dans le réseau social de l'entrepôt des trajectoires peut également changer. Par conséquent, la séquence des valeurs de centralités permettra la visualisation de l'évolution de l'importance des objets mobiles, des régions et des activités au cours de temps. D'autre part, l'application de la méthode d'analyse simpliciale sur l'entrepôt des trajectoires permettra de répondre à plusieurs questions du mining des trajectoires. En effet, la génération des classes d'équivalences permet de trouver les trajectoires similaires des objets mobiles, les régions similaires dans un intervalle de temps et de déduire par la suite le model de trajectoire d'un objet mobile. Exploiter le résultat concernant les trajectoires similaires permettra de trouver des solutions à plusieurs problèmes, par exemple, proposer le covoiturage aux personnes ayant des trajectoires similaires pour éviter les embouteillages et diminuer la pollution. Spatialement et temporellement, les groupes sociaux proches ont tendance à partager l'information et avoir un comportement homogène dans l'espace et le temps. L'identification des tendances intéressantes peuvent fournir des indications importantes dans de nombreux domaines d'application tels que l'écologie et les affaires.

- **Environnement d'analyse des trajectoires**

MongoDB utilise une technique appelée fragmentation (sharding) pour mettre à l'échelle sa performance sur un cluster de serveur. Il s'agit d'un processus permettant de fractionner les données de manière uniforme sur le cluster pour paralléliser l'accès. Le traitement des données avec MongoDB est effectué selon les options suivantes : (i) Traitement dans MongoDB en utilisant Map / Reduce ; (ii) Traitement en utilisant le Framework d'agrégation de MongoDB ; (iii) Traitement externe à MongoDB en utilisant Hadoop et d'autres outils externes.

Les requêtes MongoDB examinent un enregistrement à la fois, ce qui signifie que les requêtes sur des documents multiples doivent être mises en œuvre sur le client ou utilisent le paradigme MapReduce (MR) intégré de MongoDB. Bien que MapReduce de MongoDB puisse être exécuté en parallèle dans chaque fragment, il a deux inconvénients majeurs :

1) Le langage de programmation de Map Reduce de MongoDB est JavaScript, qui est lent et a de bibliothèques analytiques pauvres ;
2) JavaScript utilisé par MongoDB, n'est pas thread-safe, donc seulement un seul programme de MapReduce peut fonctionner à la fois (Monkey, 2013).

En résumé, MongoDB fournit de hautes performances pour le stockage et la récupération à grande échelle et dispose d'une interface de requête robuste permettant des opérations intelligentes. Certes, il fournit des fonctionnalités de traitement mais n'est pas un système de traitement de données.

Traitement des données des trajectoires avec Hadoop

L'analyse des trajectoires nécessite beaucoup de calculs qui exigent une puissance de traitement. Pour répondre aux nouveaux enjeux de traitement de très hautes volumétries de données, l'accent a été mis sur les solutions suivantes :

1. Utilisation des machines de capacités de stockage, de puissance de traitement et de mémoire élevés. Mais avec l'augmentation rapide des données, l'utilisation des machines simples a échoué à l'échelle.
2. Utilisation des systèmes distribués permettant de répartir les taches sur plusieurs machines. Mais, les solutions analytiques des données sont souvent complexe, sujette à l'erreur, ou tout simplement pas assez rapide.

Au cours des 10 dernières années, une solution appelé « Hadoop » émerge. C'est un framework Open Source conçu pour réaliser des traitements sur des volumes de données massifs. Il support les applications destinées au Big Data, en particulier l'analytique, sur le système de fichier distribué HDFS (Hadoop Distributed File System). L'infrastructure de Hadoop applique le principe bien connu des grilles de calcul « MapReduce », consistant à répartir l'exécution d'un traitement sur plusieurs nœuds, ou grappes de serveurs. De cette façon, les données non structurées peuvent faire l'objet d'un traitement analytique distribué et en parallèle pour accélérer le traitement, jusqu'à se rapprocher du temps réel.
Pour l'utilisateur, tout est transparent partant du découpage de la donnée en blocs, leur répartition sur les nœuds qui composent le cluster à l'exécution des tâches (parallélisassions). Dede et al. (Dede et al., 2013) ont comparé l'implémentation native de MapReduce de MongoDB avec mongo-hadoop MapReduce. La configuration du système de test est que MongoDB tourne dans un serveur et utilise 3 nœuds cluster Hadoop. Ils ont trouvé que mongo-hadoop plugin est cinq fois beaucoup plus performant en termes de temps de traitement.

Traitement des données des trajectoires en utilisant Hadoop dans un environnement cloud

Le cloud computing, littéralement « l'informatique dans les nuages », est un concept apparu récemment, mais de plus en plus à la mode dans le secteur de l'informatique. Pour traiter des volumes de données massives, le déploiement de la plateforme Hadoop doit se faire sur

plusieurs nœuds. Il nous faut une plateforme en architecture « distribuée », et donc plusieurs machines qui servirons de DataNode / TaskTracker..
La flexibilité du Cloud apporte une agilité dans la gestion de l'infrastructure et l'affectation des ressources. La couche virtuelle d'abstraction est également centralisée, et se révèle ainsi plus facile à gérer. Kang et al. (Kang et al. 2013) ont comparé l'utilisation des clusters physiques et des machines virtuelles dans le Cloud. Le résultat de l'étude montre que l'exécution de Hadoop dans des machines virtuelles d'un Cloud privée permet d'obtenir plus que 110.76% de performance du serveur physique. Par conséquent, nous avons adopté l'architecture pour l'analyse des trajectoires qui consiste à traiter les données avec Hadoop en utilisant MapReduce du Framework Hadoop dans des machines virtuelles dans le cloud. La communication entre MongoDB et Hadoop est faite en utilisant le connecteur MongoDB-Hadoop-Connector (Boulmakoul et al., 2014).
D'autre part, vu que l'analyse des trajectoires moyennant l'analyse des réseaux sociaux est effectué dans le système R et le paquetage Statenet et les données entreposées sont dans la dimension des big data, nous avons intégré le système R avec l'environnement Hadoop à travers un ensemble de paquets pour le langage R « RHadoop » afin que le système R puisse profiter du paradigme MapReduce et le système de fichier HDFS de l'écosystème Hadoop.

5. **Couche des applications des trajectoires**

Contient des services Web spécialisés, autonomes, auto-descriptifs, pour l'exploitation, le suivi, la visualisation et l'interrogation des trajectoires, ces services peuvent être publiés et invoqués à travers le Web, en utilisant un large éventail de machines connectées au web et les appareils mobiles.

5. Étude de cas « Système de suivi des chemins pris par les clients dans les espaces commerciaux»

Le principe de cette étude de cas se base sur le suivi en temps réel des chemins des clients à l'intérieur d'un espace commercial. Grace à l'entrepôt de données « trajectoire-ticket » crée, nous pouvons comprendre le comportement spatio-financier des clients et répondre à plusieurs requêtes des décideurs : (i) Évaluer l'impact d'ajout des produits/ changement d'emplacement des produits / animations (le taux des personnes attirées par un style..) ; (ii) Trouver le profil d'ambiance adéquat ; (iii) Visualiser les chemins d'une date donnée ; (iv) Trouver la corrélation entre la disposition géographique et le CA généré par les chemins des clients ; (v)Identifier les caisses lentes.
Chaque centre commercial possède une base de données « Ticket de caisse » pour enregistrer les achats des clients avec la date, l'heure et le numéro de caisse de la transaction. Nous identifions la marge de bénéfice générée par un chemin par un rapprochement automatique entre le détail du CA d'une transaction et le chemin sous forme de régions d'intérêt d'un client. L'identification du client est faite sur la base de sa position géographique, la date, l'heure et le numéro de caisse du paiement. Une fois la position géographique d'un client se trouve dans une région de type Caisse dans une date d1 et heure t1, nous récupérons le numéro de caisse N1 correspondant à sa position géographique à partir de la base de données spatiale de l'espace commercial. En parallèle, nous récupérons la transaction de paiement enregistrés dans la caisse N1 dans la date d1 et l'heure t1. A ce moment, nous pouvons

construire l'entrepôt de données recherchés 'trajectoire-ticket' en faisant une consolidation des informations de la base de données des chemins clients présentées avec les régions et les informations sur le détail de paiements se trouvant dans la base de données « Ticket de caisse ». Grâce au système proposé, nous pouvons détecter en temps réel le ou les caisses/caissières responsables de la lenteur des files d'attente et ce en requêtons sur le nombre de clients se trouvant dans la région géographique des files d'attente d'une caisse donnée. Le Tableau 1 présente la matrice d'incidence des trajectoires de 10 clients dans l'espace commercial :

	R1	*R2*	*R3*	*R4*	*R5*	*R6*	*R7*	*R8*	*R9*	*R10*	*R11*	*R12*	*R13*
T1	1	1	1	0	1	0	1	0	0	0	0	1	0
T2	1	0	1	1	0	0	0	0	1	0	0	0	0
T3	1	0	1	1	1	1	0	0	0	1	1	0	0
T4	1	0	1	1	1	1	1	1	1	1	1	1	1
T5	1	1	1	0	1	1	1	0	0	1	1	0	0
T6	1	1	0	0	0	0	1	0	0	0	0	1	0
T7	1	0	1	1	1	1	1	1	1	0	1	1	1
T8	1	1	1	0	1	0	0	0	0	1	1	1	0
T9	1	0	1	1	0	1	1	1	1	1	1	1	1
T10	1	1	1	0	1	1	1	0	0	1	1	1	0

Tableau 1: Matrice d'incidence des trajectoires des clients dans le centre commercial.

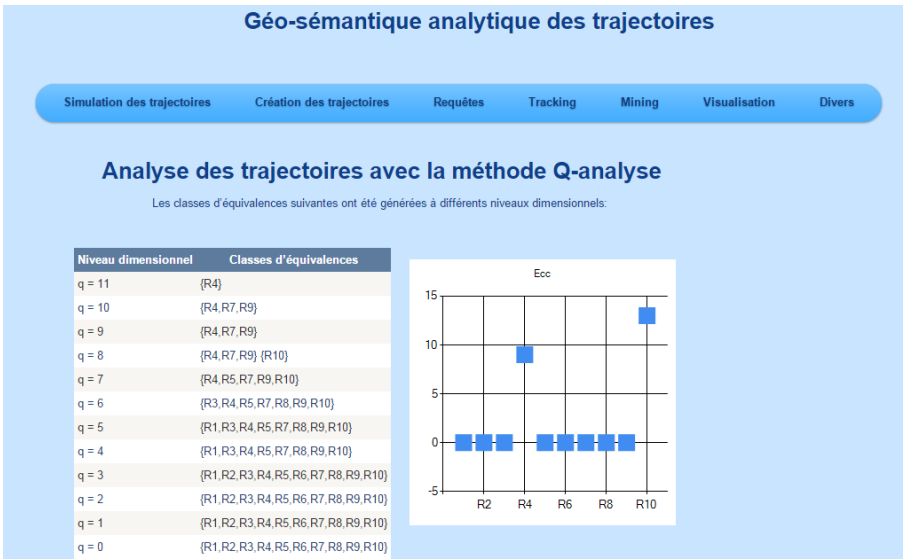


Figure 3: L'excentricité des régions de l'espace commercial.

Le résultat de l'analyse simpliciale (figure 3) montre que les régions 4 et 10 ont des excentricités maximales c.-à-d. les clients ne les visitent pas. Par contre la région 8 ne génère pas un chiffre d'affaire même si elle est connecté avec une excentricité égale à zéro et donc visité par les clients. À partir de ce résultat nous concluons à travers cette analyse la cause réelle du problème de chiffre d'affaire dans le centre commercial. Dans le cas des régions 4 et 10, le problème réside dans l'organisation spatiale des produits par contre pour la région 8, le problème réside dans la nature des produits exposés.

Les résultats de l'analyse des trajectoires des clients de l'espace commercial avec la méthodologie d'analyse des réseaux sociaux avec le système R et le paquetage STATNET. La mesure de betweenness indique si un sommet est partagé dans un grand nombre de trajectoires non redondants. Dans les valeurs suivantes de betweeness des sommets, les valeurs obtenues montrent la présence des sommets qui n'existent pas sur les chemins non redondants entre les sommets. Ainsi, nous concluons que la région 1 est la plus importante des régions et la trajectoire 4 présente la trajectoire la plus importante en termes de nombre des régions visitées.

3. Conclusion

Nous avons présenté l'architecture d'un système scalable conçu pour gérer les objets trajectoires du méta modèle unifiée des trajectoires des objets mobiles. Le système proposé offre des composants de collecte des données spatio-temporelles des dispositifs de géolocalisation en utilisant les sockets. L'utilisation des interfaces des sockets vise à faire une collecte de données en temps presque réel entre les sources de données et le serveur de collecte de données. Les appareils de positionnement des objets en mouvement exécutent un programme de socket pour envoyer les traces spatio-temporelles et le serveur de collecte exécute un programme pour recevoir les données géographiques massives en temps réel à partir de différents programmes des clients et de différents capteurs. En ce qui concerne le stockage et l'entreposage des objets trajectoires, nous utilisons une base de données NoSQL de type document « MongoDB ». Hadoop, HDFS et MapReduce, RHadoop forment l'infrastructure logicielle Big Data, de son architecture en cluster avec des nœuds, à sa capacité de traitement au service de l'analyse des trajectoires. Les trajectoires des objets mobiles sont affichées en utilisant les fichiers json, dans les bibliothèques de visualisation.

Réferences

Agrawal, D., Das, S., & Abbadi, A. (2011). Big data and cloud computing : current state and future opportunities. In 14th International Conference on Extending Database Technology, EDBT/ICDT '11, ACM (pp. 530–533.). New York, USA.

Atkin, R. (1974). Mathematical Structure in Human Affairs. Mathematical Structure in Human Affairs. London, Heinemann.

Boulmakoul, A., Karim, L., & Lbath, A. (2012). Moving Object Trajectories Meta-Model and Spatio-Temporal Queries. International Journal of Database Management Systems (IJDMS), 4(2), 35–54.

Boulmakoul, A., & Karim, L. (2012). A Scalable Data Collector Framework for the Unified Moving Object Trajectories' Meta-Model. In Innovation et Nouvelles Tendances dans les Systèmes d'Information (pp. 19–24).

Boulmakoul, A., Karim, L., Elbouziri, A., & Lbath, A. (2012). A System Architecture for Heterogeneous Moving Objects Trajectory Models Using Different Sensors. In SoSE

in cooperative and competitive distributed decision making for complex dynamic systems. Genova, Italy.

Boulmakoul, A., Karim, L., Elbouziri, A., & Lbath, A. (2013). A System Architecture for Heterogeneous Moving-Object Trajectory Metamodel Using Generic Sensors: Tracking Airport Security Case Study. IEEE System Journal. doi:10.1109

Boulmakoul, A., & Karim, L. (2013a). Dispositif, système et procédé de suivi des chemins pris par les clients dans les espaces commerciaux.

Boulmakoul, A., & Karim, L. (2013b). Space Time Path Data Warehouse Mining based on Simplicial Complex Analysis. In Innovation et Nouvelles Tendances dans les Systèmes d'Information (pp. 19–24).

Boulmakoul, A., Karim, L., & Lbath, A. (2013). A High Performance Scalable Data Collection System for Moving Objects. International Journal of Computer Applications, 36–43. doi:10.5120/11424-6769

Boulmakoul, A., & Karim, L. (2014). Construction et entreposage des trajectoires. In 4ème Edition du Workshop Internationnal sur l'Innovation et Nouvelle Tendances dans les Systèmes d'Information. Rabat, Maroc.

Boulmakoul, A., Karim, L., Laarabi, M. H., Sacile, R., & Garbolino, E. (2014). MongoDB-Hadoop Distributed and Scalable Framework for Spatio-Temporal Hazardous Materials Data Warehousing. In « International Congress on Environmental Modelling and Software (iEMSs 2014). San Diego, California, USA.

Dede, E., Govindaraju, M., Gunter, D., Canon, R. S., & Ramakrishnan, L. (2013). Performance evaluation of a MongoDB and hadoop platform for scientific data analysis. In Science Cloud '13 Proceedings of the 4th ACM workshop on Scientific cloud computing (pp. 13–20.).

Giannotti F, Nanni M, Pedreschi D, Pinellin F (2007). Trajectory Pattern Mining. Int. Conf. Knowl. Discov. Data Min. pp 330–339.

Kang, Y., & Kang, K.-W. (2013). An Empirical Study of Hadoop Application running on Private Cloud Environment. Advanced Science and Technology Letters, 35, 70–73.

Monkey, S. (2013). https://developer.mozilla.org/en/SpiderMonkey.

Ramsey, P. (2012). OpenGeo. Http://www.postgis.fr/chrome/site/docs/workshop-foss4g /doc /geometries.html.

Spaccapietra S, Parent C, Damiani MD, Macedo JA, Porto F, Vangenot C (2008). A Conceptual view on trajectories. Data Knowl Eng 26–146.

Yu H, Shaw S (2007). Revisiting Hägerstrand's time-geographic framework for individual activities in the age of instant access. 103–118.

Summary

To support location-based services to meet the market demands, we propose a general architecture of ascalable and performing system to manage trajectories of moving objects. We also present prototypes

of developed software components. Indeed, we evaluate the scalability and performance of each component proposed to collect, process, store and analyze trajectories of moving objects. The collection module is used to collect different types of geographic data from the positioning devices using asynchronously sockets with object pooling. Then we present the storage module in a NoSQL database, adapted to the volume of trajectories' data. Finally, to have a scalable and accessible analysis system in the cloud, we offer trajectory analysis

module in Hadoop system and evaluate the proposed system through the case study «Customers trajectories analysis in commercial spaces».

Système d'Aide à la Décision Multicritères pour le Rangement des Zones Industrielles (RPRO4SIGZI)

Aissa Taibi, Baghdad Atmani
Laboratoire d'Informatique d'Oran – LIO
Université d'Oran 1 Ahmed Benbella
taibiissa@yahoo.fr
atmani.baghdad@gmail.com

Résumé. L'intégration des Systèmes d'Information Géographique(SIG) et de l'analyse multicritère constitue une voie privilégiée et incontournable pour faire évoluer les SIG vers de véritables systèmes d'aide à la décision. Le système RPRO4SIGZI proposé dans cet article permet, à partir d'une étude détaillée des critères géographiques, environnementaux et socioéconomiques, de faire coopérer un SIG et une méthode d'analyse multicritère pour le choix géographique du site adéquat pour l'aménagement et l'installation d'un projet industriel. Le résultat obtenu par RPRO (Ranking PROMETHEE) pour le rangement des zones industrielles candidates de l'ouest Algérien est affiné par une visualisation SIGZI (Système d'Information Géographique pour les Zones Industrielles). Le module RPRO procède au rangement des zones industrielles candidates en utilisant la méthode de sur classement PROMETHEE et le module SIGZI à la visualisation de ces zones sur la carte géographique. Le système RPRO4SIGZI a été conçu pour l'évaluation d'une nouvelle méthodologie d'analyse multicritères guidée par datamining. L'objectif est de montrer comment le datamining est utilisé pour la modélisation des préférences du décideur et la génération des tables de performances. Seul le système RPRO4SIGZI est présenté dans ce papier.
Mots clefs : Système d'Information Géographique(SIG), Analyse multicritère, Zone industrielle, intégration SIG-AMC, Cartographie.

1 Introduction

L'étude géo-décisionnelle d'aptitude zonale pour le choix de l'emplacement géographique de nouveaux sites d'habitations, d'industries et de services s'avère primordiale et constitue un vrai problème de décision à référence spatiale. Les décideurs doivent agir précocement en se basant sur des analyses approfondies des critères (facteurs, contraintes) environnementaux, socioéconomiques et autres pour mener soigneusement leurs décisions à fin de diminuer les risques.

Ce travail consiste à ranger les zones industrielles de l'ouest Algérien en utilisant la méthode de sur classement PROMETHEE, J.P. Brans et Ph.Vincke (1985). Il enchaine un choix préliminaire basé sur une analyse d'aptitude zonale utilisant des Méthodes d'agrégation non compensatoires. Chaque zone est une action spatiale puisqu' une action à prendre est spatiale si elle est définie par sa localisation géographique, sa forme et/ou ses relations spatiales, Raffaella(2012). La majorité des critères de jugement ont un caractère géographique. Les spécificités de ce genre de problème est en faveur d'une intégration entre SIG et AMC d'où l'adoption de cette approche. Les chercheurs se sont penchés sur ce type d'approche

depuis 1999, des centaines d'articles ont été publiés, Jacek Malczewski (2006). L'idée conceptuelle sur laquelle se base les travaux d'intégration SIG-AMC consiste à utiliser les fonctionnalités du SIG pour préparer les entrées (*inputs)* nécessaires à l'application d'une méthode multicritère et d'exploiter les potentialités de présentation du SIG pour visualiser les résultats de l'analyse, Chakhar (2006). Dans la littérature il y a une multitude de définitions pour les SIG, une définition cohérente avec cette étude est celle de Marc (2002) :

Un système d'information géographique (SIG) est avant tout un système de gestion de base de données capable de gérer des données localisées, et donc capable de les saisir, de les stocker, les extraire (et notamment sur des critères géographiques), de les interroger et analyser, et enfin de les représenter et les cartographier. L'objectif affiché est essentiellement un objectif de synthèse, permettant à la fois la gestion des données comme l'aide à la décision. L'entrée de la méthode PROMETHEE (INPUT) est une table des performances qui regroupe les valeurs (score) de chaque action (zone industrielle) par rapport à l'ensemble des critères ainsi que la pondération intra critères. L'évaluation des actions par rapport aux critères Géographiques se base sur une importante fonctionnalité des SIG : la cartographie, cette discipline constitue la première étape de l'analyse spatiale, une carte est un modèle de la réalité contenant la représentation géométrique des objets et des catégories d'objets avec une logique graphique et sémiologique, Régis et collet (2011)**.** La valeur sismique d'une zone par exemple découle de sa position géographique sur la carte sismique d'Algérie. En résultat un rangement total de la meilleure zone à la plus pire est obtenu avec visualisation cartographique avant et après rangement.

L'adoption de l'approche SIG-AMC pour ce cas nous a confrontés à plusieurs problématiques telles que le choix de la méthode adéquate ? La subjectivité des préférences du décideur ?... Pour résoudre le deuxième problème nous allons impliquer le datamining a fin de modéliser les préférences du décideur et la génération des tables de performances. La suite de cet article est présentée comme suit. La section 2 est consacrée à la problématique et quelques travaux connexes, La section 3 est consacrée au modèle proposé, Ensuite dans la section 4 une étude de cas est illustrée, enfin, nous terminons par une conclusion et des perspectives.

2 Problématique et travaux connexes

Les problèmes liés à l'évolution des tissus urbains, à la construction de nouvelles villes, et à la création de nouvelles zones industrielles sont des problèmes d'analyse d'aptitudes zonales dans un contexte plus vaste d'aide à la décision. L'étude géo-décisionnelle d'aptitude zonales pour le choix de l'emplacement géographique de nouveaux sites d'habitations, d'industries et de services s'avère primordiale et constitue un vrai problème de décision à référence spatiale. Un zonage anarchique pour résoudre de tels problèmes peut causer une mutation épidémiologique et une détérioration de la santé des citoyens. Le modèle linéaire de Simon et ses extensions sont insuffisantes pour répondre à la complexité de ce type de problèmes, Fatima et al. (2012).

Les systèmes d'information géographiques (SIG) jouent un rôle important pour l'analyse des problèmes décisionnels ou la composante géographique des données est prise en considération. Les deux domaines de recherche SIG et AMCD quoi qu'ils sont distincts ils s'entraident pour arriver aux meilleures solutions des problèmes géo-décisionnels.

Les travaux d'intégration SIG-AMC se sont multipliés depuis 1990. La plus part de ces travaux depuis 1990 jusqu'a 2004 sont recensés et catégorisés dans Jacek Malczewski

(2006), Dans Randal et al. (2011) il y a constatation de la variété et la complexité des méthodes d'analyse multicritère des problèmes spatiaux, pour remédier a cela les auteurs ont fait un balayage et une classification de toutes les méthodes. La classification a aboutit aux classes suivante :

- Méthodes d'agrégation non compensatoires
- Méthode de pondération
- Méthode d'agrégation compensatoire
- Méthodes de sur classement (Electre, Prométhée …)
- Méthode de programmation mathématique.
- Méthodes heuristiques (MOLA, GA, SA…)

D'autres travaux plus récents comme dans Valentina et al. (2003) ou l'objectif était d'estimer les valeurs écologiques de la région de Piedmont au nord de l'Italie et de générer des cartes géographiques pour les utiliser comme variables d'aide a la décision dans le domaine de la planification et de la gestion du territoire pour des fins de protection de l'environnement et des écho systèmes. Dans S'habou et al. (2011) L'objet de l'étude était de trouver une géographie adéquate pour jeter les margines (eaux résiduaires issues de la trituration d'olives). Dans L.pugnet et al. (2013) l'objectif est de mesurer la vulnérabilité des interfaces habitat foret, les auteurs ont utilisés la méthode AHP (Saaty 1980) pour traiter six critères vulnérables de décision (aménagement, Topographie, Structure de la végétation, structure de l'habitat, Propriétés des constructions, Structures socio-économiques). Ils ont procédé à la cartographie de la vulnérabilité de chaque critère a part en utilisant ARCGIS. Dans Carlo et al. (2003), l'objectif est d'alléger le mécontentement de certains groupe de citoyens au Québec lors de la planification d'une section de parc linéaire de moins de 15 km par la région de port neuf du Québec. Un autre travail qui rentre dans le cadre de la diversification énergétique consiste à concevoir un modèle AMC-SIG pour guider un projet sur l'énergie éolienne au canada, Vazquez et al (2011).

3 Approche proposée

Le système global est composé de trois modules (Figure 1) :

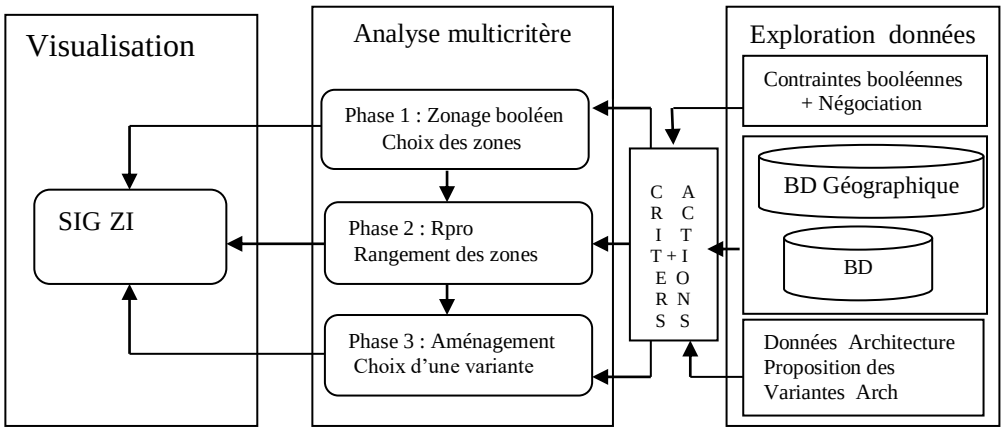


Figure 1 : Architecture générale du système

a- Le module « visualisation » : Puisque les actions dans cette étude sont toutes spatiales, SIGZI assure l'affichage de ces zones sur la carte géographique de l'Algérie avant et après chaque phase décisionnelle. Pour accomplir cette tache le mode vecteur est adopté, chaque zone industrielle est une entité géographique du type abstrait spatial « POINT », elle est implémentée a l'aide de ces deux composantes de position géographique (Latitude et longitude).

b- Le module « Exploration des données » : Les principales entités de notre démarche décisionnelle multicritère sont les critères et les actions, les données qui les caractérisent sont recueillis à partir des bases de données géographiques, socioéconomiques et de climat ainsi que les archives des régions, les cartes critères sont construits. L'utilisation des données qui modélisent les préférences des décideurs et les poids octroyés sont assuré aussi par ce module.

c- Le module « Analyse multicritère » : C'est le module principal, il agit pour la solution du problème décisionnel globale en trois phases (L'étude d'aptitude et le choix géographique de la zone, le rangement des zones choisies, le choix d'une variante en matière d'aménagement). Seule la phase 2 est explicitée dans ce papier.

- La première phase réalisée par l'ANIREF (2013) constitue une recherche d'aptitude zonale. Une telle recherche est au cœur des processus intervenant dans l'aménagement du territoire et constitue une composante majeure d'aide à la décision, Régis et Collet (2011). Le zonage par sélection booléenne utilisé appartient à la classe des méthodes d'agrégation non compensatoire Randal et al. (2011) qui s'opère selon des règles comme : Si (transport électrique HT > 10 m de la zone) alors (zone apte) sinon (zone inapte). L'aptitude d'une zone est calculée avec l'intersection de plusieurs indices. $Iap_{t,j} = C_{1,j} \cap C_{2,j} \cap \ldots \cap C_{,k\ ,j}$, $C_{k,i}$: valeur binaire du critère k pour la zone j. Le résultat est discuté, commenté et complété par négociation.
- La deuxième étape est le rangement total des zones en utilisant les valeurs qualitatives et quantitatives des critères retenus. La méthode de sur classement PROMETHE est utilisée, le comportement de chaque action par rapport aux autres est apprécié par trois flux J.P. Brans et Ph.Vincke (1985) ;

Le flux de surclassement sortant: $\varphi^{+}(a) = \sum_{x \in A} \pi(x, a)$,

Le flux de surclassement entrant : $\varphi^{-}(a) = \sum_{x \in A} \pi(a, x)$

Le flux de surclassement global : $\varphi(a) = \varphi^{+}(a) - \varphi^{-}(a)$

$\pi(a,b) = (1/m)\sum_{j=1}^{k} P_j(a,b) w_j$ est l'indice de préférence d'une action **a** par rapport a une autre **b**, **A** : ensemble des actions, **m** : nombre de critères, $P_j(a,b)$:fonction de préférence de l'action **a** par rapport à **b** vis-à-vis du critère ***j*** , w_j : Poids du critère j. La valeur du flux global determine le rang d'une action.

- La troisième phase consisite à choisir une parmi trois variantes proposés, les critères de choix étant, l'architecture, le cout d'aménagement et le nombre d'ilots morcelés ainsi que les types d'investissements prevus.

4 Etude de cas

4.1 L'ensemble des actions :

Sur les 39 zones industrielles créés a travers l'ensemble du territoire nationale par ANIREF (2013) notre étude s'est portée sur les zones industrielles de l'ouest Algérien. Chaque zone constitue une action (**A1 :** Maghnia, Tlemcen. **A2 :** Sidi Bel Abbes. A**3:** Ras Elma, Sidi Bel Abbes. **A4:** Sidi Ahmed, Saida. **A5 :** Horchaia, Naama. **A6 :** Tamazzoura, Ain Témouchent. **A7 :** Oggas, Mascara. **A8 :** El Haciane, Mostaganem. **A9 :** Sidi khettab, Relizane).

4.2 Les critères :

Les critères utilisés dans cette étude sont classés en trois catégories: les contraintes naturelles, Les critères socio-économiques et juridiques et les contraintes environnementales. Selon ces catégories, 8 critères d'évaluation différents sont définis. La figure 2 présente la hiérarchie des critères de jugement.

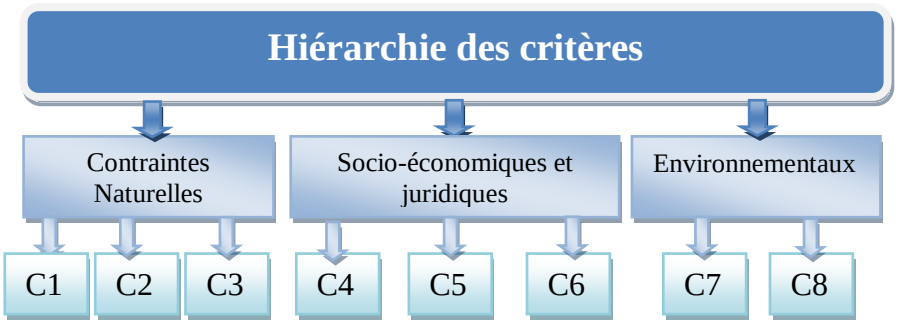


Figure 2: hiérarchie des critères de jugement.

- **Critère (C1) :** Sismicité.
- **Critère (C2) :** Contrainte climatique : Pluviométrie.
- **Critère (C3) :** Contrainte climatique : Température.
- **Critère (C4) :** Superficie.
- **Critère (C5) :** Cout d'aménagement.
- **Critère (C6) :** Proximité des réseaux de transport.
- **Critère (C7) :** Contrainte bioclimatique
- **Critère (C8) :** Proximité au centre urbain d'habitation.

Pour évaluer les différentes Zones à ranger sur la base des critères qualitatives, on a associé à chaque critère qualitatif un barème de notation (échelle de 1 à 5) de façon à en faire une dimension mesurable.

Le principe utilisé est d'évaluer les zones (actions) par rapport aux critères sur la base de la cartographie. Le procédé d'évaluation consiste à analyser la position géographique des zones industrielles sur les cartes thématiques correspondantes à chaque critère géographique (Sismicité, humide,…).

- Contraintes naturelles :

  C1- Sismicité : Le zonage sismique du territoire algérien (Figure 3) révèle cinq zones sismique.

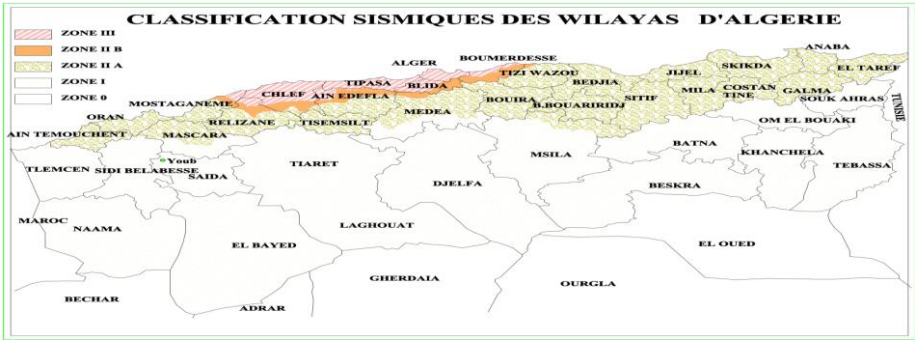


Figure3: classification sismique des wilayas d'Algérie (ANIREF, 2013).

La table 1 si dessous regroupe les différentes valeurs utilisées pour l'évaluation des actions selon le critère de sismicité à partir de la carte sismique des wilayas d'Algérie, selon l'échelle de mesure proposée.

Les actions	Sismicité	Valeur Numérique
A1	faible à modère	2
A2	faible à modère	2
A3	faible à modère	2
A4	faible à modère	2
A5	Faible	1
A6	Modère	3
A7	Modère	3
A8	fort à modère	2
A9	Modère	2

Table1: évaluation des actions selon le critère de Sismicité.

C2, C3- Contraintes climatiques *:* Les valeurs numériques moyennes de ces deux critères sont issues des stations climatiques installées sur le territoire national.

Les actions	Pluviométrie (mm)	température C°
A1	350	19
A2	310	24
A3	410	17
A4	380	19
A5	190	17
A6	400	18
A7	320	21
A8	350	20
A9	370	19

Table2: évaluation des actions selon les deux critères, Pluviométrie et température.

- Critères socio-économique et juridique **:**

C4: Superficie : C'est une information quantitative représentant la superficie des zones industrielles.

C5 : Cout d'aménagent : C'est une information quantitative représentant le cout d'aménagement. Il faut remarquer que la situation géographique du site (sol, pente, altitude …) influe directement sur le mentant et indirectement sur le poids de ce critère.

C6 - Proximité des réseaux de transport (routes, chemin de fer, aéroport) : L'évaluation de ce critère se fait par une analyse spatiale qui consiste à comparer cartographiquement les deux cartes thématiques, celle de la situation géographique des zones en question avec celle des réseaux de transport de proximité (ANIREF, 2013)

Les Actions	Coutd'aménagement (DA)	Superficie(Ha)	Proximité
A1	900592576	104	2500
A2	867750000	100	4100
A3	523765223	60	5000
A4	867750000	100	6500
A5	1301625000	150	3500
A6	1778911797	205	3000
A7	851772119	98	8100
A8	1735585907	200	6500
A9	4338750000	500	3000

Table 3: Évaluation selon cout d'aménagement, superficie et proximité réseaux de transport.

- Critères environnementaux: La carte si dessous présente les étages bioclimatiques de l'Algérie.

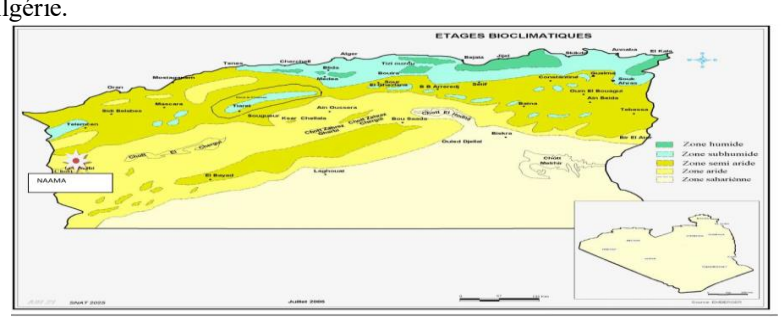


Figure 4: Etages Bioclimatiques des wilayas d'Algérie. (ANIREF, 2013).

C8 : Proximité aux centre urbain d'habitation : C'est la cause des nuisances sonores liées à l'intensification des flux de circulation notamment des poids lourds, de propagations des gazes nocives pour la santés respiratoires des citoyens et des rejets industrielles liquides et solides.

Les actions	Description(V Num)	Proximité (m)
A1	Zone semi aride (2)	14000
A2	Zone semi aride (2)	17000
A3	Zone semi aride (2)	13500
A4	Zone semi aride (2)	15000
A5	Zone aride (1)	18000
A6	Zone semi aride (2)	16500
A7	Zone semi aride (2)	18300
A8	Zone semi aride (2)	13000
A9	Zone aride (3)	17800

Table4: Contraintes bioclimatiques des zones et Proximité aux centres urbains

Les poids des critères sont définies par L'équipe technique de la direction générale de L' ANIREF qui procédé selon les étapes suivantes :

- Classement des huit critères par ordre d'importance décroissant selon un jugement unanime issu d'une consultation entre tous les membres de l'équipe (ingénieurs, techniciens et gestionnaires).
- La deuxième étape consiste à répartir un ensemble de 100 points entre les différents critères. Les valeurs des poids finaux sont données dans le tableau (Table 5) :

Critère	Description de Critère	Poids (%)	Poids (point)
C1	Sismicité.	10%	10
C2	Contraintes climatiques : Pluviométrie	5%	5
C3	Contraintes climatiques : Température.	5%	5
C4	Superficie	20%	20
C5	Cout d'aménagent	15%	15
C6	Proximité des réseaux de transport	20%	20
C7	Contraintes bioclimatiques	5%	5
C8	Proximité au centre urbain d'habitation	20%	20
		100%	100 point

Table 5: Table des poids intra critères.

Notons qu'un critère peut être un facteur à maximiser pour converger vers l'optimisation de la décision ou une contrainte à minimiser. Le sens de chaque critère à été adopté selon l'avis de l'expert (table6).

Après l'évaluation des zones par rapport aux différents critères, la pondération intra critère et la détermination du sens de chaque critère nous avons obtenu la table de performance suivante :

Critère/Action	C1	C2	C3	C4	C5	C6	C7	C8
A1	2	350	19	104	900592576	2500	3	14000
A2	2	310	24	100	867750000	4100	3	17000
A3	2	410	17	60	523765223	5000	3	13500
A4	2	380	19	100	867750000	6500	3	15000
A5	1	190	17	150	1301625000	3500	2	18000
A6	3	400	18	205	1778911797	3000	3	16500
A7	3	320	21	98	851772119	8100	3	18300
A8	4	350	20	200	1735585907	6500	3	13000
A9	3	370	19	500	4338750000	3000	2	17800
Sens de critère	Min	Min	Min	Max	Min	Min	Min	Max

Tableau 6: Table des performances

4.3 Seuils d'indifférence et de préférence

Le seuil d'indifférence est fixé à 5% de la différence entre le plus haut score et le plus bas tandis que le seuil de préférence est fixé à 10% de la même différence.

Critère	C1	C2	C3	C4	C5	C6	C7
Préférence	2	22	2	44	134503680	560	2
Indifférence	1	11	1	22	67251384	280	1

Table 7 : seuils d'indifférence et de préférence de tous les critères.

4 .4 Résultat

Quoique la signification signalée du résultat provienne de l'utilisation d'une méthode validée et d'un noyau de SIG spécifique aux données réelles de ce cas d'étude, reste une analyse de sensibilité sur les seuils de préférences et d'indifférences pour valider la stabilité de la solution. La table 8 si dessous présente le rangement des zones obtenu.

Les Zones	Flux positif (φ^+)	Flux négatif (φ^-)	Flux Globale (φ)	Rang
A1	0.38531917	0.3-	0.08531916	4
A2	0.2971698	0.5250107	-0.2278409	8
A3	0.36762434	0.35000002	0.017624319	5
A4	0.23181818	0.4375	-0.20568182	7
A5	0.59375	0.2363376	0.3574124	1
A6	0.4321429	0.3094263	0.122716606	3
A7	0.3124035	0.37500003	-0.06259653	6
A8	0.22386363	0.58466977	-0.36080614	9
A9	0.5196429	0.24578992	0.27385297	2

Table 8: Le rangement obtenu.

4.5 Visualisation :

- **Prévisualisation :** Visualisation des ZI sur la carte géographique de l'Algérie avant l'analyse décisionnelle multicritère, les rangs sont affectés aléatoirement (Fig 5) :

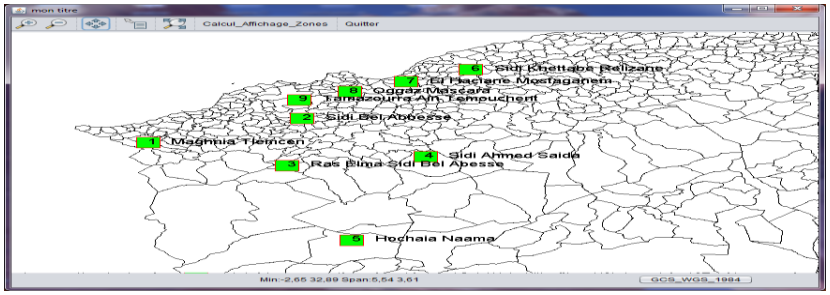


Figure 5 : Carte administrative avec rangement aléatoire des zones

- **Post visualisation** : Visualisation des ZI sur la carte géographique de l'Algérie après l'analyse décisionnelle multicritère avec des rangs issus de l'analyse (Fig6) :

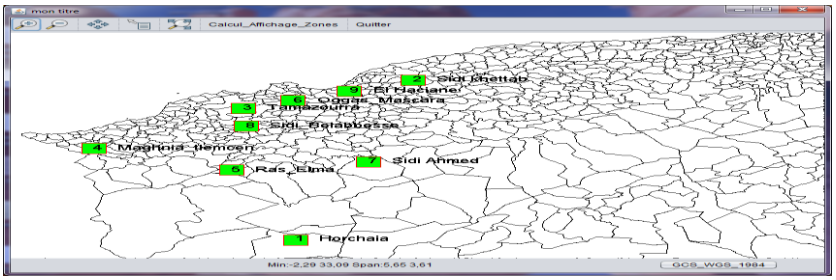


Figure 6: Visualisation des zones rangées après l'analyse.

5 Conclusion et perspectives

Ce travail enchaine une analyse d'aptitude zonale basé sur des méthodes multicritères non compensatoires booléenne. Les contraintes de choix des zones ont été fixées par négociation et par la législation en vigueur. La proposition faite dans ce papier est d'entamer une deuxième phase décisionnelle multicritère pour consolider ces choix. Le caractère quantitatifs et qualitatif des informations recueillis concernant chaque zone a réconforté le décideur et a instauré chez lui une confiance à l'approche d'intégration SIG-AMCD. Cette étude nous a permis de constater l'utilité de l'approche pour beaucoup de secteur ou la décision est importante et dangereuse et se croise avec la géographie et même avec l'histoire. C'est une contribution pour faire sortir l'approche de l'aspect académique vers le terrain. Le rang d'une zone industrielle ainsi obtenu est un indice qui peut :

- Remettre en cause le choix de cette zone.
- Alerter les aménagistes et les constructeurs de la zone.
- Affecter la zone au projet d'investissement adéquat.

Nos perspectives est de prolonger l'étude sur toutes les zones industrielles à l'échelle national ce qui nous amène à revoir le choix de la méthode multicritère à utiliser. Le choix de la méthode est une étape cruciale surtout dans notre cas d'étude, quatre approches peuvent être utilisés : Ad hoc, arbre de classification, méthode multicritère ou les systèmes experts Chakhar .(2006) .

Pour remédier à l'inconvénient des méthodes d'AMCD en matière de modélisation des préférences des décideurs nous avons décidé de suivre l'approche proposée par Kary (1996), pour expérimenter plusieurs autres techniques de data mining.

Réference

ANIREF : Agence Nationale d'Intermédiation et de REgulation Fonciére (2013). *Etude preliminaire d'amenagement du parc industriel* . Alger .

B.Roy (1985).*Méthodologie multicritère d'aide à la décision* - Paris : Economica,.

B.taibi (2010). *L' analyse multicritère comme outil d'aide à la décision Aplication de la methode PROMETHEE etude de cas:l'entreprise SEROR*. These de Magister ,Université de Tlemcen,Algerie.

Carlo.Prévil, Marius.Thériault et Joël.Rouffignat (2003). *Analyse multicritère et SIG pour faciliter la concertation en aménagement du territoire : vers une amélioration du processus décisionnel*. Les Cahiers de géographie du Québec. : p. 35-61,. - n° 130 : Vol.. 47. DOI: 10.7202/007968ar

Chakhar Salem (2006). *Cartographie Décisionnelle Multicritère : Formalisation Et Implémentation Informatique* . Thèse de doctorat , Université paris dauphine.

Deliverable 4B(2010). Multi-criteria analysis and ranking of alternative waste technologie/management system. Rapport de recherché, Faculté des sciences Eljadida, National technical university of Athens, Municipality of the urban communoty of Azemmour, Maroc.

Egenhofer M.J A.(1989). *Formal Definition of Binary Topological Relationships*, Proceedings of the 3th International Conference on Foundations of data Organization and Algorithms. - Paris, France Lecture Notes in Computer Science 367, 1989. - pp. 457-472 .

H.Laurent (2000). *Systémes d'évaluation et de classification multicritères pour l'aide à la décision*. Thése de doctorat Université Paris Dauphin .

J.Malczewski (2006). *GIS-based multicriteria decision analysis: a survey of the literature*, International Journal of Geographical Information, 20:7, 703-726, DOI: 10.1080/13658810600661508

J.P. Brans, Ph.Vincke (1985). A preference Ranking Organisation Method:(The PROMETHEE Method for Multiple Criteria, Management Science, Vol. 31, No.6(Jun., 1985), PP.647-656 Published by: INFORMS.

L. Pugnet a, E.Maillé (2013). *Analyse multicritères pour l'évaluation de la vulnérabilité des interfaces habitat-foret, international conference on forest* « fire risk modelling and mapping » Aix en Provence, France.

Marc.souris (2002*). Les principes de systèmes d'information géographique* : *Principes algorithmes et architecture du système SAVANE*, These de doctorat IRD*(*(Institut de recherche pour le développement) France.

Pugnet L. et E.Maillé (2013). *La modélisation et la cartographie des risques de forets*, International conférence on Forest risk modelling and mapping.

S'habou R ,Zairi M , Kallel A, Neji J. Ben Dhia H. (2011). *Intégration du SIG et des méthodes d'analyse multicritère pour la gestion de la pollution*: cas de stockage des margines Sfax Tunisie, Séminaire International, Innovation & Valorisation en Genie civil & Materiaux de construction, Rabat Maroc: INVACO2 N° : 5O-309,

Raffaella Balzarini, Paule-Annick Davoine et Muriel Ney (2012). *Evolution et développement des méthodes d''Analyse spatiale multicritère pour des modèles d''aptitude : L'exemple des applications en Géosciences*. Laboratoire d'Informatique de Grenoble (LIG) équipes Steamer et Metah.ESRI France, Département Education et Recherche.

Randal Greene, RodolpheDevillers, Joan E.Luther and BriaG.Eddy (2011) . *Gis-Based Multiple-Criteria Decision Analysis* Department of geography, Memorial University of Newfoundland, Canadian Forest service, Natural resources Canada. Geographycompass 5/6.

Régis.Caloz Claude collet (2011). *Analyse spatiale de l'information géographique* , Presse polytechniques et universitaires romandes.

S.BenMena (2000). *Introduction aux méthodes multicritères d'aide à la décision* - Gembloux , B A S E *Biotechnol. Agron. Soc. Environ.* Unité de Mathématique. Faculté universitaire des Sciences agronomiques de Gembloux.

Valentina Ferretti , Silvia Pomarico , *integrating Multicriteria Analysis and Geographic Information Systems for studying ecological corridors in the Piedmont Region* 74th Meeting of the European Working Group "Multiple Criteria Decision Aiding"

Fatima Zohra Younsi , DjamilaHamdadou, Bouziane Beldjilali (2012). *Proposition d'un Système Interactif d'Aide à la Décision Spatiale : Télédétection, SIG et Analyse Multicritère*, Université d'Oran Es-Senia .

Vazquez, Maria de L. WAAUB, Jean-Philippe CHAUMEL, Jean-Louis. (2011), *Analyse spatiale et approche d'aide multicritères et multi-acteurs à la négociation pour évaluer des scénarios d'implantation des parcs éoliens*, publié dans "1ère Conférence Intercontinentale d'Intelligence Territoriale " Interdisciplinarité dans l'aménagement et développement des territoires", Gatineau : Canada.

Michel grabisc (2004). *Une approche constructive de la décision multicritère*, manuscrit Université Paris I.

Gregory A. Kiker, Todd S. Bridges, Arun Varghese, Thomas P. Seager, and Igor Linkov (2005). *Application of Multicriteria Decision Analysis in Environmental Decision Making* , Integrated Environmental Assessment and Management - Volume 1, Number 2 - pp. 95-108.

Kary Frakling (1996). *Modélisation et apprentissage des préférences par réseaux de neurones pour l'aide à la décision multicritère*. Thèse doctorat, Ecole nationale des mines de Saint-Etienne.

Abstract: Integration of Geographic Information Systems (GIS) and multi-criteria analysis is a privileged and indispensable way to evolve GIS into real decision support systems. RPRO4SIGZI, the system proposed in this article allows, from a detailed study of geographical, environmental and socioeconomic criteria, to cooperate GIS and multi-criteria analysis method for geo-graphical choosing of the right site for installing an industrial projects. The result obtained by RPRO (Ranking PROMETHEE) for ranking industrial areas in western Algeria is refined by a viewing SIGZI (Geographic Information System for Industrial Zones). The RPRO unit proceeds to rank industrial areas using the rankings PROMETHEE method issue from European school and SIGZI module to the visualization of these areas on the map. RPRO4SIGZI system was designed for the evaluation of a new methodology of multi-criteria analysis guided by data mining. The objective is to show how data mining is used to model the preferences of the decision maker and generate performance tables. Only RPRO4SIGZI system is presented in this paper.
Keywords: Geographic Information System (GIS), multi-criteria analysis, Industrial Area, MCA-GIS integration, Cartography.

Contribution à la visualisation décisionnelle : Problème des perdus de vue de la vaccination

Fatima Zohra Benhacine*, Baghdad Atmani *, Fouzia Abedlouheb *

* Laboratoire d'Informatique d'Oran – LIO
Université d'Oran 1 Ahmed Benbella
BP 1524, El M'naouer 31000 Oran, Algerie.
benhacine.fatima@gmail.com
atmani.baghdad@gmail.com
fzabdelouheb@gmail.com

Résumé. La visualisation de connaissances est l'utilisation de représentations visuelles à des fins de création et de partage. Elle exploite les techniques de visualisation d'informations et consiste à traduire l'information retenue dans une forme visuelle pertinente. Il s'agit de construire un schéma adéquat, par rapport à l'information, qu'il faut communiquer ainsi que les relations logiques sous-jacentes.
Notre contribution dans ce domaine est une nouvelle démarche d'intégration de la visualisation dans l'aide à la décision en utilisant les techniques de datamining et du web sémantique.
En collaboration avec les services SEMEP d'ORAN nous avons expérimenté notre démarche dans la détection par visualisation des perdu de vue du programme élargi de vaccination et ceci en trois étapes: 1) la construction de l'ontologie de vaccination. 2) l'extraction de l'information pertinente par les règles d'association et enfin 3) la visualisation pour l'aide à la décision. Seules les deux premières étapes sont présentées dans ce papier.

1 Introduction

Dans le monde en général et en Afrique en particulier la vaccination est devenue, à travers les différents programmes élargis de vaccination, l'une des stratégies efficaces de lutte contre les maladies virales et bactériennes, elle a été érigée en un programme de santé par l'Organisation Mondiale de la Santé (OMS) à la fin des années 1970. Ce programme a pour mission de protéger les enfants de 0 à 11 mois contre les six (6) maladies les plus mortelles. Il s'agit de la tuberculose, de la diphtérie, du tétanos, de la coqueluche, de la poliomyélite et de la rougeole. En effet, ces maladies sont responsables de plus de deux (2) millions de décès par an dans le monde selon l'OMS. Pourtant, les moyens efficaces pour protéger les enfants contre ces maladies ont été largement vulgarisés dans le monde. Malheureusement, un des problèmes qui se pose que tous les enfants qui se présentent pour la première vaccination une semaine après leur naissance ne reçoivent pas au bout des 11 mois la totalité des vaccins pour

de multiples raisons. Ce sont les perdus de vue. Un vrai problème pour les autorités sanitaires du pays qui cherchent incessamment à réduire les causes qui mènent aux problèmes des perdus de vue. C'est pourquoi nous nous sommes fixés comme objectif la mise en place d'une démarche d'aide à la décision pour la détection des enfants perdus de vue à Oran en Algérie. Le domaine de l'Aide à la décision est particulièrement vaste, source de nombreuses propositions, aussi bien dans le milieu académique qu'industriel. Au milieu des années quatre-vingt sont apparus des méthodes et des outils d'aide à la décision fournissant à un décideur ou une équipe de décideurs des indicateurs et des analyses. Ces outils, appelés système d'information décisionnel, permettent de faciliter l'accès aux données en ouvrant la possibilité à des analyses plus complètes (Little, 1970). Un bon système d'information décisionnel doit être capable, selon des règles et modes opératoires prédéfinis, d'acquérir, évaluer et traiter des données par des outils informatiques ou organisationnels, de distribuer en particulier via le Web les informations à tous les partenaires internes ou externes de l'établissement, notamment d'une manière visuelle. La visualisation décisionnelle consiste en la présentation visuelle d'informations décrivant un phénomène connu, de façon à prendre une décision. Plusieurs possibilités peuvent être envisagées et faire l'objet d'études (simulations numériques, réalisation de maquettes, d'expériences, etc.) avant d'être comparées lors d'un processus décisionnel. La visualisation permet alors de présenter et de comprendre les conséquences de tel ou tel choix, de façon à prendre une décision éclairée. Dans ce contexte, la visualisation a essentiellement pour but d'exprimer de façon claire et synthétique un problème initialement complexe et difficile à appréhender dans son intégralité. La visualisation doit donc mettre en valeur les éléments pertinents pour la décision à prendre, et à leur donner du sens de façon à faire apparaitre de façon claire et compréhensible une structure souvent sous-jacente ou masquée au sein de données nombreuses et complexes (Coninx, 2012).

La visualisation de l'information consiste à traduire, à transcrire ou encore à coder l'information retenue dans une forme visuelle pertinente : il s'agit de construire un schéma (graphique, une image, tableaux, schémas, organigrammes, graphiques, cartes, plans, dessins figuratifs, photographies, etc.) adéquat par rapport à l'information, qu'il faut communiquer ainsi que les relations logiques sous-jacentes. Il faut donc considérer chaque type de représentation visuelle comme une forme de communication spécifique, comme un langage possédant ses règles et son code. Il s'agit de formes d'expression relativement conventionnelles qui doivent être respectées si l'on veut garantir l'efficacité de la communication (Karouach, 2001).

Notre contribution dans ce domaine est la proposition d'une nouvelle démarche de visualisation pour l'aide à la décision guidée par un processus d'extraction de connaissances à partir d'une source ontologique. Le processus Extraction de connaissance à partir de données (ECD) est apparu pour explorer la grande quantité de données et d'en extraire des nouvelles connaissances utiles, aidant à prendre des décisions à propos de divers sujets qui touchent différents domaines. L'ECD utilise la visualisation pour exposer la composante informationnelle profonde contenue dans les données brutes et pour faciliter le processus de découverte de connaissances associé dans ses phases d'analyse et d'interprétation par les experts du domaine (Karouach, 2001). La visualisation de l'information et l'exploration visuelle des données peut aider à faire face aux flux important d'informations. L'avantage de la visualisation, considérée comme une technique de fouille de données (Bonczek et al ,1980), est que l'utilisateur est directement impliqué dans le processus d'exploration des données. Il ya un grand nombre de techniques de visualisation des informations qui ont été développés au cours des dix dernières années pour soutenir la visualisation décisionnelle. La suite de l'article est organisée comme suit : Dans la deuxième section nous faisons un tour d'horizon des travaux déjà réalisés sur

Les systèmes d'aide à la décision Médicale. Dans la section 3 nous décrivons l'approche proposée. Et enfin nous terminons par une conclusion et quelques perspectives.

2 Système d'aide à la décision Médicale

Dans son sens courant, le terme décision renvoie au choix d'une action à faire. Etant donné la grande complexité des problèmes de décision, il est souvent utile de faire appel à une aide extérieure pour la prise de décision. L'aide à la décision n'a pas pour but de remplacer le décideur en lui proposant des solutions « toutes faites ». Elle cherche plutôt à le guider vers des décisions qu'il aura à prendre sous sa responsabilité. L'aide à la décision se rencontre dans de nombreux domaines applicatifs tels que l'économie, les mathématiques, l'informatique, la médecine, etc.

Dans le domaine médical, la décision est le centre ou la raison d'être de l'acte médical, c'est l'acte fondamental de la médecine. Le processus de la décision médicale, en présence d'un patient ou d'une collectivité, consiste à choisir un mode d'investigation, poser un diagnostic puis proposer un traitement ou le différer. De là, l'aide à la décision médicale est définie comme étant l'ensemble des techniques de gestion de l'information susceptibles d'aider, partiellement ou globalement, le médecin dans son processus de décision (Darmoni, 2003).

Plusieurs définitions concernant les systèmes d'aide à la décision médicale ont été proposées. Un système d'aide à la décision médicale est un ensemble organisé d'informations, conçu pour assister le praticien dans son raisonnement en vue d'identifier un diagnostic et de choisir la thérapeutique adéquate, en opérant un dialogue entre l'homme et la machine (Darmoni, 2003). Les systèmes d'aide à la décision médicale (SADM) sont « des applications informatiques dont le but est de fournir aux cliniciens en temps et lieux utiles les informations décrivant la situation clinique d'un patient ainsi que les connaissances appropriées à cette situation, correctement filtrées et présentées afin d'améliorer la qualité des soins et la santé des patients» (Lobach et al., 2007).

Il existe trois grandes catégories de systèmes d'aide à la décision médicale. La première catégorie concerne les systèmes d'aide indirecte à la prise de décisions ou d'assistance documentaire dont l'objectif est de faciliter l'accès aux informations pertinentes en un temps record mais ces systèmes n'ont pas de méthode de raisonnement à proprement parler (Cleret et al., 2001). La deuxième catégorie concerne les systèmes d'alerte ou de rappels automatiques qui permettent de rappeler au médecin des erreurs à ne pas commettre ou des éléments importants à prendre en compte pour la décision. Ils sont plus actifs et plus directement impliqués dans la décision médicale. L'assistance fournie n'est pas une aide au raisonnement ou à l'appréhension globale du cas du patient, mais plutôt un aide-mémoire fournissant une information utile et pertinente dans une situation facile à définir a priori (Degoulet & Fieschi, 1991). Ces systèmes, comme les précédents, ne raisonnent pas véritablement (Cleret et al., 2001). Enfin la troisième catégorie appelés systèmes consultants qui fournissent à l'utilisateur des conclusions argumentées selon les méthodes de raisonnement employées. La conception est intellectuellement plus satisfaisante que celles des systèmes n'utilisant pas de véritables processus de raisonnement. Donc les développeurs s'intéressent principalement à ce type de système où l'on note le plus de réalisations en matière de système d'aide à la décision (Cleret et al., 2001) : on parle de système interactif d'aide à la décision SIAD.

De nombreuses définitions des SIAD ont été proposées ((Little,1970) ; (Gorry et al,1971) ; (Alter, 1980) ; (Bonczek, 1980) ; (Keen, 1980); (Moore et al, 1980)). Ces diverses définitions

ont mis l'accent soit sur le type de problèmes, soit sur les fonctions du système, soit sur ses composants ou encore sur le processus de développement. Nous reprenons ici la définition de (Turban,1993) qui porte à la fois sur les fonctions et la constitution du système : Le SIAD est un système d'information interactif, flexible, adaptable et spécifiquement développer pour aider à résoudre un problème de décision en améliorant la prise de décision. Il utilise des données, fournit une interface utilisateur simple et autorise l'utilisateur à développer ces propres idées ou points de vue.
Un SIAD basé sur l'ECD (Ltifi et al., 2008) est un système qui permet de détecter les stratégies de résolution d'un problème de décision par le biais d'un processus de fouille de données. Dans ce processus, l'analyse des besoins des décideurs, les différentes activités réalisées en rapport avec la préparation et la manipulation des données pertinentes, de même que la visualisation des résultats constituent des étapes très importantes. C'est sur de telles étapes que repose l'acceptation ou le refus par l'utilisateur final de l'outil d'aide à la décision visé. Les interactions homme-machine au niveau de ce système devraient permettre de guider les utilisateurs tout au long des étapes d'ECD ; il est important aussi d'adapter au mieux l'IHM à chaque classe de décideur et/ou à chaque décideur. (Ltifi et al , 2008) propose le schéma visible en figure 1 présentant le déroulement du processus d'un SIAD basé sur un processus ECD, ou encore SIAD basé Data Mining.

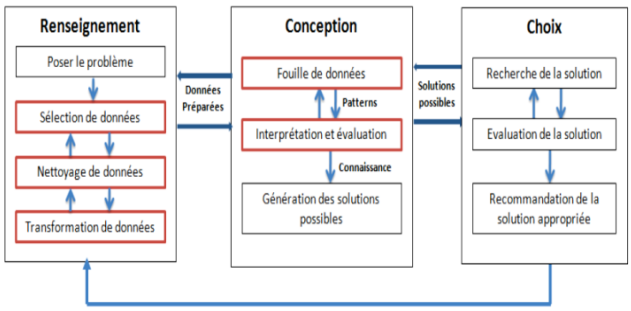


FIG. 1 – *Architecture du SIAD /ECD* **Ltifi et al (2008)**

La figure 1 nous montre l'intégration des étapes de processus d'ECD dans le processus d'aide à la prise de décision. En effet, l'identification du problème, du processus d'ECD, permet de cerner les objectifs et définir les différents objectifs principaux du futur système. Les étapes de prétraitements consistent à construire des corpus de données spécifiques ainsi qu'à faire le nettoyage des données, le traitement des données manquantes, la sélection d'attributs ou la sélection d'instances puis la transformation de ces données. Ces étapes sont cruciales pour la recherche des informations pertinentes du processus de prise de décision (Ltifi et al,2008).
La fouille de données peut alors être opérée pour aboutir à des connaissances mises sous la forme de modèles qui doivent être validés. Des post-traitements sont nécessaires pour rendre ces modèles intelligibles soit par un humain soit par une machine (Ltifi et al, 2008). D'où la génération, l'analyse et le développement des solutions possibles au problème posé, basées sur les connaissances découvertes par le processus d'ECD. (Ltifi et al, 2008) envisagent une décomposition du SIAD/ECD en quatre modules (en se référant aux étapes du processus d'ECD) : saisie et stockage de données, fouille de données, évaluation des données et gestion de connaissances. Chacun de ces modules est considéré lui-même comme un système interactif.

3 Approche proposée

En médecine, la décision est considérée comme étant le centre de l'acte médical. Le processus de la décision médicale consiste entre autres à poser un diagnostic, proposer un traitement ou le différer, etc. Ainsi, de très nombreuses applications d'aide à la décision ont été développées dans ce domaine. Ces applications sont destinées à soutenir le personnel de santé dans leurs prises de décisions. Cela implique l'utilisation de divers outils d'aide à la décision, tel que la visualisation qui peut être vue comme un instrument de gestion permettant de prendre des décisions. Dans le cadre de ce domaine nous nous sommes fixés l'implication des techniques de visualisation dans le processus de prise de décision guidé par fouille de données (datamining). La visualisation pour l'aide à la décision consiste en l'extraction de connaissances dans de grands volumes de données pour en extraire des informations pertinentes. L'approche que nous avons proposée se décompose en trois phases importantes constituant un véritable processus d'ECD (FIG 2) :

1. Phase de prétraitement : cette partie consiste à collecter les données brutes médicales, les sélectionner, prétraiter ensuite les représenter par une ontologie.
2. Phase de traitement de fouille de données : cette partie consiste à extraire des connaissances à partir des données sémantiques ECDS et à lancer une modélisation guidée par datamining et ontologie.
3. Phase de validation : cette partie consiste à développer une interface de visualisation des informations pertinentes pour l'aide à la décision médicales.

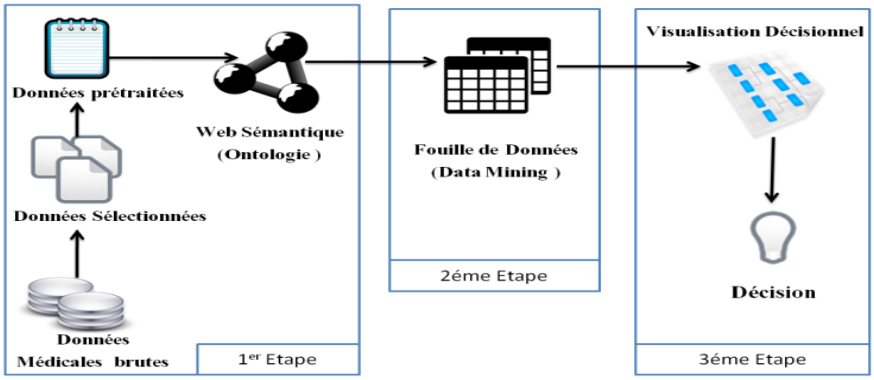


FIG. *2 – Processus de l'approche proposée*

3.1 Phase de Prétraitement

L'une des raisons majeures qui a poussé la recherche sur les ontologies, c'est que ces dernières fournissent un vocabulaire commun d'un domaine pour les chercheurs qui ont besoin de partager l'information, elles retracent le champs de recherche de l'information et elles permettent le partage et la compréhension commune de la structure de l'information entre les personnes impliquées dans le domaine. De plus, elles permettent la réutilisation du savoir sur un domaine.

La conception d'ontologies est une tâche difficile qui nécessite la mise en place de procédés élaborés afin d'extraire la connaissance d'un domaine, manipulable par les sys-

tèmes informatiques et interprétable par les êtres humains. Par le biais de la Fusion automatique nous avons construit notre Ontologie finale pour le SEMEP appelée VaccinOnto dans le cadre du projet PNR. La construction de VaccinOnto est réalisée par la fusion progressive et itérative des différentes ontologies créées séparément sur les vaccins, la vaccination, la couverture vaccinale, les maladies etc. les figures 3 et 4 présentent quelques concepts de l'ontologie VaccinOnto

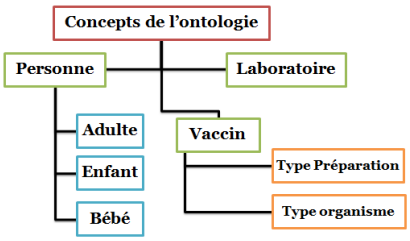


FIG 3– *quelques concepts de l'ontologie VaccinOnto*

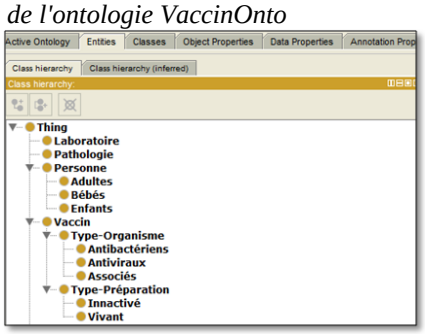


FIG 4 – *Classes de l'ontologie de domaine avec l'outil Protégé*

3.2 Phase d'Extraction des connaissances

La fouille de données (Datamining) utilise une variété de méthodes pour traiter de grandes quantités d'informations afin de découvrir les connaissances utiles pour la prise de décision. Elle constitue donc un support d'aide à la décision dans différents secteurs dont les organismes de santé qui font face à des pressions croissantes pour améliorer la qualité des soins tout en réduisant les coûts. En raison de l'important volume de données générées dans les organismes de santé, il n'est pas surprenant que les organismes de santé ont été intéressés par la fouille de données pour améliorer les pratiques des médecins, la gestion des maladies et l'utilisation des ressources. D'où l'utilisation progressive de la fouille de données dans le domaine médical. La fouille de données est au cœur du processus de l'Extraction de Connaissances des Données (ECD) ou *Knowledge Discovery in Databases* (KDD). Elle se présente comme un ensemble de techniques permettant d'extraire des informations utiles et nouvelles à partir de grandes quantités de données (données de nature médicale, financière, ou économique, expériences scientifiques, ...). Ces techniques permettent de découvrir des modèles prévisionnels, des règles de classification et d'autres types de connaissances qui serviront à l'aide à la décision (Hanifi, 2009). L'ECD se trouve au carrefour de nombreuses disciplines comme l'apprentissage automatiques, la reconnaissance de formes, les bases de données, les statistiques, la repré-

sentation de la connaissance, l'intelligence artificielle, les systèmes experts, etc (Hervouet, 2012). On s'intéresse dans le cadre de ce projet aux règles d'associations. Le processus de fouille de données par la recherche de règles d'association (Agrawal et al, 1993) se base sur une classe particulière de motifs appelés itemsets fréquents ou conjonction d'items fréquents. En s'appuyant sur la particularité de fréquence des itemsets, la technique consiste à mettre en évidence des règles de la forme prémisse (antécédent) -> conclusion (conséquence). Les règles d'association expriment alors à partir des données contenues dans une base de données relationnelle les tendances implicatives entre les attributs de la prémisse, et ceux qui apparaissent dans la conclusion. Le processus de développement de la 2éme étape est décrit par la figure 5

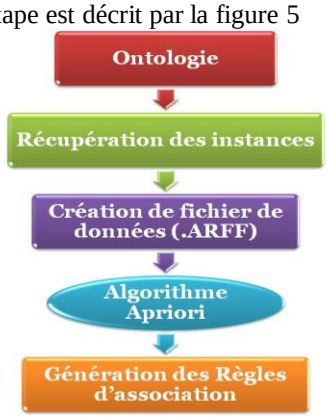


FIG **5** – *Phase de traitement de fouille de données*

La première partie consiste à récupérer les instances de l'ontologie VaccinOnto ensuite à créer un fichier de donnée (.ARFF) pour l'extraction des règles d'associations. Pour illustrer notre approche nous allons utiliser un extrait de l'ontologie suivant (voir figure 6):

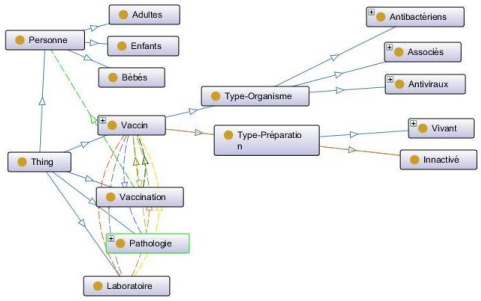


FIG **6**– *Extrait de l'ontologie*

Nous avons pris un échantillon de 100 individus vaccinés de manières différentes, 82 totalement vaccinés (TV) et 18 partiellement vaccinés (PV). Un individu TV est un bébé qui a eu les 6 vaccins obligatoires par contre un individu PV a raté au moins un vaccin parmi les 6 (figure 7).

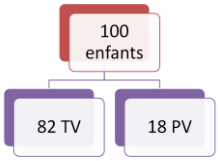
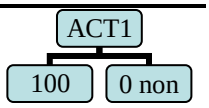
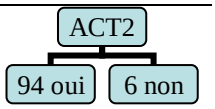
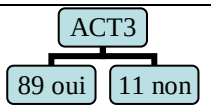
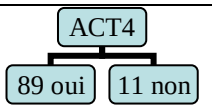
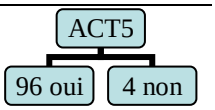
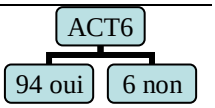


FIG **7**– *description de l'échantillon global*

Pour simplifier la nomenclature nous avons utilisé une indexation pour désigner les différents vaccins en actes comme suit :

Description	
Act 1 : BCG VPO Vaccin contre Tuberculose-poliomyélite	
Act 2 : D.T.C.P 1 Vaccin contre diphtérie, tétanos, coqueluche, poliomyélite	
Act 3 : D.T.C.P 2 Vaccin contre diphtérie, tétanos, coqueluche, poliomyélite	
Act 4 : D.T.C.P 3 Vaccin contre diphtérie, tétanos, coqueluche, poliomyélite	
Act 5 : V.A.R Vaccin contre la rougeole	
Act 6 : D.T.C.P 4 Vaccin contre diphtérie, tétanos, coqueluche, poliomyélite	

TAB. 1 – *description de l'échantillon détaillé*

Chacun des 6 concepts représente une vaccination (TAB1) et chacune des 100 transactions représente un bébé avec ses 6 vaccins (TAB 2) dont on peut lui associer un id.

Le tableau suivant (TAB 2) décrit un extrait du tableau de données.

Id	**Act 1**	**Act 2**	**Act3**	**Act4**	**Act5**	**Act6**	**Classe**
1	oui	oui	oui	Oui	oui	oui	TV
2	oui	non	oui	Oui	oui	non	PV
3	oui	oui	oui	Oui	oui	oui	TV
4	oui	oui	oui	Oui	oui	oui	TV
5	oui	oui	oui	Non	oui	oui	PV
6	oui	oui	oui	Non	non	oui	PV
7	oui	non	non	Non	non	non	PV

TAB. 2 – *description de l'échantillon détaillé*

Nous avons utilisé *WEKA* pour extraire les règles d'associations à partir de notre fichier de données en appliquant l'algorithme Apriori et nous avons fixé un seuil de support de 8% et

Règle	Confiance
Act5=oui ==> Act1=oui	1
Act2=oui ==> Act1=oui	1
Act6=oui ==> Act1=oui	1
Act5=oui Act6=oui ==> Act1=oui	1
Act2=oui Act5=oui ==> Act1=oui	1
Act2=oui Act6=oui ==> Act1=oui	1

TAB. 3– *extraits des réglés d'association*

un seuil de confiance de 80%. L'algorithme a extrait 200 règles d'association. Un extrait des ces règles est présenté dans le tableau TAB. 3

3.3 Phase de validation

La visualisation de règles d'associations est un sujet qui a fait l'objet de nombreux travaux en Data Mining. Dans (Chakravarthy et Zhang, 2003), les auteurs visualisent des ensembles de règles d'association sous forme de tableaux, chaque ligne correspond à une règle et chaque colonne représente une caractéristique des règles. Ils proposent une interface qui permet de filtrer les règles en spécifiant les items que l'utilisateur désire avoir dans la règle ou en fixant un seuil de support ou de confiance. Elle permet aussi de trier les règles par ordre croissant ou décroissant de support ou de confiance. D'un autre coté les graphes sont utilisés pour représenter des relations entre les items. Ils sont donc bien adaptés pour visualiser les règles d'association en reliant leur tête et leur corps par un arc. Les arcs peuvent être orientés (Klemettinen et al, 1994) ou non orientés (Bruzzese et Buono, 2004). Passons aux outils géométriques qui utilisent principalement la distance entre objets, la surface ou encore les coordonnées d'objets pour exprimer des mesures d'intérêt de motifs ou des relations entre les motifs. Les diagrammes en mosaïque ont été introduits dans (Hartigan et Mosaics, 1981) pour visualiser des tables de contingence. La table de contingence est un moyen permettant de représenter simultanément deux caractéristiques observées sur un ensemble de données. Les coordonnées parallèles ont été introduites en 1981 par Inselberg (Inselberg, 1981) comme une méthode de représentation de données multidimensionnelles. Elles sont adaptées dans (Bruzzese et Buono, 2004). pour visualiser des règles d'associations. Blanchard et ses collègues proposent une métaphore visuelle pour représenter des ensembles de règles d'association (Blanchard et al, 2003). Les règles sont visualisées par classe. Chaque classe est identifiée par un itemset et contient deux types de règles : les règles spécifiques et les règles générales. Les règles spécifiques sont celles dont le corps est l'identifiant de la classe. Les matrices constituent certainement le moyen le plus utilisé pour représenter les règles. Elles permettent essentiellement de mettre en évidence les mesures de qualité des motifs. Nous distinguons deux types de matrices : les matrices à deux dimensions (2D) [(Ben Yahia et Nguifo, 2003), (Zhao et al, 2005), (Couturier et Rouillard , 2006)] et les matrices à trois dimensions (3D) [Ben Yahia et Nguifo, 2003), (Zhao et al, 2005)]. Boulicaut et ses collègues ont proposé une représentation matricielle d'ensembles de règles d'association (Boulicaut et al, 1999). Le principe est de placer des itemsets sur chaque axe de la matrice. Une règle est représentée par une cellule. La tête de la règle est l'itemset affiché en ligne et son corps est l'itemset affiché en colonne ou inversement. Dans chaque cellule correspondant à une règle, Dans (Ben Yahia et Nguifo, 2003), les auteurs expriment la confiance des règles par une couleur graduelle des cellules. Plus la confiance est grande, plus la couleur de la cellule est sombre. La visualisation d'ensembles de règles d'association avec des matrices 3D a été introduite dans (Brunk et al, 1997) avec MineSet 6. Ces matrices appelées matrices item à item sont composées de trois axes dont deux sont utilisés pour afficher les items se trouvant dans la tête et le corps des règles. Le troisième axe sert à exprimer les mesures d'intérêt des motifs. Une règle est schématisée par une barre dont la couleur graduelle

exprime son support et la hauteur correspond à sa confiance. La probabilité d'avoir l'item de la tête sans celui du corps de la règle est aussi représenté par une portion de la barre. La visualisation d'information est un domaine de recherche qui se développe et qui a montré son utilité pour faire des découvertes et prendre des décisions. Lorsque les représentations visuelles sont générées, elles ne sont pas forcément statiques ; elles peuvent être dynamiques et l'utilisateur peut les manipuler pour prendre une décision. Seule la partie visualisation des règles d'associations pour l'aide à la détection des perdus de vue reste à détailler et fera l'objet d'un autre article.

Conclusion

Les méthodes de fouille de données ont été bien adoptées et adaptées pour fournir une aide à la décision médicale, permettant au praticien d'améliorer la prise en charge des patients et de rendre l'acte médical plus performant. D'une part, les systèmes d'aide à la décision médicale se définissent comme une suite d'étapes décisives dont la finalité est l'amélioration de la qualité des soins apportés aux patients. De ce fait, les étapes de ce processus doivent être bien planifiées. Certains systèmes d'aide à la décision utilisent la visualisation pour recommander des actions à prendre (Beaudry, 2011). La visualisation est essentiellement un instrument de gestion qui permet aux responsables de prendre des décisions. C'est donc un processus d'aide à la décision qui vise à prévoir des ressources et des services requis pour atteindre des objectifs déterminés, selon un ordre de priorité établi, permettant ainsi le choix d'une solution préférable parmi plusieurs alternatives. Ce choix prend en considération le contexte et les contraintes internes et externes connues actuellement ou prévisibles dans le futur (Jourdain & Frossard, 1995). Ce papier présente une solution à une problématique complexe qui est celle de la visualisation décisionnelle. Nous avons essentiellement essayé, dans cet article, de mettre en valeur l'utilisation combinée du web sémantique et du datamining afin de proposer une approche bien définie, basée sur des données médicales dans un système interactif d'aide à la décision.

Références

Inselberg A. N-dimensional graphics, part i - lines and hyperplanes. IBM LASC Tech. Rep. G320-2711, page 140, 1981.

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

Alter, S. L., Ed. Decision Support Systems: Current Practices and Contuinuing Chalennges, Addison-Wesley 1980.

Bimonte S., Intégration de l'information géographique dans les entrepôts de données et l'analyse en ligne : de la modélisation à la visualisation. Thèse de doctorat, L'Institut National des Sciences Appliquées de Lyon, 2007.

Blanc F.-X., Baneyx A., Charlet J., Housset B., Représentation des connaissances en pneumologie : l'ontologie doit pouvoir aider au codage Revue des Maladies Respiratoires 27, 741—750 , 2010.

Beaudry, E. : Planification d'actions concurrentes sous contraintes et incertitude, Thèse de Doctorat, Université de Sherbrooke, 2011.

Bonczek, R. H., C. W. Holsapple, et al. The evolving roles of models in Decision Support Systems. Decision Sciences 1980

Boulicaut J., Marcel P, and Rigotti C. Query driven knowledge discovery in multidimensional data. In DOLAP '99 : Proceedings of the 2nd ACM international workshop on Data warehousing and OLAP, pages 87–93, New York, NY, USA, 1999. ACM Press.

Bruzzese, D., Buono, P.: Combining Visual Techniques for Association Rules Exploration. In: Proceedings of the International Conference Advances Visual Interfaces, Gallipoli, Italy, May 25-28 (2004)

Brunk C., Kelly J., and Ron Kohavi. Mineset : An integrated system for data mining. In KDD, pages 135–138, 1997.

Burgun A., Rosier A., Temal L., Jacques J., Messai R., Ducheminf L., Deleger L., Grouin C., VanHille P., Zweigenbaum P., Beuscart R., Delerue D., Dameron O., Mabob P., Henry C. Aide à la décision en télécardiologie par une approche basée ontologie et centrée patient. IRBM 32, 191–194, 2011

Choquet R., TeodoroD., Mels G., Assele A., Pasche E., Ruch P., Lovis C., Jaulent M., Partage de données biomédicales sur le web sémantique, Proceedings of the IC'2010, Nîmes, France , juin 2010 .

Cleret, M. _ Le Beux, P. _ Le Duff, F.: Les systèmes d'aide à la décision médicale, Les Cahiers du numérique, vol. 2, N° 2/2001, p. 125-154, 2001.

Coninx A., Visualisation interactive de grands volumes de données incertaines: pour une approche perceptive. Thèse de doctorat, Université Joseph Fourier Grenoble 1, 2012

Couturier Olivier, Rouillard J., Chevrin V., Une approche hybride pour une meilleure visualisation de grands ensembles de règles d'association, 10ème Conférence Francophone ERGO-IA (ERGOIA'06), octobre 2006

Darmoni, S. J.: Titres et travaux, Informatique de santé, Sciences et Technologies de l'Information et de la Communication, 2003.

Degoulet, P. _ Fieschi, M.: Traitement de l'information médicale. Méthodes et applications hospitalières, Masson, 1991.

Gorry, G. A. and S. Scott-Morton, A framework for management information Systems. Sloan management Review 1971

Hervouet D. Visualisation des règles d'association en environnement virtuel 3D interactif. Information Retrieval. 2011.

Hanifi M.: Extraction de caractéristiques de texture pour la classification d'images satellites, Thèse de Doctorat, Université de Toulouse, 2009.

Hervouet D. Visualisation des règles d'association en environnement virtuel 3D interactif. Information Retrieval. 2011.

J. Hartigan and B. Kleiner. Mosaics for contingency. In 13th Symposium on the Interface, pages 268–273, 1981.

Jourdain, A. Frossard, M.: Les nouveaux outils de planification sanitaire, Actualité et Dossier en Santé Publique, n° 11, 1995.

Blanchard J, Guillet F, and Briand H. Exploratory visualization for association rule rummaging. In KDD-03 Workshop on Multimedia Data Mining (MDM-03, pages 107–114, 2003.

Karouach S., Visualisation interactive pour la découverte de connaissances : GeoECD. Dans : Veille Stratégique, Scientifique et Technologique, VSST'2001,Barcelone, 15/10/2001-19/10/2001, Actes I, p. 301-311, octobre 2001.

Keen, P. W. Adaptative Design for Decision Support Systems 1980.

Krzysztof J.C. et Moore G.W. Uniqueness of medical data mining. Artificial Intelligence in Medicine, Vol. 26 (1-2), pp. 1-24, 2002

Klemettinen M, Mannila H, Ronkainen P, Toivonen H, and A. Inkeri Verkamo. Finding interesting rules from large sets of discoveredassociation rules. In CIKM, pages 401–407, 1994.

Laublet P., Charlet J., et Reynaud C., Introduction au Web sémantique, journées scientifiques Web sémantique, Paris , 10 et 11 octobre 2002 .

Little, J. D. Models and Managers: The Concept of a Decision Calculus. Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 1970

Lobach, D.F. _ Kawamoto, K. _ Anstrom, K.J. _ Russell, M.L. _ Woods, P. _ Smith, D.: Development, deployment and usability of a point-of-care decision support system for chronic disease management using the recently-approved HL7 decision support service standard, Studies in Health Technology and Informatics, vol. 129, p. 861-865, 2007.

Ltifi H., Ben Ayed M., Kolski C., Alimi A-M. Prise en compte de l'utilisateur pour la conception d'un SIAD basé sur un processus d'ECD. D. Galaretta, P. Girard, J.C Tucoulou, M. Wolff (Ed.), Actes de la Conférence ERGO'IA 2008, ESTIA, pp. 85-92, octobre, 2008.

Ben Yahia S. and Mephu Nguifo E. Emulating a cooperative behavior in a generic association rule visualization tool. In ICTAI '04 : Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'04), pages 148–155, Washington, DC, USA, 2004. IEEE Computer Society

Moore, J. H. and M. G. Chang. Design of Decision Suport Systems. 1980

Turban E. Decision Support and Expert Systems. Macmillan, New York, 1993.

Chakravarthy S. and Zhang H. Visualization of association rules over relational dbmss. In SAC, pages 922–926, 2003.

Zhao K, Liu B, Thomas M. Tirpak, and Weimin Xiao. Opportunity map : a visualization framework for fast identification of actionable knowledge. In CIKM '05 : Proceedings of the 14th ACM international conference on Information and knowledge management, pages 60–67, New York, NY, USA, 2005. ACM Press.

Summary

The visualization of knowledge is the use of visual representations for purposes of creation and division. It exploits the techniques of visualization of information and consists in translating the information retained in a relevant visual shape. It is a question of building an adequate plan, with regard to the information, which it is necessary to communicate as well as the underlying logical relations. Our contribution in this domain is a new approach of

integration of the visualization in the decision support using the techniques of datamining and of semantic Web.

In collaboration with SEMEP services of ORAN, we experimented our approach in the detection and visualization of lost sight of an expanded immunization program and this in three steps: 1) the construction of the ontology of vaccination. 2) The extraction of relevant information by association rules and finally 3) visualization for decision support.

Microarray Data Integration for Efficient Decision Making

Fadoua Rafii*, M. Aït Kbir * and B. D. Rossi Hassani **

* LIST Laboratory, CED Sciences and Techniques of Engineers,
Tangier, Morocco
fadoua.rafii1@gmail.com, m.aitkbir@fstt.ac.ma
** LABIPHABE Laboratory, CED Sciences and Techniques of Engineers,
Tangier, Morocco
bd.rossi@fstt.ac.ma

Abstract. The Microarray data are organized as large matrices of expression levels of genes, each row representing a gene, and each column representing a condition of a specified experiment. The fact that these data are accumulated in multiple different datasets represents an impediment for researchers specifically the biologists; in fact, they are confused on getting the right and adequate information. Integrating Microarray data is regarded to be of high importance to increase the reliability and the generalization of the results. By exploring the most used Microarray databases, we have found variations among the data generated by different Microarray experiments, which make the data integration a big challenge. This paper presents an integration scheme that aims getting the valuable results and making pertinent decisions for the hypothesis and crucial questions of the investigators.

1 Introduction

With the benefits of high-throughput technology, Microarray is a promising tool that allows biologists to measure hundreds of thousands of gene expressions simultaneously (C.-R. Chen et al., 2012). While Microarrays are an extremely powerful technique and have produced an enormous amount of new knowledge, they have several limitations (M. Mooney and S. McWeeney, 2014). Handling and analyzing the large volume of data generated by Microarray experiments are not trivial tasks. Thus, we focus on exploring the wealth of Microarray data by integrating data obtained from different datasets, with the same biological issue. The integration of heterogeneous Microarray data will allow better analysis of the complexity in gene expression data. Indeed, the process of integrating Microarray data should be adapted in the sense of leading the investigators taking momentous decisions. To integrate multiple Microarray datasets, two meta-learning approaches have been reported in the literature: (a) merging multiple studies through the combination of raw data of primary studies (F. Hong and R. Breitling, 2008); (b) functional integration of the resulting signature identified in different published studies (P. Cahan et al., 2007). The adopted scheme of integration seek to improve the decision making such as molecular classification, identification

of diagnostic, prognostic signatures and robust biomarkers for early detection. By Microarray data integration from different sources we try to give à visualization in a comprehensive and simplified manner. Our strategy of integration consists on increasing the number of conditions by integrating generated Microarray datasets that have the same biological objectives, in order to maximize the use of the abundant Microarray data being stockpiled by different research groups. To improve the integration accuracy, we will combine gene expression datasets related to a specific human disease for finding informative Microarray data. In this work, we highlight the worthwhile of the information obtained by Microarray data integration. It's followed by exploring the adopted approach and the steps for combining Microarray data. The decision of choosing the method of combining the data on the gene expression level in different Microarray data sets is a challenging problem because gene expression levels generated by different experiments are not necessarily directly comparable. In this paper, we propose a method to integrate two different Microarray datasets. At the first stage, we apply the filtering methods for selecting the common set of genes that exist on the datasets before combining them. In the second stage of the process, we implement the PCA technique for visualizing the efficiency of integrating gene expression of different Microarray experiments. Since this second stage produces different results, our approach is capable of increasing information accuracy. The obtained results showed that our approach was able to provide high accuracy where the sample size of the combined datasets grew larger.

2 Microarray Technology

Microarray is a technology which allows quantitative, simultaneous monitoring and expression of thousands of genes (Afshari, 2002). In a Microarray, many thousands of spots are placed on a rectangular grid with each spot containing a large number of pieces of DNA from a particular gene (C. Naidu and Y. Suneetha, 2012).

2.1 Microarray Experiment

Once the biological questions are fixed, the experimental planning begins on taking into consideration the questions, the conditions or materials involved into the experiment, the sources of variation, and the used methodology.

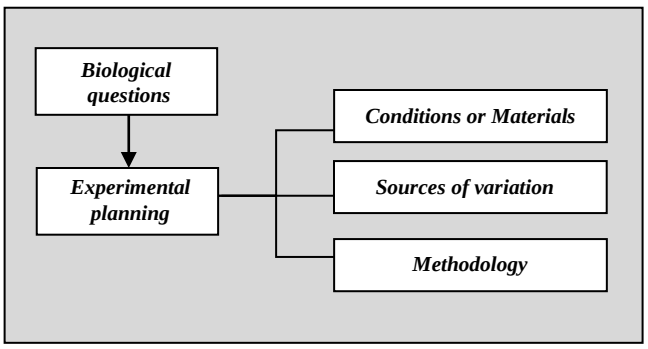


FIG. 1 – *Experimental planning*

For a specific question which is addressed by Microarray experiment, most laboratories don't take just one-time experience. They followed the first experiment by others experiments to verify and to test additional factors. Microarray experiments address scientific tasks such as:

- The identification of co-expressed genes;
- The discovery of gene groups that have similar expression patterns;
- The identification of genes whose the patterns of expression are highly differentiating (e.g., tumor types);
- The study of the activity of the gene patterns under different stress conditions (e.g., chemical treatment).

The objects of DNA Microarray experiments implicate the comparison of gene transcription which is the gene expression in two or more types of cells like cardiac muscle versus prostate epithelium, and in cells that are exposed to various conditions, for example biological conditions such as normal versus disease. The various steps of the Microarray experiment are:

1. The preparation of the probe by preparing one DNA Microarray per patient.
2. The preparation of the target sample by obtaining, purifying, and making the dye of the target mRNA samples.
3. The preparation of the reference sample by obtaining and preparing reference or control mRNA and labeling it.
4. The reaction of hybridization by hybridizing target and reference mRNA with the cDNA on the array.
5. Washing up of the dishes.
6. The detection of the red and the green intensities by scanning the array to determine how much target and reference mRNA is bound to each spot.
7. Determining and recording the abundances of relative mRNA.

2.2 Microarray Data

The Microarray data are stored as a matrix of gene expression where each row represents a gene and each column represents a condition.

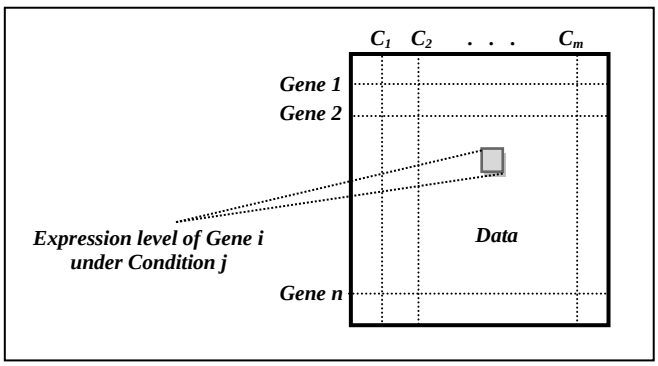


FIG. 2 – *Microarray data matrix of n genes and m conditions*

The gene expression profile describes the expression values for a single gene across many experimental conditions, and the array profile describes the expression values for many

genes under a single condition. The Microarray matrix of gene expression is described by *n x m*:

- n is the dimension of genes of the Microarray experiment;
- m is the dimension of conditions implicated on the specified Microarray experiment.

2.3 Microarray Databases

Microarray experiments produce tremendous quantities of raw data: depending on the platform and the specific array being used, a single experiment can generate millions of data points (Sherlock G., 2001). The Microarray technology has resulted huge amounts of information that we can categorize into:

- The individual samples;
- The processing of the samples;
- The hybridization of the arrays.

The important issue in the Microarray process is to manage the raw data and the associated experimental information. For this reason, the Microarray databases were created, in the object of organizing and retrieving the raw data. These databases are facilitating the storage, management, and retrieval of experimental Microarray data. Microarray data is managed in a way that we can retrieve subsets of data stored in the format of expression values of genes exposed under the specified conditions. The Microarray resources have efficient role in keeping track of experimental information and making it accessible by different kind of users. It provides many advantages such as comparing results, visualizing the data, making further analysis and responding to crucial issues like extracting gene markers of diseases. We have explored some of the Microarray repositories that are the most used:

Repository	Reference
Gene Expression Omnibus (GEO)	(Barrett et al., 2013)
ArrayExpress	(Parkinson et al., 2011)
RNA-seq Atlas	(Kapushesky et al., 2012)
Expression Atlas	(Kodama et al., 2012)
ReCount	(Krupp et al., 2012)
Sequence Read Archive (SRA)	(Frazee et al., 2011)

TAB. 1 –*Public repositories containing Microarray data.*

3 PROBLEMATIC

Microarray technology has known an extensive use as a potential prognostic and diagnostic tool. The results obtained from Microarray experiments represent wealth information for reproducibility and comparisons. The main obstacle facing researchers interested on Microarray fields is the decision making by analyzing the resulted data. The integration of the datasets can yield more efficient information, for that, we have focused on integrating Microarray gene expression data that are derived by different studies. The suggested methodology transforms the values of gene-expression on a common scale to obtain comparable gene signatures. We have applied our model to publicly available breast cancer gene-expression

datasets. Our study shows that the adopted methodology can relieve the limited data size problem by reporting high accuracies with adopting the integration of multi-experiment data.

4 MICROARRAY DATA INTEGRATION

Generally data integration involves some process that combine data from heterogeneous of data sources in order to provide a unify view of data in a structured form (M. Friedman et al., 1999) (M. R. Genesereth et al., 1997). The data integration process is operates by applying a global data model and by detecting and resolving schema and data conflicts so that a homogeneous, unify view can be provided (M. Lenzerini, 2002) (Ziegler and Patrick, 2004). However, secondary analysis and reproducibility of publicly archived high-throughput gene expression studies still have challenges given gaps in meta-data and study annotation needed to appropriately process and analyze the data (Rung, 2013). Integration of multiple studies that are based on the same technological platform, or, combining data from different array platforms carries the potential towards higher accuracy, consistency and robust information mining (Kumar Sarmah and Samarasinghe, 2010). Thus, the integrated data generated allows constructing a more complete and accurate picture. The sample size within each individual experiment is generally small comparing to the size of the genes. To increase the reliability of the results generated by Microarray experiments, the integration may produce better information for analysis on a specific field. In the recent researches, we find works exploring the advantages of integrating Microarray data, such as:

- Data Integration for Microarrays: Enhanced Inference for Gene Regulatory Networks (Sîrbu et al., 2015)
- Bayesian correlated clustering to integrate multiple datasets (Kirk et al., 2012)
- Integrated Analysis of Multiple Microarray Datasets Identifies a Reproducible Survival Predictor in Ovarian Cancer (Konstantinopoulos et al., 2011)

4.1 The integration methodology

Integration of the expression values consists on combining the results of the individual studies. The Microarray expression data from various studies are integrated, after that some transformations are applied on the expression values to obtain measures that are numerically comparable. The generated data from the individual studies are combined on the object of enlarging the sample size. The further analysis is implemented on the new merged dataset. Our methodology explore the importance of specifying the biological issue by introducing the adaptable query for the used Microarray database, and getting as a result set of Microarray datasets that had the same experimental objectives. After that, we choose the Microarray datasets that look like interesting for studying and analyzing collectively, and we extract only the genes common to the selected datasets that were generated independently by implementing an algorithm for filtering and verification. When combining the datasets, we should take into consideration the correspondence between the genes and their values on each sample. And the values of gene expression in the combined dataset should be verified specifically their variation within all the samples.

4.2 Adopted schema of integration

The proposed methodology is illustrated on the Figure.3 which involves the following steps:

- Retrieving public available Microarray data related to a specific human disease by using the annotations recommended for GEO database
- Selecting the datasets responding to the biological questions and issues
- Verifying the same genes in the selected datasets
- Assuring that the values of gene expression in the datasets are varying on the same rank in order to have comparable values
- Combining the datasets and implementing the technique of PCA in order to get the valuable Microarray information

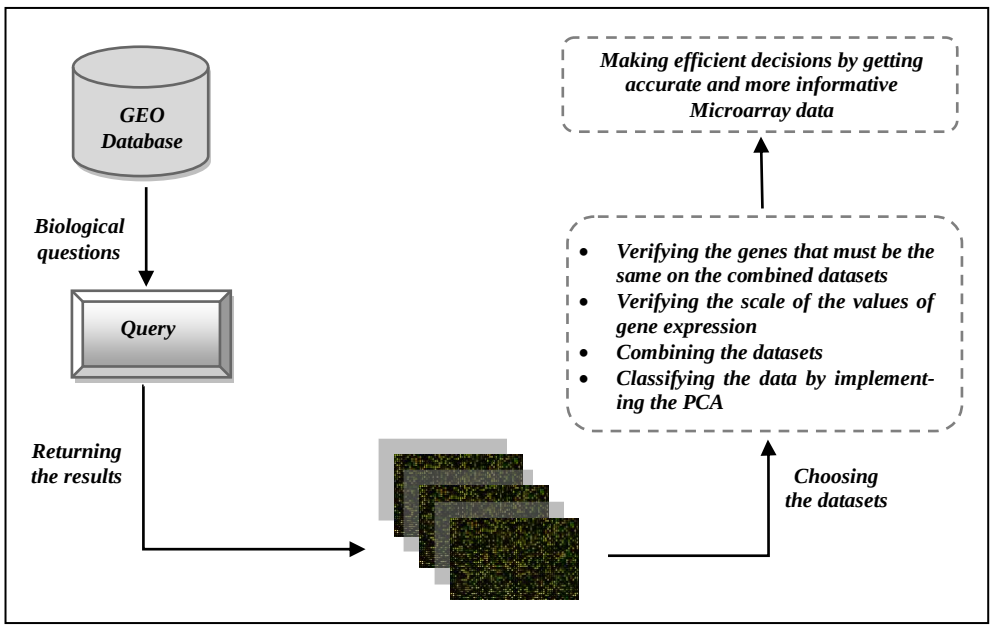


FIG. 3 – *Methodology of integrating Microarray data*

4.3 Microarray Datasets

There are different databases containing Microarray data. For integrating Microarray data, we have chosen GEO (Tanya Barrett and Ron Edgar, 2006), which is one of the largest Microarray database .GEO was originally built to archive the burgeoning volumes of high-throughput gene expression data beginning to be produced by the research community at that time (T. Barrett et al., 2011). The GEO database architecture is designed for the efficient capture, storage, and retrieval of heterogeneous sets of high-throughput molecular abundance data.

To illustrate our method, we first considered two Microarray gene expression data sets collected for Breast cancer studies, and selected from GEO database. The two data sets are described on the following tables:

Feature	Value
GEO accession	GSE2990
Title	Gene Expression Profiling in Breast Cancer: Understanding the Molecular Basis of Histologic Grade To Improve Prognosis
Submission date	July 25, 2005
Last update date	September 11, 2014

TAB. 2 –*First Dataset selected from GEO database*

Feature	Value
GEO accession	GSE3529
Title	Profiling of three different ER+ breast cancer cell lines grown in the presence and absence of estrogen
Submission date	October 28, 2005
Last update date	September 11, 2014

TAB. 3 –*Second Dataset selected from GEO database*

We have chosen these two data sets that have common objective of studying human Breast cancer; the two datasets explore the profiling of gene expression in Breast cancer.

4.4 Results

Before the implementation of the PCA, we had applied the techniques of pretreatments on the selected datasets in order to have efficient Microarray data and to eliminate the genes representing biases on the data. The types of gene expression profiles depicting the biases on data are:

- Gene expression profiles that are empty: the genes that have EMPTY as a noun are not representing valuable information.
- Gene expression profiles containing NaN (Not a Number): some values corresponding for a gene profile and a condition or sample are filled by NaN.
- Gene expression profiles where the variance is less than the 10th percentile of the variance: for each a gene profile, if the variance is less than the 10th percentile of the variance, the gene profile is not important for further analysis
- Gene expression profiles containing absolute values less than a critical value: we fix a value which allow us to detect the gene profiles that aren't representing pertinent data.
- Gene expression profiles containing low entropy values that are common on all the samples.

And we have center and rescale the integrated Microarray data to have comparable values.

Furthermore, the quality of individual Microarray data can directly influence the result of integrated analysis (M. J. Zaki, 2014). It is known that not all of Microarray raw data deposited in public databases provides the same level of meaningful information (O. Larsson et al., 2006). For this reason, we have applied the PCA on each of the two datasets, and after the combination of the two datasets.

Index of Eigen values	Cumulative percentage of the Variance
1	99,36
2	0,21
3	0,09
4	0,04

TAB. 4 –*Values of cumulative percentage of the variance explained by implementing the PCA on the Dataset 1*

Index of Eigen values	Cumulative percentage of the Variance
1	99,69
2	0,15
3	0,11
4	0,02

TAB. 5 –*Values of cumulative percentage of the variance explained by implementing the PCA on the Dataset 2*

Index of Eigen values	Cumulative percentage of the Variance
1	99,30
2	0,20
3	0,10
4	0,08
5	0,03
6	0,02
7	0,01
8	0,01
9	0,01
10	0,01

TAB. 6 –*Values of cumulative percentage of the variance explained by implementing the PCA on the integrated Microarray data*

We have integrated two different datasets that have the same genes to explore the behavior of genes on breast cancer. By comparing the three results after applying the PCA on each dataset and on the integrated data, it shows a remarkable difference where on the datasets, we have found that the information is located on the first principal component; as it is illustrated on table 4 and table 5 where we had over than 99% for the two datasets. On the other hand,

for the integrated datasets as it is depicted on the table 6, it shows that after the integration of the two datasets, we had significant results where the valuable information is distributed on more than the three first principal components to achieve 99,65% of the information.

The results show that the integration of the two datasets is more interesting than the individual datasets, specifically when we want to get information about genes within several samples or conditions that were not explored on the same experiment. For crucial fields such as the detection and diagnostic of cancer, it's more worthwhile to explore the values of gene expression under different samples or conditions to make the right decision and interpretations. In this work, we remark that the combination of independent datasets; generated by individual experiments and treating the same biological question; remains an important key challenge in the genomic medicine and the biology systems. By benefiting from the Microarray technology, which generates huge different data types providing distinct information, the integration provides rich information for the investigators.

5 CONCLUSION

Microarray technology has rapidly provided an extensive source of data for better responding to complex biological questions. The increasing number of studies of gene expression of human and other organisms has given birth to a strong motivation to analyze and study the huge generated data. For these reasons, we have focused on the integration of Microarray data in order to offer valuable informative genetic data. The accuracy and reproducibility of Microarray datasets represent a useful issue to increase the reliability of the data produced by different experiments. To address this challenge, this paper presents the process followed in integrating two datasets that were selected from the GEO database. The two datasets represent the results of Microarray experiments about Breast cancer. Using the gene expression of publicly available Microarray data, we have followed a process for integrating independently the generated Microarray data with the same experimental objectives to obtain informative data. We have discovered more reliable information by increasing the size of samples through combining two independent Microarray datasets. Moreover, the integration approach has made the classification more accurate by providing different results while implementing the PCA technique. In order to visualize the pertinence of integrating gene expression data, the classification of the individual and the combined datasets has produced comparable results. Basing on the obtained results, we look forward to generalize the integration of gene expression data specifically on the crucial fields like cancer, in the object of giving the researchers accurate information.

References

Afshari CA. (2002). Perspective: microarray technology, seeing more than spots. Endocrinology; 143(6): 1983-1989.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: Archive for functional genomics data sets—Update. Nucleic Acids Research, 41(Database issue), D991–D995.

C. Kumar Sarmah and S. Samarasinghe (2010). Microarray data integration: frameworks and a list of underlying issues, *Current Bioinformatics*, vol. 5, no. 4, 280–289.

C. Naidu and Y. Suneetha (2012). Review Article: Current Knowledge on Microarray Technology - An Overview, Tropical Journal of Pharmaceutical Research, vol. 11, no. 1.

C.-R. Chen, W.-Y. Shu, M.-L. Tsai, W.-C. Cheng, and I. C. Hsu (2012). THEME: A web tool for loop-design microarray data analysis, Computers in Biology and Medicine, vol. 42, no. 2, 228–234

F. Hong, R. Breitling (2008). A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. Bioinformatics 24: 374-382.

Frazee, A. C., Langmead, B., & Leek, J. T. (2011). ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets. BMC Bioinformatics, 12, 449.

Kapushesky, M., Adamusiak, T., Burdett, T., Culhane, A., Farne, A., Filippov, A., et al. (2012). Gene Expression Atlas update— A value-added database of microarray and sequencing-based functional genomics experiments. Nucleic Acids Research, 40 (Database issue), D1077–D1081.

Kirk, P., Griffin, J.E., Savage, R.S., Ghahramani, Z., Wild, D.L., 2012. Bayesian correlated clustering to integrate multiple datasets. Bioinformatics 28, 3290–3297.

Kodama, Y., Shumway, M., Leinonen, R., & International Nucleotide Sequence Database Collaboration (2012). The Sequence Read Archive: Explosive growth of sequencing data. Nucleic Acids Research, 40(Database issue), D54–D56.

Konstantinopoulos, P.A., Cannistra, S.A., Fountzilas, H., Culhane, A., Pillay, K., Rueda, B., Cramer, D., Seiden, M., Birrer, M., Coukos, G., Zhang, L., Quackenbush, J., Spentzos, D., 2011. Integrated Analysis of Multiple Microarray Datasets Identifies a Reproducible Survival Predictor in Ovarian Cancer. PLoS ONE 6, e18202.

Krupp, M., Marquardt, J. U., Sahin, U., Galle, P. R., Castle, J., & Teufel, A. (2012). RNA-Seq Atlas— A reference database for gene expression profiling in normal tissue by next-generation sequencing. Bioinformatics, 28(8), 1184–1185.

M. Mooney and S. McWeeney (2014). Data Integration and Reproducibility for High-Throughput Transcriptomics, in International Review of Neurobiology, vol. 116, Elsevier, 55–71.

M. Friedman, A. Y. Levy, and T. D. Millstein (1999). Navigational Plans For Data Integration. In AAAI/IAAI, 67–73.

M. R. Genesereth, A. M. Keller, and O. M. Duschka (1997). Infomaster: An Information Integration System, in SIGMOD Conference, 539–542.

M. Lenzerini (2002). Data Integration: A Theoretical Perspective, in PODS, 233–246.

M. J. Zaki (2014). Data mining and analysis: fundamental concepts and algorithms. New York, NY: Cambridge University Press.

O. Larsson, K. Wennmalm, R. Sandberg (2006). Comparative microarray analysis. Omics 10: 381-397.

Parkinson, H., Sarkans, U., Kolesnikov, N., Abeygunawardena, N., Burdett, T., Dylag, M., et al. (2011). ArrayExpress update— An archive of microarray and high-throughput se-

quencing-based functional genomics experiments. Nucleic Acids Research, 39(Database issue), D1002–D1004.

P. Cahan, F. Rovegno, D. Mooney, J.C. Newman, G. St. Laurent III, and T.A. McCaffrey (2007). Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization, GENE 401: 12-18.

Rung, J., & Brazma, A. (2013). Reuse of public genome-wide gene expression data. Nature Reviews Genetics, 14, 89–99.

Sherlock G. (2001). Analysis of large-scale gene expression data. Briefings in bioinformatics, 350-362.

Sîrbu, A., Crane, M., Ruskin, H., 2015. Data Integration for Microarrays: Enhanced Inference for Gene Regulatory Networks. Microarrays 4, 255–269.

Tanya Barrett and Ron Edgar (2006), Gene Expression Omnibus: Microarray Data Storage, Submission, Retrieval, and Analysis.

T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muertter, M. Holko, O. Ayanbule, A. Yefanov, and A. Soboleva (2011). NCBI GEO: archive for functional genomics data sets--10 years on, *Nucleic Acids Research*, vol. 39, no. Database, D1005–D1010.

Ziegler, Patrick (2004). User-Specific Semantic Integration of Heterogeneous Data: What Remains to be done? Building the Information Society, 156:3-12.

Résumé

Les données Microarray sont organisées sous forme de larges matrices contenant des expressions de gènes, chaque ligne représente un gène et chaque colonne représentant une condition d'une spécifique expérimentation. Le fait que ces données sont accumulées dans plusieurs différentes sources de données représente un obstacle pour les chercheurs et surtout les biologistes. En effet, il n'est pas simple d'obtenir la bonne information avec une simple manœuvre. L'intégration de données Microarray est de grande importance pour quant il s'agit d'accroître la fiabilité et la généralisation des résultats. En explorant les bases de données Microarray qui sont les plus utilisées, nous avons trouvé des variations entre les différentes expérimentations Microarray, ce qui rend l'intégration de données un grand défi. Cet article présente un schéma d'intégration qui vise à obtenir de précieux résultats et prendre de pertinentes décisions pour répondre aux hypothèses et aux cruciales questions des chercheurs.

Indexing-based link discovery in Linked Data

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

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

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

1 Introduction

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

Currently, many researches help to discover typed links between URIs that represent the same real word object in different data sources (Glaser et al. (2009),Raimond et al. (2008),Cervantes (2013),Scharffe et al. (2009),Volz et al. (2009),Isele et al. (2011),Ngomo and Auer (2011),Ngomo (2011),Axel-Cyrille and Ngomo (2012),Nikolov et al. (2012), Dreéler and Ngomo. (2014)). These frameworks have the problem of runtime complexity in the matching task that can be measured by the number of comparisons necessary to complete this task. In this paper,

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

we present an approach that allows optimizing the runtime of link discovery in LOD Cloud; we improve the first algorithm of LIMES (Link Discovery Framework for metric spaces) by (1) WordNet based indexing task to index each source instance in the target knowledge base by using structured inverted indices to rapidly obtain possible candidate results from entities that have been indexed in the LOD Cloud, and (2) removing each exemplar from the target dataset to delete duplicate exemplars and therefore reduce the number of comparisons in the first algorithm of LIMES.

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

2 Background

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

2.1 Linked Data

Tim Berners-Lee Berners-Lee (2006) proposed the concept of Linked Data to connect data in the web. There are four rules to publish Linked Data on the web Heath and Bizer (2011):

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

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

2.2 Indexing task

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

Linked Open Data project makes it possible to index and explore many structured data on the web to navigate through its large datasets. For example, Semantic Search workshop evaluated indexing task by extracting a set of queries from a commercial search engine in

2010 [2] and 2011 [3]. In Blanco et al. (2011), the authors takes advantage of indexing and ranking methods over rdf datasets used at the Billion Triple Challenge [4]. In Tonon et al. (2012), the authors implement a mixed architecture that combines unstructured inverted indices with a structured graph database to improve instance matching task. FLBSM Gupta et al. (2014) is an approach that used Fuzzy Logic to develop a new similarity measure based on tf/idf measure.

2.3 Matching

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

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

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

3 Related work

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

3.1 Domain-particular

RKB Knowledge base (RKBCRS) Glaser et al. (2009) allows discovering links between entities from the domain of academics. RKBExplorer extract RDF data from heterogeneous data sources, and it populates its knowledge bases with instances from the AKT ontology [5]. GNAT Raimond et al. (2008) is a tool that was developed for the music domain. It implements

2. http://km.aifb.kit.edu/ws/semsearch10/ 01.05.2015
3. https://km.aifb.kit.edu/ws/semsearch11/ 01.05.2015
4. http://challenge.semanticweb.org/ 01.05.2015
5. http://www.aktors.org/publications/ontology/ 01.10.2014

the online graph matching algorithm (OGMA) to discover equivalent resources. In Cervantes (2013),the authors present FLORA project to generate a financial dataset. It uses LIMES and SILK to discover different pertinent links in the LOD Cloud.

3.2 General

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

3.3 Discussion

The following table shows the difference between state-of-the-art approaches, they have

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

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

the problem of time complexity that is high; due to the large number of comparisons between the source and target knowledge base. In most approaches, each source instance must been compared $|T|$ times, and in the LIMES framework, it must be compared $\sqrt{|T|}$ times even if it has not a candidates matches. In order to reduce the number of comparisons of each source instance, we present an approach that allows improving LIMES framework by indexing task.

4 Proposed framework

In order to discover owl: sameAs links between URIs in an efficient time, we present an approach that improve LIMES framework by indexing task. Our system takes as input two datasets from LOD Cloud S (source) and T (target), and it generates owl: sameAs as output (see figure 1).

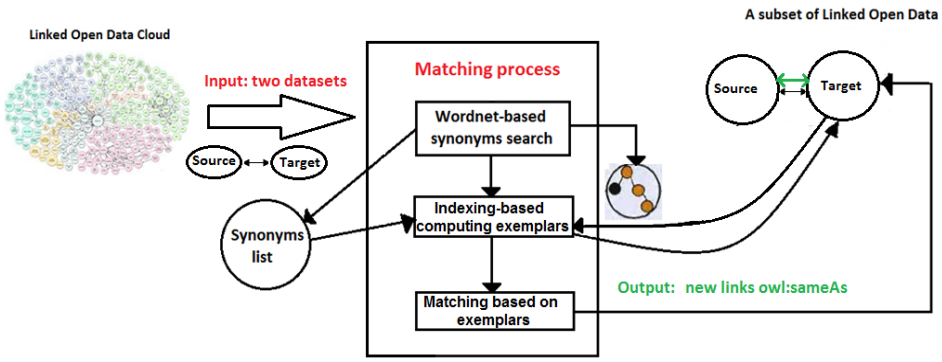


FIG. 1 – *Our system architecture*

4.1 LIMES algorithms

In order to reduce the number of comparisons, LIMES calculated the approximates of the similarity between instances by using the mathematical characteristics of metric spaces (the triangle inequality: $m(x;y) - m(y;z) > \theta \Rightarrow m(x;z) > \theta$). Then, it used these estimates to sort out the instances pairs that do not enough the mapping conditions. LIMES present two core algorithms:

Algorithm 1 of LIMES: Computing exemplars

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

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

Algorithm 2 of LIMES: The matching based on exemplars

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

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

4.2 Example

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

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

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

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

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

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

4.3 WordNet-based synonyms search

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

4.4 Indexing based computing exemplars

In order to reduce the time complexity of the matching process, we improve the first algorithm of LIMES by: (1) **WordNet based indexing:** receives as input a SPARQL endpoints or LOD dumps of the source and the target datasets and retrieves all corresponding triples as output. We use structured inverted indices Demartini et al. (2013) by using Lucene [8] search engine to obtain a ranked list of candidate matches from the target data set (see algorithm 3: step 1 and 2). Since the indexes do not contain synonyms, we use the WordNet to improve the result of indexing. (2) **Removing each exemplar from the target dataset:** when the first algorithm of LIMES adds an exemplar to the list of exemplar, it does not remove this exemplar from the target dataset and therefore, there may be a duplication of exemplars, this can increase the number of comparisons. In order to resolve this problem, we add an instruction that remove each exemplar from the target dataset to reduce the number of comparisons (the complexity become $O(|E|(|T'| - |E|))$ where $T' \subseteq T$) (see algorithm 3: step 5 and 9).

Algorithm 3 takes as input a point $s \in S$ and a target dataset T. First, we search all synonyms

Algorithm 3 Indexing based computing exemplars

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

of each instance $s \in S$ and we put them in a list $v = v \cup \{s\}$. Then we index the elements of

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

v in the target dataset to retrieve all candidate matches from the target dataset and we put them in a list T' (step 2). In steps 3, 4, 5, we initialize E by selecting a random point e_1 in the metric space (T', m) ($E = \{e_1\}$), then we remove e_1 from T'. Then, we compute the similarity from the exemplar e_1 to every other point $t \in T'$(step 4). As the size of E is less than n, we pick a point $e' \in T$ such that the sum of the distances from e' to the exemplars $e \in E$ is maximal. Then, we add this point to E and we remove it from T' (step 6 and 7). In step 8, we compute the distance from e' to all other points in T. Finally (step 9), we map each point in T' to the exemplar to which it is most similar. The complexity of this algorithm is $O(|E|(|T'| - |E|))$.

4.5 Matching based on exemplars

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

5 Experimentation

In this section, we experimentally evaluate the performance of our system on real datasets (DBpedia [9], LinkedCT [10], Drugbank [11], Bio2RDF [12]) by comparing the number of comparisons and the runtime in our approach, LIMES version 0.6 and SILK version 2.6. We run all experiments on a 64 bits system with a 2.5GHz Intel Core i5 machine with 8 GB RAM. LIMES

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

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

framework used all instances of the target knowledge base to calculate the number of exemplars, for example, if the size of our knowledge base is 1000 then the number of exemplars is 32 and therefore each source instance will be compared 32 times. While our framework uses a portion of the target knowledge base, which was calculated by indexing task to reduces the search space of candidate matches. As the result indexing is ranked, we take the top 20 instances and therefore each source instance will be compared 4 times (see table 3). We calculate the total runtime of our approach is calculated like that:

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

1. The time to retrieve the instances from the source knowledge base;
2. The time to index the sources instances in the target knowledge base by using WordNet;
3. The time to compare the instances of the source with their of the target knowledge bases;
4. the time to put up the results.

Table 4 shows the different sizes of the sources and targets knowledge bases and the runtime of LIMES, SILK and our approach.
In order to evaluate the effective of our approach for instance matching, we select a subset

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

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

from DBpedia, drugbunk, LinkedCT, and we validate the result of the matching process by an expert. In order to compute the precision and the recall, we compare the set of matching and non matching data for each source instance provided by our system and by the expert. The precision is defined as: $P = \frac{tp}{(tp+fp)}$, and the recall is defined as: $R = \frac{tp}{(tp+fn)}$. Where:
True positive (tp): In the case where the expert and our approach select the matches.
False positive (fp): In the case where our system selects a match while the expert does not.
False negative (fn): In the case where the expert selects a match while our approach does not.

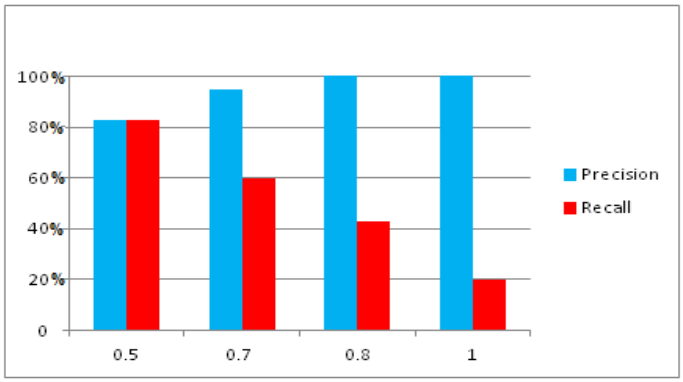


FIG. 2 – *Precision and recall as compared to the threshold $\theta \in [0, 1]$. The x-axis shows the threshold θ, the y-axis the values of precision and recall.*

In our evaluation, we will focus on the precision since it is the most useful metric in practice. Figure 2 shows how precision and recall could vary according to the threshold th. So, the precision is maximum if the threshold nearing to the value 1.0, while the recall is low (because

while we select a high threshold, the number of the matches retrieved by the system is low). This means that our system returns substantially more results in the matching process due to the use of wordnet and string similarity measure.

6 Conclusion and future work

In this paper, we have presented an approach to discover owl: sameAs links between data sources in the LOD Cloud in very efficient time. Our approach used the WordNet to improve indexing task and the algorithms of LIMES. We evaluated our approach with real data and we showed that it outperforms state-of-the-art approaches with respect to the number of comparisons. In our future works, we use this framework for facilitate integrating ontologies in the LOD Cloud.

References

Axel-Cyrille and N. Ngomo (2012). Link discovery with guaranteed reduction ratio in affine spaces with minkowski measures. In *International SemanticWeb Conference*, pp. 378–393.

Berners-Lee, T. (2006). Linked data - design issues. `http://www.w3.org/DesignIssues/LinkedData.html`.

Blanco, R., P. Mika, and S. Vigna (2011). Effective and efficient entity search in rdf data. In *Proceedings of the 10th International Conference on The Semantic Web - Volume Part I*, ISWC'11, Berlin, Heidelberg, pp. 83–97. Springer-Verlag.

Cervantes, J. L. S. (2013). Discovering and linking financial data on the web.

Demartini, G., D. E. Difallah, and P. Cudré-Mauroux (2013). Large-scale linked data integration using probabilistic reasoning and crowdsourcing. *The VLDB Journal 22*(5), 665–687.

Doan, A., A. Y. Halevy, and Z. G. Ives (2012). *Principles of Data Integration.*

Dreéler, K. and A.-C. N. Ngomo. (2014). On the efficient execution of bounded jaro-winkler distances.

Euzenat, J. and P. Shvaiko (2013). *Ontology Matching, Second Edition*. Springer.

Glaser, H., I. C. Millard, W.-K. Sung, S. Lee, P. Kim, and B.-J. You (2009). Research on linked data and co-reference resolution. Technical report, University of Southampton.

Gupta, Y., A. Saini, A. K. Saxena, and A. Sharan (2014). Fuzzy logic based similarity measure for information retrieval system performance improvement. In *Distributed Computing and Internet Technology - 10th International Conference, ICDCIT 2014.*, pp. 224–232.

Heath, T. and C. Bizer (2011). *Linked Data: Evolving the Web into a Global Data Space*.

Isele, R., A. Jentzsch, and C. Bizer (2011). Efficient multidimensional blocking for link discovery without losing recall.

Ngomo, A.-C. N. (2011). A time-efficient hybrid approach to link discovery.

Ngomo, A.-C. N. and S. Auer (2011). Limes: A time-efficient approach for large-scale link discovery on the web of data. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, IJCAI'11, pp. 2312–2317.

Nikolov, A., M. d'Aquin, and E. Motta (2012). Unsupervised learning of link discovery configuration. In *Proceedings of the 9th International Conference on The Semantic Web: Research and Applications*, ESWC'12, Berlin, Heidelberg, pp. 119–133. Springer-Verlag.

Pedersen, T., S. Patwardhan, and J. Michelizzi (2004). Wordnet::similarity: Measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, HLT-NAACL–Demonstrations '04, Stroudsburg, PA, USA, pp. 38–41.

Pound, J., P. Mika, and H. Zaragoza (2010). Ad-hoc object retrieval in the web of data. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, New York, NY, USA, pp. 771–780. ACM.

Raimond, Y., C. Sutton, and M. Sandler (2008). Automatic interlinking of music datasets on the semantic web. In *Proceedings of the 1st Workshop about Linked Data on the Web*. en

Scharffe, F., Y. Liu, and C. Zhou (2009). Rdf-ai: an architecture for rdf datasets matching, fusion and interlink. In *Proc. IJCAI 2009 workshop on Identity, reference, and knowledge representation (IR-KR), Pasadena (CA US)*.

Tonon, A., G. Demartini, and P. Cudré-Mauroux (2012). Combining inverted indices and structured search for ad-hoc object retrieval. In *Proceedings of the 35th International ACM SIGIR Conference*, New York, NY, USA, pp. 125–134.

Volz, J., C. Bizer, M. Gaedke, and G. Kobilarov (2009). Discovering and maintaining links on the web of data. In *The Semantic Web - ISWC 2009*, Volume 5823, pp. 650–665.

Zezula, P., G. Amato, V. Dohnal, and M. Batko (2006). *Similarity Search: The Metric Space Approach*, Volume 32 of *Advances in Database Systems*. Springer.

Résumé

Linked Data est un paradigme de publication pour rendre les données et pas simplement des documents accessibles et inter-cliquables sur l'internet. Cela permet la création d'un espace de données mondiale fondée sur des normes ouvertes- le web des données. Dans ce contexte, les différents types de liaisons sémantiques peuvent être établies entre les données. Un certain nombre de sources de données liées mettre des liens owl: sameAs pointant vers d'autres sources et d'autres ne le font pas. Afin de faciliter la mise en place de ces liens, les cadres de découverte des liens ont été élaborés. Mais ils ont le problème du temps d'exécution qui est très élevé en raison du grand nombre d'instances dans la source de donnée source et cible. Dans cet article, nous présentons une approche pour trouver des liens typés entre les ensembles de données dans Linked Open Data Cloud dans un temps très efficace; on améliore le premier algorithme de LIMES avec la tâche d'indexation pour réduire le nombre de comparaisons et par conséquent réduire le temps d'exécution. Nous évaluons notre travail sur des ensembles de données réelles et nous montrons que notre approche a un plus petit nombre de comparaisons dans le processus de matching. En outre, nous comparons le temps d'exécution de notre approche avec celle de LIMES et SILK et nous montrons que notre approche est la plus rapide.
Keywords: Linked Data, la découverte des liens, l'indexation, le matching.

Identification des communautés dans des réseaux complexes basée sur des nœuds influents

Sara AHAJJAM*, Hassan BADIR*, Azedine BOULMAKOUL**, Mohamed EL HADDAD*
*Laboratoire des Technologies d'Informations et de Communcation- ENSAT
ahajjamsara@gmail.com
hbadir@gmail.com
elhaddad.mohamed@gmail.com
**Université Hassan II - Casablanca
azedine.boulmakoul@gmail.com

Résumé. Les systèmes complexes sont omniprésents, ils apparaissent dans une grande variété de scénarios, allant de système social à des systèmes biologiques et technologiques. Les réseaux complexes sont un outil puissant pour comprendre les différents mécanismes des systèmes complexes. Le clustering où la détection des communautés est l'une des caractéristiques les plus pertinentes des graphes représentant les systèmes complexes. Une communauté est un ensemble de nœuds fortement connectés entre eux, et relativement moins connectés avec les autres nœuds du réseau. L'inconvénient majeur de la plupart des approches proposées est qu'ils exigent une connaissance préalable du nombre des communautés à détecter. Dans ce papier, nous proposons une nouvelle approche pour la détection de communautés dans des réseaux complexes basée sur les nœuds influents où les nœuds leaders qui sont susceptibles de diffuser l'information et de propager l'influence, sans une connaissance a priori de k-nœuds influents.

1 Introduction

Durant ces dernières décennies, de grands progrès ont été accomplis dans le développement de la science des réseaux surtout après la proposition du modèle « Petit-monde » (Watts et Strogatz, 1998) et du modèle « Scale-Free » (Barabasi et Albert, 1999). Les réseaux complexes est une nouvelle approche qui permet de décrire et de modéliser les différents systèmes réels complexes. En effet, tous les systèmes complexes peuvent être modélisés par des graphes, où les nœuds représentent les acteurs du phénomène et les liens (arcs) représentent les interactions entre ces acteurs. Le réseau complexe est devenu le modèle fondamental qui permet de comprendre les relations topologiques complexes et le dynamisme du comportement dans divers domaines, tels que l'Internet, la diffusion de l'influence dans les réseaux sociaux, etc. La propriété clé d'un véritable réseau est sa structure de communauté. Une communauté est un ensemble de nœuds qui sont fortement connectés entre eux qu'avec le reste du réseau. D'après des études récentes, la manière dont les nœuds sont organisés joue un rôle fondamental dans les processus de propagation de l'information (de Arruda et al., 2014). Dans la propagation de l'épidémie, nous aimerions trouver les nœuds importants pour comprendre les processus dynamiques, qui pourraient aboutir à une méthode efficace pour immuniser les réseaux modulaires (Wang et al., 2011).

Ces stratégies bénéficieraient d'une caractérisation quantitative de l'importance du nœud à la structure de la communauté.

Étudier les modèles de l'influence peut nous aider à mieux comprendre pourquoi certaines tendances ou innovations sont adoptées plus rapidement que d'autres et comment nous pouvons aider les annonceurs et les marqueteurs à concevoir des campagnes plus efficaces. Les sociétés commerciales et les banques pourraient être intéressées à trouver des agents actifs ou influents dans leur environnement et de leur offrir de nouveaux produits (Gliwa et al., 2015). Ce constat a lancé de nombreux chercheurs à développer une méthode efficace pour trouver les membres les plus influents à travers les réseaux sociaux. Récemment, les réseaux sociaux, autant qu'une catégorie spécifique des réseaux complexes, ont commencé à être étudiés d'une manière similaire, où les nœuds désignent les utilisateurs du réseau et les liens désignent les relations entre ces utilisateurs. Ils jouent un rôle fondamental dans la diffusion de l'information, des idées et de l'innovation. Identifier d'une manière efficace et efficiente les membres influents de ces réseaux reste un grand défi de nos jours. Dans les réseaux des médias sociaux à grande échelle, tels que twitter, le résultat du classement de nœuds basé sur des méthodes de centralité est utilisé pour déterminer les personnes les plus populaires ou importantes dans un domaine. De cette manière, les utilisateurs, en particulier les nouveaux, peuvent trouver des sources d'information rapidement (Xia et al., 2014). Nous allons concentrer notre attention sur une caractéristique topologique de type centralité. Un nombre important des mesures de centralité a été proposé afin de remédier ce problème, tels que la centralité de degré, la centralité de proximité, la centralité de voisinage, et la centralité de vecteur propre (Chikhi, 2010).

Ce papier traite la problématique de la détection des communautés dans les réseaux complexes. L'inconvénient majeur de la plupart des approches proposées est qu'elles exigent une connaissance préalable du nombre des communautés à détecter. Nous proposons un nouvel algorithme pour identifier les nœuds leaders pour détecter les communautés dans le réseau sans une connaissance préalable de k-nœuds leaders d'un réseau. Le reste de cet article est organisé comme suit : la section 2 introduit quelques préliminaires de ce travail. La section 3 présente un état de l'art sur la détection des nœuds influents et la détection des communautés dans les réseaux complexes. Dans la section 4, nous présentons nos approches pour la détection des communautés en se basant sur les nœuds influents du réseau. Les expérimentations et les résultats des méthodes proposées sont illustrées dans la section 5. La conclusion est proposée dans la section 6.

2 Préliminaires

La science de réseau est un domaine académique interdisciplinaire qui étudie les réseaux complexes (Xia et al., 2014), et se fonde sur les théories et les méthodes, y compris la théorie des graphes des mathématiques, la mécanique statistique relevant de la physique, l'exploration de données et de visualisation de l'information issus de l'informatique, la modélisation d'inférence à partir de statistiques (Newman, 2010). Elle fournit un modèle unifié pour étudier différents problèmes complexes. Pour l'analyse des réseaux sociaux, la diffusion de l'influence entre les différents nœuds est toujours étudier en utilisant les mesures de la centralité.

2.1 La centralité basée sur la localité

Les mesures de centralité locale tendent, en général, à capturer les caractéristiques du nœud par l'intermédiaire des informations partielles autour de lui. Le nombre de voisins joue un rôle essentiel dans ces méthodes, soit des voisins directs ou indirects, ou les deux.

2.1.1 La centralité de degré

Le principe de cette mesure est que le nœud qui possède plus de voisins c.à.d. plus de liens est le plus influent. Elle est définie comme le nombre de liens incidents d_i ou le degré d_i du sommet v_i. La centralité de degré $C_d(i, g)$ est donnée par :

$$C_d(i,g) = \frac{d_i(g)}{n-1} = \frac{|N_i(g)|}{n-1}$$

2.1.2 La k-shell décomposition

Cette méthode suggère que l'emplacement du nœud dans le réseau est un aspect important. Autrement dit, les nœuds dans la zone centrale du réseau ont tendance à diffuser de l'information à plus de gens que le nœud dans la zone périphérique avec le même degré. Le processus de la décomposition du réseau est le suivant, tout d'abord, supprimer tous les nœuds dont le degré est égal à 1 et tous les liens attachent eux. Et puis le degré de certains nœuds réservés peut devenir à nouveau 1. Jusqu'à présent, tous les nœuds ont été retirés seront alloués à la couche la plus extérieure, tandis que leur valeur de centralité est $k_s = 1$. (Fig.1).

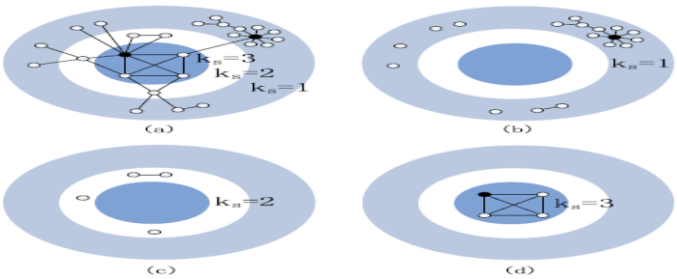


FIG. 1 - *les trois shell du réseau (Kitsak et al., 2010)*

2.2 La centralité basée sur la globalité

Ces mesures tiennent compte de l'information globale et donnent un meilleur classement des résultats, tels que la centralité d'intermédiarité et la centralité de proximité.

2.2.1 La centralité d'intermédiarité

La centralité d'intermédiarité peut être définie comme la fraction de plus courts chemins entre des paires de nœuds qui passent à travers le nœud d'intérêt. L'intermédiarité est, dans un certain sens, une mesure de l'influence d'un nœud sur l'information diffusée par le réseau ou la charge prévue d'un nœud dans un réseau de transport (Guimerà et al., 2002). La centralité d'intermédiarité est:

Identification des communautés basée sur les nœuds influents

$$C_b(i,g) = \frac{2}{(n-1)(n-2)} \sum_{k \neq j, i \notin \{j,k\}} \frac{P_i(kj)}{P(kj)}$$

Où : $\frac{P_i(kj)}{P(kj)}$ est la probabilité de tomber sur une géodésique sélectionné au hasard reliant k et j.

2.2.2 La centralité de proximité

La proximité de nœud v_i est définie comme l'inverse de la somme des distances géodésiques à tous les autres nœuds de v_i. Supposons que d_{ij} est la longueur d'un chemin géodésique de v_i à v_j. La centralité de proximité est donnée par:

$$C_c(i,g) = \frac{n-1}{\sum_{i \neq j} d_{ij}}$$

3 Etat de l'art

La détection des communautés et l'analyse de la nature des relations et des liens entre les entités sont une clé véritable vers la compréhension d'une variété de phénomènes. La détection des communautés a fait l'objet de plusieurs travaux de recherche. La plupart des études classent les articles et les travaux de recherche en fonction du type de l'algorithme. Les algorithmes de détection des communautés appartiennent à deux principaux types d'approches à savoir le partitionnement de graphe et la classification. L'inconvénient majeur de méthodes basées sur le partitionnement des graphes est qu'ils nécessitent une connaissance préalable du nombre et de la taille des groupes à déterminer (P. Pons, 2010). Cependant, il y a toujours des nœuds qui jouent un rôle plus important que d'autres dans ces groupes, d'où … des nœuds influents ou leaders. Dans les sciences sociales, l'influence peut être définie comme le pouvoir de négociation ou du contrôle de l'information (Tsai et al., 2014). Plusieurs méthodes ont été proposées pour l'étude de l'influence des utilisateurs d'un réseau et la découverte du leader de la communauté à partir d'un réseau social en ligne. Les approches de détection de Leader sont divisées en deux types : les méthodes globales et locales. La méthode globale porte sur toute la topologie du réseau, tandis que les méthodes locales portent sur la position locale du nœud. Khorasgani et al. suggèrent une nouvelle approche pour la détection des nœuds leaders. Elle prend en compte les nœuds qui n'appartiennent à aucun leader (outlier). Cet algorithme est inspiré de k-means, les k nœuds leaders seront sélectionnés aléatoirement. Les autres nœuds seront affectés aux plus proches leaders pour former des communautés, et pour chaque étape, on cherche à trouver de nouveaux leaders pour chaque communauté autour duquel se réunissent des suiveurs (followers) jusqu'à ce qu'aucun nœud ne se déplace. Pour chaque communauté, la centralité de degré de chaque membre est calculée et le nœud avec le plus haut degré est choisi comme le nouveau leader (Khorasgani et al., 2013).

D'autres études proposées utilisent la classification. La classification a été introduite pour analyser les données et les partitions en se basant sur une mesure de similarité entre les partitions. Le problème de la détection des communautés peut être considéré comme un problème de classification de données pour lequel nous devons sélectionner une distance appropriée (S. Fortunato, 2011). Le résultat obtenu par ces procédés dépend du choix de la mesure de similarité qui a été utilisé initialement. Zhang et al. Ont proposé un nouvel algorithme pour la mesure de similarité basé sur l'entropie relative afin de former des

partitions. Ils se basent sur le principe que les nœuds avec une structure commune ont une forte similarité avec d'autres, et lorsque la similarité entre ces nœuds est égale à 1, cela signifie que les deux nœuds ont la même structure dans des réseaux complexes. Les nœuds marginaux ont une faible similarité à d'autres (Zhang et al. 2015). Zhou et al proposons un algorithme glouton en fonction des préférences de l'utilisateur (GAUP) pour faire activer les k-haut utilisateurs influents, basés sur le modèle Extended indépendant Cascade (EIC a dit que un nœud actif v est actif en t-1, a une seule chance pour activer tous les voisins inactifs) (Zhou et al., 2014). Des recherches récentes ont trouvé que l'emplacement du nœud dans la topologie de réseau est un facteur important dans l'estimation de la capacité de la propagation de l'information. Une nouvelle approche pour identifier l'emplacement du nœud par la méthode de décomposition k-shell par laquelle le réseau est divisé en plusieurs couches a été proposée par (Xia et al., 2014) (Li et al. 2015) qui indique que l'intérieur de la couche est importante plus que le nœud. Chaque nœud correspond à une couche et l'ensemble du réseau forme la structure noyau-périphérie. Une nouvelle approche pour détecter les communautés et les nœuds importants des communautés en utilisant le spectre du graphe définit les nœuds importants à la communauté comme les changements relatifs dans les plus grandes valeurs propres c de la matrice d'adjacence du réseau après enlèvement de ces nœuds. L'inconvénient majeur de cette approche, c'est que, pour avoir un meilleur résultat, on a besoin de connaître le nombre de partitions dans le réseau et aussi on ne peut pas identifier les nœuds importants dans les petites communautés (Wang et al., 2011). Les approches de détection des nœuds leaders et des communautés sont diverses. Chaque algorithme proposé apporte une nouvelle idée ou une amélioration des algorithmes existants.

4 Notre contribution

Il y a un certain nombre d'idées et de théories sur la façon dont les tendances et les innovations se répandent et se sont adoptées. La vue traditionnelle suppose qu'une minorité de membres dans une société possède des qualités qui les rendent exceptionnellement convaincants dans la diffusion des idées à d'autres membres. Ces individus exceptionnels conduisent les tendances au nom de la majorité des gens ordinaires. Ils sont vaguement décrits comme étant informés, respectés, et bien reliés ; ils sont appelés les leaders, les innovateurs dans la diffusion de la théorie des innovations, les influents et les connecteurs (Wang et al., 2014). La théorie des dirigeants est intuitive et convaincante. En identifiant et en convaincant un petit nombre d'individus influents / Leader, une campagne virale peut atteindre un large public à un coût modique. La théorie propage bien au-delà du monde universitaire et a été adopté dans de nombreuses entreprises de marketing, par exemple, RoperASW et Tremor (Khorasgani et al., 2013). Identifier l'influence sociale dans les réseaux est essentiel pour comprendre comment les comportements se propagent. L'analyse des ensembles de données des réseaux sociaux révèle que, dans chaque communauté, il y a généralement un certain membre (ou leader) qui joue un rôle clé dans cette communauté (Ahajjam et al., 2015). Nous nous intéressons à détecter les nœuds / influents Leader qui sont responsables sur la diffusion de l'information, et de former des communautés autour de ces nœuds afin de faciliter la propagation de l'influence. En fait, la centralité est un concept important (Kernighan and Lin, 1970) au sein de l'analyse des réseaux sociaux, qui mesure l'importance relative d'un sommet dans le graphe. Différent des autres méthodes, notre approche a pour but de détecter les membres influents du réseau, et de construire des

communautés autour de ces influents sans une connaissance a priori du nombre des membres influents du réseau et de taille des communautés.

Soit un graphe non orienté et non pondéré G=(V,E). V est l'ensemble des nœuds. Chaque nœud en V représente un élément de l'ensemble des données. Et chaque lien E représente une relation entre deux nœuds. Nous allons présenter deux algorithmes de détection des communautés en utilisant les nœuds leaders, qui se sont basés tous les deux sur les étapes suivantes : Classement des nœuds, détection des nœuds influents et la détection des communautés.

Avec l'émergence des réseaux sociaux, nous avons face à des données massives « big data » qui deviennent difficiles à traiter avec des outils classiques de gestion de base de données ou de gestion de l'information. Pour faire le traitement de ces données, nous nous orientons vers les bases de données NoSQL afin de gérer l'ensemble gigantesque des données issues de ces réseaux. Nous allons stocker notre ensemble de données dans une base de données de graphe « graph-database » NoSQL, à savoir Neo4j qui permet de représenter les données connectées naturellement, en tant qu'objets reliés par un ensemble de relations, chacun possédant ses propres propriétés. Les traitements se fassent sur RHadoop comme il est montré dans la figure (Fig2.).

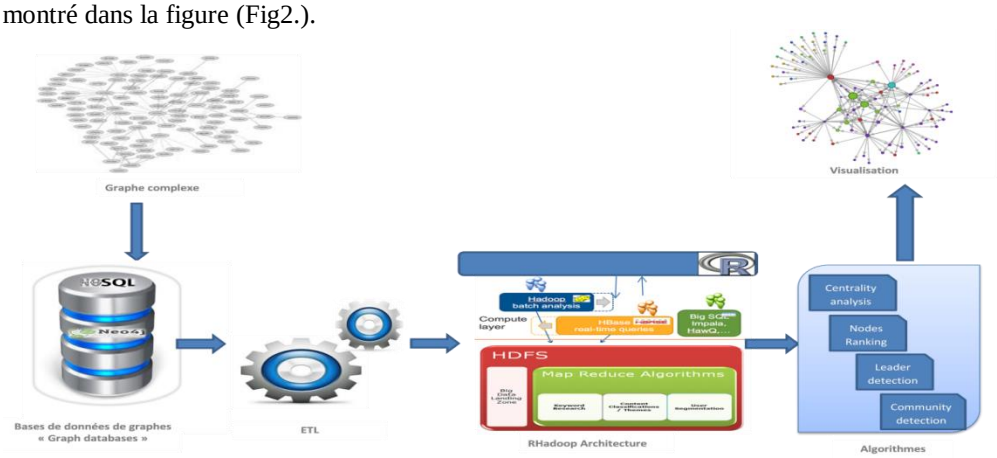


FIG. 2 - *Architecture de la solution proposée*

Algorithme 1:

Notre algorithme passe par trois importantes étapes (Fig3.) :

- **Etape A: La centralité des nœuds.** Pour chaque nœud v dans le réseau G, on va calculer la centralité de vecteurs propres. La centralité de vecteur propre ou l'index Gloud d'accessibilité (SPIZZIRRI ,2011) décrit comment un individu bien connecté repose sur des relations directes et indirectes (par exemple, il prend en compte les connexions dierctes et indirectes des individus) (Ruhnau, 2000). La centralité de vecteur propre d'un nœud est proportionnelle aux centralités de ces voisins (Newman,2004), les nœuds les plus influents seront plus connectés avec d'autres personnes influentes. Enfin, l'encastrement quantifie comment un individu est isolable ou comment il est impliqué dans la structure du réseau (Moody et al.,2003).

$$A\,x = \lambda\,x$$

Avec: A est la matrice d'adjacence du réseau et λ est la valeur propre.

- **Etape B: Classement des nœuds "Nodes ranking".** Les nœuds sont classés par ordre décroissant de la valeur de centralité. Le nœud V_1 avec la plus grande centralité est choisi comme nœud influent.
- **Etape C: Détection des communautés.** on calcule la fonction de voisinage afin de déterminer les voisins du nœud influent/leader sélectionné. On affecte les voisins au nœud leader afin de former des communautés.
- On supprime la communauté créée c.à.d. le nœud leader et les nœuds voisins du réseau, et on va répéter les étapes précédentes (Etape B, Etape C) à ce que tous les liens seront supprimés. Cependant, avec la suppression des nœuds et donc des liens à chaque étape, nous aurons des nœuds qui ne seront affectés à aucune communauté. Dans le but d'attribuer ces nœuds à des communautés, nous allons chercher les voisins de ces nœuds sur le graphe en étude, et par intuition, nous attribuons ce nœud au plus fort leader auquel appartient l'un de ces voisins.

Input: undirected, unweighted graph G=(V,E)
Output: Set C=(C_1,C_2,…,C_n)
1: $i = 0$
2: While $Q \neq 0$
3: Calculate the centrality score of each vertex $V \in G$,
4: Loop
5: **Nodes ranking:** Order V via their centrality scores, such that $Q = (V_1, V_2, \ldots, V_n)$ with Cent (V_1) ≥ Cent (V_2) ≥ · ·· ≥Cent (V_n).
6: $i = i + 1$
7: **Select** where V_{i1} is the first vertex in the vertex list Q.
8: Create a new group $Ci = \{V_{i1}\}$,
9: New $Q = Q - \{V_{i1}\}$
10: Q = New Q
11: **Community detection:** Calculate the neighborhood function of V_{i1} to find the **candidate neighbors set "neighbors N(V_j)"**.
12: **insert into** N(V_j).
13: New Q=Q-N(V_j)
14: End loop

FIG 3. – *Pseudo-code de l'algorithme1*

Algorithme 2:

Cet algorithme passe par les mêmes étapes : Calcul de la centralité des nœuds, classement des nœuds par ordre décroissant et détection des communautés. Cependant, pour cette étape à chaque fois qu'on détecte une communauté, on va la supprimer du réseau, autrement dit supprimer le nœud leader détecté et ses nœuds voisins du graphe et recalculer la centralité des nœuds pour le nouveau graphe (qui est le graphe initial moins les nœuds de la communauté détectée) (Fig. 4). On répète ces étapes à tous les nœuds du réseau.

Identification des communautés basée sur les nœuds influents

Input: undirected, unweighted graph G=(V,E)
Output: Set C=(C_1,C_2,…,C_n)
1: $i = 0$
2: Loop
3: While $Q \neq 0$
4: Calculate the centrality score of each vertex $V \in G$,
5: **Nodes ranking:** Order V via their centrality scores, such that $Q = (V_1, V_2, \ldots, V_n)$ with Cent (V_1) ≥ Cent (V_2) ≥ · ·· ≥Cent (V_n).
6: $i = i + 1$
7: **Select** where V_{i1} is the first vertex in the vertex list Q.
8: Create a new group $Ci = \{V_{i1}\}$,
9: New $Q = Q - \{V_{i1}\}$
10: Q = New Q
11: **Community detection:** Calculate the neighborhood function of V_{i1} to find the **candidate neighbors set "neighbors N(V_j)"**.
12: **insert into** N(V_j).
13: New Q=Q-N(V_j)
14: End loop

FIG. 4 – *Pseudo-code de l'algorithme 2*

5 Résultats et expérimentations

5.1 Datasets

Nous avons élaboré deux ensembles de données pour évaluer nos algorithmes qui servent à la détection des communautés en utilisant les nœuds influents : Le réseau social « Zachary karaté club » qui est un réseau de référence bien connu pour les algorithmes de détection de la communauté de test. Le réseau est constitué de 34 nœuds et 78 bords. En raison d'un litige entre l'administrateur et l'instructeur du club, le club est finalement scindé en deux factions centrées sur l'administrateur et l'instructeur, respectivement. Et le deuxième, c'est un réseau linguistique « adjectives and noun adjacencies », le réseau d'adjectif et de nom communs pour le roman «David Copperfield» de Charles Dickens, comme décrit par M. Newman. Les nœuds représentent les adjectifs et les noms les plus fréquents dans le livre. (Tableau 1.). En raison de la difficulté de visualisation d'un grand réseau, les réseaux adoptés sont principalement de petite taille.

Datasets	***Nœuds***	***Liens***	***Communautés réelles***
Zachary Karaté Club	*34*	*78*	*2*
Word adjacencies	*112*	*425*	*2*

TAB. 1 - *Propriétés des datasets*

5.2 Résultats obtenus

Les résultats présentés sur la Fig. 5et la Fig. 6 montrent la structure des communautés dans le réseau de "Zachary karaté club", "Adjectives and noun adjacencies" respectivement. Nous avons comparé nos algorithmes de détection des communautés basés sur les nœuds influents avec d'autres algorithmes de détection de la communauté, à savoir LPA (Label

Propagation Algorithm» (Raghavan et al., 2007) et LEA (Leading Eigenvector Algorithm) (Newman, 2006) en utilisant des paramètres différents.

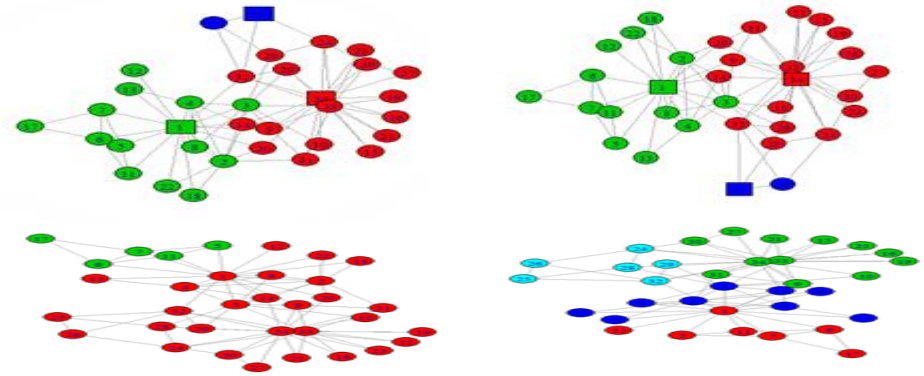


FIG. 5 - Structure *des communautés des algorithmes 1, 2, LPA et LEA respectivement. Les nœuds en carré sont les nœuds influents*

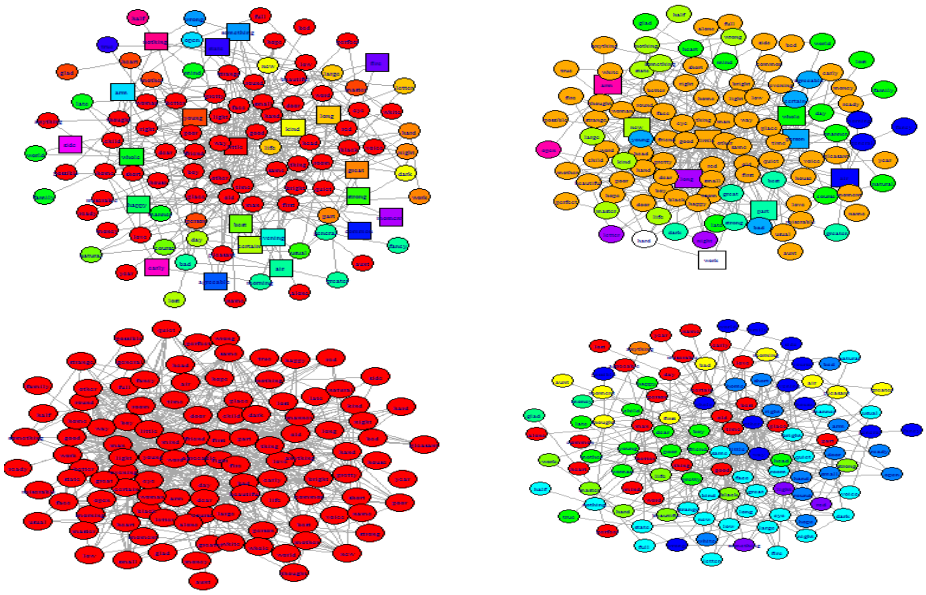


FIG. 6 - *Structure des communautés des algorithmes 1, 2 , LPA et LEA respectivement. Les nœuds en carré sont les nœuds influents*

Pour évaluer le résultat obtenu de nos algorithmes, nous sommes basés sur trois paramètres d'évaluations : La modularité Q, NMI et ARI.

- La modularité Q évalue la qualité des communautés obtenues.

$$Q = \sum_{i=1}^{k} (e_{ii} - a_i^2)$$

Où : e_{ii} est la proportion des liens à l'intérieur des communautés, et le second terme représente la valeur attendue de la même quantité dans un réseau aléatoire construit en gardant le même ensemble des nœuds et la même distribution des degrés, mais reliant les liens entre les nœuds au hasard.

- ARI (Adjusted Rand Index) : la mesure pénalise les faux négatifs et les faux positifs. Soient a, b, c et d désignent le nombre de paires de nœuds qui sont respectivement dans la même communauté à la fois G et R, dans la même communauté à G, mais dans différentes communautés de R, dans les différentes communautés à G, mais dans la même communauté R, et dans différentes communautés en G et R. Alors l'ARI est calculé par la formule suivante :

$$ARI = \frac{\binom{n}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{n}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)]}$$

- NMI (Normalized Mutual Information) est calculée comme suit :

$$NMI(X,Y) \coloneqq \frac{2I(X,Y)}{H(X)+H(Y)}$$

Où I (X, Y) l'information mutuelle correspond à la quantité d'information partagée par les variables.

Les résultats d'évaluation de nos algorithmes par rapport aux algorithmes cités : LPA et LEA sur la base des trois critères (modularité, NMI et IRA) sont présentés dans le tableau 2. Pour le premier réseau "Zachary Karaté Club" notre algorithme 1 fournit le meilleur résultat pour ARI et NMI comparant à LPA et LEA, tandis que pour la modularité qui présente la qualité des communautés, notre résultat est assez bon par rapport à LEA qui fournit le plus élevé score. Pour le deuxième réseau, notre algorithme 1 présente le meilleur résultat pour les trois critères d'évaluations.

Réseau	**Algorithme**	**Communautés**	**Modularité**	**NMI**	**ARI**
Zachary Karaté club	Label Propagation Algorithm	2	0.13280	0.0022325	-0.02747
	Leading Eigenvector Algorithm	4	**0.39340**	0.0067954	-0.03726
	Algorithm 1	3	0.31854	**0.2166125**	**0.255949**
	Algorithm2	3	0.30144	**0.2166125**	**0.255949**
Adjectives and nouns adjacencies	Label Propagation Algorithm	1	0	0	-1.10171
	Leading Eigenvector Algorithm	10	0.24260	0.0088321	-0.01261
	Algorithm 1	22	**0.58444**	**0.1095189**	**-0.00021**
	Algorithm 2	9	0.55254	0.0204514	-0.00653

TAB.2 – *Résultats des evaluations*

6 Conclusion

Cet article présente une étude des différents algorithmes de détection de communautés. Elle est particulièrement focalisée sur l'exploitation des nœuds leaders dans les réseaux complexes. L'intérêt manifesté par la recherche dans ce domaine est le fait que la diffusion de l'information, la distribution de l'influence dans les réseaux complexes sont des éléments stratégiques et particulièrement sensibles à leur utilisation. Ainsi, nous avons proposé de nouveaux algorithmes pour détecter les communautés en utilisant les nœuds leaders. Notre approche permet de classer les réseaux en groupes et de connaître facilement le nœud responsable de la diffusion de l'influence au sein de son groupe. Comme perspectives de ce travail, nous devons tester ces algorithmes sur des grands réseaux.

Références

De Arruda, G.F., Barbieri, A.L., Rodríguez, P.M., Rodrigues, F.A., Moreno, Y., and Costa, L. da F. (2014). *Role of centrality for the identification of influential spreaders in complex networks.* Phys. Rev. E *90*, 032812.

Wang, Y., Di, Z., and Fan, Y. (2011). *Identifying and Characterizing Nodes Important to Community Structure Using the Spectrum of the Graph*. PLoS ONE *6*, e27418.

Gliwa, B., Zygmunt, A., and Bober, P. (2015). *Analysis of Content of Posts and Comments in Evolving Social Groups.* In Advances in ICT for Business, Industry and Public Sector, M. Mach-Król, C. M. Olszak, and T. Pełech-Pilichowski, eds. (Cham: Springer International Publishing), pp. 35–55.

Xia, Y., Ren, X., Peng, Z., Zhang, J., and She, L. (2014). *Effectively identifying the influential spreaders in large-scale social networks*. Multimed Tools Appl 1–13.

Chikhi, N.F. (2010). *Calcul de centralité et identification de structures de communautés dans les graphes de documents.* phd. Université de Toulouse, Université Toulouse III - Paul Sabatier.

Newman, M. (2010). Networks: *An Introduction* (New York, NY, USA: Oxford University Press, Inc.).

Kitsak, M., Gallos, L.K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H.E., and Makse, H.A. (2010). *Identification of influential spreaders in complex networks*. Nat Phys *6*, 888–893.

Guimerà, R., Díaz-Guilera, A., Vega-Redondo, F., Cabrales, A., and Arenas, A. (2002). Optimal Network Topologies for Local Search with Congestion. Phys. Rev. Lett. *89*, 248701.

Pons, P. (2010). *Détection de communautés dans les grands graphes de terrain* (Paris 7).

Tsai, M.-F., Tzeng, C.-W., Lin, Z.-L., and Chen, A.L.P. (2014). *Discovering leaders from social network by action cascade.* Soc. Netw. Anal. Min. *4*, 1–10.

Khorasgani, R.R., Chen, J., and Zaïane, O.R. (2013). *Top leaders community detection approach in information networks*. In Proceedings of the 4th Workshop on Social Network Mining and Analysis, 2010. ISSN : 2319-7323, p. 228.

Ahajjam, S., Badir, H., El haddad, M., *Detection of leader's nodes in complex networks.* In Global Journal of Engineering Science and Researches (GJESR).ISSN: 2348 – 8034.

Fortunato, S. (2011). Community detection in graphs. Physics Reports 486, 75–174.

Zhang, Q., Li, M., Deng, Y., and Mahadevan, S. (2015). *Measure the similarity of nodes in the complex networks.* arXiv:1502.00780 [physics].

Zhou, J., Zhang, Y., and Cheng, J. (2014*). Preference-based mining of top- influential nodes in social networks.* Future Generation Computer Systems *31*, 40–47.

Li, R.-H., Qin, L., Yu, J.X., and Mao, R. (2015). *Influential Community Search in Large Networks.* Proc. VLDB Endow. *8*, 509–520.

Wang, P., Yu, X., Lu, J., and Chen, A. (2014*). Identification of important nodes in artificial bio-molecular networks.* In 2014 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1267–1270.

Kernighan, B.W., and Lin, S. (1970). *An Efficient Heuristic Procedure for Partitioning Graphs.* Bell System Technical Journal *49*, 291–307.

SPIZZIRRI, Leo. (2011). Justification and application of eigenvector centrality. Algebra in Geography: Eigenvectors of Network.

Ruhnau B. (2000). *"Eigenvector-centrality—a node centrality?"* Soc Network, 22:357–365.

Newman, M.E.J. (2006). *Finding community structure in networks using the eigenvectors of matrices.* Physical Review E *74*.

Moody J, White DR.. (2003). "*Structural cohesion and embeddedness: a hierarchicalconcept of social groups".* Am Sociol Rev. 68:103–127.

Summary

Complex systems are ubiquitous in nature; they appear in a wide variety of scenarios ranging from social and biological to technological systems. Complex networks are a potent tool for understanding the mechanisms of these systems and it can be considered as the skeleton of complex systems. One of the most relevant features of graphs representing real systems is clustering, or community structure. The communities are clusters (groups) of nodes, with more edges connecting to nodes of the same cluster and comparatively fewer edges connecting to nodes of different clusters. The major drawback of most of the proposed algorithms is that they require knowledge of number of communities to detect. In this paper, we propose a new approach for community detection in complex networks based in influential nodes, without a priori knowledge of k influential nodes. Influential nodes, central nodes, hub nodes or leader nodes have been identified as good influential. There are more able and susceptible to diffuse information and propagate influence.

Nouvelle approche de détection de communautés dans un réseau social : Application à une plateforme d'entreprise 2.0

Seddik Reguieg*, Noria Taghezout**
Abdelwahid Elmaghit***

*Laboratoire LIO, Département d'Informatique, Université Oran1 Ahmed Benbella, BP. 1524 El Mnaouer, Oran, Algeria
rseddiko@gmail.com
** Laboratoire LIO, Département d'Informatique, Université Oran1 Ahmed Benbella, BP. 1524 El Mnaouer, Oran, Algeria
taghezout.nora@gmail.com
***Département d'Informatique, Université Hassiba Benbouali, Hay Salem, route nationale N° 19, Chlef, Algérie
el_meghit_abdelwahid@yahoo.fr

Résumé. Aujourd'hui, nous sommes entrés dans une nouvelle ère où la plupart des utilisateurs du web sont connectés à un réseau social. Il devient alors indispensable de faciliter les communications et les échanges entre ces utilisateurs. Pour se faire, il est nécessaire de les étudier sous leurs divers aspects structurels à savoir une représentation graphique où les utilisateurs sont vus comme des nœuds reliés entre eux dans un graphe. Dans ce travail, nous nous intéressons à l'aspect dynamique de ces réseaux et proposons une approche de détection de communautés dans ces derniers. Dans ce contexte d'idée, s'inscrivent plusieurs algorithmes, l'un d'eux est très utilisé et très connu, il s'agit de l'algorithme Louvain. Néanmoins, cet algorithme ne s'applique que sur les graphes statiques. Notre but est alors, d'améliorer cet algorithme en lui donnant un contour sémantique, qui pourra être exploité pour la gestion dynamique des communautés. La particularité de notre approche réside dans l'attribution de nouveaux nœuds aux communautés les plus proches avec un gain de temps et cela tout en gardant un maximum de modularité.

1 Introduction

Récemment, le modèle des entreprises a changé à partir d'un ancien qui est basé sur les commandes et la gestion de l'information, pour un nouveau qui intègre un nouveau paradigme de la collaboration qui est activé et assuré par l'émergence des technologies Web 2.0, ce dernier a fait naitre le concept d'entreprise 2.0.

L'entreprise 2.0 prend en charge l'informelle organisation comme un ensemble de technologies à savoir les réseaux sociaux (Cross & Parker, 2004). Ces réseaux sociaux sont combinés avec d'autres ressources selon les axes de collaboration, de connexion et de communication afin d'améliorer l'organisation des connaissances et des compétences des utilisateurs.

Ainsi, selon Lazega (Lazega, 1994), Un réseau social est généralement défini comme un ensemble de relations d'un type spécifique (par exemple de collaboration, de soutien, de conseil, de contrôle ou d'influence) entre un ensemble d'acteurs. Chaque acteur du réseau est amené à créer des liens avec d'autres acteurs, l'analyse de ces liens peut permettre de prédire

les caractéristiques des acteurs ou l'apparition des liens entre eux. On peut aussi chercher à détecter des communautés d'acteurs fortement connectés, d'où l'objectif de notre travail.

La détection de communautés dans les réseaux sociaux devient, depuis quelques années un domaine de recherche en pleine effervescence, vu qu'elle peut faciliter la gestion de certaines applications, par exemple, pour les réseaux sociaux professionnels elle peut améliorer la politique de collaboration en relevant des communautés de talents, composées d'individus compétents et ayant les mêmes centres d'intérêt. En effet, elle permettra une meilleure collaboration et une communication plus efficace entre les collègues.

La détection de communautés dans les réseaux a récemment pris une importance considérable étant donné les enjeux scientifiques et industriels qu'elle représente. Une particularité de ces réseaux sociaux est que ce sont souvent des graphes dynamiques, ils évoluent au cours du temps, ce qui se traduit par l'ajout et la disparition de nœuds et de liens.

La dynamique de ces réseaux est donc un facteur important à prendre en compte, et de plus en plus de travaux sont consacrés à ce sujet et les recherches n'arrêtent pas d'évoluer. Plusieurs études ont été faites pour intégrer de la dynamique dans la détection de communautés.

Dans ce contexte, notre choix d'algorithme de détection de communautés s'est porté sur l'algorithme de Louvain (Blondel, 2008). Malheureusement cette méthode est dite statique puisqu'elle s'applique sur des graphes qui ne changent pas au cours du temps.

Pour cela, nous proposons une amélioration de cet algorithme et cela en prenant en compte les nouveaux nœuds sans ré-exécuter de nouveau la méthode.

Cette considération consiste dans le fait de trouver la communauté la plus proche à ce nouveau nœud en utilisant l'algorithme Dijkstra selon un critère de distance dans le graphe.

2 Travaux connexes

Le premier aspect de la gestion des communautés à étudier est la capacité des utilisateurs à s'agglutiner en communautés. Ceci ouvre les portes à un domaine de recherche en plein essor qui est la détection de communautés.

Il existe plus de 250 algorithmes de détection de communautés publiés, et ce nombre continue à croître. Comme il existe de nombreuses approches proposées, nous allons retenir celles ayant reçu le plus d'intérêt de la part de la communauté scientifique. Ces méthodes se classent en trois grandes catégories (S. Fortunato, 2010).

La première catégorie contient les méthodes de classification hiérarchiques qui permettent de choisir une structure de communauté parmi plusieurs niveaux hiérarchiques représentant différentes structures possibles. Parmi ces méthodes, nous citons l'algorithme à marches aléatoires *Walktrap* (Pons and Latapy, 2006). L'idée de base de cet algorithme est que *si* deux communautés sont loin l'une de l'autre, elles auront de faibles liens entre elles mais par contre plus de liens à l'intérieur. Elles ont de ce fait une grande distance et donc elles ne seront pas fusionnées. Pour dégager les communautés finales walktrap fait appel à la modularité de Newman (Newman and Girvan, 2002) pour traiter le dendrogramme obtenu.

Selon Newman (Newman and Girvan, 2002), la modalité est une métrique qui peut être considérée comme une définition de ce qu'est une «bonne communauté» d'où l'idée de chercher directement le découpage en communautés correspondant à la valeur maximale de la modularité pour un graphe donné.

De plus, un autre algorithme de classification hiérarchique nommé *EdgeBetweenness* (M.J.M Newman, 2004) est basé sur l'identification des liens se trouvant entre les communautés, ainsi que leur élimination, permettant ainsi d'identifier les différentes communautés dans un réseau. Afin de trouver les liens inter-communautés, l'algorithme *Edge*Betweenness accorde à chaque lien une mesure de centralité d'intermédiarité basée sur le calcul du plus court chemin.

La deuxième catégorie quant à elle englobe les méthodes d'optimisation d'une fonction objective, qui identifient les communautés en maximisant une fonction de qualité. Dans cette catégorie, il existe un certain nombre d'algorithmes basés sur des heuristiques pour définir la structure de communauté des réseaux. Ce type d'algorithme consiste à définir une fonction objective dont la valeur varie selon les communautés dégagées. La fonction est maximale pour la meilleure structure de communauté. Un exemple de ce type d'algorithme est *Fast-Greedy*de Newman (Newman, 2004). La présente approche commence par considérer chaque nœud comme une communauté à part, pour ensuite fusionner les communautés en paires, en choisissant à chaque étape la paire de communautés dont la fusion donne la plus forte augmentation de la modularité par rapport à l'itération antérieure. La valeur maximale de la modularité correspond à la meilleure structure de communautés. Cette méthode a une complexité en temps total d'exécution en *O(mn)* à cause du calcul exhaustif de la modularité, de ce fait cette méthode peut être utilisée pour les réseaux de taille faible (une vingtaine ou trentaine de nœuds) pour avoir un temps d'exécution raisonnable.

Finalement, la dernière catégorie porte sur les méthodes à base du modèle, dont des formalismes (définis à l'avance) s'appliquent sur les nœuds du réseau d'une manière itérative jusqu'à avoir une structure de communautés stable. Cette catégorie englobe des algorithmes de classification non supervisée utilisant des méthodes à base de prototypes exprimés dans un formalisme de modèles. Pour chaque type de données, un modèle d'apprentissage adapté à la nature des données traitées est proposé à savoir l'algorithme *Label Propagation* décrit en (Raghanva, 2007). Cet algorithme permet à chaque nœud qui se trouve au début dans une seule communauté de changer sa communauté tout en respectant un modèle d'apprentissage. Avec cette méthode, un groupe de nœuds, fortement liés entre eux, finit par se trouver dans la même communauté. L'algorithme choisit le nœud à traiter aléatoirement et avec une condition d'arrêt (non pas une mesure), plusieurs résultats finaux peuvent être obtenus par exemple lorsque les nœuds ne changent plus de communautés même si ce n'est pas toujours le cas. Notons que cet algorithme est non déterministe puisque son exécution avec le même réseau (graphe) peut donner plusieurs résultats.

Force est de constater qu'il est impossible d'évaluer toutes les méthodes publiées, néanmoins nous avons choisi des méthodes récentes et reconnues comme efficaces. Mais il est important de noter que ces méthodes sont des méthodes statiques et sans recouvrement. De plus, il ne faut pas oublier que Fortunato et al. (Fortunato and Barthelemy, 2007) ont prouvé la limite de résolution de la modularité sur laquelle se base la plupart de ces algorithmes.

Si le domaine de la détection de communautés statiques sans recouvrement semble aujourd'hui arrivé à maturité, avec plusieurs méthodes se révélant à la fois rapides, performantes et élégantes dans leur concept, nous pensons que le problème de la détection de communautés avec recouvrement n'a pas encore atteint ce stade (Rémy, 2013).

Quant aux communautés dynamiques, nous pensons que l'on est encore dans une phase d'exploration, et il faudra encore attendre quelques années avant d'arriver à un stade de maturation. Ceci explique le nombre restreint d'algorithme existant dans cette catégorie et

l'abondance de l'utilisation des méthodes de détection de communautés statiques dans la plupart des applications actuelles.

Cependant, la plupart des méthodes de détection de communautés dynamique proposées se base sur des méthodes de détection de communautés statiques, malgré les efforts fournis ces méthodes ne couvrent pas toutes les opérations de communautés.

3 Approche proposée

Avant de détailler notre approche, nous allons décrire notre contexte d'application qui est la détection de communautés pour un réseau social professionnel Algérien dans notre plateforme d'entreprise 2.0.

3.1 Notre réseau social professionnel

Notre réseau social professionnel (figure 1) offre un soutien aux entreprises algériennes à savoir des nouveaux modèles et des outils de collaboration basés sur les technologies web 2.0.

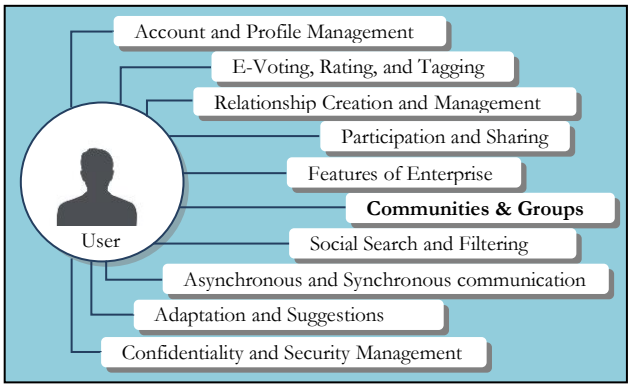


FIG. 1 – *Modules de base de notre réseau social professionnel.*

Notre réseau social permet l'attribution d'une identité professionnelle numérique à des profils différents. Il permet la mise en place de la collaboration tout en enrichissant la relation entre les différents profils; ceci afin de maintenir la continuité de leurs liens professionnels.

En termes de communication, notre réseau est considéré comme un territoire d'expression pour les entreprises où un espace virtuel qui permet à la fédération de leurs employés et permettra par la suite de travailler tout en partageant des informations et des idées sur leur activité du secteur.

Dans notre article , nous allons nous concentrer sur le module Communautés. Ce module permet la détection des communautés basée sur les liens entre les utilisateurs. Nous avons adopté dans notre approche le principe du graphe social qui désigne le réseau de connexions et de relations entre les utilisateurs de notre réseau social, ce qui permet la diffusion et le filtrage efficaces de l'information. Dans ce graphe, les acteurs de notre plateforme sont représentés par des nœuds et les liens entre eux par des arêtes pondérées. La détection de

communautés dans les réseaux sociaux analyse donc l'état d'un graphe social en se basant sur l'étude des relations entre les utilisateurs. Dans ce qui suit nous allons faire le point sur le notre algorithme de détection de communautés que nous avons nommé DynLouv.

3.2 Algorithme DynLouv

Vu le travail colossal qui reste à accomplir pour la détection de communautés dynamiques. La conclusion selon laquelle aucun algorithme n'est le meilleur dans tous les cas, mais qu'au contraire les solutions se complètent semble corroborer l'intuition selon laquelle chercher un algorithme qui serait le meilleur pour tout type de communauté dans tout type de réseau est une utopie. Néanmoins, il existe une méthode, dite méthode de Louvain (Blondel, 2008) qui est plus efficace dans le domaine (Thomas, 2011), suite à cela, l'idée nous est venue de proposer une stratégie qui pourrait bénéficier de ces avantages.

L'exécution de l'algorithme de Louvain nous ramène à un problème NP-difficile (Brandes et al., 2006) (Thomas, 2011). Dans la suite de notre article, nous avons utilisé principalement la méthode de Louvain qui produit des résultats parmi les modularités les plus élevées tout en étant très rapide. Néanmoins l'algorithme de Louvain s'applique spécialement aux graphes statiques qui ne changent pas de structure au cours du temps. Comme nous l'avions déjà mentionné dans les travaux connexes, certains chercheurs ont utilisé les méthodes statiques pour les réappliquer à chaque instant "t" du changement de la structure des graphes dynamiques. Cela pose quelques problèmes à savoir le temps d'exécution qui sera très important, ainsi que le parcours de tout le graphe à chaque moment.

Dans ce qui suit nous allons proposer une méthode (nommé algorithme DynLouv) assez satisfaisante qui permettra un gain considérable de temps tout en maximisant le critère de la modularité et cela en prenant en considération les nouveaux nœuds.

A un instant t, lorsqu'un nouveau nœud arrive dans le graphe, nous cherchons et classons les communautés les plus proches. La distance entre le nouveau nœud et les communautés est calculée via la somme des pondérations des liens reliant le nouveau nœud et les centres de gravité de ces communautés. Visiblement, Cette réflexion permettra un gain de temps du calcul (au lieu de ré-exécuter toute la méthode de Louvain) et on essayera dans tous les cas d'avoir une modularité maximale.

Dans ce qui suit, nous proposons le pseudo code de notre algorithme DynLouv.

Algorithme 2. Pseudo-code de l'algorithme de DynLouv

```
1:  G le graphe initial; M modularité initiale
2:  Si M != -1 && N Nouveau Nœud alors
3:  Calculer le plus court chemin aux centres des communautés existantes.
4:  Classement des communautés Ci..n selon le critère distance
5:  Tant que Ci n'est pas la dernière communauté faire
6:  Si il y a un gain de modularité par rapport à Ci
7:  Placer N dans la communauté la plus proche Ci
8:  Aller à 31
9:  Finsi
10: Fin tantque
11: Sinon
12: Répéter
13: Placer chaque nœud de G dans une unique communauté
14: Sauver la modularité de cette décomposition
```

15: Tant que il y a des sommets déplacés faire
16: Pour tout nœud n de G faire
17: Chercher c la communauté voisine de n maximisant le gain de modularité
18: Si n est déplacé dans C
19: si c induit un gain strictement positif alors
20: déplacer n de sa communauté dans c
21: finsi
22: fin pour
23: fin tantque
24: si la modularité atteinte est supérieure à la modularité initiale alors
25: fin faux
26: Afficher la décomposition trouvée
27: Transformer G en le graphe entre les communautés
28: sinon
29: fin vrai
30: finsi
31: jusqu'à fin

Dans le pseudo code, nous remarquons qu'une première partie sert à traiter le nouveau nœud selon les communautés déjà créées ou pas.

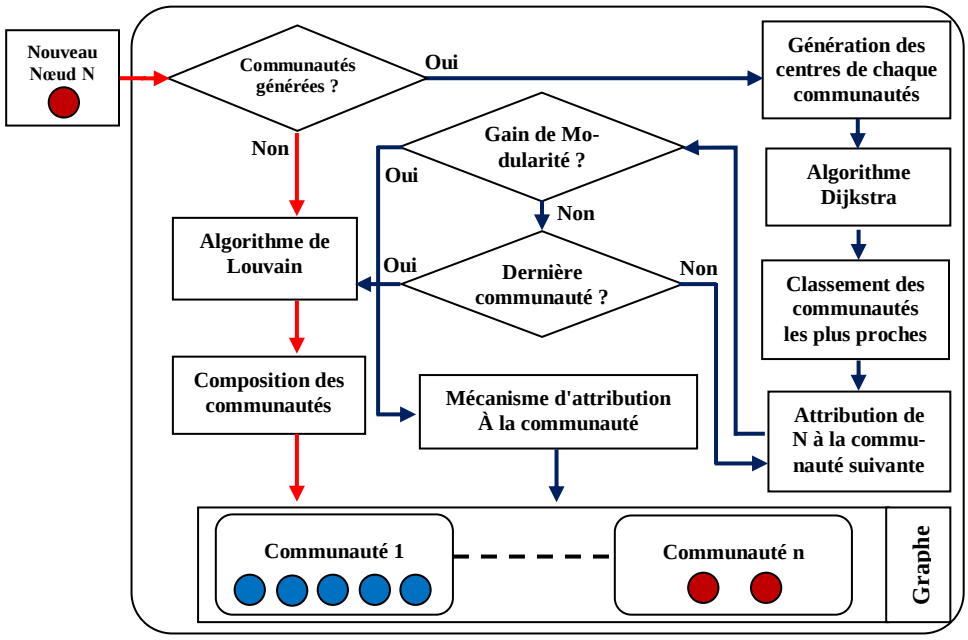


FIG. 2 – *Schéma descriptif de notre méthode.*

La figure 2 présente un schéma descriptif de notre méthode. Dans notre approche, la méthode de Louvain sert à créer les premières communautés d'une façon statique, d'où son réexécution pose un problème de temps de calcul. Notre motivation est de proposer une méthode qui évite cette réexécution. Pour cela, lorsqu'un nouveau nœud arrive sur le graphe, nous essayons de choisir la communauté adéquate qui maximisera le gain de modularité.

Dans le cas ou le gain ne sera pas maximisé, le nœud est supposé ‘pas bien’ classé et donc on relance l'exécution de Louvain ce qui va nous donner forcément une nouvelle composition de communautés.

3.3 Générateur de graphes

Nous avons développé dans notre approche un générateur de graphes aléatoires complexes (figure 3).

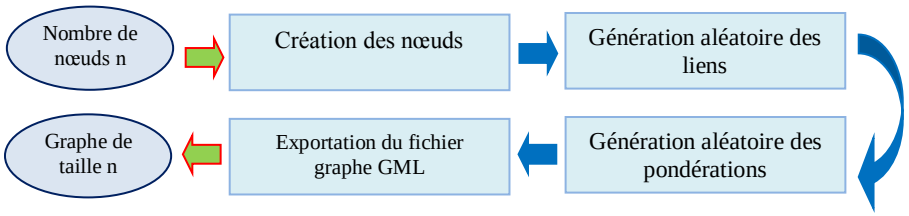


FIG. 3 – *Schéma descriptif de notre méthode DynLouv.*

Afin de sauvegarder le même graphe pour une utilisation ultérieure, nous avons utilisé comme sortie un fichier de type GML[1].

4 Illustrations

Nous avons développé notre application sur une machine Intel Pentium(R) Dual-Core CPU, avec une vitesse de 2.00 GHZ, dotée d’une capacité mémoire de 4 GB RAM.

L’application est développée en utilisant le langage de programmation PHP afin de l'intégrer directement dans notre plateforme d'Entreprise 2.0.

Dans ce qui suit, nous présentons le déroulement de notre méthode à travers un exemple, afin de comprendre son fonctionnement. La figure 4, illustre l'arrivée d'un nœud 62, sur le réseau. Après le classement des communautés les plus proches, notre méthode a attribué ce nœud à la communauté dont la couleur est bleue et cela en maximisant le gain de la modularité.

Les figures 5, 6 et 7 décrivent l'interface d'accueil de notre plateforme d'entreprise 2.0 ainsi que quelques fonctionnalités à savoir la suggestion de relations, l'administration d'un compte d'entreprise et un aperçu du graphe social qui reflète les liens entre les membres de notre réseau social professionnel.

[1] GML : Graph Modeling Language

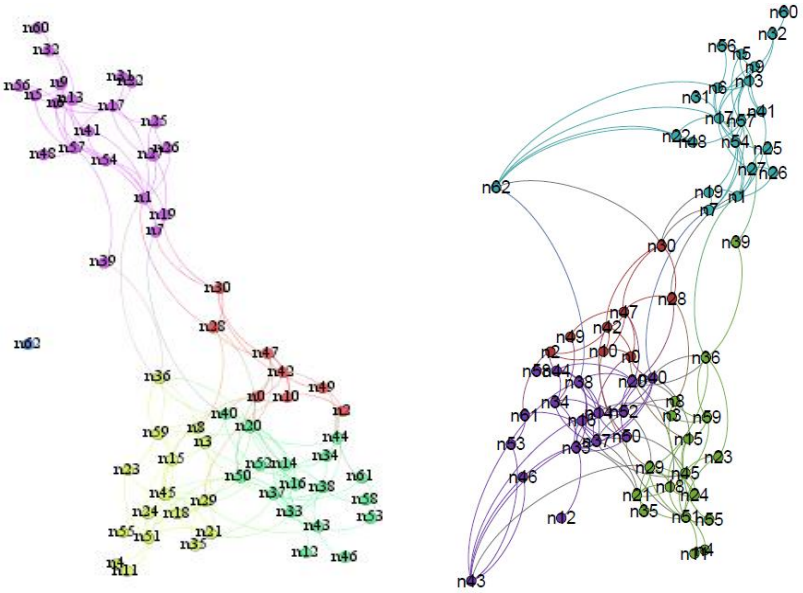


FIG. 4 – *Arrivée et attribution d'un nouveau nœud à une communauté sur le réseau.*

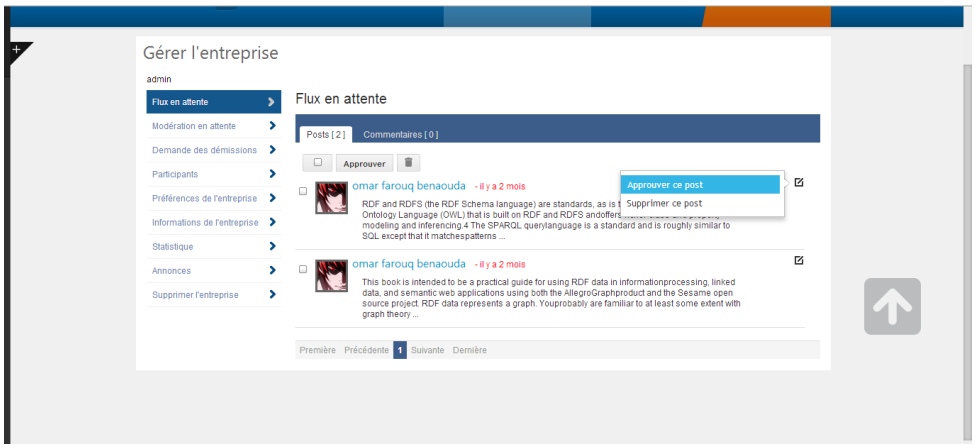


FIG. 5 – *Gestion du compte d'une entreprise.*

Notre plateforme d'entreprise 2.0 possède un module de réseaux social tout comme les autres réseaux sociaux existants, permettant l'échange et la collaboration entre les utilisateurs, la figure 5 présente un aperçu de la fonctionnalité de gestion de compte d'entreprise, cette fonction permet la modération du compte, la gestion totale de ses publications, de sa sécurité, de sa confidentialité, ses employés et d'avoir des statistiques sur son compte.

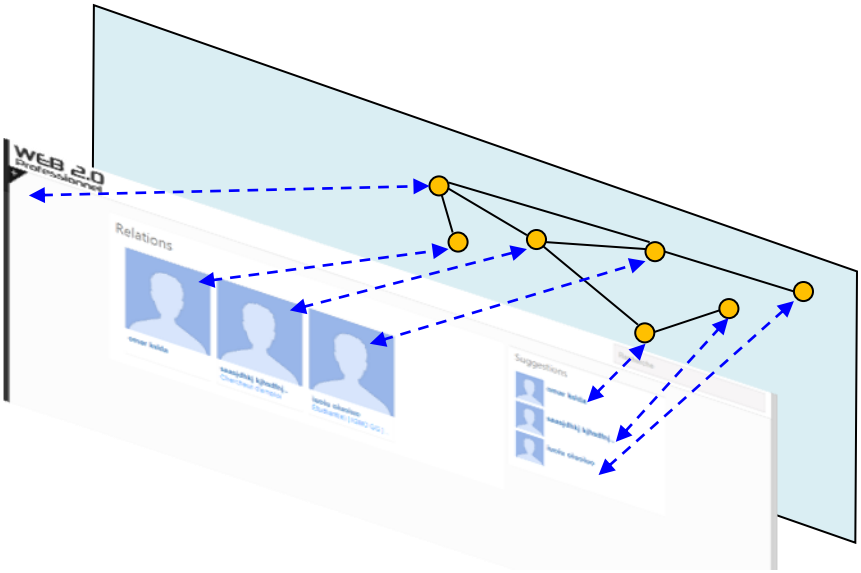


FIG. 6 – *Suggestion de relations à un utilisateur et superposition sur le graphe.*

Dans notre plateforme, l'utilité de la détection des communautés ne réside pas seulement à la création des groupes ou des communautés mais aussi une première phase pour les systèmes de recommandation. Ainsi, notre approche permet la recommandation des utilisateurs similaires qui appartiennent à la même communauté (figure 6).

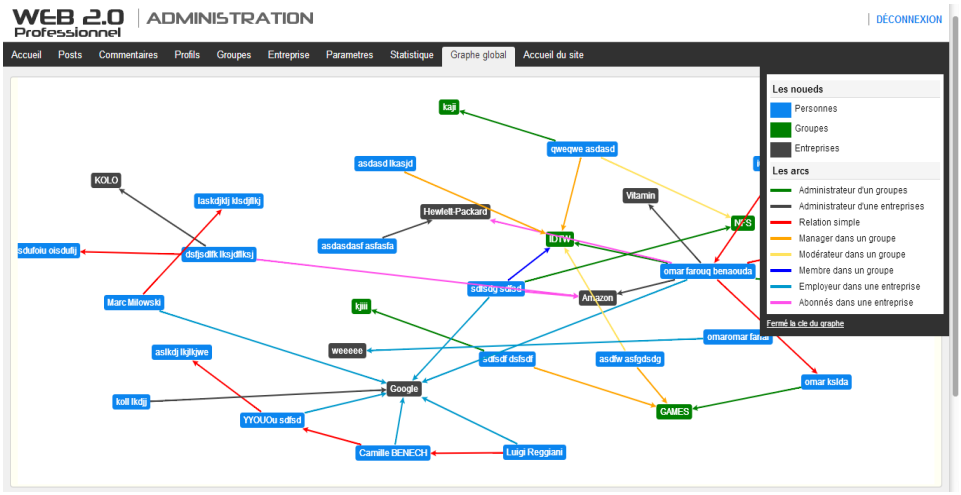


FIG. 7 – *Aperçu global du graphe social de notre plateforme d'entreprise 2.0.*

5 Résultats

Les résultats obtenus sont basés sur 17 graphes de différentes dimensions. Dans ce qui suit quelques résultats qui paraissent satisfaisants de point de vue modularité et temps de calcul. Nous avions adopté la démarche suivante afin de récolter ces résultats (figure 8, figure 9) :

1. Générer un graphe G (Nœuds et liens)
2. Extraction des communautés avec Louvain
3. Ajouter un nœud N au graphe
4. Attribution du nœud à la communauté la plus proche selon le classement
5. Calculer la modularité M' de notre méthode DynLouv
6. Calculer la modularité Louvain M sur le nouveau graphe G' tel que G' = G+{N}
7. Comparaison entre les deux modularités.

Nous cherchons dans notre approche à maximiser le gain de la modularité et donc nous avons jugé que le fait d'attribuer le nouveau nœud à la communauté la plus proche provoquera un gain de temps assez considérable (figure 9) et une augmentation de la modularité (figure 8).

Cependant, parfois le temps d'exécution de notre méthode est plus lent que la méthode de Louvain et cela est induit par le fait qu'il n'existe pas de communauté qui maximise le gain, et donc faut ré-exécuter tout l'algorithme de Louvain.

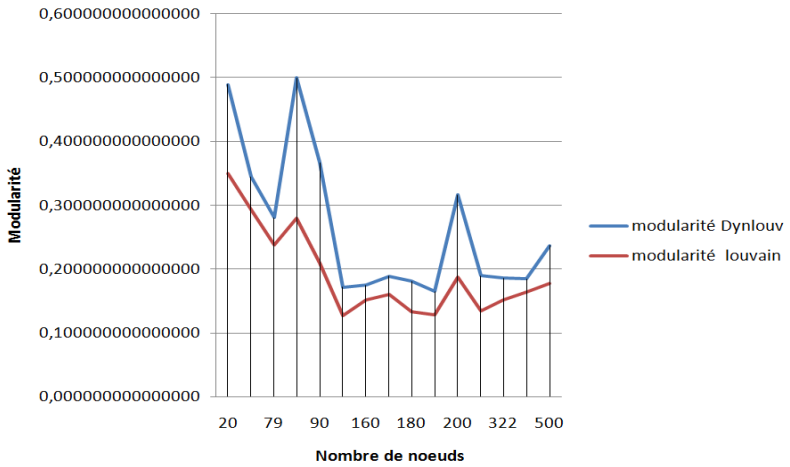


FIG. 8 – *Comparaison entre les modularités de la méthode Louvain et DynLouv.*

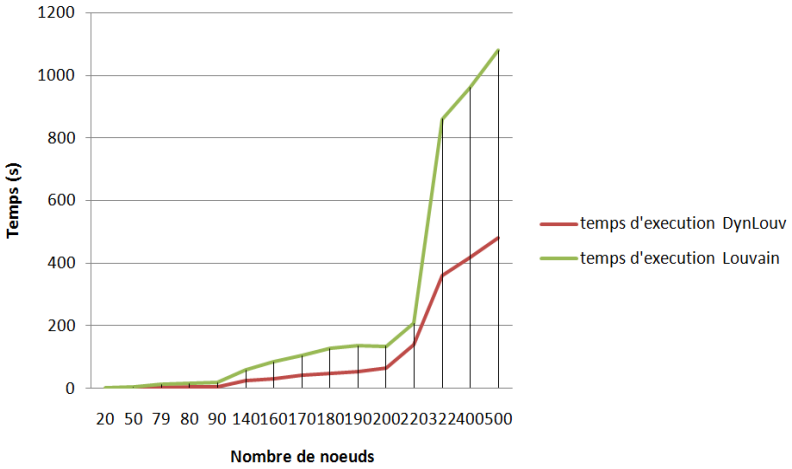


FIG. 9 – *Comparaison entre les temps de calcul des méthodes Louvain et Dynlouv.*

6 Conclusion

Nous avons trouvé passionnant de se confronter à un problème pour lequel il reste beaucoup à inventer. Nous proposons une approche riche en perspective pour l'ajout des nouveaux nœuds. Il se trouve que nous étions en quête d'une méthode efficace qui nous permettra de réaliser cette opération d'ajout. C'est pourquoi nous nous sommes basés sur l'algorithme de Louvain reconnu comme le plus efficace du domaine malgré qu'il soit une méthode de détection de communautés statiques.

La méthode de Louvain utilise la modularité comme fonction de qualité d'une partition, cette fonction est optimisée à travers un processus itératif enfin de trouver la meilleure partition. Or il s'avère que la modularité n'est pas exempte de limitations, ce qui nous a poussé à inclure une démarche de distance afin de bien faire l'attribution de nouveaux nœuds sans même ré-exécuter l'algorithme de Louvain et ainsi améliorer le temps de calcul tout en maximisant le gain de la modularité.

A notre connaissance, cette approche est la première dans le domaine de la détection de communautés pour la classification des nouveaux nœuds, apportant ainsi un aspect dynamique à l'algorithme de Louvain.

Nous avons obtenu des résultats assez satisfaisants de point de vue modularité et temps de calcul. Notre méthode possède quelque limites à savoir, lors de choix de la communauté la plus proche, il s'avère que parfois il n y avait pas de gain de modularité ce qui nous a poussé à réappliquer la méthode de Louvain afin d'avoir une nouvelle composition des communautés.

Références

Cross, R.L. and Parker A. (2004). The Hidden Power of Social Networks: Understanding How Work Really Gets Done in Organisations. Harvard Business Press. Boston, MA: USA

Lazega, E (1994). Analyse de réseaux et sociologie des organisations. *Revue française de sociologie 35*, 293.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiaotte et Etienne Lefebvre (2008). Université Catholique de Louvain-la neuve- Belgique, Université Pierre et Marie Curie. Département de Génie mathématique. Fast unfolding of Communities in large networks. *Travaux de recherches*.

S. Fortunato (2010). Community detection in graphs. *Physics Reports*, 486(3) :75{174}.

P. Pons and M. Latapy (2005). Computing communities in large networks using random walks. *Computer and Information Sciences-ISCIS 2005*, pages 284{293}.

M. Girvan and M.E.J Newman (2002). Community structure in social and biological networks. *Procedings of the National Academy of science*, 99(12).

M.E.J. Newman and M. Girvan (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2) :026113.

M.E.J. Newman (2004). Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(6) :066133.

U.N. Raghavan, R. Albert, and S. Kumara (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3) :036106,

S. Fortunato and M. Barthelemy (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1) :36{41}.

Rémy Cazabet (2013). Détection de communautés dynamiques dans les réseaux temporels. Université Toulouse3 Paul Sabatier, France. *Thèse de doctorat*.

U. Brandes, D. Delling, M. Gaertler, R. Goerke, M. Hoefer, Z. Nikoloski, and D. Wagner (2006). Maximizing Modularity is hard. *ArXiv Physics e-prints*. (Cité en pages 16 et 70)

Thomas Aynaud (2011). Détection de communautés dans les réseaux dynamiques, Université pierre et marie curie , France. *Thèse de doctorat.*

Summary

Today, we have entered a new era where most web users are connected to a social network (local or global). It then becomes essential to facilitate communication and exchanges between these users. To do this, it is necessary to study them in their various structural aspects namely a graphical representation where users are seen as interconnected nodes in a graph.

In this work, we focus on the dynamic aspect of these networks and we propose a new community detection approach.

In this context idea enroll several algorithms, one of which is widely used and well known is the Louvain algorithm. However, this algorithm is applicable only on static graphs.

Our goal is to improve the algorithm by assigning a giving a semantic outline, which can be exploited for dynamic management of communities. The particularity of our approach lies in the allocation of new nodes to the nearest communities with saving time and keeping maximum modularity.

User Profile Extraction Based on Social Tagging Case Study: Handicrafts women in emerging countries

Saida Kichou, Abdelkrim Meziane

Research Center on scientific and Technical Information (CERIST) Benaknoun, Algiers, Algeria
{skichou, ameziane}@cerist.dz

Abstract. Social tagging systems are based on assigning keywords freely by users, which promotes resources sharing and organization and improves information retrieval. The tags allocation by users is illustrated particularly in sites sharing photos or videos (Flickr, YouTube). As navigations and clicks, tags can be good indicators of the user's interests. In this paper, we examine the case of handicrafts women in emerging countries where we study the usefulness of social tagging to improve profile extraction. Knowing the profile will then help to improve collaboration and ease contact between handicrafts women. Two types of tagging are considered: User-User tagging and User-Product tagging.

Keywords: Social Tagging, user profile, profile extraction, handicraft woman

1 Introduction

Social Tagging has become a very popular way to share, annotate, and discover online resources in Web 2.0. In this way, the user is becoming active; he is involved in the information production where he can enrich the content of these resources.

Collaborative or social Tagging is recently recognized for its potential to leverage collaborative production of information that support a wide range of mechanisms such as social search (Yahia, 2008), and recommendation (Sigurbjörnsson, 2008) (Bellongin, 2013), although tagging was originally thought as a technique to improve personal content management.

Tagging system has emerged as a support to organize shared resources. It allows users to participate in content enrichment by adding key words (tags) to describe the resources, for a better categorization. Its simplicity and usefulness to improve information retrieval have attracted a high number of users (Michlmayr, 2007).

As the social media are growing in terms of number of users, resources and interactions, the user may be lost or unable to find useful information. Social elements could avoid this disorientation. Tags as an example of social elements, become more and more popular and contribute to avoid the disorientation of the user (Mezghani, 2012). Users on the Social Web interact with each other, create/share content and express their interests through chosen tags. Tags are new information to create or enrich the user profile (Carmagnola, 2007).

Tags are tools to mark resources, on the one hand for guiding other users to have information (Helic, 2010), and on the other hand to receive information about a user due to the history of tagging (Gupta, 2010).

Researches try to represent the user as accurate as possible through different techniques. Several studies are conducted in this area and propose approaches for profile extraction.
Our work is a part of a global Algerian-Tunisian study [1]: "Towards a new Manner to use Affordable Technologies and Social Networks to Improve Business for Women in Emerging Countries", in which artisan profile is defined using ontology. We have so adapted our model presented in (Kichou, 2011).

In this paper, we present briefly social tagging systems, in which a considerable number of users annotate shared resources, (text document, image, video) to which several and divergent tags are already assigned. These tags can enrich the user profile associating them. Then, we present a set of works in tag-based profile extraction that propose different manners to construct profile. We present then our approach used for handicrafts women in emerging countries.

In this work we propose two types of tagging, (User-User) Tagging in which users are tagged by others and (User-Product) Tagging allowing us to extract users preferences and skills.

According to this, we propose an approach to extract user profile by building users interest's vectors using method proposed in (Kichou, 2011).

The paper is organized as follows, Section 2 shows briefly what the Social Tagging is, in Section 3 we present the related works in our area. We present our approach in Section 4. Finally, we finish with a conclusion.

2 The Social Tagging

Social Tagging denotes the process of free associating one or more "tags" to a resource (web page, photo, video, blog ...etc.) by a set of users. The term tagging is often associated with folksonomy, it refers to a classification (taxonomy) made by users (Folks) (Mathes, 2004), (Vanderwal, 2005), and defined by (Broudoux, 2006) as a series of metadata created by a collective users to categorize and retrieve online resources.

Many tagging systems are present on the web such as Delicious for web pages, Flickr for images, YouTube for videos, Technorati for blogs and CiteULike for scientific papers.

[1] http://projetat.cerist.dz/

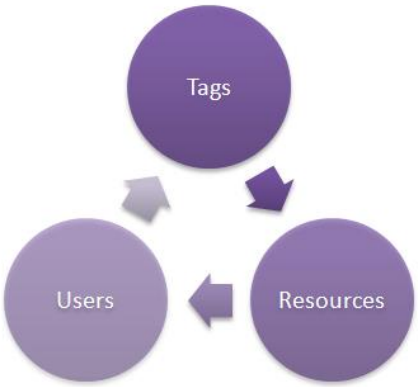


FIG.1- *Tripartite structure of tagging system*

In a tagging model, there is mainly three entities, users, tags and resources (FIG.1). Links can be found between resources (such as links between web pages), and between users (social network) (Marlow, 2006).

In our context, social tagging enables actors such as end-users, handicraft women, customers and suppliers to tag content. This action may enrich initial descriptions of the products, and the set of tags handled by a given user is considered as an important indicator on user preferences (Carmagnola, 2008).

3 Related Work

In recent years, several studies are conducted in profile extraction. (Carmagnola, 2008) Considers that tags are a new type of user feedback, and can be a very important indicator of his preferences. Different approaches of profile construction based on tagging are presented.

(Cayzer, 2009) Presents the naïve and the co-occurrence approaches that are used to construct a user profile. The first one results in a generic profile, the second results on a better profile but tags are not weighted. He presents also a new approach based on creation of a user tags graph, which takes into account the tag age.

A hybrid approach was implemented in (Kichou, 2011), it is a combination of naïve and co-occurrence approaches. It seems more efficient in that it results a more specific profile with weighted tags.

To create a specific and dynamic user profile basis on tags, (Huang, 2008) introduced the concept of tag capacity to represent a resource based on two factors, the order of tagging (ie. the position of the tag in the list cited by the user) and popularity. According to (Golder, 2005) the first tag given by a user for a resource is more representative than the following. Huang in (Huang, 2008) used this theory to calculate the tag strength to describe a person (user).

Abel in (Abel, 2011) represents tag-based profile as a set of weighted tags for cross-system user profile, as well as (Firan, 2007) creates a user profile based on tags used for a better music recommendation on Last.fm, based on a logarithm function. (Schöfegger,

2012) tests in addition to the popularity binary values indication whether or not the user has used the tag.

According to (Kichou, 2013), for a better result, resource tags must represent well its content. Existing systems consider 'Popularity' as the unique criterion to judge the tag effectiveness. However it does not always reflect its importance and representativeness to the resource. Authors proposed a novel approach based on tag strength to represent a user. In which they introduce weighted tags based on user expertise.

In the context of a dedicated system for handicrafts women, in addition to these we have suppliers and clients/customers. For this, we propose an approach to extract profile by using two types of tagging: (User-User) tagging and (User-Product) tagging.

4 User Interest Extraction Based on social tagging for Handicrafts Women

In this section, we present our proposition to exploit social tagging for handicrafts women in order to know their preferences, and make them in contact with other communities of handicrafts women, customers and/or suppliers.

Our motivation is to improve their business by making both them and their products visible for others.

The idea is to allow different users to tag other users and tag handicrafts women products.

We defined Two types of tagging: (i) (User-User) Tagging and (ii) (User-Product) Tagging.

(i)- The (User- User) Tagging consists in associating one or more keywords (tags) to different actors (handicraft woman, supplier, customer);

(ii)- The (User- Product) Tagging consists in associating one or more keywords to describe existing products.

We first present a model of the user profile to contain their personal information, activity and expertise, and a construction method of this profile based on tags.

4.1 User profile Modeling

To make our idea, we first define a user model that will contain his different information, and then we present the different steps of building the user profile, we give in the following a brief description.

The user profile is a structure of heterogeneous information, which covers broad aspects such as cognitive environment, social and professional users (Tamine, 2006). This heterogeneity is often represented by a multidimensional structure. Eight dimensions in the literature are defined for the user profile (Amato, 1999), (Bouzeghoub, 2005): the personal data, interests, expected quality, customization, domain ontology, the return of preferences

(feedback), the security and privacy and other information. A user profile is constructed either in a static way, by gathering information that rarely changes like name, age, etc., or in a dynamic way, by gathering information that frequently changes. Information about a user is obtained explicitly by the user himself or implicitly by observing his behavior during his session (history, clicks, pages visited, etc.). The user profile contains information such as (Zayani, 2008): 1) Basic information, which refers to the name, age, address, etc. 2) Knowledge of the user, which is extracted generally from his web page navigation. 3) Interests, which are defined through a set of keywords or logical expression. 4) History or feedback, which design, collected information form user's activity and could be deduced from numb er of clicks, time allowed in consulting resource, etc. 5) Preferences, which are characteristics of user describing presentation style, color, font etc. in web pages.

Defining the profile of a particular user for a given application is equivalent to select the dimensions considered useful (Bouzeghoub, 2005). In our case, we have two main categories: (1) handicrafts women or artisan and (2) customer or supplier. The difference is mainly due to the large quantity of information existing for the artisan (cognitive and physical capacities, improving the production-process, necessary tools for her production…etc.) which do not exist for the second category of users.

In our context, we are interested in information concerning tagging data, so we present only the related concepts.

(1) To make a tagging system, the handicrafts woman ontology profile is completed with the following concepts:

- *The interest Concept*: Is a vector Int (u_i) telling us about the user interests and preferences, based on tags he associates to different products via the User-Product Tagging. The interest vector is built using the hybrid approach Kichou (2011, 2013).

Int (u_i) = {(t_1, w_1), (t_2, w_2) ... ... (t_j, w_j)}. Where t is the tag, w is the tag weights' is a chosen threshold. For e.g.: Int (u_i) = {(céramique,6),(potterie,4)….}.

- *The Descriptor Concept:* The descriptor concept enable us to know the users thoughts on a given user, it is the result of User-User Tagging. It is a vector of tags, weighted by their frequency of appearance using the Naïve approach (Cayzer, 2009).

- *The expertise Concept*: Expert users in a given area, use specific terms to tag since they have a perfect mastery of the concepts in this area. This dimension is the degree of mastery of the user in tagged resources domain. It depends on the tag levels in the domain ontology used for this purpose. The more the expertise is great, the more the user is close to the resource context. For more details, see (Kichou, 2011).

For e.g. tags '*marne'*, '*silice'* and '*argile'* are deeper (so more specific) than tag '*terre'* in pottery domain.

(2) The customer/Supplier ontology profile is created, in addition to their personal information, activities their tagging information are similar to concepts presented in (1), however the expertise is not used for this type of users.

4.2 Profile Extraction

The user profile Extraction is building its three dimensions Interest, expertise and descriptor based on tagging operations the user performs.

4.2.1 Expertise Calculation

An expert user in one domain has a perfect mastery of specific terms in this domain. Therefore, he associates these terms with specific resources that he tags (e.g. in pottery, an expert associates exactly the noun of used raw material, whereas a novice just associates the term 'terre').
Expertise is defined as the average depth of the user tags and it is calculated as follows:

$$Expertise\ (Ui) = \frac{\sum Depth(tj)}{|Tu|} \qquad (1)$$

Where Depth (tj) is depth of tag tj, that is the number of nodes separating it from the root in a given ontology (we use in our case craft ontology); T_u is a set of the user's tags that it has associated to products, defined as follows:

$T_u = \{t_j \mid (u_i, t_j, r) \in Y\}$.

The expertise calculation concerns only handicrafts women.

4.2.2 User-Product Tagging and Interest Vector Extraction

The application of the (User- Product) Tagging allows, in one hand to have a descriptor that is a vector of weighted tag for the product (Called product descriptor), that will enrich its description among others, facilitate research and make it more visible to the community (FIG.2). On the other hand, this type of tagging offers the possibility to know on which products users are interested by identifying their interests' vectors. The interest vectors extraction is based on the hybrid approach (Kichou, 2011).

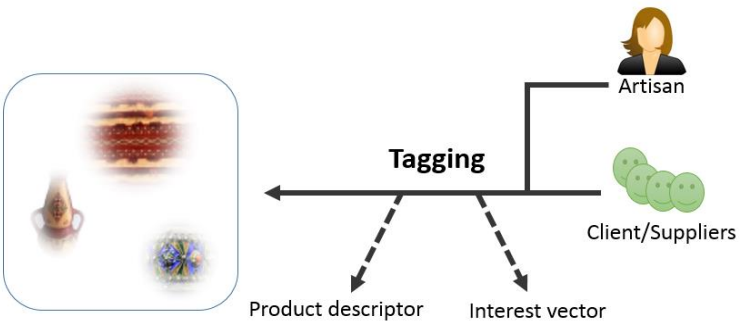


FIG.2- *The User-Product Tagging*

We extract the interest vector using the hybrid approach, which is the combination of the naïve and the co-occurrence one. It consists to create a graph where nodes represent tags cited by the user and the edges are the relations of co-occurrence between these tags (FIG.3). Tags are weighted by their popularity (naïve), arcs are weighted by the number of co-occurrences. The resulting profile is the top k nodes participating in the arcs with the greatest weight. In FIG.3 the interest vector is:{(poterie,11), (silice,9), (Argile,6), (céramique,9)}.

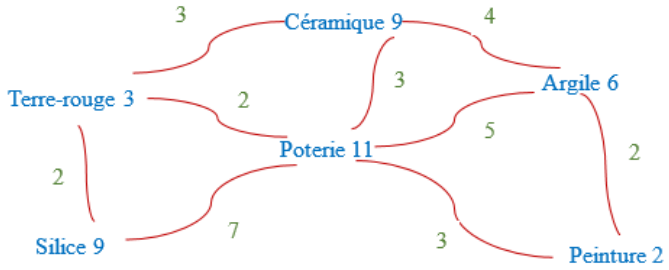


FIG.3- *A constructed graph using the hybrid approach*

4.2.3 User-User Tagging and Descriptor Extraction

Using (User- User) Tagging, the handicraft woman, the supplier and the customer will be described by a set of tags in descending order of their weight, i.e. a weighted vector of keywords (that is called user-descriptor). These sets of tags will enable us to know the thoughts of users on others.

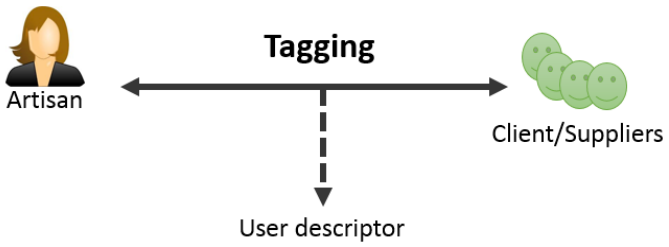


FIG.4- User-User Tagging

Tags that compose the user descriptor vector are from different other users, in this case the naive approach will be very appropriate because we just want to know what other people think without differentiation between generic or specific tag.

The result will be a list of tags in descending order of their frequency of occurrence called popularity (or tag cloud) in FIG5, the last example with the naïve approach.

FIG.5- *Tag cloud using Naïve Approach*

5 Experimentations

The user profile extraction using hybrid and naïve approaches is tested and validated in (kichou, 2011) and (kichou, 2013), however we need to show its usefulness to the handicrafts women context.
We conducted preliminary tests on a set of handicrafts women, clients and suppliers, whom tag a set of products and tag each other using French words. In the following, our results.

- Interest (u_1) = {(peinture-soie, 12), (soie,8), (broderie,4), (céramique,3), (poterie,2)} ;
- Interest (u_2) = {(ceramique,8),(peinture,7),(jarre,5),(poterie,4),(polissage,3)} ;
- Interest (u_3) = {(couture,10),(broderie,5),(peinture,3),(soie,3),(poterie,1)} ;
- Interest (u_4) = {(broderie,11),(fil-coudre,6),(coudre,5),(robe,2),(couture,2)} ;
- Interest(u_5)={(poterie,14),(polissage,11),(ceramique,10),(argile,3),(terre-cuite,1)} ;
- Interest (u_6) = {(tapis,12),(laine,8),(colorant,4),(métier-tisser,3),(polyster,3)} ;
- Interest (u_7) = {(menuiserie,9),(ébéniste,7)),(bois,5),(lissage,4),(meuble,2)} ;
- Interest (u_8) = {(broderie,7),(fetla,5),(tulle,4),(toile,1),(tissu,1)} ;
- Interest (u_9) = {(polissage,8),(argile,4),(poterie,3),(ceramique,2),(terre,1)} ;
- Interest (u_{10})= {(poterie,9),(argile,7),(terre-cuite,4),(ceramique,3),(terre,3)}.

- Descriptor (u_1)= {(sérieux, 6), (prix-bas, 4), (bonne-qualité, 3)} ;
- Descriptor (u_2)= {(prix-cher,7), (bonne-qualité, 4), (sérieux, 2)} ;
- Descriptor (u_3)= {(pas-sérieux,5), (prix-cher,4)} ;
- Descriptor (u_4)= {(sérieux,3), (moyenne-qualité,3), (prix-bas,2)} ;
- Descriptor (u_5)= {(bonne-qualité,10), (prix-bas,9), (sérieux,5)}.

5.1 Evaluation Process

We apply the hybrid approach to extract users' interest vectors and choose the first five (5) tags ordered by their weights. Interest vectors are accurate and significant enough to represent persons. Note that this vector is dynamic and may change in the time through the system use.

Descriptor vectors are very useful to know supplier reputation, to facilitate the choice of suppliers to the handicrafts woman.

6 Conclusion and Future Work

In this paper we described a technique for building a user's interests and apply two types of tagging for handicrafts women case; our first goal is to help handicrafts women communicate with each other and get in touch with costumers and suppliers. Our second goal is to extract an accurate and dynamic profile to take into account changes in preferences over time more accurately. For this, we adapted in part, our previous work (Kichou, 2013). The proposed approach was applied with preliminary tests on a set of users.

As a future work, we plan to evaluate our approach on real data (the handicrafts women network designed in the Algerian-Tunisian project). We plan to evaluate the expertise calculation for handicrafts women using the craft ontology, to know their expertise and skills for several objectives.

An important use of our approach is to exploit it for recommendation; it will be the aim of a future work.

References

Abel, F., Araújo, S., Gao, Q., &Houben, G. J. (2011). Analyzing Cross-System User Modeling on the Social Web. International Conference on Web Engineering (ICWE'11), Vol 6757, pp28-43. Springer. DOI=http://dx.doi.org/10.1007/978-3-642-22233-7_3.

Amato, G., Straccia, U. (1999). User Profile Modeling and Applications to Digital Libraries.In: Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries, Paris, France.

Broudoux, E. (2006). Folksonomie et indexation collaborative, rôle des réseaux sociaux dans la fabrique de l'information. In Collaborative Web Tagging Workshop at WWW 2006, Edinburgh, Scotland.

Bouzeghoub, M., Kostadinov, D. (2005). Personnalisation de l'information: aperçu de l'état de l'art et définition d'un modèle flexible de profils.In: Proceedings of Actes de la Conférence francophone en Recherche d'Information et Applications CORIA 2005, pp. 201–218.

Cattuto, C., Schmitz, C., Baldassarri, A., Servedio, V.D.P., Loreto, V., Hotho, A., Grahl, M., Stumme, G. (2007). Network properties of folksonomies. AI Communications Journal, Special Issue on Network Analysis in Natural Sciences and Engineering.

Carmagnola, F., Cena, F., Console, L., Cortassa, O., Gena, C., Goy, A., Torre, I. (2008). Tag based User Modeling for Social Multi-Device Adaptive Guides. Special issue on Personalizing Cultural Heritage Exploration.

Carmagnola, F., Cena, F., Cortassa, O., Gena, C., Torre, I., (2007). Towards a tag-based user model: how can user model benefit From tags? In: Proceedings of the International Conference on User Modeling.Corfù, Greece. Lecture notes in Computer Science, pp. 445–449. Springer.

Cayzer, S., Michlmayr, E.,(2009). Adaptive user profiles: Chapitre de livre Collaborative and social Information Retrieval and Access; ISBN-13: 9781605663067.

Firan.S, Nejdl.W, Paiu.R., (2007). The Benefit of Using Tag-Based Profiles. Proceedings of the 2007 Latin American Web Conference LA-WEB, page 32-41. Washington, DC, USA, IEEE Computer Society.

Golder Scott, A., Huberman, B.A. (2005). The Structure of Collaborative Tagging System. Journal of Information Science 32(2), 198–208.

Gupta, M., Li, R., Yin, Z., Han, J. (2010). Survey on social tagging techniques. In SIGKDD Explorations 12(1): 58-72.
DOI=http://doi.acm.org/10.1145/1882471.1882480.

Helic, D., Trattnery, C., Strohmaier, M., Andrews, K.; (2010). On the Navigability of Social Tagging Systems. In socialCom/PASSAT, 161-168. IEEE Computer Society, (2010).DOI= http://dx.doi.org/10.1109/SocialCom.2010.31 .

Huang, Y., Hung, C., Hsu, J. (2008). You are what you tag. In Association for the Advancement of Artificial Intelligence, http://www.aaai.org.

Kichou, S., Mellah, H., Amghar,Y,. Dahak,F. (2011). Weighting Tags Approach Based on User Profile. International Conference on Active Media Technology (AMT 2011), Lanzhou, China September 7-9.

Kichou, S., Mellah, H., Lasbeur, I., Abdelouahid, I, (2013). User Profile Extraction based on weighted tags. Webi 2013, Angers, France.

Mathes, A. (2004). Folksonomies - Cooperative Classification and Communication Through shared Metadata. Rapport interne, GSLIS, Univ. Illinois Urbana- Champaign.

Marlow, C., Mor, N., Danah, B., Marc, D. (2006).Tagging, taxonomy, flickr, article, toread. In: Collaborative Web Tagging Workshop at WWW 2006, Edinburgh, UK.

Michlmayr,E., Cayzer,S.(2007). Learning User Profiles from Tagging Data and Leveraging them for Personalized Information Access. WWW2007, May 8–12, 2007, Banff, Canada.

Schöfegger.K, Körner.C, (2012). Learning User Characteristics from Social Tagging Behavior. HT'12, June 25–28, 2012, Milwaukee, Wisconsin, USA.
Sigurbjörnsson,B., and Zwol,R,V. (2008). Flickr tag recommendation based on collective knowledge, in WWW'08. In Proceedings of the 17th international conference on World Wide Web (WWW '08). ACM, New York,USA.

Tamine, L., Zemirli, N., Bahsoun, W.: Approche statistique pour la définition du profil d'un utilisateur de système de recherche d'informations. In: Actes de la Conférence francophone en Recherche d'Information et Applications (CORIA 2006), Lyon, France (2006).

Vanderwal, T. (2005). Explaining and Showing Broad and Narrow Folksonomies, http://www.vanderwal.net/random/entrysel.php?blog=1635.

Yahia,S. et al. (2008). Efficient network aware search in collaborative tagging sites, In VLDB'08, pp. 710-721.

Zayani, C. A, (2008). Contribution à la définition et à la mise en oeuvre de mécanismes d'adaptation de documents semi-structurés. Doctoral Thesis. University of Toulouse.
Bellongin A., Cantador I. & Castells P. (2013). A comparative study of heterogeneous item recommendations in social systems. Inf. Sci., 221, 142–16.

Résumé

Les systèmes du tagging social sont basés sur l'attribution des mots clés librement choisis par les utilisateurs, ce qui favorise le partage et l'organisation des ressources et améliore la récupération de l'information. L'association des tags par les utilisateurs est illustrée en particulier dans les sites de partage de photos ou vidéos (Flickr, YouTube). À l'image des navigations et des clics, les tags peuvent être de bons indicateurs sur les intérêts de l'utilisateur. Dans cet article, nous examinons le cas des femmes artisans dans les pays en voie de développement où nous étudions l'utilité du tagging social pour améliorer l'extraction du profil. Connaitre le profil d'avantage aidera à améliorer la collaboration et le contact entre les femmes artisanes. Deux types de tagging sont considérés : le tagging User - User et le tagging User - Produit

Mots clés : Tagging Social, profil utilisateur, extraction du profil, femme artisan

Détection des intrusions : de la visualisation à l'analyse

David PIERROT*, Nouria HARBI**, Jérôme DARMONT**

*Université Lumière Lyon, Laboratoire ERIC, 69635 Lyon, Cedex FRANCE
david.pierrot1@univ-lyon2.fr
**{nouria.harbi,jerome.darmont}@univ-lyon2.fr

Résumé. La démocratisation d'Internet, couplée à l'effet de la mondialisation, a pour résultat d'interconnecter les personnes, les états et les entreprises. Le côté déplaisant de cette interconnexion mondiale des systèmes d'information réside dans un phénomène appelé "Cybercriminalité". Des personnes, des groupes mal intentionnés ont donc pour objectif de nuire à l'intégrité des Systèmes d'Information dans un but financier ou pour servir une "cause". Nous proposons une méthode d'analyse des flux permettant de détecter les comportements anormaux et dangereux afin d'appréhender les risques d'une façon compréhensible par tous les acteurs.

1 Introduction

De nos jours, il est aisé de communiquer, d'échanger, d'acquérir des biens, des connaissances via le réseau Internet. Le maintien opérationnel d'un Système d'Information et la protection des donneés sont devenus les critères essentiels pour toute entreprise, administration ou personne cherchant à délivrer, proposer un service, ou simplement souhaitant communiquer. Le côté déplaisant de l'interconnexion mondiale des Systèmes d'Information réside dans un phénomène appelé "Cybercriminalité". Des personnes, des groupes mal intentionnés ont pour objectif de nuire dans un but pécuniaire ou pour une "cause", aux informations d'une entreprise, d'une personne voire d'un État. De ce fait, des sondes de détection d'intrusions utilisant soit une base de signature des attaques connues ou des modèles statistiques comportementaux appelés "profils" ont pour mission de signaler toutes tentatives d'intrusions. Il s'avère que ces dernières sont généralement consommatrices en ressources techniques et humaines (M.Ghoniem et al., 2014) pour la prise en compte des nombreuses alertes signalées. Avec l'augmentation des échanges réseau, il devient extrêmement difficile de visualiser les flux ainsi que de prendre en compte les attaques intervenant sur un Système d'Information. Il n'existe pas d'actif secondaire ou d'entrepôts de données sans importance. Même si ce un entrepôt peu être jugé comme "sacrifiable", une intrusion pourra permettre de l'utiliser comme rebond d'attaque.
L'objectif de cet article est de présenter dans un premier temps l'état de l'art en analyse de détection d'intrusions et dans un second temps d'aborder les travaux menés afin de faciliter la détection des attaques dites de premier niveau [1]. Le but final étant de fournir une appréciation

1. Prise de renseignements, tentative d'accès une compte utilisateur par bruteforcing, mauvaises configurations de services

du risque dynamique fondée sur des indicateurs de compromission pour limiter l'exploitation des vulnérabilités en combinant la représentation graphique des flux et l'exploration des données. La première partie de cet article sera consacrée à l'étude de l'existant, dans laquelle nous présenterons les différentes approches de détection d'intrusions et leurs limites. Ensuite, nous nous intéresserons à la motivation de nos travaux et nous proposerons une solution. Nous détaillerons par la suite, la seconde phase de nos travaux ainsi que les résultats par la création d'un modèle de données et nous terminerons par une conclusion et les perspectives.

2 Etude de l'existant

Les différentes avancées technologiques des dernières décennies ont pour effet une mise à jour pratiquement obligatoire de tout Système d'Information. Ainsi les "machines", les systèmes d'exploitation, les logiciels, les progiciels n'ont cessé d'être renouvelés. Il en va de même pour les "Hommes" qui ne peuvent échapper à un constant renouvellement de leurs connaissances et de leurs compétences pour maîtriser les nouvelles fonctionnalités des Systèmes d'Information. Depuis les premiers travaux du Docteur Denning Denning (1987), les systèmes de détection n'ont cessé d'évoluer. Sumeet et Xian (2011) démontrent les avantages du Data Mining face à une augmentation des données dans les Cyber-infrastructures et du nombre croissant de cyberattaques. le Data Mining propose aussi différentes solutions pour détecter et analyser les attaques informatiques comme le précise Deepa et Kavitha (2012).

Les système de détection d'intrusions ont déjà été modélisés par de nombreuses méthodes. Ligun et al. (1995) proposent l'utilisation de méthodes à base de règles désignées par un expert. Lee et al. (1999) utilisent des méthodes du Data Mining, qui ne nécessitent pas l'intervention d'un expert. Ces méthodes génèrent énormément de règles d'association et rendent ainsi le problème complexe. Chan et al. (2008) ont vérifié les performances d'un IDS combinant plusieurs algorithmes (k-plus-proches voisins, fuzzy clustering, damper-shafer). Panda et Patra (2009a) proposent des méthodes construites à partir de règles d'association qui utilisent des mesures du support minimum et des mesures d'intérêt pour tenir compte des problèmes de la précédente approche des méthodes par clustering hybride et clustering FTT ; et des méthodes par combinaison de bayésien naïf et d'arbre de décision. Mahmud et al. (2009) propose un modèle d'IDS basé sur plusieurs algorithmes en parallèles. Zhang et Feng utilisent un modèle basé sur un séparateur à vaste marge (SVM) et colonie de fourmis Zhang et Feng (2009). Panda et Patra (2009b) ont construit des modèles par optimisation d'ACP et ACI. Siraj et al. (2009) utilisent une approche hybride intelligente. Un état global de l'utilisation pratique des IDS a été dressé par Tavallaee et al. (2010). Wang et al. (2010) proposent une méthode utilisant le fuzzy clustering et un réseau de neurones. Nguyen et al (2011) ont défini un modèle basé sur les arbres de décisions complété par une classification (k-plus-proches voisin). E.Bahri et al. (2013) montrent que l'utilisation des méthodes supervisées et non- supervisées permettent une meilleure prise en compte des comportements et abus sur un réseau informatique. L'ensemble de ces travaux a été réalisé en s'appuyant sur le base KDD99 (1999).

Cette base a été historiquement constituée à partir de données extraites d'une sonde de détection d'intrusions. Comme précisé dans l'introduction, la mise en place et surtout la surveillance d'une sonde d'intrusion par son positionnement [2] demande une analyse pointue. De

2. Écoute et enregistrement total ou partiel d'un réseau

plus la gestion des faux positifs s'avère importante et chronophage. D'autres recherches ont été menées sur un équipement de sécurité largement répandu, le "Firewall" (voir section 4). Ces études sont principalement basées sur l'amélioration du filtrage des flux réseau. Costantina et al. proposent à partir d'un modèle non-supervisé de signaler les dérivations des comportements quotidiens. K.Golnabi et al. (2006) démontrent par l'utilisation de règles d'association une amélioration de règles de filtrage. Enfin, l'étude de Ghoniem et al. M.Ghoniem et al. (2014) permet par l'utilisation de deux algorithmes (k-plus-proches voisins, unnormalized spectral clustering) la détection des anomalies et une représentation graphique plus compréhensible. Comme nous le verrons dans la section 4, nous devons être en mesure de détecter des intrusions sans disposer d'une sonde de détection d'intrusion mais à partir d'un Firewall. Ce choix peut être justifié de la façon suivante. L'implémentation d'une sonde de détection d'intrusions nécessite des compétences particulières en matière d'interprétation des événements (faux positifs, nombre important de signalements). Le notion de coût n'est pas à négliger, même si une sonde issue du monde "Open-Source" est utilisée, il convient de prendre en compte la formation des utilisateurs, la gestion des absences (congés, maladie) mais aussi le coût de la supervision durant les heures non ouvrées et les week-end. De plus, l'étude du CLUSIF (2012) montre que 95% des entreprises de plus 200 de salariés sont équipées d'un Firewall contre 34% d'une sonde de détection d'intrusions.

3 Motivations et propositions

La sécurité des données ne devant pas être réservée aux seuls experts, il convient de "vulgariser" les événements afin de les rendre compréhensibles par tous et d'automatiser au maximum les actions en découlant. Néanmoins, nous avons besoin d'un expert lors de la première itération afin de corroborer nos résultats. La mise en place d'une supervision du Système d'Information et une aide à la décision fondée sur des comportements (utilisateurs, services, serveurs) doivent permettre d'analyser les flux de données de type évènement et attaque, d'être en mesure de réagir en temps réel, et de corriger si nécessaire. L'addition de ces éléments doit permettre de prévoir les risques encourus non seulement par les différents actifs connus et suivis, mais également lors d'évolution des composants du Système d'Information. Au final, il s'agit d'inscrire la surveillance du Système d'Information dans un véritable cycle de vie, comme illustré dans la figure 1.

Notre étude portera sur quatre phases et s'inscrivent dans un cycle d'amélioration continue souvent illustré par la Roue de Deming [3] et qui peut s'appliquer à la détection des intrusions. Ces phases se décomposent de la façon suivante :

- Phase 1 : "Monitoring et visualisation" des données réseau, représentation graphique des activités d'un réseau informatique via un modèle de données.
- Phase 2 : Analyse des comportements et alertes, phase qui s'appuiera sur des méthodes de Data Mining.
- Phase 3 : "Scoring" des risques et phase d'évaluation.
- Phase 4 : Détermination d'un plan d'actions.

Les travaux présentés dans la précédente section sont basés sur des flux provenant de KDD99 ou fondés sur les événements "Firewall" et n'offrent que la possibilité d'améliorer

3. Cycle PDCA, https ://deming.org/theman/theories/pdsacycle

visualisation à l'analyse

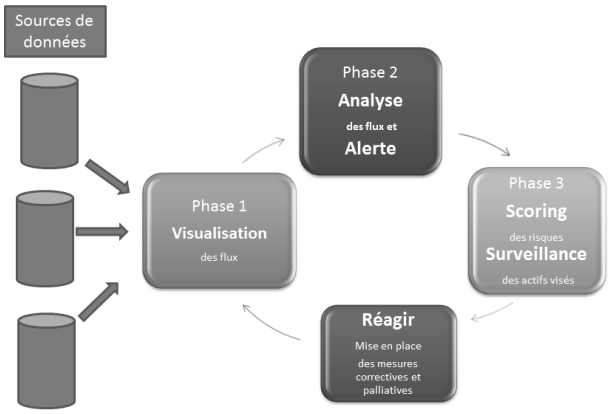


FIG. 1: Représentation schématique des différentes phases

la politique de filtrage (ce qui est déjà une bonne chose). De ce fait, nous ne pouvons reproduire des résultats similaires car par défaut les variables étudiées (durant la phase 1) sont inférieures au nombre de variables utilisées par KDD99 (41 variables). Il est donc judicieux d'utiliser certaines méthodes de Data Mining sur des événements de type "Firewall" pour la détection des anomalies. Le challenge repose sur la possibilité, à partir d'un équipement de filtrage qui par sa nature ne peut délivrer autant d'information qu'une sonde d'intrusion, de détecter les tentatives d'intrusion sur un actif, d'alerter et de mettre en place les contre-mesure adéquates.

3.1 Anatomie des attaques et des comportements

Les Systèmes d'Information sont de plus en plus interconnectés et ceci afin de proposer des services (vente en ligne, messagerie, réseaux sociaux, B2B [4], B2C [5]). L'utilisation des ports d'écoute n'est pas laissée au hasard. La RFC 6335 [6] définie par l'IANA (Internet Assigned Numbers Authority) permet de prendre connaissance d'une partie de l'assignation mise en place sur les ports allant de 0 à 65535. Un attaquant aura pour mission d'obtenir la liste des services utilisés afin de nuire à un Système d'Information.

3.1.1 Principe de base d'une attaque

Une intrusion réseau repose sur cinq étapes déclinées de la façon suivante :

1. La reconnaissance, basée sur une recherche d'information à partir d'Internet (adresse ip, créateur du site et propriétaire du nom de domaine, adresse de messagerie, liste du personnel, etc...).
2. Le balayage réseau (inventaire des services et ports, des systèmes d'exploitation et des versions logiciels serveurs utilisées).

4. Business to Business : ensemble des relations commerciales entre les entreprises et les professionnels
5. Business to Consumer : des entreprises aux particuliers
6. Request For Comments, http ://tools.ietf.org/html/rfc6335

3. L'obtention d'accès (exploitation des vulnérabilités et obtenir un accès).
4. Maintenir l'accès (rendre l'accès permanent).
5. Couvrir les traces (effacer et réduire les traces).

La première étape est difficilement détectable car dépendante de la vie numérique d'une entreprise ou de son personnel. Nos travaux seront axés sur les étapes deux et trois. Toute attaque informatique commence généralement par une prise de renseignements. Cette étape (2) consiste à réaliser un balayage des services proposés par un serveur. Une prise de renseignement fructueuse permet d'obtenir la liste des services disponibles (http, https, ftp, messagerie, etc...), la version des services utilisés (Microsoft IIS [7] ou Apache pour un serveur Linux) et le socle système "OS" [8]. Par la suite, il n'est pas difficile de se procurer des outils malveillants pouvant exploiter une vulnérabilité connue pour l'actif ciblé. Il convient de détecter rapidement la phase "renseignement". Les possibilités de prise de renseignements ayant aussi évolué, il est possible d'utiliser plusieurs adresses IP fictives [9], ou de réaliser une prise de renseignements étirée dans le temps afin de limiter l'arrivée massive d'accès à différents ports d'un serveur. La possibilité d'utiliser un ordinateur distant et vulnérable comme relai permet aussi de masquer l'adresse de l'émetteur. De plus, l'apparition de systèmes d'anonymisation comme "Tor" démontré par l'étude de Ujjaneni et Achutha 2013 et des outils comme "Shodan" [10] Sélégny (2013) ou "Google Hacking"Lancor et Workman (2007) limitant les traces laissées lors de la phase de reconnaissance Akanksha (2012), donne un sentiment d'impunité auprès de ses utilisateurs. Même si les actifs visés sont protégés par un élément de sécurité (voir section 4), la prise de renseignement peut s'avérer positive sur certains services. L'objectif pour l'attaquant est de réduire au maximum les traces laissées par son activité, ceci se traduit par le terme "attaque de signaux faible". Il convient d'identifier les prises d'informations qui peuvent servir à des préparation d'attaques, et de détecter les anomalies techniques (mauvaises manipulations, configurations non mises à jour, erreurs humaines). L'addition de ces différents vecteurs d'attaques est plus communément appelé APT [11]Advanced Persistent Threat : Menace Permanente Avancé). Un APT agissant sur plusieurs couche du modèle OSI, la détection peut être jugée comme difficile.

4 Réalisation de la première phase : Visualisation des flux réseaux

Cette phase est un préambule à la "fouille de données" qui sera effectuée dans les phases suivantes. Un des principaux équipements de sécurité est le "Pare-Feu" [12] ou plus communément appelé "Firewall". Il a pour mission comme le décrit Al-Shaer et Hamed 2003 de filtrer les flux autorisés à pénétrer dans un réseau par rapport leurs provenances, leurs destinations et les services souhaités (navigation internet, transfert de fichiers, etc...). Par son positionnement, il donne une visibilité totale de l'ensemble des flux. Cet équipement offre aussi la possibilité

7. Microsoft Internet Information Server
8. Operating System : système d'exploitation
9. Firewall/IDS Evasion and Spoofing, http ://nmap.org/book/man-bypass-firewalls-ids.html
10. Systèmes informatiques exposés en ligne, http ://www.shodanhq.com
11. (
12. Équipement de sécurité basé sur le filtrage des entrées/sorties des flux réseau.

visualisation à l'analyse

"d'historiser" vers des journaux les flux ayant été autorisés ou interdits. Il est opportun de capturer les flux réseau à partir de cet équipement et d'exporter les traces de connexion vers un conteneur de données.
Le nombre de flux passant par ce dernier étant relativement élevé, une représentation graphique dite "classique" [13] pourra permettre une vision globale mais sera sans doute inutile pour visualiser une attaque ou un comportement anormal. La figure 2 illustre parfaitement la difficulté de l'analyse des flux transitant sur un réseau. Cette dernière montre les flux tcp reçus par un "Firewall" sur une durée d'une heure soit plus de 52911 transactions. Dans cette première phase, nous obtenons une vision graphique des flux transitant sur le réseau afin d'établir un diagnostic ou d'émettre une alerte selon les points suivants :

— Prise de renseignement non discrète étendue ou ciblée.
— Suivi d'une transaction, de sa source à sa destination, selon les protocoles et les services utilisés.
— Réalisation d'une cartographie des flux afin de satisfaire les exigences d'une norme de sécurité [14] ou d'une politique de sécurité interne.

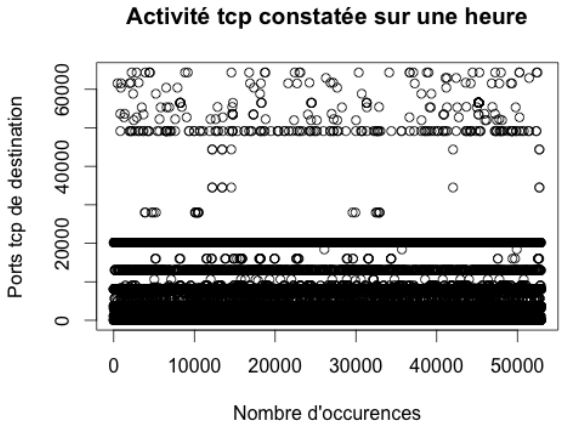


FIG. 2: Représentation des transactions réseaux TCP durant 60 minutes

4.1 Description des architectures

Il a été possible de tester la phase 1 sur plusieurs architectures et équipements de filtrage.

- Architectures de filtrage de second niveau

Cette architecture concerne une entreprise publique dans le domaine de la santé composée de 92 000 employés. Notre étude porte sur 3 réseaux interconnectés au sein d'un reseau

13. Histogramme, graphique circulaire, etc..
14. Management de la sécurité de l'information, ISO/IEC 27001 http ://www.iso.org/.

étendu (WAN : Wide Area Network) distant géographiquement et pourvu d'équipements de filtrages. L'objectif est d'être en mesure d'analyser les événements liés aux règles de filtrage via une exportation vers un conteneur de données. Afin de simplifier les références aux différents réseaux, le nommage suivant sera utilisé :

— Site de production : **SP1**
— Site de qualification : **SQ1**
— Site d'administration distante et bureautique : **SAB1**

Ces trois sites sont opérationnels, c'est à dire que les données traitées et analysées dans les sections suivantes correspondent à des données de production. Pour des raisons de **confidentialité** les adresses IP ont été anonymisées. Le réseau SP1 est doté de son propre conteneur de données issues des événements envoyés en temps réel par le "Firewall". Les réseaux SAB1 et SQ1 mutualisent un même conteneur. Les données brutes envoyées par l'ensemble des équipements de filtrage sont traitées selon une extraction de motifs.

4.2 Description des données

Le réseau SP1 propose des services à destination de **14 millions** de personnes. Les données peuvent être considérées comme sensibles et portent sur une quantité de 9.2 Teraoctets et plusieurs dizaines de millions d'euros par jour. Ces données sont hétérogènes et proviennent de plusieurs sources différentes. Les transactions réseau filtrées par le "Firewall" représentent plus de 6000 lignes par minute en matière d'événements. Le second et le troisième sites sont respectivement utilisés pour des taches dites de "bureautique-administration distante" et de qualification (reproduction des infrastructures de production). Les modalités des variables listées ci-dessous sont exportées vers les conteneurs de données. La phase 1 se focalisera uniquement sur l'analyse et la représentation graphique de ces dernières.

— adresse ip source, adresse ip de destination
— port de destination, protocole (udp et tcp)
— date et heure de la connexion
— numéro de la règle du pare feu correspondant aux flux
— action effectuée par le pare feu, flux accepté ou rejeté

Le tableau 1 synthétise le volume en nombre de lignes traitées par les équipements de filtrage.

	flux traités par journée	moyenne par minute
SP1	9 886 928	6 865
SAB1	572 272	397
SQ1	20 670	14

TAB. 1: Flux traités par SP1, SQ1, SAB1 en nombre de lignes

4.3 Traitement et résultat

4.3.1 Traitement des données

Les différents moyens décrits dans cette section permettent de réaliser le pré-traitement. Ces derniers dispensent à partir des événements bruts envoyés par les équipements de filtrage

un format exploitable via le conteneur de données. La figure 3 présente en détail les différentes étapes de conversions des données.
Les différentes traces de connexions des équipements de filtrage (option LOG) sont envoyées au serveur SYSLOG-NG [15]. Ce dernier, via ses options de reconnaissance PCRE [16] et de filtrage, dépose l'ensemble des flux traités sur un serveur de bases de données (MYSQL [17]). Par la suite, un traitement est réalisé via un script Perl [18], créé par nos soins, ayant pour objectif de préparer le résultat des différentes requêtes. Enfin, les programmes issus de la suite Graphviz [19] et du script Afterglow [20] sont utilisés pour la création des graphiques comme démontré par Marty 2008. L'utilisateur [21] n'a en fait besoin que d'un navigateur Internet pour être en mesure de visualiser les flux. Afin d'avoir une vision critique et surtout externe par rapport aux travaux entrepris, un avis a été demandé à plusieurs experts de différentes entreprises. Le constat est plus que positif. Il est facile de visualiser les événements et les diagnostics n'en sont que plus simples à déduire.

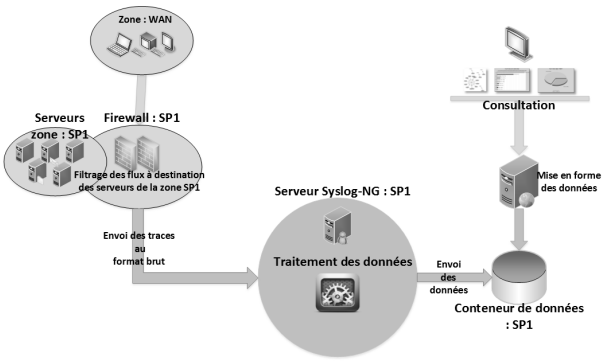


FIG. 3: Schéma traitement des événements

4.3.2 Résultat

A l'issue de la phase 1, l'ensemble des événements liés au filtrage est exporté en temps réel vers des conteneurs de données. Le traitement des informations recueillies sur les différents équipements de filtrage permet une représentation des flux. La lecture graphique permet de visualiser rapidement les tentatives de connexion depuis plusieurs sources vers plusieurs destinations. Il peut s'agir d'une tentative d'intrusion ou d'une prise de renseignement ou encore d'une mauvaise configuration d'un script. Ce constat permet de soulever des interrogations sur cette transaction et de mettre en place une action de surveillance. D'autres options ont été créées afin de permettre la visualisation des règles de filtrage les plus utilisées.

15. Gestion des journaux, http ://www.balabit.com/network-security/syslog-ng
16. Perl Compatible Regular Expression, http ://www.pcre.org
17. Base de données open Source, http ://www.mysql.com
18. https ://www.perl.org/
19. Logiciel de visualisation graphique, http ://www.graphviz.org
20. AfterGlow, outil de génération graphique, http ://afterglow.sourceforge.net/
21. Responsables de la sécurité du système d'information, ingénieurs sécurité, administrateurs réseau

5 Phase 2 : Analyse des flux réseau et alertes

La première phase étant axée sur la représentation des flux, la seconde phase a pour objectif de permettre d'analyser et de détecter les comportements anormaux et d'alerter si besoin.

5.1 Initialisation de la seconde phase

Suite à la présentation des données étudiées dans la phase 1 de visualisation, il convient de réaliser un pré-traitement afin de réduire les modalités de certaines variables issues des journaux d'événements. En effet, la variable "port destination" comporte 65535 possibilités. Nous avons opté pour un regroupement selon les trois catégories suivantes :

- Ports de destination inférieurs à 1024 (inf1024) donc les ports "connus".
- Ports supérieurs à 1024 (sup1024).
- Ports d'administration (portadm), permettant d'avoir une vision sur l'activité des ports destinée à l'administration des serveurs, des bases de données.

Le pré-traitement effectue un agrégat des adresses IP source et réalise la somme des occurrences selon les classes citées ci-dessus. De plus, il serait sans doute judicieux d'inclure le nombre total des transactions réalisées par cette même adresse IP source ainsi que le nombre de flux rejetés (actiond) et autorisés (actionp) par le Firewall. Cette opération a permis, à partir d'un échantillon de 52911 lignes d'événement, d'obtenir au final 115 individus.

	nombre	actiond	actionp	inf1024	sup1024	portadm	risque
1	16	0.0	100.0	0.0	0	0.0	non
2	12	0.0	100.0	0.0	0	0.0	non
3	**3296**	**99.3**	**0.7**	61.9	38	0.1	oui
4	36	100.0	0.0	0.0	100	0.0	non

TAB. 2: Exemple de flux agrégés ayant une définition de risque

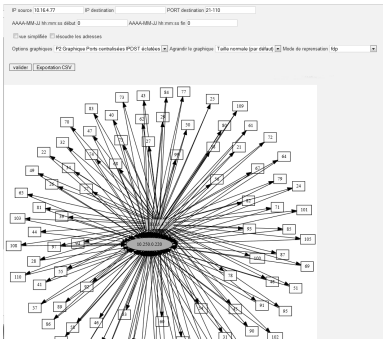


FIG. 4: Représentation graphique d'un balayage de ports (prise de renseignement)

visualisation à l'analyse

5.2 Résultat de la première itération

A l'issue de ce pré-traitement, nous avons une seconde fois demandé l'avis d'experts (cinq au total) afin de définir si le comportement constaté pouvait être considéré comme à risque. Cette variable se révèle exogène. Le choix d'un apprentissage supervisé s'est donc naturellement imposé. Nous avons pu estimer le taux d'erreurs de mauvais classement à 8% par l'utilisation d'une validation croisée (méthode du leave-one-out). Nous avons par la suite intégré de nouvelles données et il a été possible détecter des prises de renseignements comme illustré dans la ligne 3 du tableau 2. Cette ligne montre 3296 connexions dont 99.3% ont été stoppées par le "Firewall", mais les 0.7% (23 connexions) de flux autorisés par la politique de filtrage ont sans doute permis d'obtenir des renseignements sur l'actif visé. Dès lors, une surveillance sur ce dernier devrait être mise en place. L'activité réalisée sur cet actif peut être définie comme un **indicateur de compromission** [22] (activité malveillance constatée, suite à une prise de renseignement positive). La figure 5 montre l'arbre de décision issue de notre phase d'apprentissage et permet d'établir la prédiction sur le comportement à risque des flux transitant par le "Firewall".

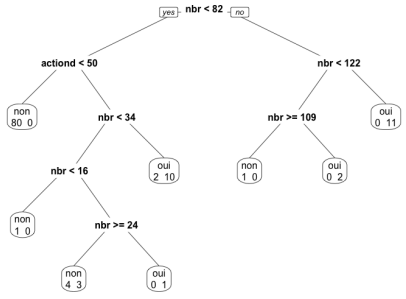


FIG. 5: Arbre de décisions obtenu lors l'initialisation de la seconde phase

6 Conclusions et perspectives

A l'issue de la phase 1, l'ensemble des événements liés au filtrage est exporté en temps réel vers des conteneurs de données. Ceci permet de visualiser rapidement les tentatives de connexion depuis plusieurs sources vers plusieurs destinations. Il peut s'agir d'une tentative d'intrusion ou d'une prise de renseignement ou encore d'une mauvaise configuration d'un script. L'inconvénient de la solution réside dans le fait que la surveillance est encore manuelle, mais la compréhension des graphiques proposés est simple et surtout accessible. Il convient de poursuivre les travaux menés sur la phase 2 afin d'obtenir des résultats plus probants et de commencer une réflexion pour la détection sur la couche applicative ou sur des prises de renseignements réalisées sur une échelle temporelle importante voir camouflées. Ceci permettra d'aborder par la suite les phases 3 et 4, à savoir établir un "scoring" et la création d'un plan

22. IoC : Indicator of Compromise

d'actions. Ces phases permettront de générer des règles d'association en fonction des actifs visés selon plusieurs vecteurs d'attaques. Une prise en compte de techniques de Data Mining permettrait une meilleure considération des attaques avec une définition automatique des seuils de signalements. Il conviendrait de créer un système évolutif et adaptatif en temps réel permettant, en fonction des changements intervenant sur un Système d'Information, d'offrir une véritable aide à la décision.

Références

Akanksha, B. (2012). Ethical hacking and social security. *Journal of Radix International Educational and Research Consortium 1*.

Al-Shaer, E. et H. Hamed (2003). Firewall policy advisoor for anomaly detction and rules. *Integrated Network Management, 2003. IFIP/IEEE Eighth International Symposium on*. english

Chan, T. et al. (2008). Network intrusion detection design using feature selection of soft computing paradigms. *International journal of computational intelligence*, 196–208.

CLUSIF (2012). Menaces informatiques et pratiques de sécurité en france. Technical report, CLUSIF (https ://www.clusif.asso.fr/).

Deepa, A. J. et V. Kavitha (2012). A comprehenseive survey on approaches to intrusion detection system. *ICMOC-2012*.

Denning, D. (1987). An intrusion-detection model. *IEEE Transactions on Software Engineering*.

E.Bahri et al. (2013). Combining unsupervised and supervised learning for better intrusion detection.
english

KDD99 (1999). Kdd cup 1999 data. URL: `http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html/` [accessed: 2015-02-27].
english

K.Golnabi et al. (2006). Analysis of Firewall Policy Rules Using Data Mining Techniques. *Network Operations and Management Symposium, 10th IEEE/IFIP*, 305 – 315.

Lancor, L. et R. Workman (2007). Using google hacking to enhance defense strategies. *ACM SIGCSE Bulletin 39*.
english

Lee, W. et al. (1999). A data mining framework for building intrusion detection model. *IEEE symposium on security and privacy*, 120–132.
english

Ligun, K. et al. (1995). State transition analysis: a rule based intrusion detection approach. *IEEE Transaction on software engineering*, 181–99.
english

Mahmud, W. et al. (2009). Intrusion detection using rough sets based parallel genetic algorithm hybrid model. *Proc. of the world congress on Engineering and computer Science WCECS*.

visualisation à l'analyse

Marty, R. (2008). *Applied Security Visualization*. ISBN: 0-321-51010-0.
english

M.Ghoniem et al. (2014). VAFLE: Visual Analytics of Firewall Log Events. *Visualization and Data Analysis 2014,*.

Nguyen, H. et al (2011). An effcient local region and clustering-based ensemble system for intrusion detectiond. *15th International Database Engineering Applications Sympo-sium 185-191*.
english

Panda, M. et K. Patra (2009a). Mining association rules for constructing a network intrusion detection model. *International journal of applied engineering research 4*, 381–98.
english

Panda, M. et K. Patra (2009b). Semi Naïve Bayesian method for anomaly based network intrusion detection. *Proc. of ICONIP, Lecture Notes in Computer Science*.

Sélégny, G. (2013). Shodan, un moteur de recherche alimente des peurs justifiées. Technical report, Centre National de ressource et d'information sur l'Intelligence économique et stratégique (http ://www.portail-ie.fr/article/834/Shodan-un-moteur-de-recherche-alimente-des-peurs-justifiees).
english

Siraj, M. et al. (2009). A hybrid intelligent approach for automated alert clustering and filtering in intrusion alert analysis. *Journal of computer theory and engineering*.

Sumeet, D. et D. Xian (2011). *Data Mining and Machine Learning in Cybersecurity*.
english

Tavallaee, M. et al. (2010). Towards credible evaluation of anomaly based intrusion detection methods. *IEEE Transaction on System, Man and Cybernetics, Part-c, Applications and Reviews 5*.

Ujjaneni, S. et S. V. Achutha (2013). A novel cell reckoning intrusion against tor. *International of Computer Trends and Technology IJCTT 4*.
english

Wang, G. et al. (2010). A new approach to intrusion detection using ANN and fuzzy clustering. *Expert systems with application journal*.
english

Zhang, Q. et F. Feng (2009). W. Network intrusion detection by support vectors and ant colony. *Proc. of 2009 Intl. workshop on information security and applications IWISA*.

Summary

The democratization of the Internet , coupled with the effect of globalization, resulting interconnect individuals, states and businesses. the side unpleasant this global interconnection of information systems is in a phenomenon called " Cybercrime ." Individuals, groups badly meaning therefore aim to undermine the integrity of Information Systems in a financial goal or to serve a "cause" . We offer flow analysis method to detect abnormal behavior and dangerous to understand the risks in an understandable way by all players.

cet article est en double,
voir page 163

Expansion Sémantique de requêtes basée sur la similarité Cosinus ou les Modèles de Langue

Btihal El Ghali*, Abderrahim El Qadi**

*LRIT Unité associé au CNRST - URAC n°29 Faculté des Sciences
Université Mohammed Rabat, Maroc
btihal.elghali@gmail.com
**Equipe TIM, EST- Université Moulay Ismaïl
Meknès, Maroc
elqadi_a@yahoo.com

Résumé. L'expansion de requêtes est une approche permettant d'améliorer la performance du système de recherche d'information (RI) sur le web. Elle consiste à suggérer à l'utilisateur des termes extraits à partir de documents pertinents de la requête initiale. Dans cet article, nous proposons une méthode d'expansion de requêtes basée sur l'Analyse Sémantique Latente (ASL), et le contexte autour la requête utilisateur. Nous avons utilisé deux modèles de recommandation : le premier à base de Similarité Cosinus (SC) et le deuxième en utilisant les Modèles de Langue (ML) pour l'extraction du contexte de la requête. Et pour améliorer la précision du système, nous avons enrichi le modèle de langue par des informations contextuelles supplémentaires en se basant sur WordNet. Les résultats d'expérimentations sur la collection CISI Smart, montrent une amélioration de l'efficace du système de RI par 48,1% en utilisant le modèle contextuel basé sur SC et 19,2% en utilisant celui basé sur ML.

1 Introduction

Entre la requête de l'utilisateur et les documents contenus dans le web, se trouve un écart dû au fait que les utilisateurs ne forment pas leurs requêtes en utilisant les mêmes termes que ceux utilisés dans les documents qui répondent à leur besoin en information. Ajoutant à cette problématique le fait que, tant que les requêtes deviennent plus longues, tant que la possibilité que des termes importants co-apparaissent dans la requête et ses documents pertinents augmente (Xu et Croft, 2000). Cependant, une étude a été faite sur les requêtes soumises sur un moteur de recherche par Wen et al. (2001) et il a été observé que le plus souvent les utilisateurs soumettent des requêtes très courtes. La longueur moyenne des requêtes sur le Web est de deux mots. Par ailleurs, ces requêtes la plupart du temps contiennent des termes ambigus. Ainsi, récupérer des documents pertinents en utilisant la requête initiale soumise par l'utilisateur est une tâche presque impossible selon Baziz (2005), en raison de l'augmentation continue du volume des bases d'information.

Dans le but de minimiser l'écart entre la requête initiale de l'utilisateur et son besoin en information, on propose dans cet article deux modèles contextuels de recommandation de requêtes en se basant sur les termes et les documents partagés entre les requêtes. Le premier modèle proposé est un algorithme qui se base sur le calcul de similarité par l'expression connue de similarité Cosinus (Algorithme de Recommandation de requête: ARQ). Le deuxième est basé sur les modèles de langue en calculant le score de recommandation en utilisant la divergence de Kullback-Leibler (KL-Divergence) (Imran and sharan, 2010). Et en se

basant sur le contexte extrait par l'un des modèles de recommandation, on propose d'appliquer la méthode ASL pour l'expansion de requête des utilisateurs.

Cependant, les Modèles de Langue (Zhai, 2008) sont exploités traditionnellement dans le domaine de la Recherche d'Information dans le but de représenter la relation de pertinence entre un document et une requête (Bouchard et Nie, 2006), en estimant la probabilité de génération de la requête par le modèle de langue du document. En revanche, la méthode d'analyses sémantiques latentes (ASL) (Manning et al., 2009) donne en résultat, une matrice qui relie les documents à leurs termes, pour pouvoir calculer la similarité entre deux termes ou deux documents.

Les expérimentations ont été effectuées sur la base de données CISI de la collection de tests SMART. Des expérimentations intensives ont été menées sur chaque modèle pour sélectionner les valeurs appropriées à utiliser en tant que nombre de documents, nombre de termes d'expansion et nombre de requêtes proposées à utiliser pour élargir les nouvelles requêtes des utilisateurs.

Ce papier est structuré comme suite: La section 2 dresse l'état l'art du sujet traité. La section 3, décrit les deux modèles de recommandation de requêtes proposés. La méthode d'expansion de requêtes proposée est présentée en section 4. Section 5, montre les résultats d'expérimentations les plus importants, tandis que la section 6 donne les principales conclusions et introduit nos travaux futurs.

2 Travaux connexes

Pour combler l'écart entre la requête initiale de l'utilisateur et son besoin en information, de nombreuses méthodes ont été proposées. Les méthodes les plus communément utilisées sont les techniques d'expansion et de suggestion de requêtes. Dans le cas de l'expansion, les termes utilisés pour l'expansion peuvent être sélectionnés à partir de ressources externes ou à partir du corpus lui-même. Parmi les méthodes utilisées pour sélectionner les termes du corpus, nous citons l'analyse globale, où la liste des termes candidats d'expansion est générée à partir de l'ensemble de la collection, tandis que d'autres sont basées sur une analyse locale (Lin et Huang, 2006) utilisant les techniques de retour de pertinence (Gupta et al., 2013), les termes d'expansion sont choisis parmi les termes des documents les mieux classés jugées pertinent par l'utilisateur.

Les méthodes d'analyses globales sont très coûteuses en calcul, et leur efficacité n'est pas généralement meilleure et parfois pire que les méthodes d'analyses locales. Le problème de l'analyse locale, est que l'utilisateur doit intervenir pour fournir leur jugement de pertinence concernant les documents les mieux classés.

L'implication de l'utilisateur rend difficile le développement des méthodes automatiques pour l'expansion de requête. Pour éviter ce problème une pseudo-approche de retour de pertinence est préférée, où les documents sont récupérés à l'aide d'une fonction d'appariement efficace et les documents les mieux classés sont supposés être pertinents (Saint-Réquier et al., 2010) (Cui et al., 2002).

Cependant, ces méthodes d'expansion sont limitées dans l'extraction des termes d'expansion à partir d'un ensemble de documents, et n'utilisent pas d'informations concernant les interactions entre les utilisateurs et le système; tel est le cas de l'expansion basée sur les fichiers logs (Bai et al., 2007) (Cui et al., 2002). Les fichiers archives du moteur de recherche

représentent une mine d'informations, donnant une idée à propos de l'interaction entre les utilisateurs et le système de recherche d'information.

A titre d'exemple, l'approche proposée par Fonseca et al. (2005) est une méthode de recommandation de concepts pour étendre la requête originale de l'utilisateur avec un contexte supplémentaire en générant des concepts à partir d'un journal de requête et en appliquant une méthode d'expansion basée sur ces concepts. Ces approches sont simples, intuitives et efficaces selon les expérimentations faites. Mais, elles n'utilisent les requêtes d'utilisateurs passées, pour l'expansion de la nouvelle requête, et ne réduisent pas l'écart entre les termes des requêtes et les termes des documents.

Dans les dernières décennies, la notion du contexte a été introduite, elle inclue à la fois le contexte de l'utilisateur (ses domaines d'intérêt, ses préférences et son historique de recherche) et le contexte autour de la requête, qui signifie l'environnement de la requête (ses documents pertinents, ses termes, son domaine, ...). Le premier contexte nécessite une recherche basée sur des profils utilisateurs, où un seul profil peut regrouper une grande variété de domaines et de préférences, qui ne sont pas toujours pertinents pour une requête particulière (Bai et al., 2007). La création de plusieurs profils est aussi une solution possible à ce problème selon Liu et al. (2002), un profil pour chaque domaine d'intérêt. Ensuite, pour chaque nouvelle requête, un seul domaine est identifié. Ainsi, la solution d'utiliser le second contexte comme un contexte approprié permet d'améliorer la précision de la requête.

3 Méthodes de recommandation de requêtes

La recommandation de requêtes est le fait de suggérer à l'utilisateur des requêtes similaires le plus possible à sa requête initiale. C'est une façon d'améliorer la performance de la recherche, et cela est dû au fait qu'elle résout des problèmes important dans le domaine de Recherche d'Information. Dans cette section, on présente deux modèles contextuels de recommandation de requêtes en se basant sur les termes et les documents partagés entre les requêtes.

3.1 Algorithme de Recommandation de requêtes (ARR)

L'algorithme de Recommandation de requêtes (ARR) que l'on présente dans cette section permet de trouver le groupe approprié des requêtes associées à la nouvelle requête de l'utilisateur et les classifie en fonction de leur pertinence à celle-ci.

Notre contribution dans ce travail, se présente en tant que la modification radicale de l'algorithme présenté dans (Zahera et al. 2011). On a éliminé les étapes 1 et 2 de l'algorithme, qui concerne la segmentation (Clustering) des requêtes passées et l'identification du cluster approprié de la nouvelle requête quand elle est soumise. On a également changé la fonction de pondération et la mesure de similarité basé sur plusieurs comparaison des expressions les plus connus et les plus utilisées dans la littérature. Ajoutant à cela la suppression de la 3ème étape qui concerne le calcul du support de la requête, car on estime qu'elle ne donne aucune information à propos de la relation entre deux requêtes.

L'algorithme de recommandation de requête que l'on propose en résultat se compose des quatre étapes suivantes :

1. Construction d'un vecteur de documents $Q_j^D = \{D_1^{(j)}, D_2^{(j)}, \ldots, D_n^{(j)}\}$ et d'un vecteur de termes $Q_j^T = \{T_1^{(j)}, T_2^{(j)}, \ldots, T_m^{(j)}\}$ pour chaque requête Q_j, où $D_i^{(j)}$ représente le

poids du ***i*ème** document dans le vecteur requête, et $T_i^{(j)}$ représente le poids du ***i*ème** terme dans le même vecteur requête. Ces poids sont définis par la mesure de pondération classique Ltc :

$$D_i^{(j)} = \frac{\log(tf_i^{(j)}+1)\times idf_i^{(j)}}{\sqrt{\sum_{k=1}^{n}[\log(tf_k^{(j)}+1)\times idf_k^{(j)}]}} \quad (1)$$

Avec : $idf_i^{(j)} = log\ (N/n_i)$.
Où $tf_i^{(j)}$ représente la fréquence du ***i*ème** document dans le vecteur Q_j^D, N est le nombre total de requêtes dans la collection et n_i est le nombre de requêtes pour lesquels le ***i*ème** document est pertinent.
On calcule chaque $T_i^{(j)}$ en utilisant la même expression.

2. Mesure des similarités entre la nouvelle requête Q_n de l'utilisateur et toutes les requêtes passées Q_p contenues dans les fichiers logs, en utilisant les deux représentations de chaque requête (à base de documents et à base de termes), par la mesure de similarité Cosinus :

$$Sim(Q_n, Q_p) = \cos(\vec{q}_n, \vec{q}_p) = \frac{\vec{q}_n \times \vec{q}_p}{|\vec{q}_n| \times |\vec{q}_p|} \quad (2)$$

3. Calcule du score de recommandation (Rank) de chaque requête passée par rapport à la nouvelle requête. L'expression de score utilisée est comme suite :

$$Rank(Q_n, Q_p) = \gamma Sim_T(Q_n, Q_p) + (1-\gamma) Sim_D(Q_n, Q_p) \quad (3)$$

Avec Sim_D est la similarité en utilisant la représentation à base de documents, Sim_T la similarité en utilisant la représentation à base de termes et la constante $\gamma \in [0,1]$ est un paramètre utilisé pour la normalisation.

4. Finalement, on classifie les requêtes passées par rapport à la nouvelle requête, en se basant sur les valeurs de scores calculées pendant la 3ème étape.

3.2 Recommandation de requêtes par Modèles de Langues (RRML)

Sachant qu'en Recherche d'Information traditionnellement, les modèles de langues étaient utilisés pour ordonner les documents d'une collection selon leur capacité à générer la requête de l'utilisateur (L'Hadj, 2009). Donc, la pertinence du document pour une requête est liée au fait que le modèle de langue (ML) du document peut générer le ML de la requête. On propose de représenter les relations de recommandation entre deux requêtes en utilisant les modèles de langue. On ordonne les requêtes passées extraites des fichiers logs selon leur capacité de génération de la nouvelle requête utilisateur.

On propose d'utiliser la fonction de score typique (Bai et al., 2007) définie par KL-divergence dans le cadre où les modèles de langue sont utilisées comme une fonction de classement. Dont l'expression est la suivante (Asfari et al., 2010):

$$Score_{ML}(Q_n, Q_p) = \sum_{t\epsilon V} P(t|\theta_{Q_n}) \log(P(t|\theta_{Q_p}) \approx -KL(\theta_{Q_n}||\theta_{Q_p}) \quad (4)$$

Avec θ_{Q_n} le modèle de langue de la nouvelle requête, θ_{Q_p} le modèle de langue d'une requête passée, et **V** le vocabulaire de termes.
$P(t|\theta_Q)$ représente la probabilité d'un terme **t** dans le modèle de langue de la requête et elle est calculée selon l'Estimation de Maximum de Vraisemblance (**EMV**), comme suit :

$$P(t|\theta_Q) = \frac{f(t)}{\sum_{t_i \epsilon Q} f(t_i)} \quad (5)$$

Où ***f(t)*** est la fréquence de t dans la requête.

Le problème principale qui se pose pour les modèles de langue, revient au fait que la taille d'un corpus d'apprentissage, malgré sa grandeur, ne peut atteindre la taille d'une langue. Cela cause une « sous-représentassions des données », car les mots absents du corpus d'apprentissage sont estimés par une probabilité nulle. Par conséquent, une probabilité nulle est affectée à toute séquence de mots contenant ce mot.

Pour résoudre le problème de « sous-représentassions des données », les chercheurs se sont orientés vers ce que l'on appelle : le Lissage. Le Lissage revient à affecter une probabilité non nulle aux mots absents du corpus d'apprentissage, et ce en redistribuant la masse des probabilités observées.

D'après (Cao et al., 2005), le choix de la technique de lissage dépend de l'environnement d'expérimentation. L'une des méthodes de lissage couramment utilisées dans la recherche d'information est le lissage par interpolation connu sous le nom « Jelinek-Mercer Smoothing» (JM) :

$$P\left(t\middle|\theta'_{Q_p}\right) = (1-\lambda)P\left(t\middle|\theta_{Q_p}\right) + \lambda P(t|\theta_C) \quad (6)$$

Où $\boldsymbol{\lambda}$ est un paramètre d'interpolation et θ_C le modèle de langue de la collection de requêtes extraites des fichiers logs du moteur de recherche.

Nous utilisons le lissage pour les requêtes passées seulement. Le modèle de la nouvelle requête θ_{Q_n} est estimé par l'EMV sans lissage.

On calcule le ***Rank*** de chaque requête passée par rapport à la requête d'entrée en utilisant les deux représentations de chaque requête. Le score de classement de la requête $\boldsymbol{Q_p}$ pour la nouvelle requête $\boldsymbol{Q_n}$ est mesuré en utilisant l'expression :

$$Rank_{ML}(Q_n, Q_p) = \gamma Score_{ML\_T}(Q_n, Q_p) + (1-\gamma) Score_{ML\_D}(Q_n, Q_p) \quad (7)$$

Avec $\gamma \in [0,1]$ est un paramètre qu'on utilise pour la normalisation du score. Le $\boldsymbol{Score_{MLT}}$ de modèle de langue pour la recommandation en utilisant les vecteurs qui représentent la présence ou l'absence d'un terme dans la requête. Tandis que, le $\boldsymbol{Score_{MLD}}$ est calculé en utilisant les vecteurs qui représentent la présence ou non d'un document parmi les documents cliqué d'une requête.

4 Analyse Sémantique Latente pour l'expansion de requêtes

L'analyse sémantique latente (en anglais : Latent Semantic Analyses : LSA), connu également en tant que l'indexation sémantique latente (ISL), est une méthode qui tente de surmonter les problèmes de correspondance lexicale par la récupération des informations sur la base d'une signification conceptuelle au lieu de mots individuels pour la recherche. ASL suppose qu'il existe une structure sous-jacente à l'usage des mots qui est masquée par la variabilité dans le choix des mots (Manning et al., 2009).

Appliqué sur un ensemble de documents et une requête à étendre, la méthode ASL construit d'abord une matrice terme-document pondérée $A_{(t\times d)}$, avec t le nombre de termes et d le

nombre de documents en plus d'une colonne représentant le vecteur requête. Une décomposition en valeurs singulières (DVS) tronquée est utilisée pour estimer la structure dans l'usage des mots dans les documents par décomposition de la matrice en trois matrices $T_{t\times n}$, $S_{n\times n}$ et $D_{d\times n}$:

$$\mathbf{A_{(t\times d)} = T_{t\times n}\, S_{n\times n}\, D'_{n\times d}} \quad (8)$$

Où n = min (t, d) est le nombre de dimension auquel la matrice A est décomposée, nommé également le rang de A, et D' est la transposé de D.

Ensuite, les matrices résultantes sont tronquées en réduisant le rang n à un espace de dimension inférieur k, tout en minimisant la distance entre les deux matrices A et A' tels que mesurée par la 2-norme :

$$\Delta = \ ||A - A'||_2 \quad (9)$$

Avec A' est la matrice tronquée: $A'_{(t\times d)} = T_{t\times k}\, S_{k\times k}\, D'_{k\times d}$

La DVS tronquée, capte la majorité de la structure sous-jacente importante dans l'association des termes, et en même temps supprime le bruit ou la variabilité dans l'usage des mots. Par exemple, les termes qui surviennent dans des requêtes ou des documents similaires seront près l'un de l'autre dans l'espace de dimension k même s'ils ne coexistent pas dans les mêmes documents. En fait, certains termes qui ne coïncident jamais avec les termes de la nouvelle requête, peuvent être semblables à eux dans le k-espace.

Dans ce travail, on propose d'appliquer la méthode ASL en utilisant:

- La nouvelle requête à étendre ;
- Les documents les mieux classés pour elle ;
- Les requêtes recommandé les plus similaires à elle ;
- Et les documents les mieux classés pour chaque requête recommandée.

Ensuite, les vecteurs des termes de la nouvelle requête peuvent être comparés à tous les vecteurs des termes candidats d'expansion, et ils peuvent être classés par leur similitude par rapport à chaque terme de la requête à élargir à l'aide de la mesure commune de similarité cosinus, dont l'expression est comme suite :

$$Simc(\vec{t_i}, \vec{t_j}) = \cos(\vec{t_i}, \vec{t_j}) = \frac{\vec{t_i} \times \vec{t_j}}{||\vec{t_i}|| \times ||\vec{t_j}||} \quad (10)$$

En combinant les similarités de chaque terme candidat d'expansion t_j pour tous les termes de la nouvelle requête t_i^Q, nous pouvons calculer le poids de cohésion d'un terme candidat, qui représente la relation (corrélation) entre ce terme et la requête à étendre en entier.

Le poids de cohésion du terme t_j pour la requête utilisateur Q est mesuré par l'expression (Cui et al., 2002):

$$CoWeight(Q, w_j^{(d)}) = \ln\left(\prod_{t_i^Q \in Q}\left(P(t_j | t_i^Q) + 1\right)\right) \quad (11)$$

Cette méthode renvoie une liste de termes pondérés. Les termes les mieux classés peuvent être sélectionnés comme termes d'expansion de la nouvelle requête de l'utilisateur.

5 Expérimentations

Comme collection de test, nous avons utilisé la base de données CISI de la collection standard SMART. Cette collection offre 111 requêtes, 1460 documents et une matrice représentant la pertinence ou non-pertinence de chaque document par rapport à chaque requête.

Nous avons utilisé uniquement les requêtes courtes (contenant moins de cinq termes), et pour évaluer la performance du système, nous avons utilisé la précision moyenne non-interpolée (UAP). Au court de nos expérimentations, nous avons cherché des documents pertinents jusqu'au 20ème document récupéré, et on s'est limité seulement aux deux premières requêtes recommandées, sauf dans le cas de la dernière expérimentation qui concerne une variation du nombre de requêtes recommandées (Figure 1).

Concernant la méthode ASL, nous avons utilisé le logiciel R et le package proposé par Wild (2005) pour construire la matrice DVS tronquée avec le nombre approprié de dimensions k de sorte qu'elle puisse capturer la majorité de la structure importante latente dans l'association entre les termes, toute en supprimant le bruit (section 3).

Afin de vérifier les performances des méthodes proposées ci-dessus, nous avons comparé les meilleurs résultats des deux modèles de recommandation (ARR et RRML) après plusieurs expérimentations où nous avons variés le nombre de documents utilisés et le nombre de termes d'expansion. Nous avons comparé aussi l'UAP des requêtes initiales (avant expansion) avec les meilleures valeurs de chaque cas d'expansion en utilisant le thésaurus WordNet uniquement, ASL basée sur RRML et ASL basée sur ARR. On a testé également le cas d'expansion par la méthode ASL basée sur RRML en cherchant les requêtes recommandées par rapport à des requêtes déjà étendues par les synonymes extraits de WordNet. Le Tableau 1 présente ces résultats.

Méthode d'expansion	---	WordNet	Analyse Sémantique Latente		
Modèle de Recommandation	---	---	RRML basée sur WordNet	RRML	ARR
UAP	0,52	0,52	0,62	0,53	0,77
Nombre de RR	---	---	2	2	2
Nombre de documents	---	---	5	9	9
Nombre de termes d'expansion	---	Tous les synonymes termes de la requête	2	4 ou 5	4 ou 5

TAB. 1 – *Comparaison des meilleures valeurs de chaque méthode d'expansion avec ses conditions appropriées.*

Le tableau 1 montre que l'approche ASL basée sur le modèle ARR donne la valeur la plus élevée de l'UAP, et cela en améliorant l'UAP de 48,1% par rapport à la requête initiale de l'utilisateur. En second lieu, nous remarquons que l'expansion basée sur ASL, et RRML en utilisant des requêtes déjà étendus avec des synonymes extraits de WordNet a un UAP amélioré de 19,2%.

Nous remarquons qu'en comparant l'expansion utilisant le modèle RRML appliqué directement sur les requêtes initiales avec l'expansion utilisant le modèle RRML sur les requêtes étendue en utilisant WordNet; nous avons trouvé une différence de 17% de la valeur de

l'UAP. Dans le but de prouver cette conclusion, on a comparé ces cas en utilisant les mêmes conditions (5 documents, 5 termes d'expansion et 2 requêtes recommandées) et les résultats sont donnés sur le tableau 2.

Modèle de recommandation	RRML basée sur WordNet	RRML
UAP	0,56	0,45

TAB. 2 – *Comparaison de la méthode d'expansion, utilisant directement le modèle RRML et le modèle RRML basée sur WordNet sur la base des mêmes conditions: 5 documents, 5 termes d'expansion et 2 requêtes recommandées.*

Le tableau 2 montre que lorsqu'on utilise les mêmes conditions d'expérimentations, on a également une valeur plus élevée pour le cas du modèle RRML basée sur une première expansion de la requête en utilisant WordNet.

Ainsi, on estime que peut être le modèle ARR également pourra donner de meilleurs résultats s'il est appliqué sur des requêtes déjà étendues initialement par WordNet. Les résultats trouvés en testant cette hypothèse sont présentés dans la figure 1.

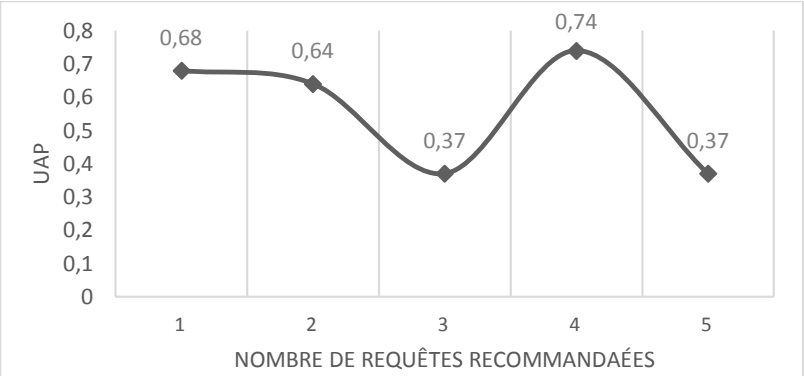


FIG. 1 – *Variation du nombre de requêtes recommandées (RR) utilisées pour l'expansion de requêtes courtes utilisant le modèle de recommandation ARR basée sur des requêtes étendues initialement par le thésaurus WordNet.*

Sur la figure 1, nous remarquons que lors de l'utilisation de deux RR la valeur de l'UAP diminue de 20,3% par rapport au cas utilisant le modèle ARR sans WordNet. Tandis que la valeur la plus élevée de l'UAP dans le cas d'utilisation des synonymes WordNet est de 0,74 avec quatre RR qui est également inferieure à la valeur donnée sur le tableau 1 (0,77 utilisant ARR). Ainsi, nous concluons qu'il n'est pas approprié de fusionner l'expansion par les synonymes WordNet avec le modèle de recommandation statistique ARR, même s'il améliore nettement les résultats de recommandation en utilisant le modèle de recommandation de requêtes par Modèles de Langue (RRML).

6 Conclusion

Dans cet article, nous avons proposé une méthode d'expansion de requêtes basée sur la l'Analyses sémantiques latente (ASL), et sur le contexte autour de la requête. Ce contexte est extrait à partir des requêtes historiques, par le biais de deux modèles de recommandation de requêtes qu'on a proposé. On a réalisé nos expérimentations en utilisant les requêtes courtes de la base de données textuelle CISI de la collection standard de test SMART.

Les résultats montrent que les meilleures valeurs de la précision moyenne non-interpolée sont donnés par l'approche d'expansion basée sur ASL et l'Algorithme de Recommandation de Requête (ARR). On remarque également que l'ajout de l'étape d'expansion de la requête avec WordNet, au modèle RRML améliore la performance du système par 16,98%, tandis que la performance diminue de 20,3% pour le modèle ARR. Nous concluons que pour les requêtes courtes, si nous utilisons ASL comme une méthode d'expansion, on extrait le contexte le plus approprié à l'aide de l'algorithme de recommandation de requête (ARR), sans faire une expansion initiale des termes de la requête en utilisant les synonymes extraites de WordNet.

Nous proposons comme futur travaux, d'améliorer notre approche et de l'appliquer sur d'autres collection de test, comme un journal de requête extrait d'un moteur de recherche réel.

Références

Asfari, O.,Doan, B-L., Bourda, Y., Sansonnet, J-P. (2010). *Context-based Hybrid Method for User Query Expansion*. In Proceedings of the fourth international conference on Advances in Semantic Processing (SEMAPRO), Italy, Florence, 69-74.

Bai, J., Nie, J-Y. Bouchard, H., Cao, G. (2007). *Using query contexts in information retrieval*. SIGIR '07 Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. New York, USA, 15-22.

Baziz, M. (2005). *Indexation conceptuelle guide par ontologie pour la recherche d'information*. PhD thesis, Institut de Recherche en Informatique de Toulouse, Université Paul Sabatier de Toulouse, December.

Bouchard, H., Nie, J.Y. (2006). *Modèles de langue appliqués à la recherche d'information contextuelle*. In CORIA'06. Lyon France, 213-224.

Cao, G., Nie, J., Bai, J. (2005). *Integrating Word Relationships into Language Models*. In Proceedings of SIGIR'05, Salvador, Brazil, August.

Cui, H., Wen, J.R., Nie, J.Y., Ma, W.Y. (2002). *Probabilistic Query Expansion Using Query Logs*. WWW2002. Honolulu Hawaii USA, May 7-11.

Fonseca, B.M., Golgher, P., Pôssas, B., Ribeiro-Neto, B., Ziviani, N. (2005). *Concept-based interactive query expansion*. In Proceedings of the 14th ACM International Conference on Information and Knowledge Management - CIKM'05. New York USA, 696-703.

Gupta, Y., Saini, A., Saxena, A.K. (2013). *A Review on Important Aspects of Information Retrieval*. International Journal of Computer, Control, Quantum and Information Engineering, Vol. 7, No. 12:990-998.

Imran, Ha., Sharan, A. (2010). *Selecting Effective Expansion Terms for Better Information Retrieval*. International Journal of Computer Science & Applications (IJCSA), Vol. 7, No. 2:52-64.

Lin, S.M., Huang, C.M. (2006). *Personalized Optimal Search in Local Query Expansion*. In Proceedings of the 18th Conference on Computational Linguistics and Speech Processing, Hsinchu, Taiwan; September, 221-236.

Liu, F., Yu, C., Meng, W. (2002). *Personalized web search by mapping user queries to categories*. In CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management. November, 558-565.

Manning CD, Raghavan P, Schütze H. (2009). *An Introduction to Information Retrieval - Chapitre 18: Matrix decompositions and latent semantic indexing*. Online edition (c) Cambridge University Press; April, 403-419.

SaidL'Hadj L. (2009). *Recherche Conceptuelle d'Information, Modèle d'Indexation Mixte : Concepts-mots.* Mémoire de Magister en Informatique. Ecole nationale Supérieure d'Informatique (ESI), Algérie.

Saint-Réquier, A., Dupont, G., Adam, S., Lecourtier, Y. (2010). *Évaluation d'outils de reformulation interactive de requêtes*. In Proc. CORIA, 223-238.

Wen, J., Nie, J., and Zhang, H. (2001). *Clustering User Queries of a Search Engine*. In Proceedings of WWW10, Hong Kong, May.

Wild, F. (2005). *An Open Source LSA Package for R*. CRAN. November.

Xu, J., Croft, W.B. (2000). *Improving the effectiveness of information retrieval with local context analysis.* ACM Transactions on Information Systems (TOIS), Vol. 18, No. 1. January, 79-11.

Zahera, H.M., El Hady, G.F., Abd El-Wahed, W.F. (2011). *Query Recommendation for Improving Search Engine Results.* International Journal of Information Retrieval Research, Vol. 1, Issue 1, January, 45-52.

Zhai, C. (2008). *Statistical Language Models for Information Retrieval: A Critical Review*. Foundations and Trends in Information Retrieval, Vol. 2, No. 3:137-215.

Summary

Query expansion is an approach which improve the performance of the Information Retrieval System on the web. It consist on the fact of suggesting to the user, terms that were extracted from the relevant documents of the initial query. In this paper, we are proposing a query expansion method based on the Latent Semantic Analyses (LSA) and the context around the user query. We had use two recommendation models: the first one based on the Cosine Similarity (CS) and the second using the Language Models (LM) in order to extract the query context. In order to improve the system's precision, we enriched the LM model

with supplementary contextual information based on WordNet. The experimentations results on the Collection CISI Smart, shows an improvement of 48,1% of the efficiency using the SC model and 19,2% using the ML model.

Geographical Taxonomy Construction using Association Rules

Omar El Midaoui*, Abderrahim El Qadi**

*LRIT Associated Unit to the CNRST - URAC n°29
Faculty of Science Mohammed V-Agdal University
Rabat, Morocco
omarelmidaoui@gmail.com
**TIM Team, High School of Technology
Moulay Ismaïl University Meknes, Morocco
elqadi_a@yahoo.com

Summary. Due to the specificities of geographical queries, they need a special process of treatment by Information Retrieval systems (IRS). In this paper, we propose an approach which aim is to build a geographical taxonomy of adjacency automatically that can be uses for reformulating the spatial part of a geographical query. Our approach exploit the best-ranked documents retrieved when submitting the spatial entities, which are composed of the spatial relation and a noun of a city. Based on that we construct a database of transactions, considering each document retrieved as a transaction of the nouns of the cities sharing the same country of the query's city. A country's taxonomy is formed by combining many rules that we are extracting using an Association Rules techniqueWe conclude from our tests, that the proposed approach is an efficient method to interpret and improve the results of geographical queries of adjacency.

1 Introduction

Web users searching for information that are spatially located often require procedures, which are geographically specific. However, retrieval systems currently have limited support to operationalize a user's geospatial queries. Geographic information deals with physical objects that are linked with spatial relationships which are in some cases hard to express with words and that contain most of the time ambiguous terms.

Most of the current search engines handle queries by adopting a keywords matching approach without inferring the geographical scope of the spatial terms. Thus, when the name of a place is typed into a typical search engine associated with a spatial preposition (e.g. "near"), web pages that include that name and this spatial preposition in the text will be retrieved but most likely, not places that are nearby that specified place.

Most human activities are well located in the geographical area. Thus, it is not surprising that most of the documents on the Web contain spatial references. These arguments prove the

fact that it will be very useful for search engines to take into account the spatial scope of geographical queries.

In this paper, we propose an automatic method of geographical taxonomy construction using the Apriori Algorithm. The main objective of this approach is to use the resulted taxonomies in reformulating geographical queries by interpreting the spatial relationships that they contain.

This approach is tested using a collection that we created, containing 10 queries and the documents used for the taxonomy's creation are retrieved using the Google web services whenever there is a need of returned documents.

This article is organized as follows: The section 2 shows some related works. Section 3 introduces our proposed approach for constructing a geographical taxonomy of adjacency. Experimental results are presented in section 4. Finally, we draws conclusions and future works.

2 Related work

In order to do a spatial analysis of text, the first step is the annotation of spatial named entities. Several techniques has proved their ability for carrying out this annotation, such as works of Rocío and Erick (2010) and Loustau (2008) that has elaborate it using external resources named "gazetteers". A gazetteer is a dictionary or geographic directory whose inputs are names of places. Each entry in the dictionary may be associated with information as: belonging to one or more administrative structures (town, region, country, etc.), the physical characteristic (mountain, river, road, etc.), statistical data, a geometric representation expressed in a geographic referential.

Other works proposes the categorization of these spatial named entities after identification. Such as, Bouamor (2009) that exploits document structure: for example, the collaborative encyclopedia "Wikipedia". The identification of named entities is done using the title and their categorization is based on the analysis of the first sentence of the description part or the category part at end of the article. Buscaldi and Rosso (2008) has also proposed a technique for spatial named entities categorization using the thesaurus GeoWordNet.

In the other hand, some approaches aims more particularly to the disambiguation of recognized places names (Buscaldi, 2009). The ambiguity can be understood as a word or a phrase that has more than one meaning according to Vargas et al. (2012). In this case, two type of ambiguities are to treat (Einat et al., 2004): a geo/non-geo ambiguity when the entity can have a non-geographical meaning such as the term "Turkey", and a geo/geo ambiguity that occurs when the named entity refers to two different places as Rabat in Malta and Rabat in Morocco.

A hybrid approach is proposes by Gaio et al. (2012) which, first landmark names of places but also searches for these terms in ontological resources to identify related terms, potentially geographic. Domain-Specific taxonomies are also playing an important role in many applications for improving search results (Xueqing et al., 2012) or help with query reformulation (Sadikov et al., 2010).

In this study, we introduce an automatic approach for building a geographical taxonomy of adjacency. In this aim, we exploit the best-ranked documents returned by the search engine when submitting the spatial part of the query, that contain the spatial relation of adjacency and a noun of a city for which we are constructing the taxonomy.

3 Geographical taxonomy construction

A taxonomy consists of a number of names arranged in a hierarchical system and describing a specific domain (Enghoff, 2009) by a simple structure. Starting from a general concept of a specific domain, we associate to it the terms that describe it more precisely every time we move down in the hierarchy.

The proposed approach is based on the association rules. The spatial query model used in this work (Sallaberry et al., 2007) supports Absolute or Relative Spatial Entities (ASE or RSE). The spatial named entities as the city of "Paris" are well-known named entities and are defined as an ASE (Absolute Spatial Entity). While complex spatial entities as "near Paris" are defined as an RSE (Relative Spatial Entities).

3.1 Association rules technique

One of the most popular algorithms of association rules is Apriori Algorithm. This method is used to extract frequent item sets from databases and to define the association rules for discovering the knowledge. The benefit of these rules is detecting unknown relationships between objects.

The association rules discovery is divided into two phases (Al-Maolegi and Arkok, 2014): Detection of frequent item sets and generation of association rules. In this aim, two measures of evaluation and two threshold are defined: The support of an item set, the confidence of a rule and a threshold of each measure minsup and minconf.

$$Support(A) = P(A) = \frac{f(A)}{N} \quad (1)$$

$$Confidence(A \rightarrow B) = P(A|B) = \frac{Support(A \cup B)}{Support(A)} \quad (2)$$

With A and B are an item sets, $f(A)$ is the frequency of A in the database and N is the number of transactions in the database.

In the first phase, every set of k items is called a k-item set. If they occurred together with a support that is greater than or equal to the minimum support threshold (minsup), this k-item set is called k-frequent item set. The k-frequent item sets are used to generate the (k+1)-item sets.

In order to find frequent item sets, the algorithm scan the database to count the frequency of each 1-item set, which contains only one item. The 1-frequent item sets are used to find the 2-item sets, whose frequent item sets are also used to find the 3-item sets and so on until there are no more k-item sets. If an item set is not frequent (its support is lower than the minsup), any larger subset built from it is also non-frequent, this condition minimize the time of extraction of k-frequent item sets.

In the second phase, many rules can be generated from one frequent item set. The number of rules possible to be extracted from a k-frequent item set is 2^k-2. For example, if we consider the 3-frequent item set {I1, I2, I3}, the six generated rules are $\{I_1\} \rightarrow \{I_2,I_3\}$, $\{I_2\} \rightarrow \{I_1,I_3\}$, $\{I_3\} \rightarrow \{I_1,I_2\}$, $\{I_1,I_2\} \rightarrow \{I_3\}$, $\{I_1,I_3\} \rightarrow \{I_2\}$, $\{I_2,I_3\} \rightarrow \{I_1\}$.

To validate the rule X→Y, where X and Y are item sets, its confidence is to be compared with the minconf. If $Confidence(X \rightarrow Y) \geq minconf$ the rule is validated, otherwise it is ignored.

3.2 Geographical taxonomy of adjacency

Considering a database whose transactions are documents and items are the cities of the country that contain the city of the user query. We propose to build a geographical taxonomy of adjacency using Association rules. For example, if the user query's spatial part is "near Paris", the items of the database are cities of France, the transactions are top-documents extracted from the search engine when submitting the RSE "near Paris", and the root of the taxonomy is the initial ASE "Paris".

Our method do not consider all the k-item sets, but only item sets which contain the initial ASE. The rules generated are also restrained to rules whose root is the initial ASE (ASE_0). Resulted rules have the form "$ASE_0 \rightarrow$ {Item set of ASEs}", where the item set of ASEs contains one ASE or more. Thus, we ignore the comparison of rules using their confidence because from every k-frequent item set we are searching for a specific rule which root is the ASE_0. These conditions reduce the time consumed for processing the algorithm.

The fusion of the generated rules forms a one level taxonomy whose root is the ASE_0 (figure 1).

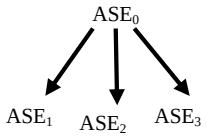


FIG. 1 – *A one level taxonomy for the ASE_0.*

Validation step. In this contribution, we propose also a step of validation of each arc of the taxonomy. To validate each arc of this taxonomy we submit the ASE resulted from this arc with the same spatial relation. Then, we apply Apriori algorithm to the top-documents returned in the same way to construct a one level taxonomy for the ASE to validate. For example, for the ASE_1 if we have ASE_0 in its taxonomy then the arc is kept and the taxonomy evolves to a two level taxonomy as shown in figure 2. Otherwise, if the arc of ASE_3 is not validated it is removed from the taxonomy of ASE_0.

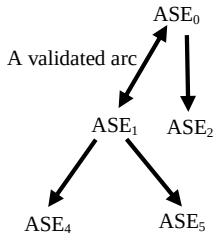


FIG. 2 – *A two level taxonomy for the ASE_0.*

The same process is applied to every new ASE that appear in the taxonomy in order to construct a geographical taxonomy of adjacency for a country, or just a specific taxonomy for a city for reformulating a user query by interpreting its RSE.

4 Experimentation results

To apply the proposed approach, we have used a lexicon of spatial relationships, and a database of ASEs associated with their countries.

To test and verify the performance of the approach presented in this paper, we propose to take our country Morocco as an example. Thus, to be able to exploit the web pages achieved by Moroccans themselves we realize our experimentations in French. Morocco's capital the ASE "Rabat" is considered as the root of the taxonomy and the construction process is beginning with it. The search engine used in our experimentations is the Google web service.

We apply our method to the five first web pages retrieved when submitting the spatial relationship and the ASE for which we are looking for the adjacent ASEs.

In order to extract the useful text for our method, we put the following conditions:

- Do not include hypertext links, which sends to another page, because most of the time they are noise (ads or proposals).
- Disable research based on search history.
- Use the Google search services without using a specific profile.

As a pretreatment step, we eliminate accents from the documents used to minimize the non-detection of ASEs, because the people who wrote the contents of websites do not write cities names in the same way, and the problem arise particularly in the case of nouns, which contain accents.

Then, we first varied the spatial relation used in spatial entity submitted to test if the spatial relation used influence the performance of our approach. The spatial relations used are presented in table 1.

Annotation	Expression
SR 1	à côté de
SR 2	à la périphérie de
SR 3	à proximité de
SR 4	aux alentours de
SR 5	aux environs de
SR 6	les environs de
SR 7	près de

TAB. 1 – *Spatial Relations.*

First, the five top-ranked documents are extracted for Rabat, associated with every spatial relationship. A database (DB) containing five transaction is constructed based on these documents. The Apriori algorithm is applied to this DB and then the association rules are generated between Rabat and every Moroccan ASE contained in the DB. Then, we varied the minsup from 0,2 to 0.8 (Table 2) without the validation step for the rules extracted using 2-frequent item sets. Later we computed the error rate and the number of rules generated in every case.

	Error rate				Number of rules			
RS\Minsup	**0,2**	**0,4**	**0,6**	**0,8**	**0,2**	**0,4**	**0,6**	**0,8**
RS 1	72,73	28,57	33,33	0	22	7	3	1
RS 2	42	0	0	0	5	2	1	1
RS 3	25	0	0	-	4	1	1	0
RS 4	40	33,33	0	0	10	3	1	1
RS 5	33,33	0	0	0	9	2	2	1
RS 6	40	50	0	0	10	4	2	2
RS 7	0	0	0	-	6	6	1	0

TAB. 2 – *The error rate and the number of rules generated while varying the minimum support threshold and the spatial relation used for the 2-frequent item sets containing the ASE "Rabat".*

From Table 2, we notice that using the minsup=0,8 the algorithm do not return any results in some cases otherwise it gives 1 or 2 answers. The same for minsup=0,6 that do not exceed 2 correct answers.

Regarding the value 0,2 it generally gives a high error rate and sometimes returns a very high number of response up to 22 resulting ASEs in the case of RS 1 with 6 correct adjacent ASEs. Thus, we favored the value of minimum support equal to 0,4 because it is the one that gives the best ratio between a minimal errors and an acceptable number of answers.

During our tests, we stopped searching for frequent item sets in k=4, because we noticed that the 3-frequent item sets and the 4-frequent item sets are not giving interesting results. In table 3 we represent the rate error and the number of rules generated by varying the number of items k in resulted frequent item sets from 2 to 4, based on the minsup=0,4.

	Error rate			Number of rules		
RS \ k	**2**	**3**	**4**	**2**	**3**	**4**
RS 1	28,57	55,55	100	7	9	3
RS 2	0	0	-	2	1	0
RS 3	0	-	-	1	0	0
RS 4	33,33	50	-	3	2	0
RS 5	0	0	-	2	1	0
RS 6	50	83,33	100	4	6	4
RS 7	0	0	0	6	6	4

TAB. 3 – *The error rate and the number of rules generated while varying the number of items in a frequent item set and the spatial relation used for item sets containing the ASE "Rabat".*

Using the minimum support threshold 0,4 we compared the results returned using 2-frequent item sets, 3-frequent item sets and 4-frequent item sets.

Table 3 shows that for k=4, most of the spatial relations do not give any result, otherwise it gives 100% of errors. Except in the case of the SR 7 for which the 2-frequent item set returns only correct rules. Thus, all the ASEs used for the following (k+1)-Item sets are correct.

We notice also that the error rate for k=3 are equal to or greater than the error rates of rules resulted using 2-frequent item sets. Therefore, the case to use for taxonomy construction is the case, which binds pair of ASEs together by association rules (k=2).

The next step of experimentation is done in order to compare the cases where we use or not the validation step for constructing the geographical taxonomy of adjacency.

	Error rate			Number of correct rules		
RS	**Without validation**	**Using validation**	**Average support**	**Without validation**	**Using validation**	**Average support**
RS 1	28,57	0	0	5	2	2
RS 2	0	0	0	2	1	1
RS 3	0	-	0	1	0	1
RS 4	33,33	50	33,33	2	1	2
RS 5	0	0	0	2	2	2
RS 6	50	33,33	33,33	2	2	2
RS 7	0	0	0	6	5	6

TAB. 4 – *The error rate and the number of correct rules generated using the validation step or not and using the average of support between the two cases, while varying the spatial relation used for 2-frequent item sets containing the ASE "Rabat".*

Comparing the results using validation with the results without validation, we note that the error rate decreases when using the validation step, with the exception of the RS 4 for which from 3 results including 2 correct ASEs, the validation has eliminated one of the correct ASEs and kept the erroneous one. Concerning the RS 3 we notice that the only ASE that was resulted without validation was eliminated with the step of validation. In general, we conclude that the validation step reduces errors sufficiently.

To minimize the error rate while keeping as much as possible of correct results (eliminate only the erroneous ASEs by the validation step). We propose to compute the average of the two supports of the opposite rules (e.g. $ASE_1 \rightarrow ASE_2$ and $ASE_2 \rightarrow ASE_1$). Table 4 shows that the result given by the case of the average support solves the problems mentioned above for the RS 3 and RS 4.

Comparing the seven spatial relations, we promote the RS 7 "près de" which gives the most interesting result with 0% error and six correct ASEs as child nodes of Rabat's taxonomy of adjacency.

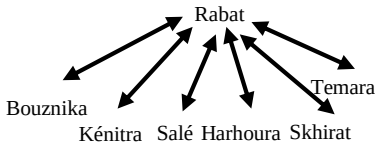


FIG. 3 – *A one level taxonomy for Rabat using the spatial relation "Près de".*

Using the favorable conditions represented above we continue the construction of Morocco's taxonomy (figure 4) with 0,4 as a minsup, 2-frequent item sets, and using the average of support for validating links.

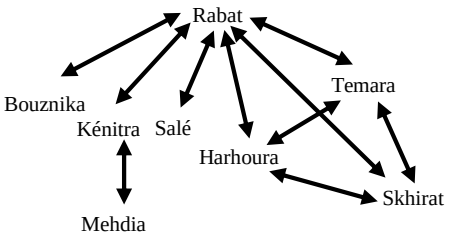


FIG. 4 – *Morocco's geographical taxonomy of adjacency.*

Morocco's geographical taxonomy of adjacency has jammed at this level. The figure 4 shows that it contain 10 correct links but it can not be developed more with the conditions that we had specified. Thus, for a more convenient method, we propose to build a 1-level taxonomy for each case of query reformulation, instead of building a global country's taxonomy and storing it in order to use it every time.

In order to evaluate the performance of our approach more precisely, we proposed 10 geographical queries of adjacency, and we have compared the Un-interpolated Average Precision of the two cases before and after reformulation of these queries. All the geographical queries used (Table 5) are from the field of the Real Estate, so we can test our taxonomy in a specific domain first before generalizing it.

Id. query	**Expression**
Q1	Appartement à vendre près de skhirat
Q2	Terrain à vendre à proximité de Rabat
Q3	Appartements à louer aux environs de Temara
Q4	Louer chalet aux alentours de Harhoura
Q5	Vente terrain à la périphérie de Marrakech
Q6	Bureau à louer dans les environs de Nouaceur
Q7	Appartement à vendre aux alentours de Kénitra
Q8	Acheter villa à proximité de Berrechid
Q9	Terrain à vendre près de Khemisset
Q10	Appartement à vendre à proximité de Casablanca

TAB. 5 – *Geographical queries of adjacency used.*

We first built Morocco's geographical taxonomy shown in figure 4, and a one-level taxonomy of each city absent in it (Marrakech, Nouaceur, Berrechid and Casablanca). Then, we submitted the queries, before reformulation (as shown in table 5) and after reformulation using the built taxonomies, to the Google search engine and we compared the 10 first web pages resulted using the Un-interpolated Average Precision (UAP) measure (Figure 5).

The judgment of relevance is done manually by verifying if the resulted web page is a good response to the query for the Thematic Entity (TE) and the Spatial Entity (SE) at the same

time. For example if the query's TE is "appartement à vendre" and the web page is proposing a Villa, a land or a real estate for rent not for sale, the web page is considered irrelevant. In the same way, if the SE is "à la périphérie de Marrakech", a retrieved web page proposing a real estate in Marrakech is considered as irrelevant.

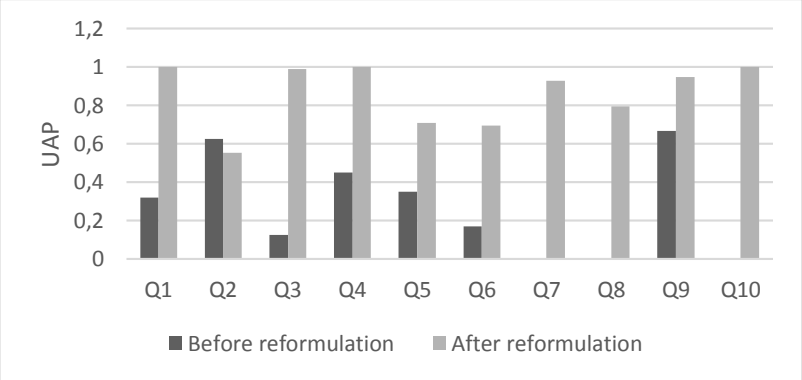


FIG. 5 – *The UAP values before and after reformulation based on the taxonomies built using the Apriori algorithm.*

We notice from figure 5 that the UAP of nine geographical queries improved significantly and reached the value 1 for the queries Q1, Q4 and Q10. The value 1 do not mean that all the 10 first web pages are relevant, but that the relevant ones are on the top of the returned web pages. While the second query shows a slight decrease in the UAP value. Knowing that, all the queries had been reformulated by correct links using real adjacent ASEs.

5 Conclusion and Future Works

In this work, we have presented an approach for building a geographical taxonomy of adjacency using association rules. We have tested our approach on 10 geographical queries, which thematic entities are from the field of the real estate that we reformulated based on the spatial taxonomies of adjacency. The results show that the reformulation based on our proposed approach has improved the value of UAP significantly for most of the queries used. We conclude from our tests, that the proposed approach is an efficient method to interpret and improve the results of geographical queries of adjacency.

As future work, we propose to improve the method developed in this work and to take into account others types of spatial relationships. Concerning the experimentations, we intend to apply this approach using a large collection of test, and to apply it to multiple countries to confirm the results presented in this paper.

Références

Al-Maolegi, M., B. Arkok (2014). An improved Apriori algorithm for association rules. International Journal on Natural Language Computing, Vol 3, No. 1, February : 21-29.

Bouamor, H. (2009). Extraction des connaissances à partir du Web pour la recherche des images géoréférencées. CORIA : 519-526.

Buscaldi, D. (2009). Toponym ambiguity in geographical information retrieval. Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, SIGIR '09, ACM, New York USA : 847-847.

Buscaldi, D., P. Rosso (2008). Using GeoWordNet for Geographical Information Retrieval. CLEF : 863-866.

Einat, A., N. Har'el, R. Sivan, A. Soffer (2004). Web-a-where: Geotagging Web Content. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, Sheffield United Kingdom : 273—280.

Enghoff, H. (2009). What is taxonomy? – An overview with myriapodological examples. SOIL ORGANISMS Volume 81 (3) : 441–451.

Gaio, M., V.-T. Nguyen, C. Sallaberry (2012). Typage de noms toponymiques à des fins d'indexation géographique. Revue Traitement Automatique des Langues, Vol. 53, n° 2 : 143-176.

Loustau, P. (2008). Interprétation automatique d'itinéraires dans des recits de voyage. PhD thesis, Université de Pau et des Pays de l'Adour.

Rocío, A.-M., L.-O. Erick (2010). Geo information extraction and processing from travel narratives. Transforming the Nature of Communication, 14th International Conference on Electronic, Helsinki Finland : 363-373.

Sadikov, E., J. Madhavan, L. Wang, A.-Y. Halevy (2010). Clustering query refinements by user intent. In WWW : 841–850.

Sallaberry, C., M. Baziz, J. Lesbegueries, M. Gaio (2007). Une approche d'extraction et de recherche d'information spatiale dans les documents textuels – évaluation. In Proceeding of Conférence en Recherche d'Information et Applications (CORIA), Saint-Etienne France : 53-64.

Vargas, R.N.P., M.-F. Moura, E.-A. Speranza, E. Rodriguez, S.-O. Rezende (2012). Discovering the Spatial coverage of the documents through the SpatialCIM Methodology. AGILE'2012 International Conference on Geographic Information Science, Avignon, April 24-27 : 181-186.

Xueqing, L., S. Yangqiu, L. Shixia, W. Haixun (2012) Automatic taxonomy construction from keywords. KDD'12 : 1433-1441.

Résumé

Les requêtes géographiques ont besoin d'un traitement spécial par le Système de Recherche d'Information due à leurs spécificités. Dans ce papier, on propose une approche dont l'objectif est de construire une taxonomie géographique d'adjacence automatiquement, qui pourra par la suite être utilisée dans la reformulation de la partie spatiale de la requête géographique. Notre approche exploite les documents les mieux classés résultants de la soumission de l'entité spatiale, qui est composée de la relation spatial et le nom de la ville. En se basant sur ces documents on construit une base de données de transactions, considérant chaque document retournées comme une transaction des villes qui partagent le pays de la ville de la requête. La taxonomie d'un pays est formée en combinant les règles extraites en utilisant une technique de Règles d'Association. A partir de nos tests, on conclut que l'approche proposé est une méthode efficace pour interpréter et développer les résultats des requêtes géographique d'adjacence.

Géo-sémantique analytique des trajectoires

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

*, ** Faculté des Sciences et Techniques de Mohammedia (FSTM),
Université Hassan II – Casablanca
*lkarim.lkarim@gmail.com
**azedine.boulmakoul@gmail.com
*** Université Joseph FourierGrenoble, France
ahmed.Lbath@ujf-grenoble.fr

RÉSUMÉ
Pour aider les services basés sur la localisation à répondre aux demandes du marché, nous proposons l'architecture générale d'un système scalable et performant pour gérer les trajectoires des objets mobiles. Nous détaillons également les prototypes des composants logiciels développés. En effet, nous évaluons la scalabilité et la performance de chaque composant proposé pour collecter, traiter, stocker, entreposer et analyser les trajectoires des objets mobiles. Le module de collecte permet la collecte des différents types de données géographiques à partir des appareils de positionnement et ce en utilisant les sockets en mode asynchrone avec pool d'objet. Ensuite, nous présenterons le module de stockage et d'entreposage dans une base de données NoSQL, adaptée au volume des données des trajectoires, de type MongoDB. Enfin, pour avoir un système d'analyse scalable et accessible dans le cloud, nous offrons un module d'analyse des trajectoires dans le système hadoop et nous évaluons le système proposé à travers l'étude de cas «Système de suivi des chemins pris par les clients dans les espaces commerciaux».

1. Introduction

Les différents domaines d'applications et de recherches ont besoin de collecter, de représenter et d'explorer les connaissances des trajectoires, tels que le domaine de transport et de la logistique, de la gestion, du marketing et des affaires, de la criminologie, de la surveillance des territoires, et de la gestion de flotte, etc. Satisfaire les attentes des utilisateurs des services de localisation en termes de vitesse de temps de réponse, performance et évolutivité est la clé du succès des entreprises. En effet, la production croissante des données spatio-temporelles engendre de très gros volumes de données disponibles et intéressantes à analyser. Ces volumes de données à très grandes échelles nécessitent des moyens de collecte, de traitement, de stockage (Agrawal et al., 2011), d'analyse et de visualisation appropriés. Les bases de données traditionnelles, support des entrepôts de données, ne sont plus adaptées pour gérer et traiter ces grandes masses de données.

Nous proposons dans le présent article une architecture globale, les composants logiciels et les technologies utilisées dans le système scalable conçu pour la collecte, le

traitement, le stockage, l'entreposage, l'analyse et la visualisation des différents types des trajectoires.

2. Concepts de base des trajectoires

Une trajectoire est une description de l'évolution au fil du temps, du mouvement physique des objets en mouvement. On cite dans ce qui suit les présentations de base des trajectoires :
(a) Trajectoire Raw ou brute est l'enregistrement des positions d'un objet dans un domaine spécifique de l'espace et du temps. Pour un objet en mouvement et un intervalle de temps donné, elle est présentée comme une séquence de lieu géométrique dans le système spatial 2D (xi, yi, ti). (b) Trajectoire structurée (Spaccapietra et al., 2008) est définie comme une trajectoire brute structurée en segments correspondant à des étapes significatives dans la trace de la trajectoire (voyages). (c)Trajectoire sémantique (Spaccapietra et al., 2008) possède une sémantique liée au domaine des applications, elle utilise les quatre composants (arrêt, déplacer, début et fin). Arrêter (S), déplacer (M), début (B) et fin (E) ne sont plus des positions spatio-temporelles, mais plutôt des objets sémantiques liés à la connaissance géographique générale et aux données géographiques de l'application. (d) Autre approche décrit les schémas de déplacement dans des contextes à la fois spatiales et temporelles basés sur la notion de région d'intérêt (Giannotti et al., 2007) en définissant la notion de voisinage spatial et de la tolérance temporelle. (e) Yu et al. 2007 ont étendu le concept du chemin espace-temps pour représenter à la fois les activités physiques (marche, conduite, etc.) et virtuelles (envoi de courrier électronique, appel téléphonique). Comme chaque activité possède un emplacement géographique et un intervalle de temps, le chemin spatio-temporel a été profilé en tant que conteneur de toutes les activités se produisant par un objet en mouvement.

3. Système proposé

Dans la littérature, il existe différentes présentations des trajectoires, chacune d'elles modélise la trajectoire d'une facette. Dans (Boulmakoul et al., 2012), nous avons présenté un méta modèle unifié pour modéliser tous les types de trajectoires des objets en mouvement, tenant compte de l'aspect structural et géo-sémantique. Le modèle « activité » a été intégré dans cette modélisation, et permet ainsi de capturer l'activité au sens sociogéographique.
Le système spatial mobile, présenté dans ce travail, permet de stocker, tracer les objets mobiles et de répondre aux requêtes spatio-temporelles destinées aux trajectoires.
Les recherches dans l'entreposage des trajectoires sont encore à un stade précoce. Des travaux préliminaires ont été trouvés pour modéliser et maintenir un entrepôt de données des trajectoires, définir un cube de données simple consistant en des dimensions spatiale et temporelle, et des mesures concernant les trajectoires numériques (Marketos et al., 2008). Cependant, les travaux existants des entrepôts des trajectoires ne prennent pas en compte, de façon explicite, les contraintes du mouvement, comme les mouvements dans un réseau routier, ni les différentes facettes des trajectoires (brutes, structurées, sémantiques, basées sur les régions d'intérêt, de plus, ces modèles rendent le système non convenable pour entreposer les données volumineuses des trajectoires.
L'analyse des trajectoires facilite le processus de prise de décision dans différents domaines d'applications. En outre, les données collectées en temps presque réel doivent être stockées dans un entrepôt de données avec une faible sinon aucune latence dans certains domaines

d'applications. Pour ce faire, nous entreposons les différents types de trajectoires pour faciliter la prise de décision instantanée en répondant à une série de questions complexes dans différents domaines (ex. Trouver instantanément, les régions où se trouvent des goulots d'étranglement). Pour des contraintes de scalabilité, de tolérance de pannes, de distribution du stockage et de traitement des données des trajectoires, l'entreposage est effectué dans des bases de données NoSQL de type cloud (Boulmakoul et al., 2014; Boulmakoul et al., 2012). Prendre des décisions, à partir des données brutes, collectées des appareils de positionnement sous forme d'une liste de points spatio-temporels en vrac s'avère coûteux et aussi pauvre en terme d'informations pour différents domaines d'application. D'où, il est nécessaire de procéder à la reconstruction de ces trajectoires(Boulmakoul et al., 2014).
Une fois les différents types de trajectoires sont entreposés une autre problématique apparaît. Cette problématique consiste à analyser et extraire des informations spatio-temporelles utiles. Nous appliquons la méthode d'analyse simpliciale et issue de la topologie algébrique, sur l'entrepôt des trajectoires, pour extraire les informations implicites et potentiellement utiles.
L'analyse simpliciale a été développé à l'origine par R. Atkin (Atkin, 1974), comme une approche issue de la topologie algébrique pour étudier les caractéristiques structurales des systèmes sociaux dans lesquels deux ensembles d'indicateurs, fonctionnalités ou caractéristiques sont liées les unes aux autres. L'utilisation des caractéristiques topologiques de la méthode d'analyse simpliciale comme connectivenness, et l'excentricité permettra de répondre à une variété de requêtes comme par exemple, trouver l'ensemble optimal des régions accroches et aussi l'excentrique. Pour extraire davantage d'informations sur l'entrepôt des trajectoires. Les données sont prétraitées et puis entreposées dans un entrepôt NoSQL dans un environnement cloud.

4. Architecture générique et composants réutilisables

L'architecture du système proposé pour notre méta-model unifié des trajectoires des objets mobiles s'appuie sur l'Architecture Orientée Service (SOA) afin d'améliorer les performances et l'interopérabilité des applications (Boulmakoul et al., 2012; Boulmakoul et al., 2013a). En architecturant les composants du système par les services web basés sur les standards publics de l'OGC en les plaçant dans un substrat de messagerie SOA, nous pouvons intégrer les différents services des trajectoires (suivi, visualisation et interrogation de l'entrepôt des trajectoires des objets mobiles) avec d'autres applications et services basés sur la localisation.
L'architecture du système, présenté sur la Figure 1, est constituée de cinq couches principales : la couche de collecte de données, couche de prétraitement, couche de génération des différents types de trajectoires, couche de base de données et entreposage des trajectoires, couche d'analyse et des applications des trajectoires.

1. Couche de collecte de données

Cette couche permet de collecter les données à partir des appareils mobiles, guichets automatiques et les caméras IP à l'aide des protocoles HTTP, FTP et SOAP. Tout dépend de la criticité des données à gérer, les services de collecte de données sont utilisés pour collecter les données en ligne ou hors ligne. L'utilisation des interfaces des sockets visent à rendre possible des applications de collecte en temps presque réel entre les sources de données et le serveur de collecte de données. Au niveau du serveur de collecte de données, nous avons utilisé l'API des sockets .Net en mode asynchrone avec un pool d'objets Sockets

(Boulmakoul et al., 2013; Boulmakoul et al. 2012) pour pouvoir les réutiliser dans plusieurs collectes des données spatio-temporelles et recevoir des notifications en cas d'erreur ou d'opération réussie. Les résultats des tests de performances de notre système de collecte des données spatio-temporelles montrent que la performance et la scalabilité de notre cadre proposé n'a pas changé et le serveur de collecte peut recueillir 600 000 messages de 600 objets en mouvement simultanés. En ce qui concerne l'évolutivité du serveur et de la performance, les essais en chiffres montrent que la variation des ressources système, tels que la mémoire, le disque physique et le processeur, est légère et contrôlée. Tandis que la durée totale consommée pour la collecte et l'enregistrement des messages dans MongoDB varie entre 9 ms lors de la collecte d'un message d'un objet en mouvement, et 67s lors de la collecte et l'enregistrement des messages de 10000 objets mobiles, et environ 4s pour collecter et enregistre 50 000 messages d'un objet en mouvement.

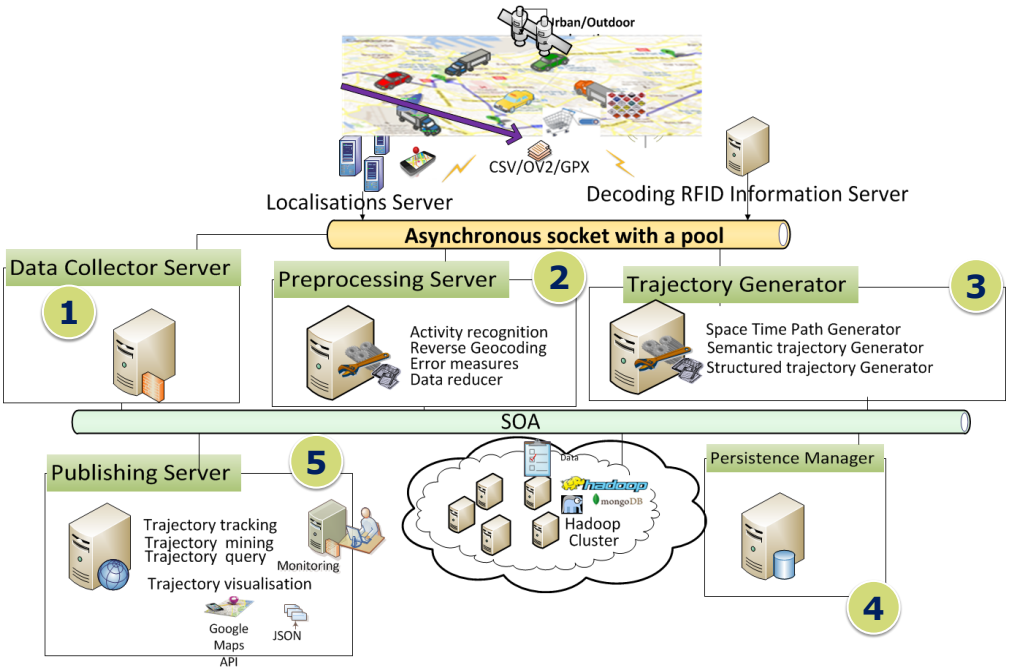


Figure 1: Architecture générale du système.

Par conséquent, notre système de collecte est évolutif et efficace pour différents types d'applications tels que: guides touristiques de la communauté; gestion des flottes, suivi des marchandises, des camions et des taxis; la publicité; protection des biens, des véhicules et antivol; et les études comportementales des êtres humains.

2. **Couche de prétraitement des trajectoires**

Elle emploie les services de réduction de données, de reconnaissance d'activité et du géocodage inverse.

3. **Couche de générateur de trajectoire**

Grâce à cette couche, un ensemble de points spatio-temporels est transformé de trajectoire brute nettoyée en structurée, sémantique, avec région d'intérêt ou chemin espace-temps. La construction des trajectoires structurées nécessite la segmentation de la trajectoire brute en trajectoire structurée, et ce en identifiant les stops (Figure 2).

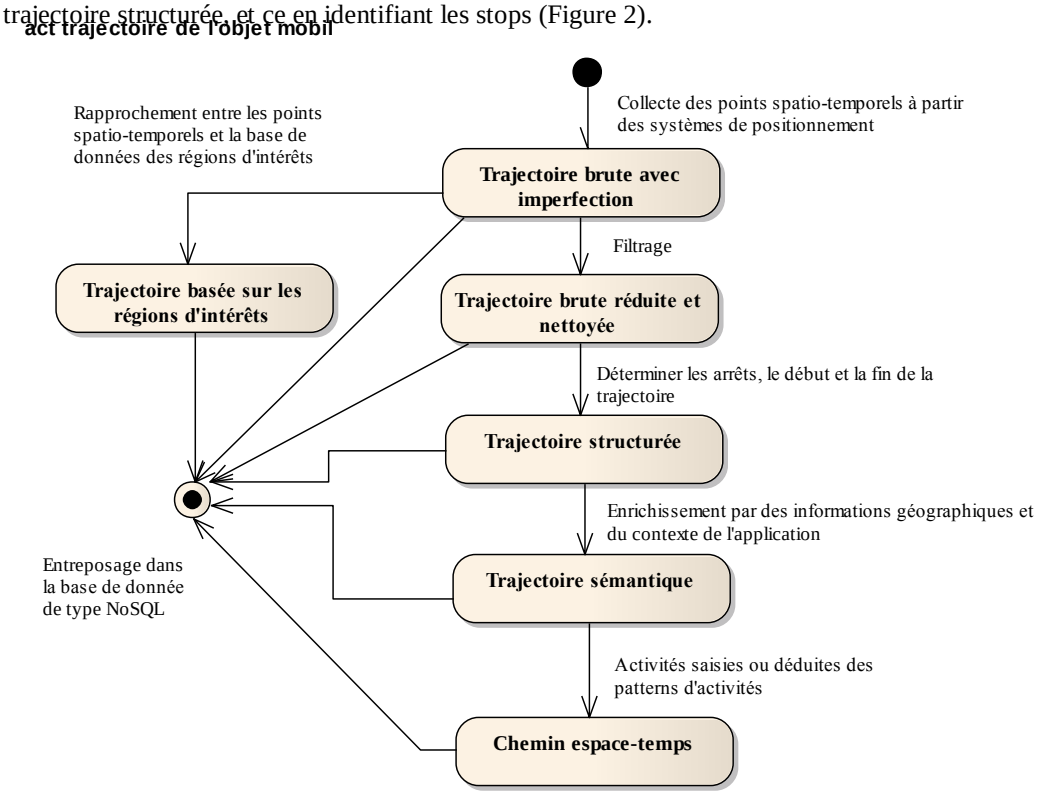


Figure 2: Diagramme d'état de transition de l'objet trajectoire.

Un stop est déterminé par un mouvement avec une vitesse qui tend vers 0 ou inférieure à un seuil paramétré, si l'intervalle temporel du stop dépasse le seuil paramétré, il est considéré comme la fin de la trajectoire, les points spatio-temporels correspondants aux changements de position entre deux stops, entre le début de la trajectoire et le premier stop, ou entre le dernier stop et la fin de trajectoire sont les déplacements (M). La construction des trajectoires sémantiques est faite en enrichissant les quatre composants début, fin et la séquence consécutive et alternée des déplacements et arrêts de la trajectoire structurée par les données de contexte de l'application pour avoir plus de sens lors de l'analyse de l'entrepôt des trajectoires. La construction des trajectoires basées sur les régions d'intérêts est faite en utilisant un rapprochement de la trajectoire de type brute ou structurée avec une table de correspondance contenant les zones géométriques et les noms des régions dans une période donnée. La construction des chemins espace-temps est faite en enrichissant les quatre composants de la trajectoire sémantique par les activités de l'objet mobile.

4. **Couche de stockage et d'analyse dans un environnement hadoop cloud**

Cette couche contient les données géographiques, et les entrepôts des trajectoires. Les trajectoires entreposées sont automatiquement réparties dans des espaces de stockage et des lots de traitement sur un ensemble de cluster grâce à la solution évolutive Hadoop qui embarque une capacité de tolérance aux pannes. En effet, le stockage s'effectue dans les bases de données MongoDB de la catégorie NoSQL, et l'analyse avec Hadoop dans un environnement Cloud.

a. **Stockage des trajectoires**

La scalabilité des bases des données relationnelles est obtenu en utilisant du matériel plus performant et en ajoutant plus de mémoires, par contre une base de données de type NoSQL profite de la montée en puissance en répartissant la charge sur plusieurs systèmes de base. Par conséquent, NoSQL est une base de données peu coûteuse pour s'approprier une base de données de trajectoires. Les deux bases de données sont évaluées sur la même machine 64 bits, en utilisant les dernières versions de PostgreSQL avec des extensions de PostGIS (version 9.2.2) et bases de données MongoDB (version 2.2.3). Le but de cette étude est de tester les performances des bases de données pour l'entreposage des données géo-spatiales lors de l'enregistrement des données brutes générées par notre simulateur de suivi des véhicules. Chaque message envoyé à partir de l'appareil GPS est un objet composé d'une position et plusieurs balises. Dans notre test, nous avons utilisé les fichiers générés par notre simulateur de suivi des véhicules ; la structure d'un message de suivi collecté est : « identifiant de l'objet mobile ; latitude ; longitude ; date ; heure ; observation »

D'un point de vue géo-spatiale, MongoDB permet nativement des points d'indexation géo-spatiales (sans extension), contrairement à POSTGIS qui permet d'utiliser non seulement des points géométriques mais aussi d'autres objets géographiques plus avancés (par exemple des lignes ou polygones). Ce benchmark, montre que MongoDB est beaucoup plus rapide dans l'insertion de données géographiques. Nous expliquons cette performance par le fait que MongoDB supporte les indexes géo-spatiales avec aucun type de données géo-spatiales dédié. Alors que PostGIS utilise deux tables en arrière-plan (Ramsey, 2012), spatial_ref_sys et geometry_columns, pour récupérer et en apprendre davantage sur les types de géométrie disponibles dans une base de données spécifique. En outre, l'architecture de MongoDB supporte l'évolutivité horizontale et la haute disponibilité grâce à l'ensemble des réplications (replication sets). L'auto-fragmentation (Auto-sharding) permet de distribuer la charge sur plusieurs serveurs et garder les données équilibrées sur ces serveurs.

b. **Analyse des trajectoires**

- **Approche d'analyse des trajectoires**

L'analyse de l'entrepôt des trajectoires permet de compléter la compréhension des dimensions spatio-temporelles du réseau social des trajectoires. Notre approche pour analyser l'entrepôt des trajectoires des objets mobiles est basée sur les démarches suivantes (Boulmakoul et al., 2013b) :

- Démarche fondée sur l'analyse structurale exploitant des outils issus de la topologie algébrique.
- Démarche faisant référence à l'analyse des réseaux sociaux.

Les valeurs des mesures obtenues en appliquant les méthodes d'analyse simpliciale et l'analyse des réseaux sociaux sur l'entrepôt des trajectoires permettent de décrire comment un acteur, une région ou une activité est intégré dans le réseau. Par exemple, le degré de centralité entrant (indegree) mesuré pour un acteur, qui pourra être un objet mobile une région ou une activité, quantifie la façon dont cet individu se rapporte avec les autres dans le réseau. A titre d'exemple, en appliquant l'approche proposée sur l'entrepôt des trajectoires dans un intervalle de temps, nous pouvons déduire l'importance d'une région R, d'un objet mobile MO ou d'une activité A dans le réseau, et par la suite aider à résoudre les problèmes de différentes applications tels que les congestions. La centralité des objets mobiles, des régions et des activités dans le réseau social de l'entrepôt des trajectoires peut également changer. Par conséquent, la séquence des valeurs de centralités permettra la visualisation de l'évolution de l'importance des objets mobiles, des régions et des activités au cours de temps. D'autre part, l'application de la méthode d'analyse simpliciale sur l'entrepôt des trajectoires permettra de répondre à plusieurs questions du mining des trajectoires. En effet, la génération des classes d'équivalences permet de trouver les trajectoires similaires des objets mobiles, les régions similaires dans un intervalle de temps et de déduire par la suite le model de trajectoire d'un objet mobile. Exploiter le résultat concernant les trajectoires similaires permettra de trouver des solutions à plusieurs problèmes, par exemple, proposer le covoiturage aux personnes ayant des trajectoires similaires pour éviter les embouteillages et diminuer la pollution. Spatialement et temporellement, les groupes sociaux proches ont tendance à partager l'information et avoir un comportement homogène dans l'espace et le temps. L'identification des tendances intéressantes peuvent fournir des indications importantes dans de nombreux domaines d'application tels que l'écologie et les affaires.

- **Environnement d'analyse des trajectoires**

MongoDB utilise une technique appelée fragmentation (sharding) pour mettre à l'échelle sa performance sur un cluster de serveur. Il s'agit d'un processus permettant de fractionner les données de manière uniforme sur le cluster pour paralléliser l'accès. Le traitement des données avec MongoDB est effectué selon les options suivantes : (i) Traitement dans MongoDB en utilisant Map / Reduce ; (ii) Traitement en utilisant le Framework d'agrégation de MongoDB ; (iii) Traitement externe à MongoDB en utilisant Hadoop et d'autres outils externes.

Les requêtes MongoDB examinent un enregistrement à la fois, ce qui signifie que les requêtes sur des documents multiples doivent être mises en œuvre sur le client ou utilisent le paradigme MapReduce (MR) intégré de MongoDB. Bien que MapReduce de MongoDB puisse être exécuté en parallèle dans chaque fragment, il a deux inconvénients majeurs :

1) Le langage de programmation de Map Reduce de MongoDB est JavaScript, qui est lent et a de bibliothèques analytiques pauvres ;
2) JavaScript utilisé par MongoDB, n'est pas thread-safe, donc seulement un seul programme de MapReduce peut fonctionner à la fois (Monkey, 2013).

En résumé, MongoDB fournit de hautes performances pour le stockage et la récupération à grande échelle et dispose d'une interface de requête robuste permettant des opérations intelligentes. Certes, il fournit des fonctionnalités de traitement mais n'est pas un système de traitement de données.

Traitement des données des trajectoires avec Hadoop

L'analyse des trajectoires nécessite beaucoup de calculs qui exigent une puissance de traitement. Pour répondre aux nouveaux enjeux de traitement de très hautes volumétries de données, l'accent a été mis sur les solutions suivantes :

1. Utilisation des machines de capacités de stockage, de puissance de traitement et de mémoire élevés. Mais avec l'augmentation rapide des données, l'utilisation des machines simples a échoué à l'échelle.
2. Utilisation des systèmes distribués permettant de répartir les taches sur plusieurs machines. Mais, les solutions analytiques des données sont souvent complexe, sujette à l'erreur, ou tout simplement pas assez rapide.

Au cours des 10 dernières années, une solution appelé « Hadoop » émerge. C'est un framework Open Source conçu pour réaliser des traitements sur des volumes de données massifs. Il support les applications destinées au Big Data, en particulier l'analytique, sur le système de fichier distribué HDFS (Hadoop Distributed File System). L'infrastructure de Hadoop applique le principe bien connu des grilles de calcul « MapReduce », consistant à répartir l'exécution d'un traitement sur plusieurs nœuds, ou grappes de serveurs. De cette façon, les données non structurées peuvent faire l'objet d'un traitement analytique distribué et en parallèle pour accélérer le traitement, jusqu'à se rapprocher du temps réel.
Pour l'utilisateur, tout est transparent partant du découpage de la donnée en blocs, leur répartition sur les nœuds qui composent le cluster à l'exécution des tâches (parallélisassions). Dede et al. (Dede et al., 2013) ont comparé l'implémentation native de MapReduce de MongoDB avec mongo-hadoop MapReduce. La configuration du système de test est que MongoDB tourne dans un serveur et utilise 3 nœuds cluster Hadoop. Ils ont trouvé que mongo-hadoop plugin est cinq fois beaucoup plus performant en termes de temps de traitement.

Traitement des données des trajectoires en utilisant Hadoop dans un environnement cloud

Le cloud computing, littéralement « l'informatique dans les nuages », est un concept apparu récemment, mais de plus en plus à la mode dans le secteur de l'informatique. Pour traiter des volumes de données massives, le déploiement de la plateforme Hadoop doit se faire sur

plusieurs nœuds. Il nous faut une plateforme en architecture « distribuée », et donc plusieurs machines qui servirons de DataNode / TaskTracker..
La flexibilité du Cloud apporte une agilité dans la gestion de l'infrastructure et l'affectation des ressources. La couche virtuelle d'abstraction est également centralisée, et se révèle ainsi plus facile à gérer. Kang et al. (Kang et al. 2013) ont comparé l'utilisation des clusters physiques et des machines virtuelles dans le Cloud. Le résultat de l'étude montre que l'exécution de Hadoop dans des machines virtuelles d'un Cloud privée permet d'obtenir plus que 110.76% de performance du serveur physique. Par conséquent, nous avons adopté l'architecture pour l'analyse des trajectoires qui consiste à traiter les données avec Hadoop en utilisant MapReduce du Framework Hadoop dans des machines virtuelles dans le cloud. La communication entre MongoDB et Hadoop est faite en utilisant le connecteur MongoDB-Hadoop-Connector (Boulmakoul et al., 2014).
D'autre part, vu que l'analyse des trajectoires moyennant l'analyse des réseaux sociaux est effectué dans le système R et le paquetage Statenet et les données entreposées sont dans la dimension des big data, nous avons intégré le système R avec l'environnement Hadoop à travers un ensemble de paquets pour le langage R « RHadoop » afin que le système R puisse profiter du paradigme MapReduce et le système de fichier HDFS de l'écosystème Hadoop.

5. **Couche des applications des trajectoires**

Contient des services Web spécialisés, autonomes, auto-descriptifs, pour l'exploitation, le suivi, la visualisation et l'interrogation des trajectoires, ces services peuvent être publiés et invoqués à travers le Web, en utilisant un large éventail de machines connectées au web et les appareils mobiles.

5. Étude de cas « Système de suivi des chemins pris par les clients dans les espaces commerciaux»

Le principe de cette étude de cas se base sur le suivi en temps réel des chemins des clients à l'intérieur d'un espace commercial. Grace à l'entrepôt de données « trajectoire-ticket » crée, nous pouvons comprendre le comportement spatio-financier des clients et répondre à plusieurs requêtes des décideurs : (i) Évaluer l'impact d'ajout des produits/ changement d'emplacement des produits / animations (le taux des personnes attirées par un style..) ; (ii) Trouver le profil d'ambiance adéquat ; (iii) Visualiser les chemins d'une date donnée ; (iv) Trouver la corrélation entre la disposition géographique et le CA généré par les chemins des clients ; (v)Identifier les caisses lentes.
Chaque centre commercial possède une base de données « Ticket de caisse » pour enregistrer les achats des clients avec la date, l'heure et le numéro de caisse de la transaction. Nous identifions la marge de bénéfice générée par un chemin par un rapprochement automatique entre le détail du CA d'une transaction et le chemin sous forme de régions d'intérêt d'un client. L'identification du client est faite sur la base de sa position géographique, la date, l'heure et le numéro de caisse du paiement. Une fois la position géographique d'un client se trouve dans une région de type Caisse dans une date d1 et heure t1, nous récupérons le numéro de caisse N1 correspondant à sa position géographique à partir de la base de données spatiale de l'espace commercial. En parallèle, nous récupérons la transaction de paiement enregistrés dans la caisse N1 dans la date d1 et l'heure t1. A ce moment, nous pouvons

construire l'entrepôt de données recherchés 'trajectoire-ticket' en faisant une consolidation des informations de la base de données des chemins clients présentées avec les régions et les informations sur le détail de paiements se trouvant dans la base de données « Ticket de caisse ». Grâce au système proposé, nous pouvons détecter en temps réel le ou les caisses/caissières responsables de la lenteur des files d'attente et ce en requêtons sur le nombre de clients se trouvant dans la région géographique des files d'attente d'une caisse donnée. Le Tableau 1 présente la matrice d'incidence des trajectoires de 10 clients dans l'espace commercial :

	R1	*R2*	*R3*	*R4*	*R5*	*R6*	*R7*	*R8*	*R9*	*R10*	*R11*	*R12*	*R13*
T1	1	1	1	0	1	0	1	0	0	0	0	1	0
T2	1	0	1	1	0	0	0	0	1	0	0	0	0
T3	1	0	1	1	1	1	0	0	0	1	1	0	0
T4	1	0	1	1	1	1	1	1	1	1	1	1	1
T5	1	1	1	0	1	1	1	0	0	1	1	0	0
T6	1	1	0	0	0	0	1	0	0	0	0	1	0
T7	1	0	1	1	1	1	1	1	1	0	1	1	1
T8	1	1	1	0	1	0	0	0	0	1	1	1	0
T9	1	0	1	1	0	1	1	1	1	1	1	1	1
T10	1	1	1	0	1	1	1	0	0	1	1	1	0

Tableau 1: Matrice d'incidence des trajectoires des clients dans le centre commercial.

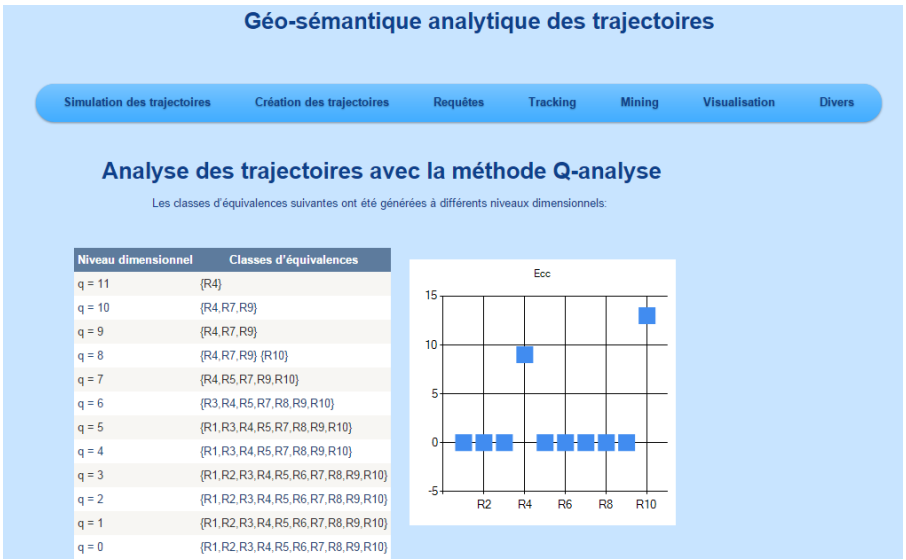


Figure 3: L'excentricité des régions de l'espace commercial.

Le résultat de l'analyse simpliciale (figure 3) montre que les régions 4 et 10 ont des excentricités maximales c.-à-d. les clients ne les visitent pas. Par contre la région 8 ne génère pas un chiffre d'affaire même si elle est connecté avec une excentricité égale à zéro et donc visité par les clients. À partir de ce résultat nous concluons à travers cette analyse la cause réelle du problème de chiffre d'affaire dans le centre commercial. Dans le cas des régions 4 et 10, le problème réside dans l'organisation spatiale des produits par contre pour la région 8, le problème réside dans la nature des produits exposés.

Les résultats de l'analyse des trajectoires des clients de l'espace commercial avec la méthodologie d'analyse des réseaux sociaux avec le système R et le paquetage STATNET. La mesure de betweenness indique si un sommet est partagé dans un grand nombre de trajectoires non redondants. Dans les valeurs suivantes de betweeness des sommets, les valeurs obtenues montrent la présence des sommets qui n'existent pas sur les chemins non redondants entre les sommets. Ainsi, nous concluons que la région 1 est la plus importante des régions et la trajectoire 4 présente la trajectoire la plus importante en termes de nombre des régions visitées.

3. Conclusion

Nous avons présenté l'architecture d'un système scalable conçu pour gérer les objets trajectoires du méta modèle unifiée des trajectoires des objets mobiles. Le système proposé offre des composants de collecte des données spatio-temporelles des dispositifs de géolocalisation en utilisant les sockets. L'utilisation des interfaces des sockets vise à faire une collecte de données en temps presque réel entre les sources de données et le serveur de collecte de données. Les appareils de positionnement des objets en mouvement exécutent un programme de socket pour envoyer les traces spatio-temporelles et le serveur de collecte exécute un programme pour recevoir les données géographiques massives en temps réel à partir de différents programmes des clients et de différents capteurs. En ce qui concerne le stockage et l'entreposage des objets trajectoires, nous utilisons une base de données NoSQL de type document « MongoDB ». Hadoop, HDFS et MapReduce, RHadoop forment l'infrastructure logicielle Big Data, de son architecture en cluster avec des nœuds, à sa capacité de traitement au service de l'analyse des trajectoires. Les trajectoires des objets mobiles sont affichées en utilisant les fichiers json, dans les bibliothèques de visualisation.

Réferences

Agrawal, D., Das, S., & Abbadi, A. (2011). Big data and cloud computing : current state and future opportunities. In 14th International Conference on Extending Database Technology, EDBT/ICDT '11, ACM (pp. 530–533.). New York, USA.

Atkin, R. (1974). Mathematical Structure in Human Affairs. Mathematical Structure in Human Affairs. London, Heinemann.

Boulmakoul, A., Karim, L., & Lbath, A. (2012). Moving Object Trajectories Meta-Model and Spatio-Temporal Queries. International Journal of Database Management Systems (IJDMS), 4(2), 35–54.

Boulmakoul, A., & Karim, L. (2012). A Scalable Data Collector Framework for the Unified Moving Object Trajectories' Meta-Model. In Innovation et Nouvelles Tendances dans les Systèmes d'Information (pp. 19–24).

Boulmakoul, A., Karim, L., Elbouziri, A., & Lbath, A. (2012). A System Architecture for Heterogeneous Moving Objects Trajectory Models Using Different Sensors. In SoSE

in cooperative and competitive distributed decision making for complex dynamic systems. Genova, Italy.
Boulmakoul, A., Karim, L., Elbouziri, A., & Lbath, A. (2013). A System Architecture for Heterogeneous Moving-Object Trajectory Metamodel Using Generic Sensors: Tracking Airport Security Case Study. IEEE System Journal. doi:10.1109
Boulmakoul, A., & Karim, L. (2013a). Dispositif, système et procédé de suivi des chemins pris par les clients dans les espaces commerciaux.
Boulmakoul, A., & Karim, L. (2013b). Space Time Path Data Warehouse Mining based on Simplicial Complex Analysis. In Innovation et Nouvelles Tendances dans les Systèmes d'Information (pp. 19–24).
Boulmakoul, A., Karim, L., & Lbath, A. (2013). A High Performance Scalable Data Collection System for Moving Objects. International Journal of Computer Applications, 36–43. doi:10.5120/11424-6769
Boulmakoul, A., & Karim, L. (2014). Construction et entreposage des trajectoires. In 4ème Edition du Workshop Internationnal sur l'Innovation et Nouvelle Tendances dans les Systèmes d'Information. Rabat, Maroc.
Boulmakoul, A., Karim, L., Laarabi, M. H., Sacile, R., & Garbolino, E. (2014). MongoDB-Hadoop Distributed and Scalable Framework for Spatio-Temporal Hazardous Materials Data Warehousing. In « International Congress on Environmental Modelling and Software (iEMSs 2014). San Diego, California, USA.
Dede, E., Govindaraju, M., Gunter, D., Canon, R. S., & Ramakrishnan, L. (2013). Performance evaluation of a MongoDB and hadoop platform for scientific data analysis. In Science Cloud '13 Proceedings of the 4th ACM workshop on Scientific cloud computing (pp. 13–20.).
Giannotti F, Nanni M, Pedreschi D, Pinellin F (2007). Trajectory Pattern Mining. Int. Conf. Knowl. Discov. Data Min. pp 330–339.
Kang, Y., & Kang, K.-W. (2013). An Empirical Study of Hadoop Application running on Private Cloud Environment. Advanced Science and Technology Letters, 35, 70–73.
Monkey, S. (2013). https://developer.mozilla.org/en/SpiderMonkey.
Ramsey, P. (2012). OpenGeo. Http://www.postgis.fr/chrome/site/docs/workshop-foss4g /doc /geometries.html.
Spaccapietra S, Parent C, Damiani MD, Macedo JA, Porto F, Vangenot C (2008). A Conceptual view on trajectories. Data Knowl Eng 26–146.
Yu H, Shaw S (2007). Revisiting Hägerstrand's time-geographic framework for individual activities in the age of instant access. 103–118.

Summary

To support location-based services to meet the market demands, we propose a general architecture of ascalable and performing system to manage trajectories of moving objects. We also present prototypes

of developed software components. Indeed, we evaluate the scalability and performance of each component proposed to collect, process, store and analyze trajectories of moving objects. The collection module is used to collect different types of geographic data from the positioning devices using asynchronously sockets with object pooling. Then we present the storage module in a NoSQL database, adapted to the volume of trajectories' data. Finally, to have a scalable and accessible analysis system in the cloud, we offer trajectory analysis

module in Hadoop system and evaluate the proposed system through the case study «Customers trajectories analysis in commercial spaces».

Expansion Sémantique de requêtes basée sur la similarité Cosinus ou les Modèles de Langue

Btihal El Ghali*, Abderrahim El Qadi**

*LRIT Unité associé au CNRST - URAC n°29 Faculté des Sciences
Université Mohammed Rabat, Maroc
btihal.elghali@gmail.com
**Equipe TIM, EST- Université Moulay Ismaïl
Meknès, Maroc
elqadi_a@yahoo.com

Résumé. L'expansion de requêtes est une approche permettant d'améliorer la performance du système de recherche d'information (RI) sur le web. Elle consiste à suggérer à l'utilisateur des termes extraits à partir de documents pertinents de la requête initiale. Dans cet article, nous proposons une méthode d'expansion de requêtes basée sur l'Analyse Sémantique Latente (ASL), et le contexte autour la requête utilisateur. Nous avons utilisé deux modèles de recommandation : le premier à base de Similarité Cosinus (SC) et le deuxième en utilisant les Modèles de Langue (ML) pour l'extraction du contexte de la requête. Et pour améliorer la précision du système, nous avons enrichi le modèle de langue par des informations contextuelles supplémentaires en se basant sur WordNet. Les résultats d'expérimentations sur la collection CISI Smart, montrent une amélioration de l'efficace du système de RI par 48,1% en utilisant le modèle contextuel basé sur SC et 19,2% en utilisant celui basé sur ML.

1 Introduction

Entre la requête de l'utilisateur et les documents contenus dans le web, se trouve un écart dû au fait que les utilisateurs ne forment pas leurs requêtes en utilisant les mêmes termes que ceux utilisés dans les documents qui répondent à leur besoin en information. Ajoutant à cette problématique le fait que, tant que les requêtes deviennent plus longues, tant que la possibilité que des termes importants co-apparaissent dans la requête et ses documents pertinents augmente (Xu et Croft, 2000). Cependant, une étude a été faite sur les requêtes soumises sur un moteur de recherche par Wen et al. (2001) et il a été observé que le plus souvent les utilisateurs soumettent des requêtes très courtes. La longueur moyenne des requêtes sur le Web est de deux mots. Par ailleurs, ces requêtes la plupart du temps contiennent des termes ambigus. Ainsi, récupérer des documents pertinents en utilisant la requête initiale soumise par l'utilisateur est une tâche presque impossible selon Baziz (2005), en raison de l'augmentation continue du volume des bases d'information.

Dans le but de minimiser l'écart entre la requête initiale de l'utilisateur et son besoin en information, on propose dans cet article deux modèles contextuels de recommandation de requêtes en se basant sur les termes et les documents partagés entre les requêtes. Le premier modèle proposé est un algorithme qui se base sur le calcul de similarité par l'expression connue de similarité Cosinus (Algorithme de Recommandation de requête: ARQ). Le deuxième est basé sur les modèles de langue en calculant le score de recommandation en utilisant la divergence de Kullback-Leibler (KL-Divergence) (Imran and sharan, 2010). Et en se

basant sur le contexte extrait par l'un des modèles de recommandation, on propose d'appliquer la méthode ASL pour l'expansion de requête des utilisateurs.

Cependant, les Modèles de Langue (Zhai, 2008) sont exploités traditionnellement dans le domaine de la Recherche d'Information dans le but de représenter la relation de pertinence entre un document et une requête (Bouchard et Nie, 2006), en estimant la probabilité de génération de la requête par le modèle de langue du document. En revanche, la méthode d'analyses sémantiques latentes (ASL) (Manning et al., 2009) donne en résultat, une matrice qui relie les documents à leurs termes, pour pouvoir calculer la similarité entre deux termes ou deux documents.

Les expérimentations ont été effectuées sur la base de données CISI de la collection de tests SMART. Des expérimentations intensives ont été menées sur chaque modèle pour sélectionner les valeurs appropriées à utiliser en tant que nombre de documents, nombre de termes d'expansion et nombre de requêtes proposées à utiliser pour élargir les nouvelles requêtes des utilisateurs.

Ce papier est structuré comme suite: La section 2 dresse l'état l'art du sujet traité. La section 3, décrit les deux modèles de recommandation de requêtes proposés. La méthode d'expansion de requêtes proposée est présentée en section 4. Section 5, montre les résultats d'expérimentations les plus importants, tandis que la section 6 donne les principales conclusions et introduit nos travaux futurs.

2 Travaux connexes

Pour combler l'écart entre la requête initiale de l'utilisateur et son besoin en information, de nombreuses méthodes ont été proposées. Les méthodes les plus communément utilisées sont les techniques d'expansion et de suggestion de requêtes. Dans le cas de l'expansion, les termes utilisés pour l'expansion peuvent être sélectionnés à partir de ressources externes ou à partir du corpus lui-même. Parmi les méthodes utilisées pour sélectionner les termes du corpus, nous citons l'analyse globale, où la liste des termes candidats d'expansion est générée à partir de l'ensemble de la collection, tandis que d'autres sont basées sur une analyse locale (Lin et Huang, 2006) utilisant les techniques de retour de pertinence (Gupta et al., 2013), les termes d'expansion sont choisis parmi les termes des documents les mieux classés jugées pertinent par l'utilisateur.

Les méthodes d'analyses globales sont très coûteuses en calcul, et leur efficacité n'est pas généralement meilleure et parfois pire que les méthodes d'analyses locales. Le problème de l'analyse locale, est que l'utilisateur doit intervenir pour fournir leur jugement de pertinence concernant les documents les mieux classés.

L'implication de l'utilisateur rend difficile le développement des méthodes automatiques pour l'expansion de requête. Pour éviter ce problème une pseudo-approche de retour de pertinence est préférée, où les documents sont récupérés à l'aide d'une fonction d'appariement efficace et les documents les mieux classés sont supposés être pertinents (Saint-Réquier et al., 2010) (Cui et al., 2002).

Cependant, ces méthodes d'expansion sont limitées dans l'extraction des termes d'expansion à partir d'un ensemble de documents, et n'utilisent pas d'informations concernant les interactions entre les utilisateurs et le système; tel est le cas de l'expansion basée sur les fichiers logs (Bai et al., 2007) (Cui et al., 2002). Les fichiers archives du moteur de recherche

représentent une mine d'informations, donnant une idée à propos de l'interaction entre les utilisateurs et le système de recherche d'information.

A titre d'exemple, l'approche proposée par Fonseca et al. (2005) est une méthode de recommandation de concepts pour étendre la requête originale de l'utilisateur avec un contexte supplémentaire en générant des concepts à partir d'un journal de requête et en appliquant une méthode d'expansion basée sur ces concepts. Ces approches sont simples, intuitives et efficaces selon les expérimentations faites. Mais, elles n'utilisent les requêtes d'utilisateurs passées, pour l'expansion de la nouvelle requête, et ne réduisent pas l'écart entre les termes des requêtes et les termes des documents.

Dans les dernières décennies, la notion du contexte a été introduite, elle inclue à la fois le contexte de l'utilisateur (ses domaines d'intérêt, ses préférences et son historique de recherche) et le contexte autour de la requête, qui signifie l'environnement de la requête (ses documents pertinents, ses termes, son domaine, ...). Le premier contexte nécessite une recherche basée sur des profils utilisateurs, où un seul profil peut regrouper une grande variété de domaines et de préférences, qui ne sont pas toujours pertinents pour une requête particulière (Bai et al., 2007). La création de plusieurs profils est aussi une solution possible à ce problème selon Liu et al. (2002), un profil pour chaque domaine d'intérêt. Ensuite, pour chaque nouvelle requête, un seul domaine est identifié. Ainsi, la solution d'utiliser le second contexte comme un contexte approprié permet d'améliorer la précision de la requête.

3 Méthodes de recommandation de requêtes

La recommandation de requêtes est le fait de suggérer à l'utilisateur des requêtes similaires le plus possible à sa requête initiale. C'est une façon d'améliorer la performance de la recherche, et cela est dû au fait qu'elle résout des problèmes important dans le domaine de Recherche d'Information. Dans cette section, on présente deux modèles contextuels de recommandation de requêtes en se basant sur les termes et les documents partagés entre les requêtes.

3.1 Algorithme de Recommandation de requêtes (ARR)

L'algorithme de Recommandation de requêtes (ARR) que l'on présente dans cette section permet de trouver le groupe approprié des requêtes associées à la nouvelle requête de l'utilisateur et les classifie en fonction de leur pertinence à celle-ci.

Notre contribution dans ce travail, se présente en tant que la modification radicale de l'algorithme présenté dans (Zahera et al. 2011). On a éliminé les étapes 1 et 2 de l'algorithme, qui concerne la segmentation (Clustering) des requêtes passées et l'identification du cluster approprié de la nouvelle requête quand elle est soumise. On a également changé la fonction de pondération et la mesure de similarité basé sur plusieurs comparaison des expressions les plus connus et les plus utilisées dans la littérature. Ajoutant à cela la suppression de la 3ème étape qui concerne le calcul du support de la requête, car on estime qu'elle ne donne aucune information à propos de la relation entre deux requêtes.

L'algorithme de recommandation de requête que l'on propose en résultat se compose des quatre étapes suivantes :

1. Construction d'un vecteur de documents $\boldsymbol{Q_j^D = \{D_1^{(j)}, D_2^{(j)}, \ldots, D_n^{(j)}\}}$ et d'un vecteur de termes $\boldsymbol{Q_j^T = \{T_1^{(j)}, T_2^{(j)}, \ldots, T_m^{(j)}\}}$ pour chaque requête $\boldsymbol{Q_j}$, où $\boldsymbol{D_i^{(j)}}$ représente le

poids du $i^{ème}$ document dans le vecteur requête, et $T_i^{(j)}$ représente le poids du $i^{ème}$ terme dans le même vecteur requête. Ces poids sont définis par la mesure de pondération classique Ltc :

$$D_i^{(j)} = \frac{\log(tf_i^{(j)}+1)\times idf_i^{(j)}}{\sqrt{\sum_{k=1}^{n}[\log(tf_k^{(j)}+1)\times idf_k^{(j)}]}} \quad (1)$$

Avec : $idf_i^{(j)} = log\ (N/n_i)$.
Où $tf_i^{(j)}$ représente la fréquence du $i^{ème}$ document dans le vecteur Q_j^D, N est le nombre total de requêtes dans la collection et n_i est le nombre de requêtes pour lesquels le $i^{ème}$ document est pertinent.
On calcule chaque $T_i^{(j)}$ en utilisant la même expression.

2. Mesure des similarités entre la nouvelle requête Q_n de l'utilisateur et toutes les requêtes passées Q_p contenues dans les fichiers logs, en utilisant les deux représentations de chaque requête (à base de documents et à base de termes), par la mesure de similarité Cosinus :

$$Sim(Q_n, Q_p) = \cos(\vec{q}_n, \vec{q}_p) = \frac{\vec{q}_n \times \vec{q}_p}{|\vec{q}_n| \times |\vec{q}_p|} \quad (2)$$

3. Calcule du score de recommandation (Rank) de chaque requête passée par rapport à la nouvelle requête. L'expression de score utilisée est comme suite :

$$Rank(Q_n, Q_p) = \gamma Sim_T(Q_n, Q_p) + (1-\gamma) Sim_D(Q_n, Q_p) \quad (3)$$

Avec Sim_D est la similarité en utilisant la représentation à base de documents, Sim_T la similarité en utilisant la représentation à base de termes et la constante $\gamma \in [0,1]$ est un paramètre utilisé pour la normalisation.

4. Finalement, on classifie les requêtes passées par rapport à la nouvelle requête, en se basant sur les valeurs de scores calculées pendant la 3ème étape.

3.2 Recommandation de requêtes par Modèles de Langues (RRML)

Sachant qu'en Recherche d'Information traditionnellement, les modèles de langues étaient utilisés pour ordonner les documents d'une collection selon leur capacité à générer la requête de l'utilisateur (L'Hadj, 2009). Donc, la pertinence du document pour une requête est liée au fait que le modèle de langue (ML) du document peut générer le ML de la requête. On propose de représenter les relations de recommandation entre deux requêtes en utilisant les modèles de langue. On ordonne les requêtes passées extraites des fichiers logs selon leur capacité de génération de la nouvelle requête utilisateur.

On propose d'utiliser la fonction de score typique (Bai et al., 2007) définie par KL-divergence dans le cadre où les modèles de langue sont utilisées comme une fonction de classement. Dont l'expression est la suivante (Asfari et al., 2010):

$$Score_{ML}(Q_n, Q_p) = \sum_{t\epsilon V} P(t|\theta_{Q_n}) \log(P(t|\theta_{Q_p}) \approx -KL(\theta_{Q_n}||\theta_{Q_p}) \quad (4)$$

Avec θ_{Q_n} le modèle de langue de la nouvelle requête, θ_{Q_p} le modèle de langue d'une requête passée, et **V** le vocabulaire de termes.
$P(t|\theta_Q)$ représente la probabilité d'un terme **t** dans le modèle de langue de la requête et elle est calculée selon l'Estimation de Maximum de Vraisemblance (**EMV**), comme suit :

$$P(t|\theta_Q) = \frac{f(t)}{\sum_{t_i \epsilon Q} f(t_i)} \quad (5)$$

Où ***f(t)*** est la fréquence de t dans la requête.

Le problème principale qui se pose pour les modèles de langue, revient au fait que la taille d'un corpus d'apprentissage, malgré sa grandeur, ne peut atteindre la taille d'une langue. Cela cause une « sous-représentassions des données », car les mots absents du corpus d'apprentissage sont estimés par une probabilité nulle. Par conséquent, une probabilité nulle est affectée à toute séquence de mots contenant ce mot.

Pour résoudre le problème de « sous-représentassions des données », les chercheurs se sont orientés vers ce que l'on appelle : le Lissage. Le Lissage revient à affecter une probabilité non nulle aux mots absents du corpus d'apprentissage, et ce en redistribuant la masse des probabilités observées.

D'après (Cao et al., 2005), le choix de la technique de lissage dépend de l'environnement d'expérimentation. L'une des méthodes de lissage couramment utilisées dans la recherche d'information est le lissage par interpolation connu sous le nom « Jelinek-Mercer Smoothing» (JM) :

$$P\left(t\middle|\theta'_{Q_p}\right) = (1-\lambda)P\left(t\middle|\theta_{Q_p}\right) + \lambda P(t|\theta_C) \quad (6)$$

Où $\boldsymbol{\lambda}$ est un paramètre d'interpolation et θ_C le modèle de langue de la collection de requêtes extraites des fichiers logs du moteur de recherche.

Nous utilisons le lissage pour les requêtes passées seulement. Le modèle de la nouvelle requête θ_{Q_n} est estimé par l'EMV sans lissage.

On calcule le ***Rank*** de chaque requête passée par rapport à la requête d'entrée en utilisant les deux représentations de chaque requête. Le score de classement de la requête $\boldsymbol{Q_p}$ pour la nouvelle requête $\boldsymbol{Q_n}$ est mesuré en utilisant l'expression :

$$Rank_{ML}(Q_n, Q_p) = \gamma Score_{ML\_T}(Q_n, Q_p) + (1-\gamma) Score_{ML\_D}(Q_n, Q_p) \quad (7)$$

Avec $\gamma \in [0,1]$ est un paramètre qu'on utilise pour la normalisation du score. Le $\boldsymbol{Score_{MLT}}$ de modèle de langue pour la recommandation en utilisant les vecteurs qui représentent la présence ou l'absence d'un terme dans la requête. Tandis que, le $\boldsymbol{Score_{MLD}}$ est calculé en utilisant les vecteurs qui représentent la présence ou non d'un document parmi les documents cliqué d'une requête.

4 Analyse Sémantique Latente pour l'expansion de requêtes

L'analyse sémantique latente (en anglais : Latent Semantic Analyses : LSA), connu également en tant que l'indexation sémantique latente (ISL), est une méthode qui tente de surmonter les problèmes de correspondance lexicale par la récupération des informations sur la base d'une signification conceptuelle au lieu de mots individuels pour la recherche. ASL suppose qu'il existe une structure sous-jacente à l'usage des mots qui est masquée par la variabilité dans le choix des mots (Manning et al., 2009).

Appliqué sur un ensemble de documents et une requête à étendre, la méthode ASL construit d'abord une matrice terme-document pondérée $A_{(t\times d)}$, avec t le nombre de termes et d le

nombre de documents en plus d'une colonne représentant le vecteur requête. Une décomposition en valeurs singulières (DVS) tronquée est utilisée pour estimer la structure dans l'usage des mots dans les documents par décomposition de la matrice en trois matrices $T_{t\times n}$, $S_{n\times n}$ et $D_{d\times n}$:

$$\mathbf{A_{(t\times d)} = T_{t\times n}\, S_{n\times n}\, D'_{n\times d}} \quad (8)$$

Où n = min (t, d) est le nombre de dimension auquel la matrice A est décomposée, nommé également le rang de A, et D' est la transposé de D.

Ensuite, les matrices résultantes sont tronquées en réduisant le rang n à un espace de dimension inférieur k, tout en minimisant la distance entre les deux matrices A et A' tels que mesurée par la 2-norme :

$$\Delta = \ ||A - A'||_2 \quad (9)$$

Avec A' est la matrice tronquée: $A'_{(t\times d)} = T_{t\times k}\, S_{k\times k}\, D'_{k\times d}$

La DVS tronquée, capte la majorité de la structure sous-jacente importante dans l'association des termes, et en même temps supprime le bruit ou la variabilité dans l'usage des mots. Par exemple, les termes qui surviennent dans des requêtes ou des documents similaires seront près l'un de l'autre dans l'espace de dimension k même s'ils ne coexistent pas dans les mêmes documents. En fait, certains termes qui ne coïncident jamais avec les termes de la nouvelle requête, peuvent être semblables à eux dans le k-espace.

Dans ce travail, on propose d'appliquer la méthode ASL en utilisant:

- La nouvelle requête à étendre ;
- Les documents les mieux classés pour elle ;
- Les requêtes recommandé les plus similaires à elle ;
- Et les documents les mieux classés pour chaque requête recommandée.

Ensuite, les vecteurs des termes de la nouvelle requête peuvent être comparés à tous les vecteurs des termes candidats d'expansion, et ils peuvent être classés par leur similitude par rapport à chaque terme de la requête à élargir à l'aide de la mesure commune de similarité cosinus, dont l'expression est comme suite :

$$Simc(\vec{t_i}, \vec{t_j}) = \cos(\vec{t_i}, \vec{t_j}) = \frac{\vec{t_i} \times \vec{t_j}}{||\vec{t_i}|| \times ||\vec{t_j}||} \quad (10)$$

En combinant les similarités de chaque terme candidat d'expansion t_j pour tous les termes de la nouvelle requête t_i^Q, nous pouvons calculer le poids de cohésion d'un terme candidat, qui représente la relation (corrélation) entre ce terme et la requête à étendre en entier.

Le poids de cohésion du terme t_j pour la requête utilisateur Q est mesuré par l'expression (Cui et al., 2002):

$$CoWeight(Q, w_j^{(d)}) = \ln\left(\prod_{t_i^Q \in Q}\left(P(t_j | t_i^Q) + 1\right)\right) \quad (11)$$

Cette méthode renvoie une liste de termes pondérés. Les termes les mieux classés peuvent être sélectionnés comme termes d'expansion de la nouvelle requête de l'utilisateur.

5 Expérimentations

Comme collection de test, nous avons utilisé la base de données CISI de la collection standard SMART. Cette collection offre 111 requêtes, 1460 documents et une matrice représentant la pertinence ou non-pertinence de chaque document par rapport à chaque requête.

Nous avons utilisé uniquement les requêtes courtes (contenant moins de cinq termes), et pour évaluer la performance du système, nous avons utilisé la précision moyenne non-interpolée (UAP). Au court de nos expérimentations, nous avons cherché des documents pertinents jusqu'au 20ème document récupéré, et on s'est limité seulement aux deux premières requêtes recommandées, sauf dans le cas de la dernière expérimentation qui concerne une variation du nombre de requêtes recommandées (Figure 1).

Concernant la méthode ASL, nous avons utilisé le logiciel R et le package proposé par Wild (2005) pour construire la matrice DVS tronquée avec le nombre approprié de dimensions k de sorte qu'elle puisse capturer la majorité de la structure importante latente dans l'association entre les termes, toute en supprimant le bruit (section 3).

Afin de vérifier les performances des méthodes proposées ci-dessus, nous avons comparé les meilleurs résultats des deux modèles de recommandation (ARR et RRML) après plusieurs expérimentations où nous avons variés le nombre de documents utilisés et le nombre de termes d'expansion. Nous avons comparé aussi l'UAP des requêtes initiales (avant expansion) avec les meilleures valeurs de chaque cas d'expansion en utilisant le thésaurus WordNet uniquement, ASL basée sur RRML et ASL basée sur ARR. On a testé également le cas d'expansion par la méthode ASL basée sur RRML en cherchant les requêtes recommandées par rapport à des requêtes déjà étendues par les synonymes extraits de WordNet. Le Tableau 1 présente ces résultats.

Méthode d'expansion	---	WordNet	Analyse Sémantique Latente		
Modèle de Recommandation	---	---	RRML basée sur WordNet	RRML	ARR
UAP	0,52	0,52	0,62	0,53	0,77
Nombre de RR	---	---	2	2	2
Nombre de documents	---	---	5	9	9
Nombre de termes d'expansion	---	Tous les synonymes termes de la requête	2	4 ou 5	4 ou 5

TAB. 1 – *Comparaison des meilleures valeurs de chaque méthode d'expansion avec ses conditions appropriées.*

Le tableau 1 montre que l'approche ASL basée sur le modèle ARR donne la valeur la plus élevée de l'UAP, et cela en améliorant l'UAP de 48,1% par rapport à la requête initiale de l'utilisateur. En second lieu, nous remarquons que l'expansion basée sur ASL, et RRML en utilisant des requêtes déjà étendus avec des synonymes extraits de WordNet a un UAP amélioré de 19,2%.

Nous remarquons qu'en comparant l'expansion utilisant le modèle RRML appliqué directement sur les requêtes initiales avec l'expansion utilisant le modèle RRML sur les requêtes étendue en utilisant WordNet; nous avons trouvé une différence de 17% de la valeur de

l'UAP. Dans le but de prouver cette conclusion, on a comparé ces cas en utilisant les mêmes conditions (5 documents, 5 termes d'expansion et 2 requêtes recommandées) et les résultats sont donnés sur le tableau 2.

Modèle de recommandation	RRML basée sur WordNet	RRML
UAP	0,56	0,45

TAB. 2 – *Comparaison de la méthode d'expansion, utilisant directement le modèle RRML et le modèle RRML basée sur WordNet sur la base des mêmes conditions: 5 documents, 5 termes d'expansion et 2 requêtes recommandées.*

Le tableau 2 montre que lorsqu'on utilise les mêmes conditions d'expérimentations, on a également une valeur plus élevée pour le cas du modèle RRML basée sur une première expansion de la requête en utilisant WordNet.

Ainsi, on estime que peut être le modèle ARR également pourra donner de meilleurs résultats s'il est appliqué sur des requêtes déjà étendues initialement par WordNet. Les résultats trouvés en testant cette hypothèse sont présentés dans la figure 1.

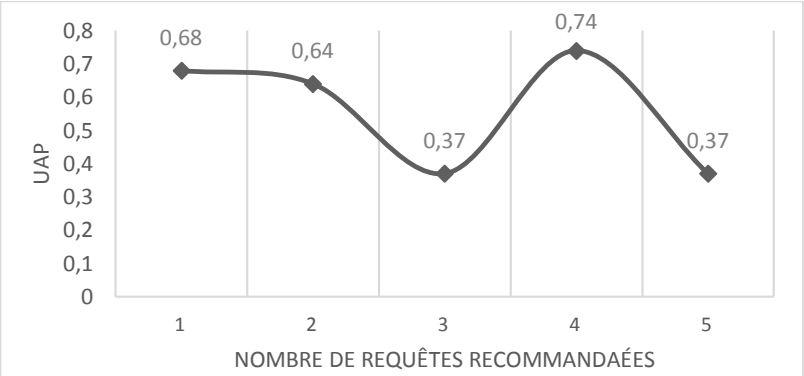


FIG. 1 – *Variation du nombre de requêtes recommandées (RR) utilisées pour l'expansion de requêtes courtes utilisant le modèle de recommandation ARR basée sur des requêtes étendues initialement par le thésaurus WordNet.*

Sur la figure 1, nous remarquons que lors de l'utilisation de deux RR la valeur de l'UAP diminue de 20,3% par rapport au cas utilisant le modèle ARR sans WordNet. Tandis que la valeur la plus élevée de l'UAP dans le cas d'utilisation des synonymes WordNet est de 0,74 avec quatre RR qui est également inferieure à la valeur donnée sur le tableau 1 (0,77 utilisant ARR). Ainsi, nous concluons qu'il n'est pas approprié de fusionner l'expansion par les synonymes WordNet avec le modèle de recommandation statistique ARR, même s'il améliore nettement les résultats de recommandation en utilisant le modèle de recommandation de requêtes par Modèles de Langue (RRML).

6 Conclusion

Dans cet article, nous avons proposé une méthode d'expansion de requêtes basée sur la l'Analyses sémantiques latente (ASL), et sur le contexte autour de la requête. Ce contexte est extrait à partir des requêtes historiques, par le biais de deux modèles de recommandation de requêtes qu'on a proposé. On a réalisé nos expérimentations en utilisant les requêtes courtes de la base de données textuelle CISI de la collection standard de test SMART.

Les résultats montrent que les meilleures valeurs de la précision moyenne non-interpolée sont donnés par l'approche d'expansion basée sur ASL et l'Algorithme de Recommandation de Requête (ARR). On remarque également que l'ajout de l'étape d'expansion de la requête avec WordNet, au modèle RRML améliore la performance du système par 16,98%, tandis que la performance diminue de 20,3% pour le modèle ARR. Nous concluons que pour les requêtes courtes, si nous utilisons ASL comme une méthode d'expansion, on extrait le contexte le plus approprié à l'aide de l'algorithme de recommandation de requête (ARR), sans faire une expansion initiale des termes de la requête en utilisant les synonymes extraites de WordNet.

Nous proposons comme futur travaux, d'améliorer notre approche et de l'appliquer sur d'autres collection de test, comme un journal de requête extrait d'un moteur de recherche réel.

Références

Asfari, O.,Doan, B-L., Bourda, Y., Sansonnet, J-P. (2010). *Context-based Hybrid Method for User Query Expansion*. In Proceedings of the fourth international conference on Advances in Semantic Processing (SEMAPRO), Italy, Florence, 69-74.

Bai, J., Nie, J-Y. Bouchard, H., Cao, G. (2007). *Using query contexts in information retrieval*. SIGIR '07 Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. New York, USA, 15-22.

Baziz, M. (2005). *Indexation conceptuelle guide par ontologie pour la recherche d'information*. PhD thesis, Institut de Recherche en Informatique de Toulouse, Université Paul Sabatier de Toulouse, December.

Bouchard, H., Nie, J.Y. (2006). *Modèles de langue appliqués à la recherche d'information contextuelle*. In CORIA'06. Lyon France, 213-224.

Cao, G., Nie, J., Bai, J. (2005). *Integrating Word Relationships into Language Models*. In Proceedings of SIGIR'05, Salvador, Brazil, August.

Cui, H., Wen, J.R., Nie, J.Y., Ma, W.Y. (2002). *Probabilistic Query Expansion Using Query Logs*. WWW2002. Honolulu Hawaii USA, May 7-11.

Fonseca, B.M., Golgher, P., Pôssas, B., Ribeiro-Neto, B., Ziviani, N. (2005). *Concept-based interactive query expansion*. In Proceedings of the 14th ACM International Conference on Information and Knowledge Management - CIKM'05. New York USA, 696-703.

Gupta, Y., Saini, A., Saxena, A.K. (2013). *A Review on Important Aspects of Information Retrieval*. International Journal of Computer, Control, Quantum and Information Engineering, Vol. 7, No. 12:990-998.

Imran, Ha., Sharan, A. (2010). *Selecting Effective Expansion Terms for Better Information Retrieval*. International Journal of Computer Science & Applications (IJCSA), Vol. 7, No. 2:52-64.

Lin, S.M., Huang, C.M. (2006). *Personalized Optimal Search in Local Query Expansion*. In Proceedings of the 18th Conference on Computational Linguistics and Speech Processing, Hsinchu, Taiwan; September, 221-236.

Liu, F., Yu, C., Meng, W. (2002). *Personalized web search by mapping user queries to categories*. In CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management. November, 558-565.

Manning CD, Raghavan P, Schütze H. (2009). *An Introduction to Information Retrieval - Chapitre 18: Matrix decompositions and latent semantic indexing*. Online edition (c) Cambridge University Press; April, 403-419.

SaidL'Hadj L. (2009). *Recherche Conceptuelle d'Information, Modèle d'Indexation Mixte : Concepts-mots.* Mémoire de Magister en Informatique. Ecole nationale Supérieure d'Informatique (ESI), Algérie.

Saint-Réquier, A., Dupont, G., Adam, S., Lecourtier, Y. (2010). *Évaluation d'outils de reformulation interactive de requêtes*. In Proc. CORIA, 223-238.

Wen, J., Nie, J., and Zhang, H. (2001). *Clustering User Queries of a Search Engine*. In Proceedings of WWW10, Hong Kong, May.

Wild, F. (2005). *An Open Source LSA Package for R*. CRAN. November.

Xu, J., Croft, W.B. (2000). *Improving the effectiveness of information retrieval with local context analysis.* ACM Transactions on Information Systems (TOIS), Vol. 18, No. 1. January, 79-11.

Zahera, H.M., El Hady, G.F., Abd El-Wahed, W.F. (2011). *Query Recommendation for Improving Search Engine Results.* International Journal of Information Retrieval Research, Vol. 1, Issue 1, January, 45-52.

Zhai, C. (2008). *Statistical Language Models for Information Retrieval: A Critical Review*. Foundations and Trends in Information Retrieval, Vol. 2, No. 3:137-215.

Summary

Query expansion is an approach which improve the performance of the Information Retrieval System on the web. It consist on the fact of suggesting to the user, terms that were extracted from the relevant documents of the initial query. In this paper, we are proposing a query expansion method based on the Latent Semantic Analyses (LSA) and the context around the user query. We had use two recommendation models: the first one based on the Cosine Similarity (CS) and the second using the Language Models (LM) in order to extract the query context. In order to improve the system's precision, we enriched the LM model

with supplementary contextual information based on WordNet. The experimentations results on the Collection CISI Smart, shows an improvement of 48,1% of the efficiency using the SC model and 19,2% using the ML model.

Aspects techniques d'un modèle de fouille de données Cloud basé sur le principe Map/Reduce de Google

Abdelfettah Idri* , Azedine Boulmakoul**

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

Résumé. L'extraction de la connaissance depuis les entrepôts de données volumineux nécessite des solutions robustes et scalables pour surpasser les contraintes temporelles et spatiales des algorithmes classiques. Notre approche de fouille de données repose sur la construction du treillis de Galois pour la génération des motifs fermés fréquents et des règles d'association. Afin d'optimiser ce processus, nous avons développédans nos travaux antérieurs une architecture parallèle distribuée basée sur CORBA qui a permis la distribution du traitement et la favorisation de la distribution de la mémoire selon notre algorithme proposé. Dans ce papier on traite la projection du modèle CORBA vers le modèle Cloud de Google Map/Reduce. On se focalise plus précisément sur la conception du modèle Map/Reduce relatif à notre approche tout en se basant sur la logique de la solution initiale.

1 Introduction

Quand on considère la relation entre les treillis de Galois et la prospection de données, on s'aperçoit qu'il existe une correspondance bijective entre les treillis de Galois et les motifs fermés du moment que l'intention du concept coïncide avec la notion de motif fermé Zaki et Ogihara (1998). D'où le besoin énorme d'algorithmes de construction de treillis de Galois. Notre travail envisage le Treillis de Galois qui est à la base de la génération des motifs fermés fréquents. Aussi, s'intéresse-t-on à la génération des règles d'association basée sur le Treillis de Galois. Notre approche s'inscrit dans l'optique d'améliorer les performances de l'algorithme séquentiel de génération du Treillis de Galois en préconisant une approche parallèle distribuéeIdri et Boulmakoul (2012). L'infrastructure déjà développée dans ce processus de fouilles de données et basée sur CORBA fera l'objet d'intégration dans une approche Cloud orientée services de traitement pour exclure dans notre cas les services de données et ceux du stockage.

La construction du treillis de Galois a fait l'objet de plusieurs recherches, spécialement dans les domaines d'analyse des concepts formels d'une part Ganter et Wille (1999), Bordat (1986), Chein (1969) et la fouille de données d'autre part Zaki et Ogihara (1998), Pasquier et al. (1999). Depuis leur apparition, l'analyse des concepts formels et la fouille de données trouvent leur voie d'application dans plusieurs domaines, surtout ceux manipulant des bases de données volumineuses.

Dans ce papier, on propose une approche de fouille de données orientée Cloud computing basée sur un algorithme parallèle distribuée pour la construction de treillis de Galois qui s'inspire des mêmes techniques des algorithmes séquentiels tels que celui de Bordat (1986) et Choi (2006). L'algorithme parallèle distribué ainsi que son architecture CORBA le décrivant ont été déjà conçus et implémentés dans nos travaux antérieurs Idri et Boulmakoul (2012).Dans ce papier, l'accent sera mis sur la projection de notre architecture existante basée sur une infrastructure CORBA vers une infrastructure Cloud computing adoptant le principe *Map/Reduce* et ceci en vue d'améliorer la performance du processus de fouille de données et plus précisément celui de la génération du treillis de Galois. Le processus détaillé de génération du treillis de Galois selon une approche parallèle distribuée a été bien élaboré auparavant Idri et Boulmakoul (2012). Cette approche concerne ladistribution intégrale du processus de la fouille de données et donc aussi bien du traitement que de l'utilisation mémoire. On s'intéresse dans ce travail à la conception du modèle de fouille de données Cloud basé sur *Map*/*Reduce*. Ce document est organisé comme suit. La section 2 présente la vision globale de l'approche. La section 3 traitela projection de l'architecture Cloud sur le modèle *Map/Reduce*. La conception du modèle *Map*/*Reduce* est détaillée dans la section 4. On conclut dans la section 5 avec nos suggestions et recommandations.

2 Vision globale

2.1 Introduction au Cloud Mining

La volumétrie et la complexité des données imposent de plus en plus des solutions robustes basées sur des infrastructures distribuées aussi bien taillée que le problème nécessite. Ceci nous a conduits vers des solutions qui doivent être scalables supportant un parallélisme au niveau du traitement. Le Cloud Computing s'inscrit en fait dans cette optique. Il offre une infrastructure de ressources et de services accessibles à partir de l'internet pour couvrir un besoin en termes de stockage, de gestion de données ou de calcul (intensif). Ces services se présentent sous forme de couches qui servent de plateforme de développement d'applicationsGrossman et Yunhong(2008).Plusieurs paradigmes existent qui s'intéressent à la fouille de données à aspect Cloud. Du moment qu'on se focalise sur le volet traitement de la fouille de données et non pas directement sur le volet stockage, on nomme particulièrement les approches qui partage ce même axe avec notre vision puisqu'elles sont les mieux positionnées pour une intégration avec notre architecture CORBA déjà réalisée (les techniques de stockages avancées telles que *BigTable* de Google ou *SimpleDB* de Amazon pourront être intégrées dans une approche de dataming globale couvrant le stockage et le traitement simultanément). Le modèle *Sector/Sphere*Grossman et Yunhong(2008) vise à minimiser le transport de données en reposant sur la composante *Sector* pour la persistance du stockage et permettre par conséquent un traitement en local à l'aide de la composante *Sphere*. Ce modèle convient le mieux à une architecture distribuée et décentralisée avec un faible couplage.Contrairement à Sector/Sphere, on trouve le modèle Map/Reduce basé sur le système de fichier de Google (GFS) et le système de fichier distribué Hadoop (HDFS), qui lui s'adapte plus aux

architectures distribuées avec un fort couplage. Similairement à la première méthode, le processus est décliné sur deux étapes : l'extraction et la répartition en parallèle des données sous forme de block (*map*), puis le traitement et la constitution du résultat final (*Reduce*).Puisque notre architecture CORBA adopte le modèle Manager/Agent, elle peutêtre considérée comme une architecture centralisée dans le sens où le Manager est responsable de la supervision intégrale du processus de fouille de données. Dans ce qui suit on adoptera Map/Reduce comme infrastructure Cloud d'intégration de notre architecture CORBA.Notre approche se focalise essentiellement sur les services à aspects traitement et donc des unités de calculs distribuées (voir Figure1).

Figure1.Services de la pile Cloud

La Figure2 schématise notre approche Cloud qui sera implémentée par l'infrastructure CORBA déjà développée dans nos travaux antérieurs Idri et Boulmakoul (2012)et qui va servir comme socle pour une fouille de données parallèle distribuée. On se limite dans cette vision à l'aspect services de traitement en mettant l'accent sur une technique de distribution de mémoire. Le fournisseur de fouille de données Cloud offre son service par le biais d'un gestionnaire de configuration « Configuration Manager » qui s'occupe de la constitution d'une instance de l'infrastructure adaptée au besoin du client demandeur de service. L'architecture qui réside derrière ce schéma repose sur le modèle Manager/Agent où les unités de traitement (Processor node) seront représentées par des agents et les unités de mémoires (Memory node) par des Tries ou sub-Tries (arbres lexicographiques). Le détail de cette architecture et l'algorithme la supportant sont décrits dans les paragraphes suivants.

2.2 Structuration de l'infrastructure Cloud pour la fouille de données

Selon notre approche, le fournisseur Cloud offre des services de traitement et donc se limite à la première couche de la Figure1. Les nœuds de cette couche sont scindés sur deux niveaux : distribution de la mémoire et celle du traitement d'après notre méthode adoptée qui est basée sur le treillis de Galois comme élément central de génération des motifs fermés et des règles d'association Idri et Boulmakoul (2012).

Distribution du traitement :

Dans notre contexte, la fouille de données commence par la génération du treillis de Galois et par la suite on en déduit les motifs fermés fréquents pour finir par la gé-

nération des règles d'association. Ce processus se caractérise par un aspect récursif qui traite les concepts en haut de la hiérarchie du treillis en premier lieu pour arriver à ceux du niveau le plus bas. Le modèle choisi est celui du Manager/Agent et par conséquent le processus de génération du treillis est subdiviséen tâches qui sont réparties sur les Agents qui représentent dans notre modèle Cloud les nœuds de traitement regroupés dans une couche de traitement comme indiqué dans la Figure2.

Distribution de la mémoire :

Pour assurer son rôle, le Manager utilise un Trie global pour consolider les résultats collectés par les Agents et pour gérer le processus entierde construction du treillis de Galois. Ce Trie représente en fait la mémoire globale du processus de fouille de données et sa taille impacte sondéroulementénormément du moment que le nombre de concepts à générer croit exponentiellement par rapport au volume initial du contexte au sens de l'AFC (Analyse Formelle des Concepts), Tan et al. (2006), Mishra et Pitt (1997). Pour remédier à cette contrainte, on a opté pour une distribution de cette mémoire (Trie) selon une technique bien précise qui vise à répartir le Trie global en sub-Tries ayant un poids optimal (le nombre total de nœuds d'une hiérarchie donnée de nœuds), Idri et Boulmakoul (2012).Les nœuds constituant cette mémoire sont regroupés dans une couche de mémoire comme schématisé dans la Figure2.

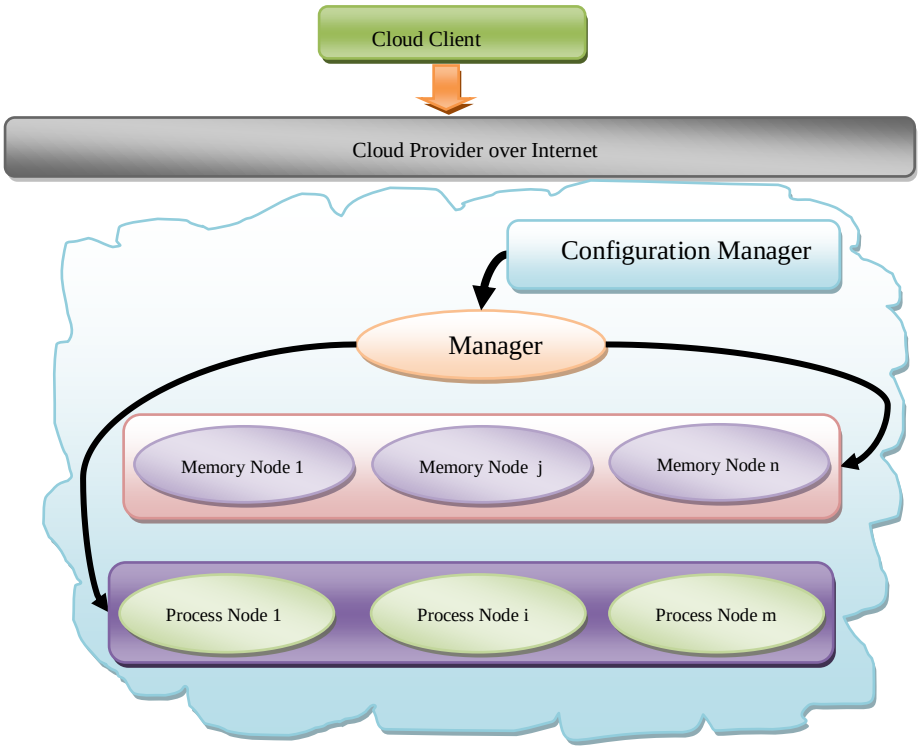


Figure2.L'infrastructure globale de l'approche Fouille de données Cloud

3 Projection de l'architecture Could proposée sur le modèle Map/Reduce

3.1 Principes de base du modèle Map/Reduce

Le modèle Map/Reduce est un paradigme de programmation qui permet une scalabilité massive à travers une distribution de données et un traitement parallèleDeanet Ghemawat (2008). Il cible la résolution des problèmes liés à la manipulation des données extrêmement volumineuses. Ce modèle a été introduit par Google en 2004 surtout pour construire et gérer son index WEB. Après avoir prouvé son efficacité, il est utilisé par d'autres sociétés telles que Yahoo et Facebook.

Son principe de base réside dans le fait de répartir les tâches de traitement (calcul) sur un nombre important de nœuds selon un modèle de programmation parallèle. Le nombre de nœuds peut atteindre facilement quelques milliers. Le modèle Map/Reduce possède une implémentation appelée Hadoop disponible sur le net. Celle-ci repose sur son propre système de fichiers HDFS (Hadoop Distributed File System) qui adopte deux types de nœuds :

- Name Node : c'est le nœud qui sauvegarde les métadonnées des fichiers et des répertoires
- Data Node : c'est le nœud qui sauvegarde les données en termes de blocks de fichier

Parmi les avantages notables de ce modèle : la scalabilité, la tolérance aux plantages, la simplicité d'utilisation, la réduction du coût.Pour pouvoir utiliser un tel modèle, l'utilisateur devrait implémenter deux fonctions principales : Map et Reduce. Ce principe est inspiré des langages fonctionnels comme LISP et Scheme.

- La fonction ***map*** : elle doit avoir lieu avant la fonction *reduce*. Elle prend les données en entrée et les formules sous forme d'un couple constitué d'une clé et d'une valeur (k, v) et les transforme en une liste de paires intermédiaires de clés et de valeurs : List (Ki, Vi).
  - **map (k, v) → List(Ki, Vi)**
- **La fonction *reduce* : elle** s'exécute après la fonction ***map*** et son rôle est la consolidationdes valeurs intermédiaires ayant une clé commune.
  - **Reduce (Ki, List(Vi)) → List (Vo)**

3.2 Fonctionnement global du modèle Map / Reduce

La figure suivante schématise le fonctionnement du modèle Map/Reduce d'une façon générale. Le nombre de blocks en entrée et le nombre de sorites est à titre indicatif. L'application du modèle à notre contexte fera l'objet de la section suivante.

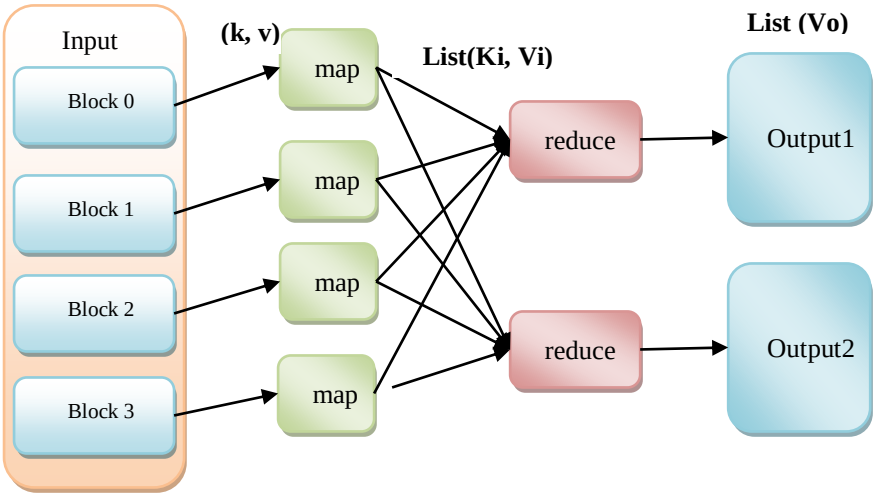


Figure 3Fonctionnement du modèle Map / Reduce

Le processus Map/Reduce répartit le fichier en entrée en n blocks uniformes et applique par la suite la fonction *map* sur chaque block en parallèle. Les blocks constituent les données en entrée pour le modèle Map/Reduce. Le résultat de cette étape est ensuite collecté par la fonction *reduce* qui va le consolider selon la fonction d'agrégation définie par l'utilisateur et fournir le résultat final. Dans notre cas, les blocks seront représentés par le contexte AFC (ensemble des objets, ensemble des attributs et la relation binaire) et le résultat final fournira le treillis de Gallois.

3.3 Projection de l'architecture proposée vers Map/Reduce

Il s'agit dans ce contexte de modéliser l'architecture basée sur l'infrastructure CORBA par les composantes du modèle Map/Reduce. Du moment que notre architecture est en soi destinée pour un algorithme parallèle distribué Idri et Boulmakoul (2012), son implémentation par des composantes Map/Reduce est presque intuitive. Le modèle Map/Reduce cible la même stratégie de distribution de traitement en mettant à la disposition de l'utilisateur toute l'infrastructure matérielle et logicielle nécessaire en plus de la prise en charge de tous les aspects de communications entre les différentes composantes de l'architecture mise en jeux. Les fonctions map et reduce réaliseront les tâches suivantes selon notre concept :

Map : cette fonction s'occupe de deux aspects principaux :

- Il prend en charge la tâche de l'agent, notamment, la génération des concepts enfants sur la base d'un concept parent et du contexte : le calcul initial

Reduce : ce processus effectue :

- La consolidation des résultats intermédiaires obtenus par le mapper et les intègre au niveau de la mémoire globale (Trie global)

- Génération du treillis de Galois

Ces fonctions peuvent être utilisées en cascade pour mieux refléter la constitution progressive du treillis de Galois reposant sur un Trie.

4 Conception du modèle Map/Reduce

On s'intéresse dans cette partie au volet traitement de notre approche, notamment le processus de génération du treillis de Galois. Ce dernier peut être décliné sur les deux phases suivantes en se référant au modèle Map/Reduce (voir le paragraphe précédent) :

1. La génération des concepts enfants
2. La construction du treillis de Galois

Par conséquent, nous proposons un modèle Map/Reduce hiérarchisé en deux niveaux conformément au découpage fonctionnel basé sur les deux phases susmentionnées.

4.1 Vision globale

La figure ci-dessous montre la projection des deux phases du processus sur le modèle Map / Reduce.Les entrées sont représentées par le contexte AFC (Objets, Attributs, Relation binaire). Comme déjà avancé, on ne s'intéresse dans ce travail qu'au processus de traitement (construction du treillis de Gallois) en laissant de côté la répartition des données pour nos travaux futurs.La première phase consiste à générer le treillis de Galois qui est représenté dans ce modèle par un ensemble de listes constituées chacune de deux éléments : le concept parent et ses concepts enfants. C'est le moyen adopté pour marquer la relation hiérarchique entre les concepts du treillis. Le concept parent est celui qui est transmis au *mapper***(C_id, (ext,int))**et il est représenté par une clé **(C_id)** qui est l'identificateur du conceptet une valeur qui est le contenu du concept, notamment son intention et son extension. Normalement on aurait pu utiliser comme clé soit l'intention soit l'extension du concept du moment que ces ensembles sont uniques par définition du concept. Mais pour des raisons de lisibilité, on a préférél'identificateur du concept généré lors de sa création.Le résultat retourné par le *mapper* est une liste de paires (parent, enfant) : **(C_id, (C_id,(ext, int)))**. Le premier identificateur **C_id** est celui du concept parent alors que le deuxième désigne le concept enfant. Selon ce schéma, la fonction du *reducer* est de consolider les résultats intermédiaires pour former les briques de base permettant la construction du treillis de Galois. Chaque brique est constitué d'un concept parent accompagné de la liste de ses concepts enfants **(C_id, list [C_id]).** Le résultat final est l'ensemble de ces briques **(List (C_id, list [C_id]))** qui n'est autre que le treillis de Galois.Il faut noter que cette première phase est itérative : elle est réexécutée jusqu'à ce que le concept minimal soit atteint (le critère d'arrêt de la construction du treillis de Galois).

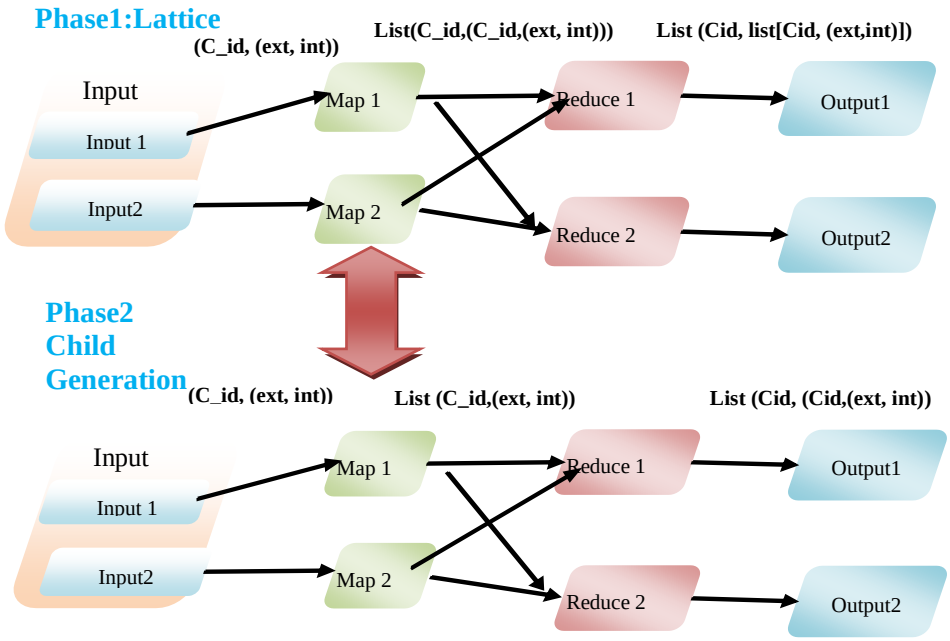


Figure 4 Vision globale Map / Reduce

La deuxième phase est déclenchée depuis la première phase au moment de l'activation du *mapper* comme le montre la figure et plus précisément lorsqu'on veut calculer les concepts enfants d'un concept parent. Le mapper reçoit dans ce cas un concept parent **(C_id, (ext, int))** et renvoie une liste de concepts candidats **List(C_id, C_id,(ext, int)))**. La fonction *reduce* permet alors de générer les concepts enfants en appliquant le test de fermeture sur les candidats et en consolidant les résultats intermédiaires.

4.2 Phase I

Nous allons spécifier dans cette partie les fonctions Map et Reduce relatives à la première phase.

```
1:Mapper
2:Method Map (int C_id, Concept C))
3:   // context = input data
4:   Variables:
5:      Ext : ObjectSet
6:      Int: ItemSet
7:      CC_id : Integer
3:   Initialse (context)
4:   Children ← generate_Children (C)
5:For each Concept CC in Children do
6:            CC_id ← CC.id
7:            Ext ← CC.extension
8:            Int ← CC.intention
9:            Emit (C_id, (CC_id,(ext,int)))
```

```
1:Reducer
2:   Method Reduce (int level, list(int C_id, Concepts Children) Res)
3:     // L is the lattice (set of pairs (parent,child))
4:    // Next job concepts NJC
5:    NJC ← Ø
6:    For each el in Res
7:      For each concept CC in el.Children do
8:         L ← L U (el.C_id, CC)
9:         NJC ← NJC U {CC}
10:   if level < max then
11:     for each CC in NJC do
12:          emit (CC.id, CC)
13:   else
14:      emit L
```

Figure 5 Spécification des fonctions Map et Reduce de la phase I

Comme le processus de la phase I est itératif, la méthode*Reduce* est responsable de la construction du treillis de Galois progressivement en insérant les nouveaux concepts générés dans la variable résultante L désignant le treillis global. Les nouveaux concepts générés par itération sont stockés dans la variable **NJC**pour qu'ils soient relancés lors de l'itération suivante.L'argument **level** représente le niveau hiérarchique du treillis et l'argument **Res** est transmis par le *mapper* au *Reducer* et il est composé d'une liste de paires (concept parent, liste (concepts enfants)). Par conséquent, chaque élément de cette liste est décortiqué de telle façon à injecter les paires (concept parent, concept enfant) dans le treillis **L** (ligne 8) et préparer le lancement récursif de ces nouveaux concepts (ligne 9). Il faut noter que le traitement des concepts générés doublement par différent parents est unique (simulation du test de l'existence d'un concept Idri et Boulmakoul (2012)) suite à l'opération d'union (ligne 9) qui réduit les concepts similaires à un seul exemplaire.

4.3 Phase II

La spécification de la phase II est décrite dans la figure suivante.

```
1:Mapper
2:Method Map (int C_id, Concept C))
3:      Initialse (context)
4:      Candidat ← {X / X C C.intention}
5:For eachXin Candidatdo
9:            Emit (C_id, Obj(X),X))
10: // Obj returns the shared objects of the
attributes belonging to X (see paper 2008)
11: we can use also (Ei, attr(Ei))
```

```
1:Reducer
2:  Method Reduce (int C_id, list (ext, int) CC)
3:     Initialse (context)
4:
7:       For each CT in CC do
10:   if isClosed (CT) then
11:      CT.id ← id
12:       emit (C_id, CT)
```

Figure 6des fonctions Map et Reduce de la phase II

La méthode Map détermine l'ensemble des parties de l'intention du concept parent et constitue les concepts candidats en appliquant la fonction « Obj »sur chaque

partie (voir Idri et Boulmakoul (2012)). Un concept est identifié lorsque le test de fermeture est vérifié.

5 Conclusions et perspectives

Notre architecture parallèle distribuée basée sur CORBA a servi comme socle pour s'orienter vers le modèle Map/Reduce qui offre une infrastructure de distribution assez robuste et scalable. Cette projection promet une fouille de données Cloud offrant des ressources supportant des traitements intensifs et complexes qui peuvent améliorer la performance de l'architecture CORBA. Par ailleurs, le modèle Map/Reduce permet la prise en charge des aspects techniques concernant la distribution et le parallélisme ainsi qu'il reste ouvert par rapport à la répartition des données grâce à son système de fichier distribué HADOOP.

Perspectives :

La nouvelle architecture basée sur le modèle Map/Reduce peut être implémentée en se basant sur la conception proposée dans ce papier et en s'inspirant de l'architecture parallèle distribuée déjà implémentée en gardant les mêmes principes. La répartition des données qui n'a pas été traitée dans ce papier, peut améliorer d'avantage la performance de la solution en optimisant le transport des données et l'utilisation de la mémoire.

Références

Bordat, J. P. : Calcul pratique du treillis de Galois d'une correspondance, Math. Sci. Hum. 96 31-47 (1986)

Chein, M. : Algorithme de recherche de sous-matrice première d'une matrice, Bull. Math. R. S. Roumanie 13 (1969)

Choi, V.: Faster Algorithms for Constructing a Concept (Galois) Lattice, Presented at SIAM Conference on Discrete Mathematics 2006, University of Victoria, Canada

Dean J. et Ghemawat S. (2008), MapReduce : Simplified Data Processing on Large Clusters, Communication of the ACM, Vol. 51, No 1

Ganter, B., Reuter, K.: Finding all closed sets : a general approach. Order, 8:283-290 (1991)

Ganter, B., Wille, R.: Formal Concept Analysis : Mathematical Foundations. Springer Verlag (1999)

Grossman, R., Yunhong, G.: Data Mining using High Performance Data Clouds: Experimental Studies Using Sector and Sphere (2008), Proceedings of The 14th ACM SIGKDD InternationalConference on Knowledge Discovery and Data Mining (KDD 2008), ACM, 2008, pages 920-927

Grossman, R., Yunhong, G.: Sector and Sphere: The Design and Implementation of High Performance Data Cloud (2008), UK e-Science All Hands Meeting 2008, September 10, 2008, Edinburgh, UK

Idri, A.F., Boulmakoul, A. : Une approche parallèle distribuée pour la génération des motifs fermés fréquents basée sur une infrastructure CORBA, ASD (2008)

Idri, A.F., Boulmakoul, A. : Une approche parallèle optimisée de distribution intégrale de la fouille de données basée sur une infrastructure CORBA, ASD (2012)

Mishra N., Pitt L.: Generating all maximal independent sets of bounded-degree hypergraphs. COLT '97 proceedings of the tenth annual conference on Computational learning theory ACM (1997)

Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L. : Efficient mining of association rules using closed itemset lattices. Information systems. 24(1), p25-46 (1999)

Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L. :, Closed set based discovery of small covers for association rules. In Actes des 15èmes journées Bases de Données Avancées (BDA'99), pages 361- 381 (1999)

Tan P., Steinbach M., Kumar V.: Introduction to Data Mining. Addison-Wesley, Chapter 6 (2006)

Zaki, M. J., Ogihara, M.:Theoretical foundations of association rules. Proc. 3rd ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, p1-7(1998)

Zaki, M., Phoophakdee, B.: MIRAGE: A Framework for Mining, Exploring and Visualizing Minimal Association Rules Rensselaer Polytechnic Institute, , RPI CS Dept Technical Report 03-04, 2003

Summary

This paper describes design issues of the Map/Reduce model derived from ourformer CORBA based parallel distributed architecture. Our approach is based on Concept Lattice as its framework for generating frequent item sets and association rules. A couple of standard algorithms exist for building the concept lattice of a binary relation.Both the distribution of the data mining process and the memory component arediscussed in this paper. Our focus is the projection of our CORBA based architecture to a *MapReduce* architecture and its design.

A Cybersecurity Model in Cloud Computing Environments for Selecting the Reliable Cloud Service Provider

Jamal TALBI*, Abdelkrim HAQIQ*, **

*Computer, Networks, Mobility and Modeling laboratory, Department of Mathematics and Computer, FST, Hassan 1st University, Settat, Morocco
**e-NGN Research group, Africa and Middle East
{talbi85, ahaqiq}@gmail.com

Abstract. Cloud computing is becoming a key factor in computer science. It represents a new paradigm of utility computing and enormously growing phenomenon in the present IT industry and economy hype. The cloud users (CUs) increase and require secure, reliable and trustworthy cloud service providers (CSPs) from the market. It's a challenge for a new customer to choose the highly secure provider. However, a number of security risks are emerging in association with cloud usage that needs to be assessed and managed for both customers and providers. Thus, there is a need to determine the selection of appropriate cloud computing services offered by different providers. In this paper, we propose a decision-making model based on cost and risk approach by measuring the mean failure costs (MFC) for the group of cloud providers which make decision of the more reliable provider among all providers and justify the business needs in terms of security and reliability of the subscribers.

1 Introduction

Cloud computing (Caron et al., 2013) is an active research subject as the information industry sees it as the new model. Many companies, enterprises and organizations outsource some of their information systems to benefit from the cloud services which are Platform as a Service (PaaS), Infrastructure as a Service (IaaS) and Software as a Service (SaaS). The main interesting features of a cloud are the cost decrease and a faster time to market. Currently there are many numbers of providers, but finding the best cloud service provider among the available cloud service providers is difficult. Thus, it is a challenge for the users to choose the reliable and secure cloud provider for fulfilling their requirements. Presently, there is a

lack of frameworks that can permit customers to evaluate cloud offerings and rank them based on their ability to meet the user's security requirements.

Like traditional computing environments, cloud computing brings risks like loss of security and loss of control (Chow et al., 2009; Hanna, 2009; Ibrahim et al., 2010; Sean et Kevin, 2011; Subashini et Kavitha, 2010; Wooley, 2011; Xuan et al., 2010). Indeed, by trusting its critical data to a service provider, a user (whether it is an individual or an organization) takes risks with the availability, confidentiality and integrity of this data: availability may be affected if the subscriber's data is unavailable when needed, due for example, to a denial of service attack or merely to a loss; confidentiality may be affected if subscriber data is inadvertently or maliciously accessed by an unauthorized user, or otherwise unduly exposed; integrity may be affected if subscriber data is inadvertently or maliciously damaged or destroyed. Thereby, it is necessary to assess and contain risk using precautionary measures that are commensurate. Thus, we have to dispose a system that measure and rank the secured cloud service providers and then, the cloud services can make a major impact and will craft a healthy competition among cloud providers to satisfy their Service Level Agreement (SLA) and improve their Quality of Service (QoS) and trustworthiness (Zibin et al., 2010).

The Mean Time to Failure (MTTF) is a commonly accepted measure for system reliability. Similar metrics of security, such as the Mean Time to Detection (of a security vulnerability), abbreviated by MTTD, and the Mean Time to Exploitation (of a security vulnerability), abbreviated by MTTE, have also been proposed (Morgan, 1998). The MTTF is by far better known than MTTD and MTTE on measuring reliability and security. The latter is one of the major issues in cloud computing and a critical requirement of the cloud provider and cloud customer. With these points in mind, the organizations must have a right comprehensive about important risks in cloud computing environment.

We propose in this paper a cyber security measure of reliability and safety as an economic function (MFC), quantified in monetary terms (dollars per hour) the loss resulted in security breaches related to the group of cloud service providers. In this way, the proposed model will help a new customer to find the most reliable and secured CP and justify the business needs in terms of security and reliability.

The remainder of this paper is organized as follows: the next section discusses related work, Section III introduces the proposed model. Section IV describes the CSP Rank Framework. Section V presents an implementation of the model. Section VI gives a conclusion.

2 Related work

Security metrics are one of criteria that play a major role in ranking service providers. A cloud user may require an efficient, cost effective and basically more secure provider for his application. Since there are many providers who will provide same type of services with different level of security, so it will be a challenge for the user to select. Our motivation in this paper is to promote a decision model for ranking providers based on measuring security metrics of cloud services.

In the same context, many researchers have proposed different approaches to help customer in this mission to select the appropriate cloud services. A collaborative filtering

approach (Linden et al., 2003) rank the items based on similar users preferences. This algorithm aggregates all the items purchased by the users and eliminate those items and ask users to rate the remaining services. In Zibin et al. (2010) cloud rank approach proposed greedy algorithm. It gives a method to rank cloud providers based on existing customer's feedback. It ranks component rather than service of providers. But there is no guarantee all explicitly rated items by customers are ranked properly. But similar users will experience the same with same cloud providers so for them this approach will be helpful.

QoS-aware web by collaborative filtering (Zheng, 2011) proposed a collaborative approach to rank providers on the basis of its web services. This method is useful for the customers who want to get an appropriate cloud provider which provides suitable web services. Thus, this method includes experience of users who used the services already and a hybrid collaborative filtering approach for evaluating web service QoS parameters.

Dhillon et Arora (2012) proposed an effective and efficient method to select best cloud service. In order to select the best provider three parameters are considered. But while ranking instead of taking all three together parameters are applied on after other. On the basis of result best provider is selected.

Zheng et al. (2013) proposed an approach for ranking equivalent cloud service providers provides similar kind of services which will help users to select suitable providers without spending much time for it. This method uses some QoS parameters for predicting best provider.

Kapgate (2014) proposed a predictive broker algorithm based on Weighted Moving Average Forecasting Model (WMAFM). It proposes a new method to balance load on data centers and also minimizes response time. So for end users can get their requested service within few seconds.

Subha et Banu (2014) had done a survey on quality of service ranking cloud computing. Here the author considered few quality of service parameters and ranked providers based on that.

Cloud Rank (Yuvarani et Sivalakshmi, 2014) approach measures and ranks cloud services for the users. It takes the feedback or rating of users who had used the services already.

An efficient approach (Amrutha et Madhu, 2014) find the best cloud provider by using a system for ranking cloud services based on QoS parameters such as service response time, cost, interoperability and suitability. It uses a broker algorithm that classify the existing providers and find out the more effective and efficient provider.

Whaiduzzaman et Gani (2013) proposed a conceptual model of federated third party cloud ranking and monitoring system (CMFCSPRS) that assures and boosts up the confidence to make a feasible secure and trustworthy market of CSPs.

3 The proposed model

We propose a broker which can act as a middleware between customer and cloud service provider. It can get the needed requirements from customer and help the customer by listing out suitable cloud providers. So our cloud broker has an important role to find out the secure

cloud service providers existing in the database of our cloud broker. The proposed model is described in the following, in terms of its architecture.

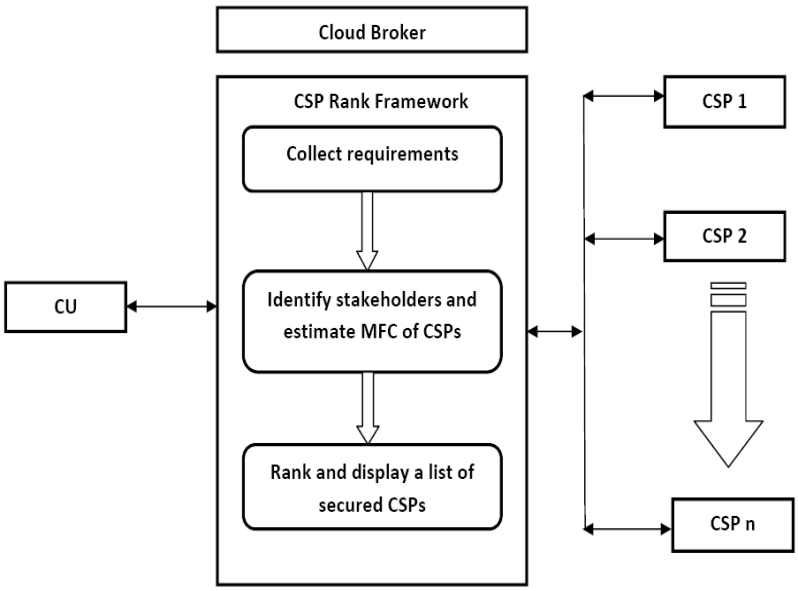


Fig. 1 – The structure of the proposed CSP Rank Framework

This system develops a model to find out the secured cloud service providers based on a security risk assessment approach by determining the vulnerabilities and computing the risks related to cloud service providers list.

3.1 Requirements requested

The broker collects requirements from user. It may be infrastructure requirements, platform requirements or software requirements.

3.2 Stakeholder's identification and computation of the mean failure cost

All the registered cloud service providers give all the services which they are providing. Cloud broker contains the level of security of cloud providers. So the client gives requirements to broker, it checks the provider's performance by quantifying a random variable in terms of financial loss per unit of operation time (e.g $/h) that represents the security failure of services.

3.3 Ranking secured cloud systems

The CSP Rank Framework using a broker provides optimal cloud service provider selection from the more numbers of CSPs based on a cyber security measure, especially mean failure cost which provides better selection of providers among many. Thus, we proposed architecture uses evaluation of MFC related to systems caused by vulnerabilities and threats for making a decision to rank and select the right provider in terms of reliability and security.

4 Description of the CSP Rank Framework

Probably all cloud service providers have a Service Level Agreements (SLA), but most of these SLAs were written to protect the vendors as opposed to being customer-centric. That has to change, and customers have to demand more with regard to service and the assurance of it. In the same time, cloud providers should protect their data or services from risk and harm. For this aim, the CSP Rank Framework will conduct vulnerability and threat scans of components and services of the existing providers. The obtained results were fed into the security ranking system that offer a list ranked of the secure providers.

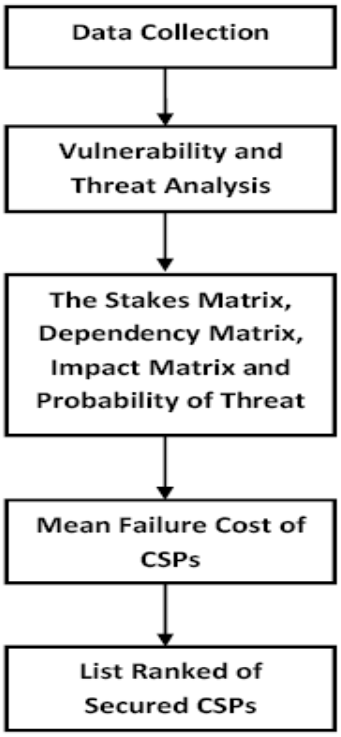


Fig. 2 – Conceptual model of CSP Rank Framework

Fig. 2 shows the conceptual model of the cloud broker for ranking secured CSPs. For this aim, some assumptions and conditions should be considered as follows (Whaiduzzaman, Gani, 2013):

- The CSP Rank Framework must maintain the trust and reliability.
- The CSP Rank Framework has enough resources to provide for processing and executing their own work.
- The broker must be maintained and regulated by strict laws and transparent policies.
- Both the broker and CSPs mutually agree before executing the software penetration test.
- We consider that a CSP provide IaaS, PaaS and SaaS of its own.
- The CSP Rank Framework is only the responsible of computing security metrics from sources and processes these measures for ranking results.
- A new cloud user looking for security and reliability should pay to the cloud broker to see the ranked results.

4.1 Vulnerability and threat analysis of CSPs

A vulnerability is a software defect or weakness in the security system which might be exploited by a malicious user causing loss or harm (Zheng et al., 2011). The identification of these vulnerabilities has been used by several approaches and researchers to estimate risks of the systems.

The Common Vulnerability Scoring System (CVSS) (Dhillon et Arora, 2012; Zheng et al., 2013) framework allows to assess the severity level of IT vulnerabilities. It associates a severity score (CVSS score) to each IT vulnerabilities, which ranges from 0.0 to 10.0. CVSS (Kapgate, 2014) is composed of three major metric groups: Base, Temporal and Environmental.

The Base metric represents the intrinsic characteristics of vulnerabilities, and is the only mandatory metric. The optional Environmental and Temporal metrics are used to augment the Base metrics, and depend on the target system and changing circumstances. The Base metrics include two sub-scores termed exploitability and impact. In the last sub-group, we find three metrics, representing the impact of the attack on the three classical security properties: Confidentiality Impact, Integrity Impact and Availability Impact.

A risk is the probability of cause of a problem when a threat is triggered by vulnerabilities. Threats are much related to the characteristics of the assets and vulnerabilities are relevant to the security controls (Foroughi, 2008). Risk can be formally defined (Stonebumer and al., 2002) as:

$$Risk = Threat \times Vulnerability \times Impact$$

4.2 Measuring the MFC as a measure of cyber security

Computing systems are characterized by five fundamental properties: functionality, usability, performance, cost, and reliability. The reliability of a computing system is the

ability to deliver service that can justifiably be trusted. Ben Aissa et al. (2010) introduce the concept of mean failure cost as a measure of reliability in general, and a measure of cyber security in particular.

Thus, once the stakes matrix ST, the dependency matrix DP, the impact matrix IM and the threat vector PT are determined by our broker, the formula (Ben Arfa et al. (2013)) of the column vector of size n that represents the mean failure costs associated to the providers can be written using the matrix product (×) as:

$$MFC = ST \times DP \times IM \times PT$$

where ST is the matrix of stakes that stakeholders have in meeting security requirement, the matrix DP is derived by the systems architect, in light of the role that each component of the architecture plays to achieve each security goal. The matrix IM is derived by the security analyst from architectural information, and the vector PT is derived from known perpetrator behavior, perpetrator models, and known system vulnerabilities by analyzing which vulnerabilities and threats can affect the system.

4.3 Final ranking of CSPs

Based on the calculation of the mean failure cost of each cloud service provider from all providers, our framework provides a list ranked of the secure CSPs starting with the providers having the minimum of the mean failure.

5 Implementation of the CSP Rank Framework

In the cloud computing systems, the confidentiality, the integrity and the availability are the important pillars of cloud security software assurance. Therefore, we based our contribution on these three principles of information security and the different levels of criticality of the data related to these requirements. In our model, we use seven generic security requirements, which are:

- AVC: Availability of critical data.
- AVA: Availability of archival data.
- INC: Integrity of critical data.
- INA: Integrity of archival data.
- CC: Confidentiality of classified data.
- CP: Confidentiality of propriety data.
- CB: Confidentiality of public data.

We illustrate the use of our CSP Rank Framework in a practical application by considering three cloud providers X, Y and Z under 14 security threats as shown in TAB. 1. which are the most often cited in relation with cloud computing systems (Security Alliance,

2010; Wooley, 2011). In our case study, we propose the following stakes matrix as shown in TAB. 2.

Each entry of this matrix represents, for stakeholder H and requirement R, the loss incurred by H if requirement R was violated and expressed by thousands of dollars ($K). In the column, we find the requirements.

Threats	Probability
Monitoring virtual machines from host (MVM)	08.063×10^{-4}
Communications between virtual machines and host (CBVH)	08.063×10^{-4}
Virtual machine modification (VMm)	08.063×10^{-4}
Placement of malicious VM images on physical systems (VMS)	08.063×10^{-4}
Monitoring VMs from other VM (VMM)	40.31×10^{-4}
Communication between VMs (VMC)	40.31×10^{-4}
Virtual machine mobility (VMM)	40.31×10^{-4}
Denial of service (DoS)	14.39×10^{-4}
Flooding attacks (FA)	56.44×10^{-4}
Data loss or leakage (DL)	5.75×10^{-4}
Malicious insiders (MI)	6.623×10^{-4}
Account, service and traffic hijacking (ASTH)	17.277×10^{-4}
Abuse and nefarious use of cloud computing (ANU)	17.277×10^{-4}
Insecure application programming interfaces (AII)	29.026×10^{-4}
No threats (NoT)	0.9682

TAB. 1 – THREAT VECTOR

	Requirements						
	AVC	AVA	INC	INA	CC	CP	CB
Stakeholders							
Provider X	200 $K	60 $K	600 $K	110 $K	1300$K	1800$K	100$K
Provider Y	500 $K	90 $K	800 $K	150 $K	1500$K	1200$K	120$K
Provider Z	700 $K	80 $K	700 $K	160 $K	1400$K	1000$K	130$K

TAB. 2 –PROVIDERS STAKES MATRIX

The cloud computing threats do not distinguish between real and virtual components. In our case, we choose a number of components independent of their types (virtual or physical) to simplify the mechanisms of operation in cloud architecture that are:

- A storage server (SS).
- A backup server (BS).
- A database server (DBS).
- An application server (AS).
- A web server (WS).
- A load balancer (LB).
- A router or firewall (R/FW).
- A proxy server (Prox), and
- A browser (Brows).

The dependency matrix and the impact matrix are shown in TAB. 3 and TAB. 4.

	Components									
	SS	BS	DBS	AS	WS	LB	R/FW	Prox	Brow	No failure
Security requirements										
AVC	1	0.01	1	0.28	0.44	1	1	1	1	0
AVA	1	0.01	0.28	0.28	0.44	1	1	1	1	0
INC	1	0.01	1	0.14	0.44	1	1	0.14	0.14	0
INA	1	0.01	0.14	0.14	0.44	1	1	0.14	0.14	0
CC	0.44	0.01	0.44	0.44	0.44	1	1	0.14	0.44	0
CP	0.44	0.01	0.44	0.44	0.44	1	1	0.14	0.44	0
CB	0.44	0.01	0.44	0.44	0.44	1	1	0.14	0.44	0

TAB. 3 –DEPENDENCY MATRIX

	Threats														
	MVH	CVH	VMm	VMS	MVV	VMC	VMM	DoS	FA	DL	MI	ASTH	ANU	IAI	NoT
Components															
SS	0.04	0.05	0.00	0.04	0.04	0.05	0.05	0.036	0.04	0.05	0.03	0.02	0.01	0.06	0
BS	0.001	0	0	0.04	0.001	0	0	0.036	0.04	0.05	0.03	0.02	0.01	0.06	0
DBS	0.001	0	0.033	0.04	0.001	0	0	0.036	0.04	0.05	0.03	0.02	0.01	0.06	0
AS	0.02	0.003	0.033	0.06	0.02	0.003	0.003	0.036	0.04	0	0.05	0.02	0.01	0.07	0
WS	0.03	0.003	0.033	0	0.03	0.003	0.003	0.02	0.04	0	0.01	0.02	0.01	0.01	0
LB	0.02	0.003	0	0.01	0.02	0.003	0.003	0.06	0.04	0	0.005	0.02	0.01	0.01	0
R/FW	0.03	0.05	0.033	0.03	0.03	0.05	0.05	0.06	0.04	0	0.005	0.02	0.01	0.01	0
Prox	0.01	0.05	0	0.01	0.01	0.05	0.05	0.02	0.01	0	0.005	0.02	0.01	0	0
Brow	0	0	0	0	0	0	0	0.02	0.01	0	0.03	0.02	0	0.03	0
No failure	0.06	0.04	0.03	0.03	0.06	0.04	0.04	0.01	0.02	0.01	0.02	0.05	0.06	0.005	1

TAB. 4 –IMPACT MATRIX

Based on the comparison of the mean failure cost for the clouds X, Y and Z as shown in TAB. 4. We conclude that the provider X is the most reliable among the three clouds.

Cloud Service Providers	MFC ($K)
Cloud provider X	13.8896
Cloud provider Y	15.1419
Cloud provider Z	14.7801

TAB. 5 –*Mean Failure Cost of Cloud Providers*

6 Conclusion

Cloud Computing became an important technology for many organizations to deliver different types of services. So, the multiple cloud service providers make a dilemma for a cloud user to choose each provider which is more secured and has the minimum security risk. Hence, in this paper, we propose a decision-making model based on cost and risk approach by measuring the mean failure costs (MFC) for the group of cloud providers which make decision of the more reliable provider among all providers and justify the business needs in terms of security and reliability of the subscribers.

References

Amrutha ,K. , B. Madhu (2014), *An Efficient Approach to Find Best Cloud Provider Using Broker*, International Journal of Advanced Research in Computer Science and Software Engineering 4(7), pp. 943-946, July.

Ben Aissa ,A., R. Abercrombie, F. Sheldon, A. Mili (2010), *Quantifying Security Threats and their Potential Impacts: a case study*, in Innovation in Systems and Software Engineering: A NASA Journal, 6(4):269–281.

Ben Arfa ,L., M. Jouini, A. Ben Aissa, A. Mili (2013), *A Cyber Security Model in Cloud Computing Environments*, International Journal of King Saud University- Computer and Information Sciences, 63-75.

Caron ,E., A. Duang Le, A. Lefray, and C. Toinard (2013), *Definition of Security Metrics for the Cloud Computing and Security-Aware Virtual Machine Placement Algorithms*, International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, IEEE.

Chow ,R., P. Golle, M. Jakobsson, E. Shi, J. Staddon, R. Masuoka, J. Molina (2009), *Controlling data in the cloud: Outsourcing computation without outsourcing control*,

Proceedings of the 2009 ACM Workshop on Cloud Computing Security (CCSW '09), USA, 1-6.

Cloud Security Alliance, Top Threats to Cloud Computing V 1.0, 2010. <https://cloudsecurityalliance.org/topthreats>.

Dhillon ,P., V. Arora (2012), *A Compositional Approach of Reliable and Efficient Cloud Service Selection*, Volume 2, Issue 8, August .ISSN: 2277 128X, International Journal of Advanced Research in Computer Science and Software Engineering.

Foroughi F. (2008), Information security risk assessment by using Bayesian learning technique. Lecture Notes in Engineering and Computer Science 2170, 91-95.

Hanna S. (2009), *Cloud Computing: Finding the silver lining*.

Ibrahim ,A. S., J. Hamlyn-Harris, and J. Grundy (2010), *Emerging security challenges of cloud virtual infrastructure*, the Asia Pacific Software Engineering Conference 2010 Cloud Workshop.

Kapgate ,D. (2014), *Weighted Moving Average Forecast Model based Prediction for Service Broker Algorithm for Cloud Computing*, International Journal of Computer Science and Mobile Computing, vol. 3, Issue. 2.

Linden ,G., B. Smith and J. York (2003), *Amazon.com Recommendations: Item-to-Item Collaborative Filtering*, IEEE Internet Computing, vol. 7, no. 1, pp. 76-80, Jan. /Feb..

Morgan C. (1998), *Programming from Specifications*. International Series in Computer Sciences. Prentice Hall, London, UK,

Sean C. et C. Kevin (2011), *Cloud Computing Security*, International Journal of Ambient Computing and Intelligence, 3(1):14-19, January-March.

Stoneburner ,G., A. Gorguen and A. Fertinga (2002), *Risk Management Guide for Information Technology Systems*, in National Institute of Standards and Technology Special Publication.

Subashini S. et V. Kavitha (2010), *A survey on security issues in service delivery models of cloud computing*, Journal of Network and Computer Applications, 34(1): 1-11.

Subha ,M., M. U. Banu (2014), *A Survey on QoS Ranking in Cloud Computing*, International Journal of Emerging Technology and Advanced Engineering, Volume 4, Issue 2, February.

Xuan ,Z., W. Nattapong, L. Hao and Z. Xuejie (2010), *Information Security Risk Management Framework for the Cloud Computing Environments*, 10th IEEE International Conference on Computer and Information Technology (CIT 2010).

Yuvarani ,R., M. Sivalakshmi (2014), *Achieve Ranking Accuracy Using Cloud Rank Framework for Cloud Services*, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Special Issue 1, March.

Whaiduzzaman ,M., A. Gani (2013), *Measuring Security for Cloud Service Provider: A Third Party Approach*, International Conference on Electrical Information and Communication Technology (EICT), pp. 1-6, IEEE.

Wooley ,P. (2011), *Identifying Cloud Computing Security Risks*, University of Oregon, Master's Degree Program,

Zheng Z., H. Ma, M. R. Lyu and I. King (2011), *QoS- Aware Web Service Recommendation by Collaborative Filtering*, IEEE Trans. Service Computing, vol. 4, no. 2, pp. 140-152, Apr.-June.

Zheng ,Z., X. Wu, Y. Zhang, M. R. Lyu, J. Wang (2013), *QoS Ranking Prediction for Cloud Services*, Parallel and Distributed Systems, IEEE Transactions on, vol.24, no. 6,pp. 1213-1222,June.

Zibin ,Z., Z. Yilei, and M. R. Lyu (2010), *Cloud Rank: A QoS-Driven Component Ranking Framework for Cloud Computing* in Reliable Distributed Systems, 29th IEEE Symposium on, pp. 184-193.

Résumé

Le Cloud Computing est devenu un facteur clé dans la science de l'informatique. Il représente un nouveau paradigme de l'informatique utilitaire et phénomène croissant énormément dans l'industrie des IT et de l'économie dans nos jours. Les utilisateurs de cloud computing (CUs) demandent et exigent des fournisseurs de service cloud du marché d'être sécurisés, fiables et dignes de confiance (CSPs). En fait, c'est un challenge pour un nouveau client de choisir le fournisseur hautement sécurisé. Cependant, un certain nombre de risqué de la sécurité est en train d'émerger en association avec l'utilisation du nuage qui a besoin d'être évalué et géré pour les clients et les fournisseurs. Ainsi, il est nécessaire de déterminer le choix des services appropriés de cloud computing proposés par différents fournisseurs. Dans cet article, nous proposons un modèle de prise de décision basé sur une approche de coût et de risque en mesurant les coûts de défaillance moyen (MFC) pour le groupe des fournisseurs de cloud, ce qui donne la décision au choix du fournisseur le plus fiable entre tous ceux existant dans la marché et justifie les besoins de l'entreprise en termes de la sécurité et la fiabilité des abonnés.

Construction d'une ressource ontologique pour l'annotation des données génomiques

Houda Fyad*, Karim Bouamrane*
Baghdad Atmani*

*Laboratoire d'Informatique d'Oran -LIO
Université d'Oran 1 Ahmed Ben Bella
BP 1524, El M'naouer 31000 Oran, Algérie
houdafyad82@gmail.com, kbouamrane@gmail.com, atmani.baghdad@gmail.com

Résumé. L'amélioration des techniques à haut débit et l'accroissement constant des connaissances en biologie ont conduit à une explosion du volume de données disponibles et accessibles aux chercheurs sur le Web. Ainsi, l'extraction et l'exploitation de ces données par le biologiste est un enjeu majeur en raison de la multiplicité des ressources, l'hétérogénéité et la variabilité des formats, la redondance des nomenclatures, etc… Une approche de fouille de données apporte une solution à cet objectif puisque ces données d'expression sont souvent exploitées pour leur partie profils d'expression mais les informations textuelles associées renseignant le protocole expérimental sont ignorées. Or, le stade de développement lors du séquençage ou les conditions de culture, modifient l'expression et cela influencerait les analyses ultérieures.
Cette communication décrit la construction d'une ontologie du domaine à partir d'un corpus de données biologiques issus de la génomique associé aux aspects expérimentaux des données d'expression de plantes supérieurs.

1 Introduction

L'aboutissement récent de plusieurs projets de séquençage issus des technologies à haut débit de la génomique et de la post-génomique a généré une quantité importante d'informations génomiques. Cette information concerne par exemple : des séquences ou des cartographies de gènes, des structures annotées, des résultats de prédictions ou d'expériences de séquençage de puces à ADN. Aussi, les connaissances qui sont nécessaires à la compréhension de ces mécanismes biologiques, sont sous forme de travaux publiés dans la littérature scientifique écrite en langage naturel.

Dès lors, l'accès à cette information documentaire est un enjeu central pour en extraire les éléments susceptibles de constituer des connaissances pertinentes de telle sorte qu'un biologiste puisse obtenir des réponses claires à une requête spécifique. Ainsi, l'objectif est de fouiller les données d'expression à la lumière des protocoles expérimentaux mis en œuvre à l'aide de termes-candidats offrant des performances intéressantes en termes de rapidité de traitement dont les résultats peuvent conduire par la suite à la construction de vocabulaires contrôlés ou d'ontologies spécifiques au domaine d'application.

Pour prendre en charge le contexte d'une expérimentation biologique, différentes caractéristiques ont été prises en considération. La première concerne l'échantillon biologique. En effet les espèces doivent être précisées, mais aussi leur stade de développement et, si besoin, l'organe ou le tissu étudié. Comme les variations spécifiques des conditions de culture agissent sur la morphologie ou le développement spatio-temporel des organismes, ces aspects « conditions de culture » doivent également être pris en charge. Enfin, un dernier biais pourrait provenir des fonctions des gènes exprimées lors de l'extraction des molécules à séquencer durant les différents protocoles expérimentaux; la spécificité de cette partie technique a donc également été enregistrée.

Dans ce contexte, une ontologie biologique a été construite à partir d'un corpus de données biologiques issus du domaine de la génomique constitué d'articles scientifiques et de commentaires textuels associé aux aspects expérimentaux des données d'expression (données de puces à ADN) de deux plantes modèles *Arabidopsis thaliana* et *Medicago truncatula* dont les données sont accessibles sur NCBI-GEO et Array Express. Pour une meilleure «couverture» du domaine, ces informations ont été complétées par plusieurs articles et thèses qui traitent du cycle de développement des végétaux d'intérêt (voir figure1. ci-dessous).

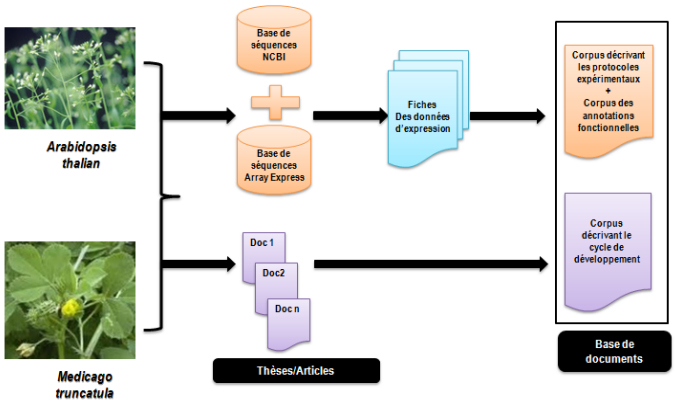


FIG. 1 – *Approche proposée.*

2 Etat de l'art

Le recours à des vocabulaires contrôlés ou à des ontologies s'est fait rapidement sentir pour capturer les concepts biologiques décrivant des objets biologiques tels que les séquences génomiques, les gènes ou encore les produits de gènes. Ces concepts sont issus de publications des résultats du séquençage de génomes et de leurs diverses annotations, par conséquent, l'utilisation de bio-ontologies devient indispensable pour faire face à l'hétérogénéité des données et des sources. Elles permettent d'unifier les différentes définitions pour ainsi améliorer la qualité des données et favoriser le partage et l'échange de données. Aussi, la construction, la fusion, l'utilisation et la réutilisation des ontologies constituent quelques-uns des défis actuels en bioinformatique. Nous présentons dans ce qui suit un état de l'art de la diversité des ontologies biologiques et bioinformatiques, notamment celles dédiées aux végétaux supérieurs et aux annotations et expérimentations des puces à ADN.

2.1 Ontologies biologiques et bioinformatiques

Le projet Gene Ontology (GO) (Berardini et al., 2010) vise à fournir un vocabulaire structuré pour des domaines spécifiques biologiques permettant de décrire les produits des gènes (protéines ou ARNm) des organismes. Il comprend trois ontologies parallèles qui sont de plus en plus utilisées par la communauté bioinformatique: fonctions moléculaires, processus biologiques et composants cellulaires. Le système eVoc (Kelso et al., 2003) associe des données d'expression (issues de puces à ADN, expériences SAGE ou ESTs) à un ensemble de quatre vocabulaires contrôlés orthogonaux appropriés pour décrire et comparer l'expression des gènes chez l'homme et la souris. Le consortium Gene Nomenclature Committee (HUGO) (Wain et al., 2002) qui est né d'une volonté d'uniformiser la désignation des gènes, propose une terminologie pour 29 000 gènes humains. Le projet Open Biomedical Ontology (OBO) (Ghazvinian et al., 2011) a pour but de créer des ontologies de référence dans le domaine biologique et biomédical. La plate-forme NCBO (Whetzel al., 2011) développe et maintient une application web appelée bioportail qui permet aux chercheurs d'accéder et d'utiliser des ontologies biomédicales. Il recense 194 ontologies. La plate-forme Ontology Lookup Service (OLS) (Côté et al., 2010) vise à intégrer des ontologies biomédicales pour les mettre a disposition a travers une base de données unique. Elle contient plus de 80 ontologies.

2.2 Ontologies des végétaux supérieurs

Les deux plantes modèles d'intérêt sont des organismes à cycle de reproduction court et faciles à gérer dans les laboratoires. Plusieurs ontologies décrivant les végétaux existent, mais les objectifs, l'espèce, ou le spectre diffèrent de nos besoins. Plant Ontology (PO) (Walls et al., 2010) est une extension du paradigme de la GO dédiée aux plantes. Son objectif est la production de vocabulaires contrôlés structurés et d'ontologies à partir utilisant comme ressources des informations contenues dans les bases de données de plantes et de la connaissance de la biologie des espèces végétales (par exemple le développement, l'anatomie, la morphologie, etc.). Plant Growth Stage Ontology (PGSO) (Pujar et al., 2006) décrit les différents stades de développement (les étapes spatio-temporelles) des plantes à fleur : de la germination à la sénescence. Il fournit aussi une plate-forme commune pour annoter la fonction du gène et l'expression génique par rapport à la trajectoire (courbe) développement des organismes. Plant Structure Ontology (PSO) (Ilic et al., 2007) est une ontologie développée pour la représentation la structure (l'anatomie et la morphologie) des plantes à fleur. Ainsi que l'annotation des profils d'expression génique et phénotypes du matériel génétique des angiospermes. Elle fournit une plate-forme pour l'annotation des profils d'expression génique et phénotypes du matériel génétique des angiospermes. The Plant Ontology Database (Avraham et al., 2008) est une base de données regroupant les informations contenues dans les deux précédentes ontologies Plant Growth Stage Ontology (GSO) et Plant Structure Ontology (PSO), elle utilise aussi comme ressources Cell Ontology (CL) et les données d'annotation d'expression des gènes. Cependant, ces ontologies sont parfois trop générales ou trop spécialisées d'où la construction de notre propre ontologie du domaine.

2.3 Ontologies d'annotation et des expérimentations des puces à ADN

Gene Ontology précédemment cité, est la ressource essentielle utilisée pour l'annotation. Elle est ainsi employée par les nombreux portails (GenBank, RefSeq, UniProt, PDB, TAIR,

etc.), mais aussi dans des portails dédiés. Gene Ontology Annotation (GOA) (Huntley et al., 2015) est un portail dédié à l'annotation des données d'expression de divers organismes d'intérêt en utilisant GO. AmiGO (Carbon et al., 2009) est un portail qui permet d'accéder aussi à GO, il contient notamment de nombreuses références croisées vers d'autres systèmes d'information. La plupart de ces portails lient des annotations de GO à leurs produits de gènes. Aussi, le téléchargement de termes d'annotation de génomes d'espèces « modèles » de la Gene Ontology est proposé avec de nombreuses références croisées correspondant à ces annotations, fait que cette ontologie est omniprésente dans le quotidien du chercheur en biologie moléculaire et cellulaire. Aussi, une description formelle des expériences est extrêmement importante pour l'organisation et l'exécution des expériences en biologie. Par exemple, les puces à ADN du projet Micro-array Gene Expression Data (MGED) (Griffin et Steinbeck, 2010) prévoient des termes pour annoter tous les aspects d'une expérience de puces à ADN de sa conception avec la définition des hybridations, à la préparation de l'échantillon biologique et des protocoles utilisés pour l'hybridation sur la puce et à l'analyse des données. Les termes MGED sont organisés sous la forme d'une ontologie. Ils permettent des requêtes structurées concernant les expériences (Griffin et Steinbeck, 2010).

3 Approche proposée

Dans cette étude, une approche bottom-up a été suivie pour l'extraction de termes issus des données afin de construire une ressource ontologique afin d'identifier les informations associées au contexte des expériences qui ont conduit aux données de puces à ADN, depuis l'extraction des molécules jusqu'à leur séquençage (Fyad et al., 2011).

Un corpus biologique issus du domaine de la génomique a été constitué de 500 documents. Ce dernier est composé de deux sous-corpus : un premier corpus décrivant les expérimentations de puces à ADN et un deuxième corpus décrivant le cycle de développement des deux végétaux. Cette extraction automatique des termes-candidats a été effectuée grâce à un extracteur appelé KEA (Automatic Keyphrase Extraction) qui par une approche statistique repose sur la fréquence d'apparition du terme candidat dans le texte.

L'ensemble de tous les termes-candidats dans un document sont identifiés à l'aide du traitement lexical, des métriques sont calculées pour chaque terme, et un apprentissage automatique est utilisé pour générer un classificateur qui détermine les termes qui devraient être assignés comme étant des termes clés. Deux métriques sont calculées dans l'algorithme : « TF×IDF » et « First occurrence ». L'évaluation de la performance de l'outil KEA à travers différentes expérimentations a montré que (Fyad et al., 2013) :

- KEA est plus performant sur les textes intégraux concernant le cycle développement des deux plantes (1.98 termes-candidats extraits) que sur résumés (0.98 termes-candidats extraits).
- KEA est plus performant avec un nombre de documents importants pour l'extraction de termes-candidats (1.41 termes-candidats extraits pour 5 documents et 1.49 termes-candidats extraits pour 20 documents).

4 Construction de la ressource ontologique

Après les étapes de standardisation des extractions et de sélection des termes, l'ontologie est construite selon la méthode proposée par l'Université de Stanford, car elle comporte des phases claires, simples et faciles à comprendre (Noy et McGuinness, 2002). L'éditeur d'ontologie « Protégé » a été également utilisé. Tout comme dans l'approche eVoc (Kelso et al., 2003), il a été décidé de créer trois ontologies afin de caractériser de manière complémentaire et quasi indépendante les aspects clés du contexte d'une expérience. Tous les termes choisis sont inclus dans l'ontologie et classés entre concept, propriété ou valeur d'instance d'un concept ou d'une propriété. Des liens de relation « is-a » entre les concepts sont établis. Cette approche, bien que comprenant des étapes manuelles, constitue néanmoins un exemple de construction d'ontologies à partir d'une extraction statistique de termes. Les termes sélectionnés ont été répartis dans les trois aspects du contexte d'une expérience d'expression des gènes : caractérisation de la plante étudiée, les conditions de culture maintenues jusqu'à l'extraction des molécules exprimées, l'organe ou le tissu extrait, et les stades de développement de l'organisme au moment de l'extraction.

Les figures 2 à 4 sont des vues schématiques des trois ontologies créées (les nœuds en jaune représentent les concepts tandis que les nœuds en orange sont des exemples d'instances des concepts). L'ontologie « cycle de développement » a cinq niveaux de ramification, représente les différentes étapes du cycle de la plante y compris les stades sexués et asexués. L'ontologie « conditions de culture », dispose de trois niveaux de ramification. Ces caractéristiques sont : Le milieu de culture qui peut être de croissance, ou de croisement, contenir des éléments nutritifs comme des vitamines, des hormones, etc. L'aspect rythme biologique est aussi pris en compte (Par exemple : le rythme circadien qui dure environ 24h), Et enfin, les stress subis par les plantes lors de leurs développements peuvent être dus à une variation de température, de la concentration d'oxydant ou de sa réponse face à un pathogène ou un parasite. L'ontologie « annotation fonctionnelle des gènes » a cinq niveaux de ramification. Ces annotations concernent l'aspect génotypique y compris le développement des plantes, l'aspect stress comme la température, de la concentration d'oxydant ou de sa réaction face à un métal lourd, ainsi que la luminosité constante ou prolongée.

FIG. 2 – *Ontologie « Cycle de développement des deux plantes ».*

FIG. 3 – *Ontologie « Mode de culture des deux plantes ».*

FIG. 4 – *Ontologie « Annotation fonctionnelle des gènes des deux plantes ».*

5 Conclusion

Beaucoup d'ontologies ont été créées pour représenter différentes bases de données d'organismes modèles et pour fournir un système d'annotation partagé afin de décrire certains aspects de la biologie des organismes. Les ontologies sont déjà utilisées par des fournisseurs

de données publiques et privées en tant que méthode pour annoter et référencer les informations des gènes et des produits des gènes. Les projets d'ontologies ont amélioré et promu le développement de stratégies efficaces pour la présentation et l'interrogation au travers de classifications étendues. Dans cet article, une contribution au domaine ontologique est présentée par la construction d'un modèle de connaissance à partir d'un corpus textuel de données biologiques. La spécificité du domaine de connaissance réside d'une part dans les deux plantes qui sont représentés, *Arabidopsis thaliana* et *Medicago truncatula*, et d'autre part dans la gestion du contexte des expériences des données d'expression. Une approche pour la construction d'un modèle de connaissance basé sur une ontologie a été définie permettant de capturer la sémantique associée au contexte expérimental des études d'expression des gènes. L'information a été extraite par une approche statistique d'extraction des termes. La représentation des connaissances qui en résulte est modulaire et se compose de trois ontologies permettant flexibilité et facilité des mises à jour. Toutefois, des améliorations doivent encore être effectuées avant qu'elles puissent être effectivement opérationnelles sur les ressources biologiques, en particulier en ce qui concerne la couverture des documents spécialisés, ou l'évaluation de l'ontologie résultante et de son utilisation.

Références

Avraham, S., Tung, C., Ilic, K., Jaiswal, P.,Kellogg, E.A et al, (2008).“The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations”, Nucleic Acids Research, Vol. 36, Database issue D449–D454.

Berardini, TZ., Li, D., Huala, E., Bridges, S., Burgess, S., McCarthy, F et al, (2010). “The Gene Ontology in 2010: extensions and refinements”, Nucleic Acids Res, Vol. 38, (Database issue): D331-D335. (cf: http://www.geneontology.org).

Carbon, S., Ireland, A., Mungall, C. J, Shu, S., Marshall, B., Lewis, S et al, (2009). “AmiGO: online access to ontology and annotation data”, Bioinformatics, Vol. 25, pp. 288–289. (cf : http://amigo.geneontology.org).

Côté, R., Reisinger, F., Martens, L., Barsnes, H., Vizcaino, J.A., Hermjakob, H., (2010). “The Ontology Lookup Service: bigger and better”. Nucleic Acids Research, Vol 38 (Web Server issue):W155-W160. (cf : http://www.ebi.ac.uk/ontology-lookup).

Despres, S.,et Szulman, S., (2008). “Réseau terminologique versus Ontologie”. Revue TOTh 2008, p. 6-7.

Fyad, H,, Bouamrane, K., Atmani, B., Toffano-Nioche, C., (2011). “Construction ontologique à partir de séquences d'expression de champignons”, Revue RNTI EGC 2011, p. 299-300.

Fyad, H,, Bouamrane, K., Atmani, B., (2013).“Keyphrase extraction from a corpus of biological data”, Journées Doctorales Laboratoire d'Informaique d'Oran JDLIO'13.

Ghazvinian, A., Noy, N.F., Musen, M.A., (2011).“How orthogonal are the OBO Foundry ontologies? Journal of Biomedical Semantics”. 2011;Vol 2(Suppl 2):S2. doi:10.1186/2041-1480-2-S2-S2. (cf : http://www.obofoundry.org/)

Griffin, J.L., Steinbeck, C., (2010).“So what have data standards ever done for us? The view from metabolomics. Genome Medicine”; Vol 2(6):38.

Huntley, R.P., Sawford, T., Mutowo-Meullenet, P., et al, (2015). “The GOA database: Gene Ontology annotation updates for 2015”, Nucleic Acids Research, Vol 43 (Database issue):D1057-D1063. (cf : http://www.ebi.ac.uk/GOA).

Ilic, K., Kellogg, E.A., Jaiswal, P., Zapata, F., Stevens, P.F., Leszek, P., et al, (2007).“Plant Structure Ontology. Unified vocabulary of anatomy and morphology of a flowering plant”, Plant Physiology, Vol. 143, pp. 587–599.

Kelso, J., Visagie, J., Theiler, G., Christoffels, A., Bardien, S., Smedley, D., Otgaar, D., Greyling, G., et al, (2003). “eVOC: A Controlled Vocabulary for Unifying Gene Expression Data”. Journal of Genome Research. 13:1223–1227.

Noy, N.F., et McGuinness, D., (2002). “Développement d'une ontologie 101: Guide pour la création de votre première ontologie”. Stanford (USA).

Pujar, A., Jaiswal, P., Kellogg, E. A., Ilic, K., et al, (2006). “Whole plant growth stage ontology for angiosperms and its application in plant biology”, Plant Physiology, Vol. 142, pp. 414–428.

Wain, H.M., Lovering, R.C., Bruford, E.A., Lush, M.J., Wright, M.W., Povey, S., (2002). “Guidelines for Human Gene Nomenclature”, Journal of Genomics, Vol. 79, Num. 4.

Walls, R. L., Cooper, L. D., Elser, J., Stevenson, D.W., Smith, B., Mungall, C., et al, (2010).“The Plant Ontology: A Common Reference Ontology for Plants”, Conférence Bio-Ontlogies 2010: Semantic Applications in Life Sciences. (cf: http://www.plantontology.org).

Whetzel, P.L., Noy, N.F., Shah, N.H., et al, (2011). “BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications”. Nucleic Acids Research, Vol. 39 (Web Server issue):W541-W545. (cf : http://www.bioportal.bioontology.org/).

Summary

The improvement of high throughout techniques and the constant increase of knowledge in biology have led to an explosion in the volume of data available to researchers and accessible on the Web. Thus, the extraction and use of these data by the biologist is a major issue because of the multiplicity of resources, heterogeneity and variability of formats, nomenclatures redundancy, etc ... A data mining approach provides a solution to this objective since these expression data is often exploited for their expression profiles but the associated text information about the experimental protocol are ignored. Now, the stage of development when sequencing or culture conditions modify the expression and this would influence subsequent analyzes.

This paper describes the construction of ontology from a biological data from genomics domain associated with experimental aspects of the expression of plant data.

A semi-automatic solution for enrichment XML query response using terminological domain ontology

Abdelmadjid Larbi [1,2] , Bachir Seddiki [2] , Smail Larbi [2], Rached Krim[1]

[1] ENERGARID Laboratory, Tahri Mohamed University, Bechar
AMDlarbi@gmail.com,
Rached.krim@gmail.com
[2] Bechar, Tahri Mohamed University
Bachir_html@gmail.com ,
Larbismail@gmail.com

Abstract The need to clarify the data semantics in various fields of science is introduced by definition data referring to ontological data base (OBDB). With the proliferation of domain ontologies, and the large volume of data to be handled, the need emerged for systems that manage data based on ontological large, such management systems are called database management systems based ontological. We can operate such system via the web if we think represent them as XML. Which allows us to:

- To exploit OBDB via the Internet.
- To enrich the answers to XML queries using domain terminology ontology.

To enable to highlight the usefulness of introducing semantics into the database and to enrich query responses, we have applied this to the pharmaceutical databases case.

1. Introduction

In recent years the database research field has concentrated on XML(eXtensible Markup Language) as a flexible hierarchical model suitable to represent huge amounts of data with no absolute and fixed schema, and a possibly irregular and incomplete structure [1].
There are two main approaches for XML as an information system where the separation data / document is not very clear. Part of a textual description of nature is "document" (Example: all text with a cinema, activities, etc.) Another part is contained in a database. The XML language is a way to make public (= publish) all or part of a database. The most important factor in the choice of a database may be whether the base is used to store data or documents. There are two broad application areas of XML technologies. The first relates to document-centric applications, and the second to data-centric applications. Because XML can be used in so many different ways, it is important to understand the difference between these two categories: Document-Centric XML and Data- Centric XML.

1.1 The data-centric

Data-centric documents are documents that use XML [2] as a data transport. They are designed for machine consumption and the fact that XML is used at all is usually superfluous. That is, it is not important to the application or the database that the data is, for some length of time, stored in an XML document. Examples of data-centric documents are sales orders, flight schedules, scientific data, and stock quotes.
Data-centric documents are characterized by fairly regular structure, fine-grained data (that is, the smallest independent unit of data is at the level of a PCDATA-only element or an attribute), and little or no mixed content. The order in which sibling elements and PCDATA occurs is generally not significant, except when validating the document.
For example, the following sales order is data- oriented :

```
<OrdreDeVentes NumeroOrdreDeVentes="12345">
<Client NumeroClient="543">
<ClientName> ABC Industries </ <CustomerName>
<Rue> 123 Main St. </ Street>
<city> Chicago </ City>
<Status> IL </ state>
<CodePostal> 60609 </ PostalCode>
</ Client>
<DateOrdre> 981215 </ DateOrdre>
<item NumeroItem="1">
<Lot NumeroLot="123">
<Description>
<p> <b> Turkey wrench: </ b> <br /> Stainless steel, one-piece construction,
        lifetime guarantee. </ p>
</ Description>
<Price> 9.95 </ price>
</ Lot>
<amount> 10 </ Quantity>
</ Item>
</ SalesOrder>
```

1.2 The document-centric documents

Document-centric XML document creation is a process [3] of marking up textual content rather than typing text in a predefined structure. It turns out that, although the final document has to be valid with respect to the DTD/Schema used for the encoding, the "in-progress" document is lmost never valid. At the same time, it is important to ensure that at each moment of time, the editor is working with an XML document that can be enriched with further markup to become valid.

The following document, for example, describes a product and it is oriented document:

```
<¨Produit>
<Intro> The <ProductName>Turkey Wrench</ProductName> from <Developer>Full Fabrication Labs,
Inc.</Developer> is <Summary>like a monkey wrench,
but not as big.</Summary>
</Intro>
<Description>
```

```
<Para>The turkey wrench, which comes in <i>both
right- and left handed versions (skyhook optional ) </i>, is made of the <b>finest
stainless steel</b>. The Readi-grip rubberized handle
quickly adapts
to your hands, even in the greasiest situations. Adjustment is
possible through a variety of custom dials.</Para>
<Para>You can:</Para>
<Liste>
<Item><Link URL="Order.html">Order your own turkey
wrench</Link></Item> </Liste>
<Para>The turkey wrench costs <b>just $19.99</b> and, if you
order now, comes with a <b>hand-crafted shrimp
hammer</b> as a bonus gift.</Para>
</Description>
</Produit>
```

Query languages used in this case are of two types:
1 – The query languages based on models: A Native XML database (NXD in English) is a database that is based on the data model provided by XML. It typically uses XML query languages such as XPath or XQuery.
2 - The query language based on SQL query languages based on SQL SELECT statements use modified the results are transformed into XML.
In this work, we are more interested in data- centric XML or database that are more or less structured with a regular structure to differentiate the XML documents that are characterized by a less regular or even irregular. The need to clarify the semantics of the data in various fields of science (biology, medicine, geography, engineering,..) is introduced by definition data referring to ontologies, also called ontological data base. With the proliferation of domain ontologies, and the large volume of data to be handled, the need emerged for systems that manage data based on ontological large, such management systems are called systems management database based ontological (OBDB). We can operate such system via the web if we think represent them as XML.
Which allows us to :

- To exploit OBDB via the Internet
- To enrich the answers to XML queries (or other OBDB BD in XML format) using domain the ontologies terminology.

1.3 Defining a native XML database:

We require a native XML database to have the following two properties [4,5] as well:

– The XML data model (either in the XML Infoset or the XQuery/XPath Data Model) is the fundamental logical data model both used internally by the database and exposed to database users when XML is the data type.

– The XML data model is the fundamental unit of physical storage of all XML data, without mapping to a different data model.

This narrowed definition means that XML is more than an externalized data type - it is how the data is handled both logically and physically. The data is represented as XML right

down to its physical storage schema on the disk. This model is the best for efficient searching of the XML data.

This vision conforms to the logic model can consider XML easily using defined standards around XML (XQuery, XPath, XSLT, XUpdate) to access and process the data in the database.

1.4 Characteristics of a native XML DB

We can briefly some characteristics of databases Native XML. This should help the reader by giving an idea of features available today and those expected in the future.

2. The concept of Ontology

Gruber [6] introduced the notion of ontology as "an explicit specification of a conceptualization" Borst. A. has slightly modify the definition by combination of the two definitions can be summarized as follows: "an explicit and formal specification of a shared conceptualization".
This definition explains: “explicit” means that the "type of concepts and constraints on their use are explicitly defined ". “formal” refers to the fact that the specification should be readable by a machine, refers to the shared notion that an ontology "captures consensual knowledge, which is not specific to an individual but by a group validated" conceptualization refers to "an abstract model of some phenomenon in the world based on identification of relevant concepts of this phenomenon".
An ontology provides a solid foundation for communication between machines but also between humans and machines in defining the meaning of First of all objects through symbols (words or phrases) that and characterized means and then through a structured representation or formal role in the field.
Ontologies are used in many fields. Areas identified in 1998 by Guarino are engineering knowledge, qualitative modeling, language engineering, databases design, information retrieval, information extraction, management and the organization of knowledge.
Since then, thanks to the growth of the Web, they are used in the domain of ecommerce and are at the heart of the Semantic Web, the future version of the current Web.
One of the biggest projects based on the use of ontologies is add to a real web layer enabling knowledge research information at the semantic level and at the simplest lexical and / or syntactical level. Ultimately, it is expected that applications deployed on the Internet may lead reasoning using the knowledge stored on the web.

2.1 Ontologies Classification

Several classifications have been defined and are based on different criteria. We retain the classifications proposed by G. Van Heijst 97 and al. They have proposed two types of ontologies classification [7] according different criteria.
The first classification is based on the types and wealth of structures used in the ontology. According to these criteria, there are three categories of ontologies:

- The terminology ontologies that are used to specify the terms of the vocabulary of a field of knowledge.
- Information ontologies that specify the structure / schema of a database to enable the storage of information.
- Ontologies that model of knowledge that offer internal structures that are richer and more defined according to their uses such information sharing. It also proposes a classification of ontologies that relies on decision account of "objectives" of modeling. It holds four categories ontologies according to this criteria:
- Ontologies applications that specify the necessary information to one or more particular applications.
- Domain ontologies that express the conceptualization of a particular domain knowledge.
- Generic ontologies that model of transverse knowledge to different areas. Typically generic ontologies define concepts such as the concepts of state, event, action, etc.
- Ontologies representation that seek to explain the conceptualizations underlying the knowledge representation formalisms. They represent the real-world entities without a priori, to "neutral" way.

This classification is illustrated in Figure 1

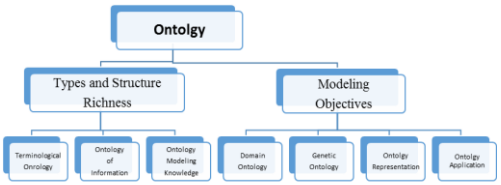


FIG. 1 : ONTOLOGY CLASSIFICATION

2.2 Ontology languages

The Ontologies play a very important role in the Semantic Web proposed by Tim Berners-Lee to specify are formally vocabularies for describing Web content. Thus, the information Web can be accessed and understood by both humans and by machines. Several languages [8] have been proposed to represent ontologies with a higher or lower expressiveness and complexity of reasoning more or less. We are interested in three languages representation of ontologies or recommended subject to recommendation by W3C (World Wide Web Consortium) RDF / RDFS , OWL [MvH04] and SWRL.

a-RDF and RDFS languages

RDF (Resource Description Framework). It is a language for describing the semantics of data in an understandable way by machines. The RDF data model is composed of three types of components:
- Topic: it is necessarily identified by a URI.
- Predicate or property: it is a property used to characterize and describe a resource. A predicate is necessarily identified by a URI.
- Re: is data or another resource (identified by a URI). Using the RDF, information resources

are described by a set of RDF statements in the form of triples: (05) (Subject, Predicate, Object).

This notation is called triplet N-TRIPLE. XML syntax has been defined to express RDF statements, it is called RDF / XML.

3. State of art

There are many works including the ontology for treating the problem of including the semantic to the query [9],

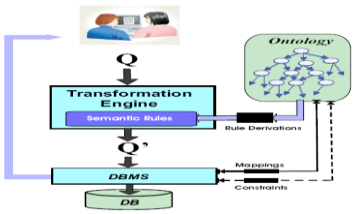


FIG. 2 : SEMENATIC QUERY TRANSFORMATION USING ONTOLOGIES

Some works include ontology for using semantic query over heterogeneous databases [10],

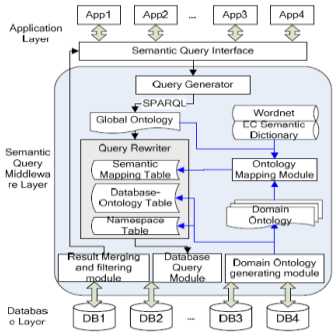


FIG. 3 : AN ONTOLOGY-BASED SYSTEM FOR SEMANTIC QUERY OVER HETEROGENEOUS DATABASES

Some others include ontology for answering queries using views[11]

4. Design of the proposed solution

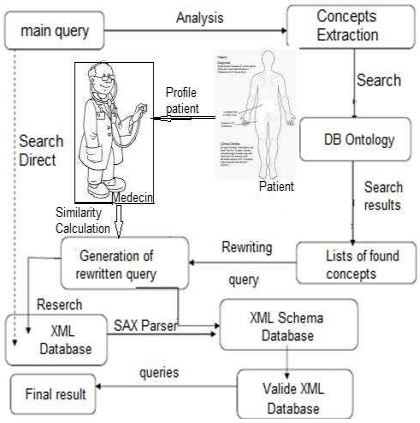


FIG 4. QUERY REWRITE MODEL BASED ON ONTOLOGY

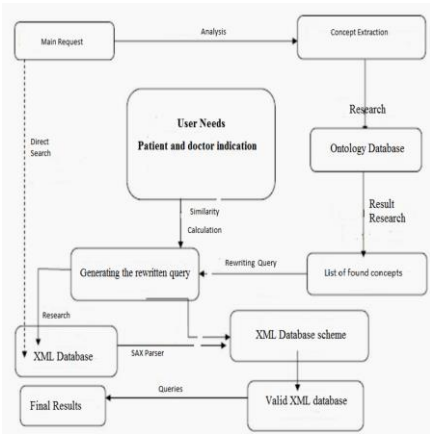


FIG.5 : GRAPH REWRITING QUERY IN AN XML DATABASE BY USE OF A PHARMACEUTICAL TERMINOLOGY ONTOLOGY

With the help of the editor Protege2000 we can create our ontology following the information given by the expert.

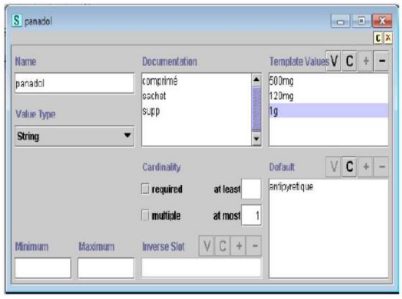


FIG.6 : CREATING CLASSES OF OUR ONTOLOGY.

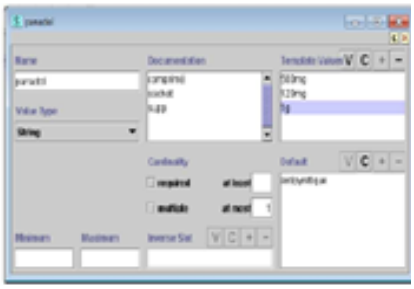


FIG.7 : CREATING INSTANCES OF OUR ONTOLOGY

The sequence diagrams are used to represent interactions between objects in a temporal point of view. The focus is on the chronology of sending messages.

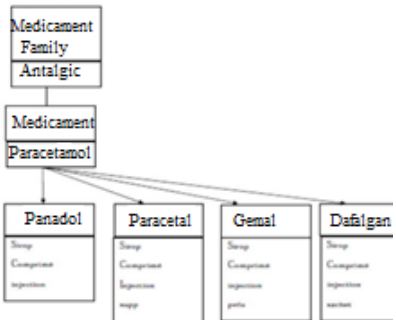


FIG. 8 : DIAGRAM OF CLASSES AND THEIR HIERARCHIES OF A DRUG (GENERAL CASE)

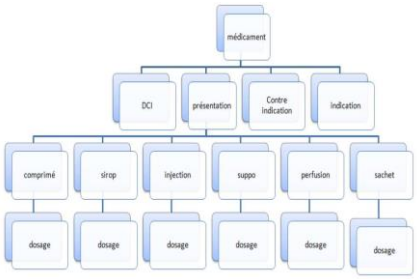


FIG. 9 : DIAGRAM OF CLASSES AND THEIR HIERARCHIES OF DRUG PANALGAN.

The application supports multiple paths to get results in this example there are two voices: 1 Direct Request. In this case the application directly accesses the database.

Semantics Query : Accesses the ontology query: Terminology and regenerate other queries against the XML database in this case the resulting query will closer to the original query in a field and in the user selection.

FIG.10 : PRESENTATION OF OUR XML-BASED ONTOLOGY DATABASE

It can be seen from the example that there is a correct result ie represents the execution of the main query directly on the basis of XML data, and other results represent the query execution generated by the use of ontology.

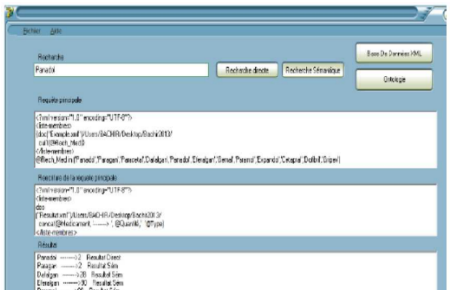


FIG. 11 : DISPLAY SEMANTIC RESULTS ON OUR APPLICATION

5. Conclusion

Our query expansion mechanism exploits the richness of relationships semantics provided by ontologies. Its adaptation to the user by the through prototypicality (represented by the terminology, by weights on the terms) allows you to customize both the extension query input the amount of results provided.
Our solution allows the time to customize the user query by patient requirements and directed by the doctor on one hand and according to the concepts provided by the ontology Terminology in the other (validated by pharmaceutical expert) in order to have a comprehensive response. The first results are encouraging and it is estimated continuity of this project in the future.

References

1. Mirjana mazuran, Elisa quintarelli, and Letizia tanca, data mining for xml query-answering support, ieee transactions on knowledge and data engineering, 2011
2. Ronald bourret , XML and databases, 1999-2003, felton, CA 95018 USA
3. IONUT E. IACOB, On potential validity of document-centric xml documents,, university of kentucy, data engineering workshops, 2006
4. petr kol´aˇr and pavel loupal, Comparison of native xml databases and experimenting with inex,dept. Of computer science and engineering fee, Czech technical university
5. W. Meier. Exist: An open source native XML database. Springer berlin heidelberg, 2003
6. Gruber, t. R. A translation approach to portable ontology specifications. Knowledge acquisition, 5, (2):199-220, 1993.
7. Guarino n. & Giaretti p. Ontologies and knowledge bases: towards a terminological clarification. In towards very large knowledge bases. N. J. I.n mars, ed., ios press: 25-32, 1995.

8. Fakhr-eddine Hachemi. Description sémantique des objets d'apprentissage à base de modèles de contenu, mémoire de magistère, université Abou Bekr Belkaid. Tlemcen-algérie, 2007.
9. Necib.C.B ,Freytag .Johann-Christoph, Semenatic Query Transformation Using ontologies . (2005)
10. Chen, Donglin, An ontology-based system for semantic query over heterogeneous databases (2009)
11. Alon Y. Levy ,Answering Queries Using Views, (2000)
12. Barbier f., Henderson- & al.. Formalization of the whole-part relationship in the unified modeling language. 2003, IEEE transactions on software engineering, 29(5), 459-470.
13. Guillermo Valente, Gomez-Carpio, & al. A query expansion methodology in a cooperation of information systems based on ontologies. 2009 - Proceedings of the fifth international conference on web information systems and technologies, pages 256{262. Insticc press,march 2009 .Guelfi, n., c.

Résumé

Le besoin d'éclaircir la sémantique des données est introduit par la définition de donnée à base ontologique. On peut l'utiliser pour enrichir les réponses de requête XML via le web. Ce travail présente le cas de base de données pharmaceutique.

Solution Logicielle Intégrée pour la Découverte et le Diagnostic de L'organisation de l'Entreprise

Zineb BESRI, Azedine BOULMAKOUL

Faculté des Sciences et Techniques de Mohammedia LIM/IDS B.P. 146 Mohammedia Maroc
z.besri@gmail.com, azedine.boulmakoul@gmail.com

Résumé. Toutes les organisations sont amenées à mettre en place une gouvernance adaptée et conforme aux lois applicables et aux bonne pratiques en la matière. L'audit organisationnel permet d'identifier les dysfonctionnements potentiels au sein des organes de directions, d'administration et des systèmes d'information. Cet article propose une approche originale et pratique centrée sur l'analyse structurale et sur l'analyse des réseaux sociaux pour l'évaluation de la structure organisationnelle de l'entreprise. Ce travail de recherche s'inscrit dans l'audit organisationnel. La problématique touche les aspects opératoires de la mise en œuvre des méthodes analytiques issues de l'analyse structurale pour le re-engineering organisationnel de l'entreprise sur la base de son système d'information existant. La méthode proposée découle des pratiques canoniques, utilisant la méthode, analyse simplicial, qui empreinte ses fondements de la combinatoire et de la topologie algébrique discrète. Ainsi que des pratiques des techniques d'analyse des réseaux sociaux pour la mesure de la centralité du réseau professionnel que l'entreprise définit. Nous proposons dans ce travail de recherche une solution logicielle intégrée pour la découverte et le diagnostic de la structure organisationnelle de l'entreprise.

Keywords: Ontologie de l'organisation, analyse structurale, topologie algébrique, analyse des réseaux sociaux, No-SQL, Base de données graphe,

1 Introduction

L'audit organisationnel a désormais un rôle majeur dans l'amélioration de la gouvernance d'entreprise. Selon les nouveaux standards de l'Institute of International Audit (IIA 2014). L'audit organisationnel permet de formuler des recommandations appropriées a fin de garantir : la promotion d'une éthique et de valeurs adaptées à l'organisation, la gestion efficace du management, la coordination efficace des activités et une communication pertinente des risques et des contrôles. L'audit organisationnel est un domaine particulier d'application des méthodes d'audit, qui sont largement utilisées dans le domaine financier, social, qualité, etc. Il est pratiqué par des auditeurs externes spécialisés ou par des auditeurs internes, parfois en coopération entre les deux.

Conceptuellement, l'audit organisationnel consiste à offrir des solutions à l'entreprise en difficulté ou souhaitant s'améliorer, par une approche concrète et technique de son organisation : ressources humaines, moyens de production, circuits de distribution, etc. Le tout en alliant les acquis et l'application de méthodologies faisant appel à l'innovation.

Concrètement, Ce procédé de diagnostic organisationnel, aura comme entrée une mission de refonte organisationnelle. Il permet au top management de préparer un plan d'évolution. Ce plan est basé sur une vision globale et objective des activités arrêtées au cours d'un exercice budgétaire.

Ce travail de recherche permet d'assister l'entreprise dans tous ses secteurs d'activité pour rechercher l'amélioration de son organisation : sa structure, ses méthodes, son emploi des ressources humaines et ses méthodes de gestion et de travail.
Il se charge d'abord d'organiser et d'inventorier les différents acquis de l'entreprise utilisant les nouvelles et les dernières tendances des systèmes d'information Big Data (O'Reilly Media, 2012).
Puis stocker l'ensemble des informations organisationnelle dans un système de gestion de base de données NoSQL (Redmond & Wilson, 2012), spécifiquement les bases de données orientées graphes. Par la suite le procédé prépare les données à analyser à travers un composant structural d'extraction, de transformation et de chargement SETL. Le procédé se base sur deux approches d'analyse (Boulmakoul & Besri, 2014). La première démarche est fondée sur l'analyse structurale exploitant des outils issus de la topologie algébrique. La seconde démarche fait référence à l'analyse des réseaux sociaux. Ce système permet de faciliter la réorganisation de la structure de l'entreprise et de déployer la refonte de l'organisation découverte. Notre solution permet aussi d'étudier l'impact de chaque réorganisation ou changements proposés. Le framework d'analyse fonctionne comme instrument de diagnostic entre les mains des gestionnaires et du comité de pilotage de l'entreprise. Leur permettant d'identifier les faiblesses structurelles et organisationnelles ainsi que des problèmes spécifiques qui peuvent en découler.

Cet articlé est structuré comme suit : section 2 présente l'architecture logicielle où nous présentons les composants de la solution logicielle. La section 3 propose la fiche produit de la solution et les différentes étapes du procédé de diagnostic. Puis un scénario de test prenant à titre d'exemple une entreprise de production. Cette section propose quelques interfaces de la solution développée. Enfin une conclusion pour synthétisé notre travail de recherche.

2 Architecture Logicielle

2.1 Architecture globale de la solution

Nous avons développé au cours de ce travail de recherche une solution logicielle "Structural Engine " pour l'analyse structural de l'organisation des entreprise. Structural Engine se compose d'une interface homme machine permettant à l'utilisateur de charger la structure organisationnelle réelle de l'entreprise, de choisir un ensemble de processus et de relation à analyser, de sélectionner les différentes métriques à calculer (Topo-Scopie ou/et SNA-Scopie), de récupérer un rapport d'analyse, puis de proposer quelque opération de réorganisation de la structure organisationnelle de l'entreprise. La figure suivante schématise cette vue globale du système développé.

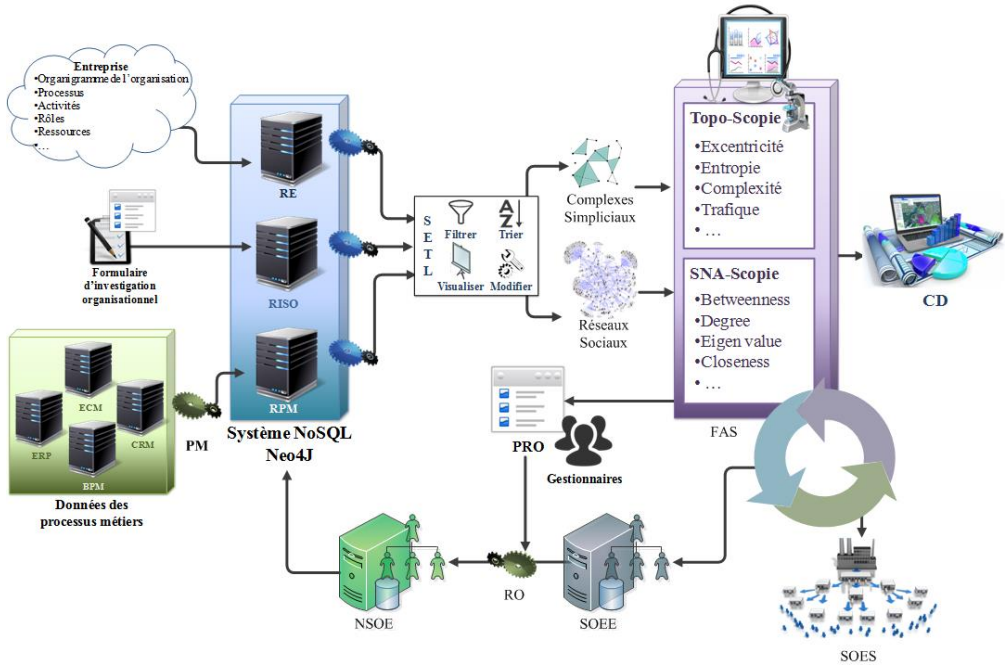


Figure 1 Architecture globale de la solution logicielle

Nous avons développé au cours de ce travail de recherche une application web « *Organizational diagnosis* » qui est sous format d'un ensemble de service web. Permettant à travers des onglets de fournir les services suivants : charger la structure organisationnelle et de sélectionner un ensemble de processus et de relations à analyser ; effectuer l'analyse Topo-Scopy et choisir les indicateurs structuraux à évaluer ; effectuer l'analyse SNA-Scopy et choisir les mesures de la centralité pour mesuré les réseaux de l'entreprise ; visualiser les résultats de l'analyse en mode graphique en utilisant les API adéquates pour chaque mesure calculée ; télécharger la carte de diagnostic de l'analyse effectuée.

2.2 Fiche produit et composants logiciels

La réutilisabilité d'un prototype logiciel constitue un des facteurs les plus concourants à sa qualité. Structural Engine s'appui sur quatre packages de bases, un qui gère la collecte des données d'analyse. Un deuxième package gère la persistance des données. Un troisième gère le traitement et l'analyse des données de l'organisation de l'entreprise. Et un quatrième qui gère la construction et l'édition de la carte de diagnostic organisationnel.

Dans ce qui suit, nous présenterons le diagramme de composants du système (figure 2) et détaillerons les composants logiciels proposés pour l'analyse et diagnostic de l'organisation de l'entreprise.

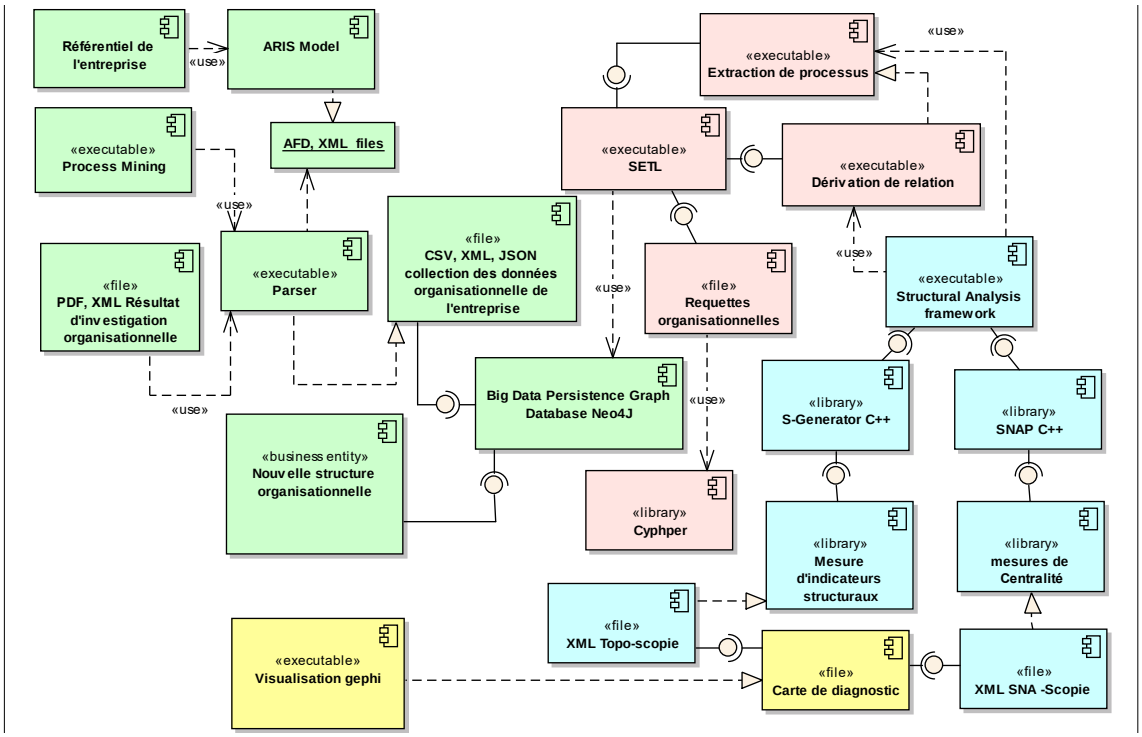


Figure 2 Composants logiciels du prototype de diagnostic organisationnel

Le processus d'évaluation et d'amélioration de l'organisation déclinée à travers six étapes est schématisé dans la figure 2. Ces composants se distinguent par leurs fonctionnalités principales. Tout d'abord :

Etape 1- Collecte des données organisationnelles de l'entreprise

En premier lieu le processus doit collecter les données liées à l'organisation et à la structure organisationnelle de l'entreprise. Ces données peuvent être récupérées à partir : du référentiel de l'entreprise, ce référentiel intègre la structure organisationnelle déclinée par le top management de l'entreprise, des formulaires de l'investigation organisationnelle qui peuvent être produites dans un support de type papier ou numérique. En particulier les données produites référent les résultats de l'investigation effectuée sur la structure organisationnelle, de l'ensemble des données des processus métier de l'entreprise (ERP, ECM, BPM, CRM, système transactionnel …), issues du Process mining. La création de la base de données de l'entreprise à diagnostiquer est une condition préalable pour le système inventé. Cette étape consiste à créer une base de données intégrant toutes les données recueillies dans un système de gestion des bases de données orientée graphes (Robinson, Webber, & Eifrem, 2013).

Etape 2 - Extraction des processus métiers et des relations dérivées

Après consolidation de toutes les informations concernant l'organisation de l'entreprise, la deuxième étape consiste à extraire un ensemble de processus. Ces processus pourront être en particulier responsables de l'inefficacité et de la congestion à l'organisation de l'entreprise. Ensuite la dérivation d'un ensemble de relations associées aux processus sélectionnés. L'extracteur structural SETL a pour mission de récupérer ces informations via l'analyse des

requêtes spécifiques aux besoins de diagnostic. Ensuite, à transformer les résultats extraits sous une forme convenable pour le framework d'analyse (des réseaux sociaux et des complexes simpliciaux). Enfin à charger les résultats dans le framework d'analyse structurale.

Etape 3 - Topo-Scopie et SNA-Scopie

Cette étape consiste à utiliser deux approches analytiques permettant de vérifier et de tester les non-conformités avec l'organisation de faite qui a été adoptée par l'entreprise. La première approche structurale est basée sur la topologie algébrique. Elle utilise la méthode canonique « Complexes simpliciaux » qui permet de projeter l'organisation dans une structure spécifique « complexe simplicial ». Cette structure modélise la communication entre ses éléments tout en conservant la sémantique des liens existants. Cette méthode nous fournit des chaines de liaisons entre simplexes (les composantes) du complexe. Le framework permet le calcul d'indicateurs structuraux tels que l'excentricité, l'entropie, la complexité, etc. La seconde approche utilise les techniques d'analyse des réseaux sociaux. Le framework prend en entrée un graphe représentant un réseau social. Cette approche permet de calculer les métriques d'analyse des réseaux sociaux telles que la centralité, la similarité, etc.

Etape 4 - Visualisation du reporting

Cette étape consiste à fournir des cartes de diagnostic. Ces cartes de diagnostic représentent une vue analytique de l'organisation de l'entreprise et sont présentées de manière synthétique. Elles comportent aussi les indicateurs de mesures à deux niveaux d'analyse : analyse structurale par les complexes simpliciaux et l'analyse des réseaux sociaux. Cette étape intègre aussi la proposition des actions et des recommandations formulées par le procédé de refonte organisationnelle.

Etape 5 - Refonte Organisationnelle

Cette étape nécessite la contribution des gestionnaires. La structure organisationnelle de l'entreprise émergente sera modifiée pour une question d'amélioration et d'alignement stratégique. Le système propose aux gestionnaires des actions à mener pour la refonte organisationnelle. Les gestionnaires produisent une nouvelle structure organisationnelle de l'entreprise afin de réduire les non-conformités par rapport à l'ancienne structure organisationnelle. La nouvelle structure répond ainsi à la stratégie du top management et aux termes du projet de l'entreprise. Une fois la nouvelle structure organisationnelle de l'entreprise est établie, cette dernière fera l'objet de l'analyse et du diagnostic pour l'évaluer à nouveau en mesurant son nouveau degré de conformité.

Etape 6 - Stabilité de la Structure

Après une suite d'amélioration de l'organisation de l'entreprise en réduisant son degré de non-conformité, cette étape fournit une structure organisationnelle stable de l'entreprise. Il convient à présent d'effectuer des tests afin d'évaluer l'efficacité et l'efficience de notre approche. Pour cela, des expérimentations ont été menée à travers l'implémentation et la mise en œuvre de prototype « organizational diagnosis ». Notre prototype de diagnostique et d'analyse structurale est composé de quatre modules principaux : la collecte : constitue l'entrée de notre prototype ; composée des différentes ressources des données organisationnelles de l'entreprise ; la préparation des données d'analyse : composée d'un SETL un composant structural permettant d'extraire les processus à analyser, de dériver les relations organisationnelles, puis de transformer en des matrices d'incidence et enfin de les charger dans le moteur d'analyse structurale ;l'analyse structurale : composer des deux modules d'analyse, analyse simplicial et ARS ; visualisation et rapport de diagnostic : le dernier module qui

permet d'afficher les résultats d'analyse et de calcul des indicateurs structuraux. Ce module fournit un rapport complet de l'analyse effectuée.

Notre prototype logiciel de diagnostic organisationnel a été implémenté sous l'environnement Eclipse Java, sur un Dell avec intel i5 CPU 2,3GHz et 4Go de RAM. Les différents diagrammes présentés dans ce rapport sont réalisés avec l'API de visualisation Gephi. Le prototype peut être utilisé dans les systèmes d'exploitation Windows et Linux. Le package de déploiement est une solution Java sous forme de services web. Elle nécessite : Serveur appache, JRE (java runtime environement, à partir de la version 6), Une connexion à la base de données orienté graphe : Serveur Neo4J (Partner, Vukotic, & Watt, 2014) (JBDC), API Gephi (visualisation et statistique de centralité).

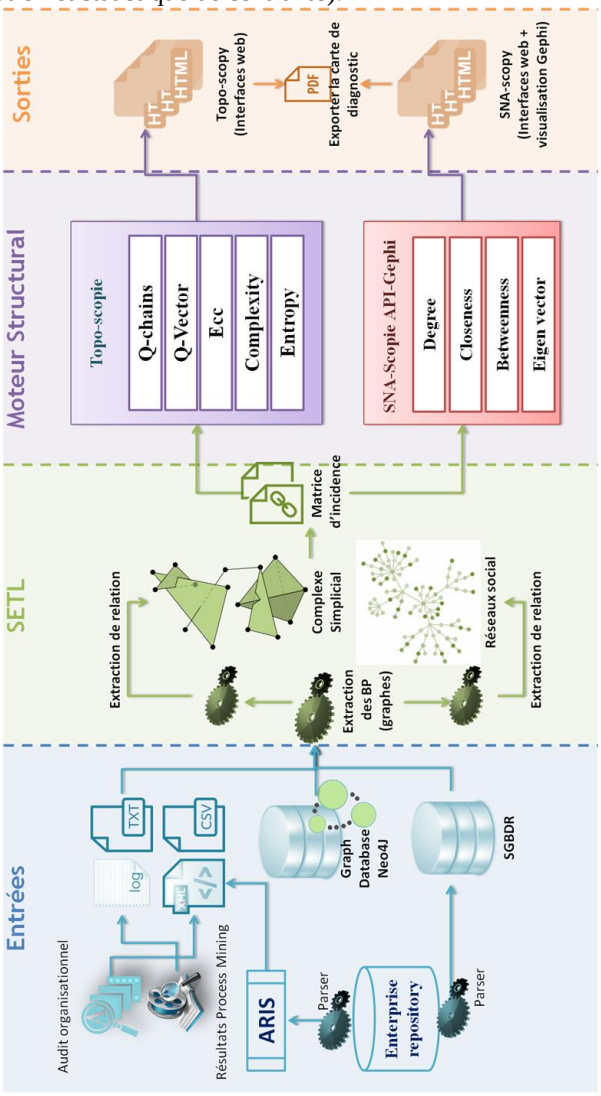


Figure 3 Fiche produit du prototype

3 Scénario de Test

L'environnement des entreprises est de plus en plus évolutif. Chaque année apporte de nouveaux défis dont l'apparition de nouveaux produits, l'évolution des technologies, l'émergence de nouvelles spécialités toujours plus pointues, les experts correspondants, les exigences et les attentes des clients, les contraintes économiques de profitabilité, de croissance ou de maîtrise des risques, la décentralisation des responsabilités…Tous ces éléments de plus en plus sévères nécessitent que l'on accorde une attention de plus en plus grande au pilotage des processus de réalisation des projets et à la structure organisationnelle orchestrée par ses même processus.

Le domaine d'application de l'entreprise étudiée, à titre d'exemple, est une entreprise de production d'automobile. Un secteur stratégique et générateur de revenus. La production automobile est un secteur clé dans le tissus industriel. Il permet de répondre à plusieurs points important à l'économique, du point de vue emplois, génération de capitale et de devise.

3.1 Capture des données de l'organisation de l'entreprise

Le référentiel de l'entreprise est capturé par le modèle ARIS. Tout d'abord nous avons capturée l'organigramme qui illustre la relation hiérarchique et attribution de rôle aux collaborateurs et employées. La figure 2 représente l'organigramme de l'entreprise à analyser.

C'est une structure organisationnelle comprenant des directions, des divisions, des rôles et des attributions à des collaborateurs.

Les différentes directions de cette structure organisationnelle sont : Direction générale ; Direction logistique et approvisionnement ; Direction manufacture/ production ; Direction Recherche et développement ; Direction Marketing ; Direction d'administration et des affaires.

Chaque direction contient des divisions spécifiques. La figure 3 présente une capture de la persistance des données de la structure hiérarchique de l'entreprise dans la base de données graphes Neo4J. Les différentes directions et division composant cette structure organisationnelle de l'entreprise à analyser. Vue la taille de l'organigramme nous affichons que les deux premier niveaux de la structure hiérarchique. Nous avons stocké aussi plusieurs processus métiers. Nous n'en choisissons que quelques un pour une fin d'analyse sur deux types de relations. Nous avons exporté l'organigramme depuis ARIS via un fichier AFD spécifique au framework. Ensuite via un parser nous avons convertis le fichier en un fichier XML pour pouvoir l'importer à la base de données Neo4J. Ainsi nous pouvons stocker les données organisationnelles de l'entreprise.

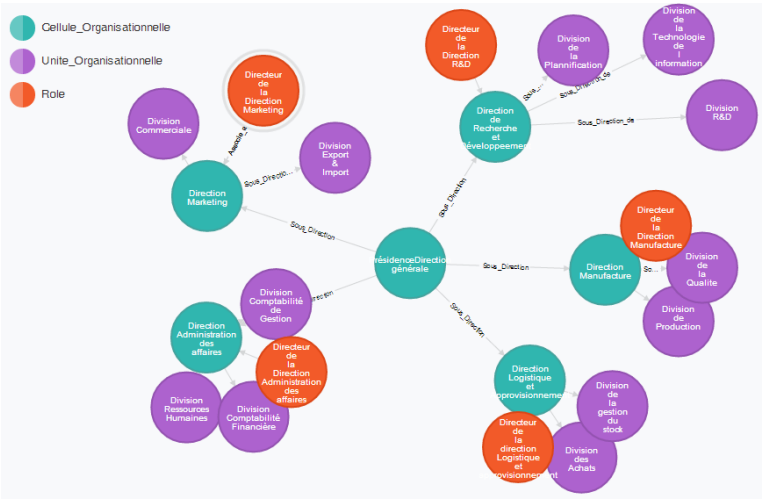


Figure 4 Organigramme de l'entreprise de production sous la base de données graphe Neo4J

Pour pouvoir évaluer l'organisation de l'entreprise (figure 2), nous devrons analyser sa structure organisationnelle et l'ensemble des interactions au sein de cette structure.

3.2 Extraction des données d'analyse

A travers le langage spécifique de requêtes « Cypher » nous interrogeons la base de données qui collecte l'ensemble d'informations organisationnelles de l'entreprise. Selon le besoin d'analyse nous extrayions les processus ainsi que les relations à analyser. Nous choisissons la relation de collaboration entre activités et ressources collaboratrices dans le processus métier « *commande de fourniture pour le service de production* » :

Tableau 1 Requête d'extraction du processus métier et de la relation de collaboration

```
MATCH (p1:`Processus_Metier`)-[*]->(p2:`Activite`)-[*]->(p3: `Performer`)
WHERE p1.name="BP1"
RETURN p3.name as Performer,p2.name as Activite, p1.name as Processus_Metier, count(p2) as Occurrences
```

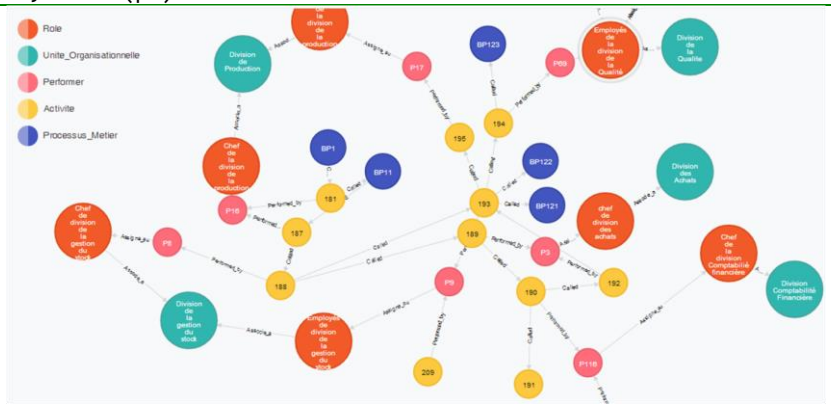


Figure 5 Extraction de relation de collaboration

Nous récupérons ainsi une matrice d'adjacence de la relation de collaboration (performer/ activity) du processus métier sélectionné.

Tableau 2 Matrice d'adjacence de la relation de collaboration Activité-Collaborateur dans le processus sélectionnémétier

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}
P_1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
P_3	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
P_8	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
P_9	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
P_{16}	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
P_{17}	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
P_{69}	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
P_{73}	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
P_{74}	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
P_{75}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
P_{78}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
P_{79}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
P_{80}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
P_{88}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
P_{92}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
P_{94}	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
P_{95}	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
P_{116}	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0

3.2.1 Analyse de l'organisation de l'entreprise à traver la méthode many-to-many-relation-processus

C'est le composant framework d'analyse qui se compose à son tour de deux composants d'analyse. ***La méthode Toposcopy :*** Nous avons calculé la matrice de faces partagées du complexe simplicial KA(λ, P) à partir de la matrice d'incidence de la relation de collaboration entre performers et activités. En appliquant la méthode d'analyse simplicial nous obtenons les chaines de connexions et le vecteur structuré pour la vue performer de la relation λ du processus métier :

Tableau 3 Matrice de face partagée vue collaborateur de la relation de collaboration Activité-Collaborateur dans le processus métier sélectionné

	P_1	P_3	P_8	P_9	P_{16}	P_{17}	P_{69}	P_{73}	P_{74}	P_{75}	P_{78}	P_{79}	P_{80}	P_{88}	P_{92}	P_{94}	P_{95}	P_{116}
P_1	0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0	–	–
P_3		1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
P_8			0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
P_9				0	–	–	–	–	–	–	–	–	–	–	–	–	–	–
P_{16}					1	–	–	–	–	–	–	–	–	–	–	–	–	–
P_{17}						0	–	–	–	–	–	–	–	–	–	–	–	–
P_{69}							0	–	–	–	–	–	–	–	–	–	–	–
P_{73}								0	–	–	–	–	–	–	–	–	–	–
P_{74}									1	0	0	0	0	0	0	–	–	–
P_{75}										0	0	0	0	0	0	–	–	–
P_{78}											0	0	0	0	0	–	–	–
P_{79}												0	0	0	0	–	–	–
P_{80}													0	0	0	–	–	–
P_{88}														0	0	–	–	–
P_{92}															0	–	–	–
P_{94}																1	–	–
P_{95}																	0	–
P_{116}																		0

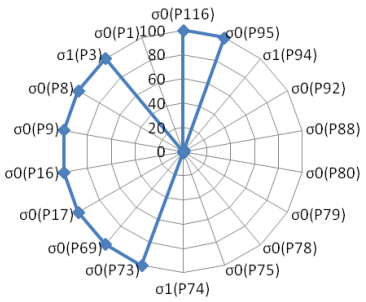
Q = 1 ; {P_3}, {P_{16}}, {P_{74}}, {P_{94}}
Q = 0 ; {P_1, P_{94}}, {P_3}, {P8}, {P9}, {P_{16}}, {P_{17}}, {P_{69}}, {P_{73}}, {P_{74}, P_{75}, P_{78}, P_{79}, P_{80}, P_{88}, P_{92}}, {P_{95}}, {P_{116}}

$$Q = \begin{Bmatrix} 1 & 0 \\ 4 & 11 \end{Bmatrix}$$

Ainsi nous pouvons mesurer les indicateurs structuraux de la méthode Topo-scopy.

Excentricité

Simplexe	Ecc	Simplexe	Ecc
$\sigma_0(P_{116})$	∞	$\sigma_1(P_{74})$	1
$\sigma_0(P_{95})$	∞	$\sigma_0(P_{73})$	∞
$\sigma_1(P_{94})$	1	$\sigma_0(P_{69})$	∞
$\sigma_0(P_{92})$	0	$\sigma_0(P_{17})$	∞
$\sigma_0(P_{88})$	0	$\sigma_0(P_{16})$	∞
$\sigma_0(P_{80})$	0	$\sigma_0(P_9)$	∞
$\sigma_0(P_{79})$	0	$\sigma_0(P_8)$	∞
$\sigma_0(P_{78})$	0	$\sigma_1(P_3)$	∞
$\sigma_0(P_{75})$	0	$\sigma_0(P_1)$	0



Nous remarquons que la majorité des employés (Performers) sont excentriques.

- ***Complexité :*** la complexité du complexe simplicial KA(λ, P) est de l'ordre de Ψ= 6,33. Ce qui justifie les valeurs de l'indicateur d'excentricité.
- ***Entropy :*** l'entropie du complexe est de l'ordre de ε = 0, 520.

3.2.2 Visualisation de résultat d'analyse

A travers les API Gephi nous visualisons le résultat d'analyse et de diagnostic effectué par le moteur structural. Voici quelques interfaces du prototype :

Figure 6 Interface de chargement de la structure organisationnelle de l'entreprise

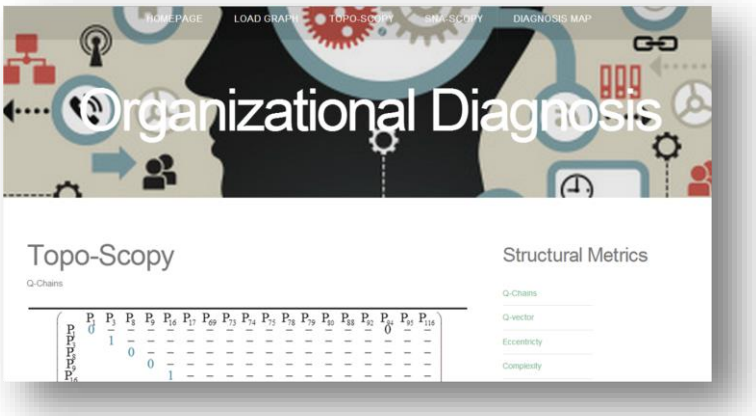


Figure 7 Toposcopy

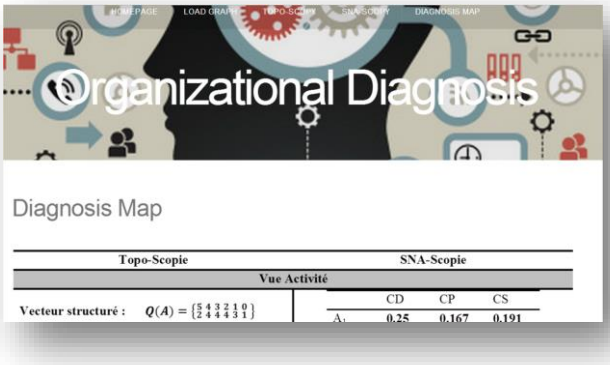


Figure 8 Interface de la carte de diagnostic

4 Conclusion

Nous proposons un prototype de la solution logicielle répondant à la problématique soulevée. Nous présentons à travers un scénario de test d'une entreprise de manufacture à titre d'exemple l'utilisation de ce prototype. Le but est d'analyser la structure de l'organisation de l'entreprise proposée. Cette analyse nous permettra d'évaluer à travers les deux approches proposées, les performances de la structure organisationnelle. Nous allons donc mesurer les indicateurs structuraux à savoir la complexité du système, son entropie et les différentes mesures d'excentricité. En ce qui concerne la deuxième approche proposée, portant sur l'analyse des réseaux sociaux, nous calculons les différents indicateurs de centralité, à savoir, la centralité par le degré, la centralité de proximité, la centralité d'intermédiarité et la centralité spectrale. Cette démarche d'analyse s'inscrit dans le cadre du procédé proposé SEPA, visant au diagnostic et à la refonte de la structure organisationnelle de l'entreprise, en vue d'un meilleur niveau de maturité, de conformité et de cohérence par rapport à la stratégie établie par le top-management. Ainsi notre prototype nous permet d'évaluer la structure organisationnelle réelle de l'entreprise. Avec les deux approches que nous avons proposées,

l'entreprise évaluée à titre d'exemple dans notre étude de cas dégage plusieurs résultats d'analyse Topo-Scopie et SNA-Scopie. Nous avons testé la proposition *many-to-many* relation processus. Elle a pour but d'étudier les interactions en termes d'organisation entre les processus selon les diverses relations.

References

Aalst, W. (1998). The Application of Petri Nets toWorkflow Management. . *The Journal of Circuits, Systems and Computers , 8* (1), 21–66.

Aalst, W., Reijers, H., Weijters, A., Dongen, B., Alves, A. M., & Song. (2007). Business process mining: an industrial application. *Information Systems, , 32* (1), 713 – 732.

Boulmakoul, A., & Besri, Z. (2014). *Patent No. 37042*. Maroc.

O'Reilly Media. (2012). *Big data Now.* by O'Reilly Media, Inc.

Partner, J., Vukotic, A., & Watt, N. (2014). *Neo4j in Action* (1 edition ed.). USA: Manning Publications.

Redmond, E., & Wilson, J. R. (2012). *Seven databases in seven weeks.* (J. Carter, Ed.) The Pragmatic Bookshelf.

Robinson, I., Webber, J., & Eifrem, E. (2013). *Graph database.* (O'Reilly, Ed.)

Song, M. (2006). *Mining Organizational relations from business Process Data.* Ph.D. Thesis.

van der Aalst, W. (2011). *Process Mining: Discovery, Conformance and Enhancement of Business Processes.* (Springer-Verlag, Berlin, & H. G. Co, Eds.) Springer.

P. A. Clayton. The Methodology of Organizational Diagnosis. Vol. 11, No. 3 June (1980).

R. Atkin, Combinatorial Connectivities in Social Systems. Basel, Birkhäuser Verlag. (1977)

A. Boulmakoul, Z. Besri. Performing Enterprise Organizational Structure Redesign through Structural Analysis and Simplicial Complexes Framework. The Open Operational Research Journal, 2013, 7.pp- 11-24. (2013)

M.P Wil, D. A. Van. Process Mining Discovery, Conformance and Enhancement of Business Processes. Springer Heidelberg Dordrecht London New York. ISBN 978-3-642-19344-6. (2011)

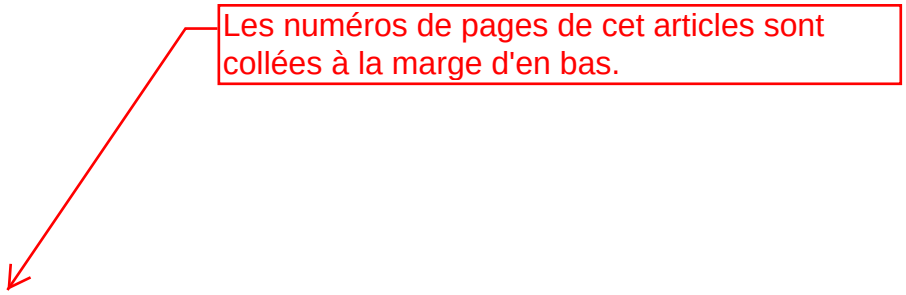
W.V.D Aalst. The Application of Petri Nets to Workflow Management. The Journal of Circuits, Systems and Computers, 8 (1): 21–66, 1998 (1998)

W.V.D Aalst, H. Reijers, A. Weijters, B. Dongen , A.M. Alves , M. Song and H. Verbeek."Business process mining: an industrial application," Information Systems, vol. 32, no. 1, pp. 713 – 732. (2007).

M. Song. Mining Organizational relations from business Process Data, Ph.D. Thesis, 2006

P. K. Kwanghoon. Discovering Activity-Performer Affiliation Knowledge on ICN-based Workflow Models. Journal Of Information Science And Engineering 29, 79-97. Pp. 79-97 (2013).

Abstract: All organizations are led to implement appropriate governance and compliance with applicable laws and best practices. The organizational audit used to identify potential dysfunctions in organs of the directions, administration and information systems. This article proposes an original and practical approach focused on the structural analysis and the analysis of social networks for the assessment of the organizational structure of the company. The problem impacts the operational aspects of the implementation of analytical methods from structural analysis to organizational reengineering of the company on the basis of its existing information system. The proposed method derives from canonical practices, using the method, simplicial analysis, which imprints the foundations of combinatorial and discrete algebraic topology. Besides the practices of social networks analysis techniques for measuring the centrality of professional network that the company defines. We propose in this research an integrated software solution for the discovery and diagnosis of the organizational structure of the company.



Confidentialité des entrepôts de données dans le Cloud Computing: Etat de l'art et Perspectives

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

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

Résumé. Un entrepôt de données est une base de données regroupant l'ensemble des données fonctionnelles et critiques employées par la haute direction des organisations pour la prise des décisions stratégiques. L'hébergement de ces entrepôts de données dans le Cloud Computing (CC) permet de surmonter l'expansion sans fin des données, en raison de sa capacité de traitement et de stockage des données. Cependant la confidentialité de ces entrepôts de données dans le CC a besoin de nombreuses améliorations et de la mise en place des normes précises dont l'objectif est d'adapter les méthodes traditionnelles de contrôle d'accès à ce nouveau paradigme CC, car ces données seront confiées à un prestataire externe. L'objectif de cet article est de donner un aperçu des aspects pertinents du contrôle d'accès aux entrepôts de données. Cet aspect qui présente un des mécanismes de confidentialité les plus importants de CC, puisque le service CC ne peut pas appliquer le modèle de contrôle d'accès traditionnel en raison de ses caractéristiques d'accès, ainsi qu'il n'y a pas un protocole standard pour gérer la connectivité des utilisateurs de CC aux ressources hébergés.

1 Introduction

Les entrepôts de données forment le socle des processus décisionnels qui constituent un support efficace pour avoir une vue clair pour les décideurs sur l'efficacité des différentes activités de l'entreprise et les aider à prendre des bonnes décisions afin d'augmenter leur profits.

D'autre part l'entrepôt de données est la concentration de toutes les bases de données d'une façon dénormalisée en accédant rapidement aux données sur mesure, restituer en différents endroits. Il regroupe les données sensibles et très pertinentes, et secrètes de l'entreprise, telle que les données médicales, financières qui ne doivent pas être accessibles sans contrôle d'accès. Dans ce contexte plusieurs gouvernements ont adopté des lois pour protéger la vie privée de leurs citoyens. Parmi ces lois, HIPAA (Health Insurance Portability and Accountability Act HHS (1996)) vise à protéger les données médicales des patients américains en obligeant les

établissements du secteur des soins de la santé de suivre des règles strictes de sécurité, de même GLBA (Gramm Leach Bliley Act GPO 1999) oblige les organisations financières américaines à protéger les données de leurs clients Fernandez-Medina et al. (2006).

Les entrepôts de données reçoivent des téraoctets des données stockées d'une façon historique, depuis les systèmes de production de l'entrepôt de données. A un certain stade, et malgré l'excellent utilitaire des entrepôts de données, le coût de maintien devient injustifié pour l'entreprise.

Aujourd'hui, la solution de l'hébergement de l'entrepôt de données dans le Cloud gagnent progressivement plus de popularité dans les entreprises, car de nombreuses entreprises se rendent compte de ses avantages. Par contre elle présente aussi des menaces de sécurité à l'égard des données des utilisateurs. Le contrôle d'accès est l'un des mécanismes de sécurité les plus importants des services de cloud computing qui garantie la confidentialité des données, cependant le service Cloud ne peut pas appliquer le modèle de contrôle d'accès traditionnel en raison de son évolutivité et son élasticité, car il n'y a pas un protocole standard pour gérer la connectivité des utilisateurs du Cloud aux ressources hébergés.

Après la présentation de la problématique dans la section 2, Le reste de cet article est structuré comme suit. La section 3 présente une vue d'ensemble des travaux connexes. La Section 4 décrit une synthèse des travaux. Enfin, la section 5 présente nos conclusions et perspectives.

2 Le contrôle d'accès aux données entreposées dans le Cloud Computing

L'analyse des risques conduit généralement à étudier cinq aspects de sécurité de l'information : la confidentialité, l'intégrité, la disponibilité, la traçabilité et la gestion des accès.

- Confidentialité : elle consiste à préciser les personnes qui ont le droit d'accès aux données sensibles, les privilèges de chaque utilisateur, les mécanismes de contrôle d'accès ainsi que le moment de chiffrement des données que ce soit lors du stockage ou bien lors de leurs mouvements.
- Intégrité : Pour faire des analyses stratégiques, l'entreprise aura besoin d'assurer l'intégrité des ces données en précisant qui peut modifier une information, de définir les mécanismes de vérification si l'information est changée, et de contrôler la cohérence des informations.
- Disponibilité : L'entrepôt de données doit être accessible aux utilisateurs autorisés à tout moment, pour garantir un service et une productivité efficace et sans interruption.
- Traçabilité : Dans un entrepôt de données, on doit garder la trace des actions effectuées sur les systèmes afin de savoir qui a effectué une tâche.
- Gestion des accès : La gestion des accès aide à protéger la confidentialité, l'intégrité et la disponibilité des actifs en s'assurant que seuls les utilisateurs autorisés peuvent y avoir accès et les modifier.

Le contrôle des accès est basé sur les informations d'identité pour permettre et contraindre l'accès. Les utilisateurs doivent avoir un accès limité aux données sensibles ce qui garantie la confidentialité de ces données, dans ce sens il faut mettre en place des procédures pour les identifier afin d'éviter les opportunités inutiles d'accès aux données des clients. Lors de l'affectation des autorisations, il faut opter pour des contrôles d'accès fondés sur

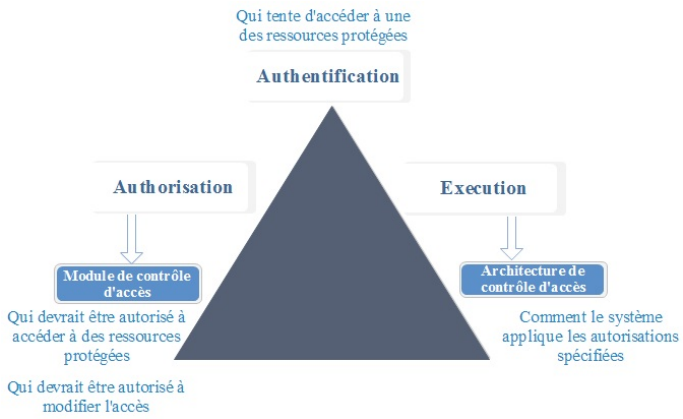


FIG. 1 – *La gestion d'accès aux données*

les rôles de manière à donner aux utilisateurs autorisés des accès qui dépendent de leurs fonctions. FIG 1 montre les trois étapes nécessaire pour l'accès aux données Ray et Ray (2014) :

- L'authentification : aborde le problème de déterminer l'identité de l'utilisateur qui essaye d'accéder à des ressources protégées.
- L'autorisation : s'exprime en terme d'un modèle de contrôle d'accès qui précise les ressources qui doivent être protégées, et qui est autorisé pour accéder à ces ressources.
- L'exécution : est un processus qui s'occupe d'appliquer les décisions prises d'autoriser l'accès ou de le refuser.

Aujourd'hui, la solution de l'hébergement de l'entrepôt de données dans le CC gagnent progressivement plus de popularité dans les entreprises, car de nombreuses entreprises se rendent compte de ses avantages. le contrôle d'accès dans un environnement CC a de nouveaux défis posés par la multi-location, l'élasticité et la dynamité de ce paradigme. Le mécanisme de contrôle d'accès doit prendre en compte les spécificités de déploiement d'un entrepôt de données dans le CC qui est très différent de celle de son déploiement sur site.

La migration des entrepôts de données vers le CC devrait améliorer la satisfaction de l'utilisateur final et induire une plus grande productivité de l'entreprise. Ce qui nécessite une haute performance qui peut être garanti par la mise en œuvre de l'intra-parallélisme de requête qui consiste à décomposer une requête complexe en sous-requêtes, et les traiter sur plusieurs processeurs, et enfin effectuer le post-traitement pour présenter une réponse à la requête principale Moussa et Badir (2013). Alors que la mise en place d'un mécanisme de contrôle d'accès ne doit pas augmenter la charge des traitements dont le but est d'avoir un système évolutif et productif avec des données qui sont bien protégées contre l'accès aux données interdites puisque ces données seront confiées à un prestataire externe.

Ce mécanisme de contrôle d'accès ne doit pas influencer l'évolutivité de l'entrepôt de données hébergé dans le CC en évaluant la charge des traitements sur l'échelle de temps, et en mesurant le nombre des requêtes traitées au cours d'un intervalle de temps. Un système évolutif, devrait maintenir le même nombre Moussa et Badir (2013).

3 Etat de l'art

Récemment, un certain nombre de solutions de contrôle d'accès aux entrepôts de données ont été proposés, nous avons organisé les travaux selon trois parties, le contrôle d'accès aux entrepôts de données depuis les autorisations des sources de données, la deuxième présente le contrôle d'accès aux données de l'entrepôt lors de la modélisation. Dans la troisième partie traite le contrôle d'accès aux données entreposées dans les Cloud Computing et la dernière partie aborde le contrôle d'accès aux entrepôts de données existants (niveau exploitation)

3.1 Le contrôle d'accès aux entrepôts de données depuis les autorisations des sources de données

Certains chercheurs ont proposé l'exploitation des autorisations définit au niveau des sources de données afin de gérer l'accès aux entrepôts de données. Dans ce sens on trouve :

Rosenthal et Sciore (2000) Proposent une approche théorique qui exploite les permissions d'accès définit au niveau des sources de données, plutôt que de créer des nouveaux mécanismes d'accès. Ils utilisent la réécriture des requêtes pour vérifier que ces dernières respectent les restrictions définit dans les sources de données, et la création des vues relationnelles afin de minimiser le risque d'inférer les données sensibles.

Saltor et al. (2002) puisqu'il y a des similitudes entre l'architecture d'une base de données fédérée et l'architecture d'un entrepôt de données, les auteurs ont proposé l'utilisation du schéma des autorisations d'accès multi-niveaux défini pour la base de données fédérée sans être modifiée pour construire un entrepôt de données sécurisé, ce schéma des autorisations décrit les règles d'accès multi-niveaux.

3.2 Le contrôle d'accès aux données de l'entrepôt lors de la modélisation

Parmi les travaux qui ont été développés sur l'intégration du contrôle d'accès dans la phase conceptuelle des entrepôts, on trouve :

Sweeney (2002) Décrit un cas réel d'inférence de données par une démonstration d'identification de l'ancien gouverneur de l'état de Massachusetts, qui figure dans les 2 listes d'inscription des élections et les dossiers d'assurances maladie. Ils ont présenté une démonstration ou une façon didentification en se basant sur l'intersection des données d'un groupe d'assurance. Ces données supposé qu'ils sont anonymes, et une liste d'inscription des électeurs, ce qui permet de détecter le nom de l'ancien gouverneur William Weld et ses dossiers médicaux, en reliant les attributs partagés.

Fernandez-Medina et al. (2006) ont développé un modèle de contrôle d'accès et d'audit(ACA) spécifique aux entrepôts de données, qui repose sur deux politiques de gestion des accès : MAC et RBAC. Ils précisent des règles de contrôle d'accès lors de la modélisation d'un modèle conceptuel, en intégrant la notion de «profil utilisateur », qui est constitué d'une table isolée contenant toutes les informations des utilisateurs (identité, niveau de classification : top

secret, secret, confidentiel ou inconnu). Ce modèle reste un modèle purement théorique car aucune solution concernant son implémentation n'a encore été proposée.

Villarroel et al. (2006) ont défini une extension OCL «Object Constraint Language» en utilisant les mécanismes d'extension UML2.0 pour résoudre les problèmes de la confidentialité, cette extension spécifie les contraintes de contrôle d'accès des éléments lors de la modélisation conceptuelle des entrepôts de données.

Soler et al. (2008) ont utilisé des mécanismes d'extension fournis par le CWM (Common Warehouse Metamodel) pour étendre le package relationnel et construire un schéma en étoile, qui représente les règles de contrôle d'accès et de vérification capturées pendant la phase conceptuelle de l'entrepôt de données.

Trujillo et al. (2009) ont développé une méthodologie comprenant quatre phases : analyse, modélisation, implémentation et validation, qui couvrent les cinq niveaux d'abstraction qui sont : analyse des besoins, niveau conceptuel, niveau logique, niveau physique et l'examen post- développement, ce dernier étant une nouvelle discipline introduite par Lujan et Trujillo (2004) . Cette méthodologie présente toutes les exigences de la contrôle d'accès tout au long du cycle de vie de l'entrepôt de données.

Blanco et al. (2010) ont Proposé une approche basée sur le diagramme états-transactions pour détecter les inférences au niveau de la conception. Cette proposition se focalise sur les requêtes sensibles et ses évolutions, mais ils ne tiennent pas en compte d'inférer les données à partir des données accessibles. L'approche est présentée sous forme d'un modèle de 3 états :

- Modèle statique : présente le profil UML spécifique aux entrepôts de données de Fernandez-médina 2007, en ajoutant un nouveau genre de règle nommée « Joint Rules », qui présente les privilèges nécessaires à certaines combinaisons.
- Modèle dynamique : ou bien le modèle états-transactions qui a l'objectif d'enrichir le modèle statique, en traitant les évolutions des combinaisons définis avec JR à travers l'application des opérations OLAP.
- Contrôle de session : cette étape rend plus d'intérêt aux sessions des utilisateurs afin de les analyser pour détecter toute possibilité d'inférence.

Rodriguez et al. (2011) ont présenté une extension d'UML 2.0 du diagramme d'activité. Cette proposition, libellée comme BPSec (Business Security Process), permet de définir un ensemble d'exigences de sécurité (contrôle d'accès, détection des risques d'attaques, non-répudiation, intégrité, confidentialité et vérification de la sécurité), ce qui améliore l'expressivité des modèles des processus métiers, et permet de sécuriser un entrepôt de données lors de son développement en prenant en considération cette exigence.

Triki et al. (2011) ont proposé un modèle pour sécuriser les données multidimensionnelles contre les inférences dans la phase conceptuelle, cette approche suppose que le schéma de DW est déjà conçu. Il permet de détecter les deux type d'inférences : Inférence précise : ou les valeurs des données déduites sont exactes. et inférence partielle : ou les valeurs des données

sont partiellement divulguées, c'est-à-dire que l'utilisateur peut déduire une idée sur la valeur des données. Cette approche se compose de trois étapes :

- Etape 1. Un expert de domaine identifie les éléments sensibles à protéger en interrogeant le concepteur de l'entrepôt de données.
- Etape 2. Construire le graphe des inférences à partir du diagramme de classe, en précisant les éléments qui présentent les inférences précises ou partielles.
- Etape 3. Présenter l'entrepôt de données avec les annotations UML en mettant en évidence les deux types d'inférences.

3.3 Le contrôle d'accès aux données entreposées dans les Cloud Computing

Les chercheurs traitent le Cloud Computing comme un nouvel espoir pour les entrepôts de données, Malgré les avantages de cette solution, la confidentialité des données dans un environnement Cloud reste un risque à traiter, parmi les travaux qui traitent cette problématique on trouve :

Bensaidi et al. (2012) ont proposé un modèle de contrôle d'accès hybride basé sur la confiance Pour la sécurité des SID (Systèmes d'Information Distribués) , et sur le modèle de recommandation d'accès. Ils ont proposé également que des agents de confiance, appelés encore les tiers de confiances ou TTP pour «Trusted Third Party ». L'approche proposée se base sur la diminution de l'indice des permissions lors de la violation des droits fixés. Ainsi, Après un nombre bien défini des tentatives malveillantes, le connecté perd tous ses privilèges au sein de l'entreprise.

Al-Aqrabi et al. (2013) se focalisent sur la sécurité des systèmes décisionnels hébergés dans le Cloud Computing. Ils traitent la gestion des risques qu'apporte cette technologie du Cloud Computing, dans ce sens ils ont proposé deux modèles :

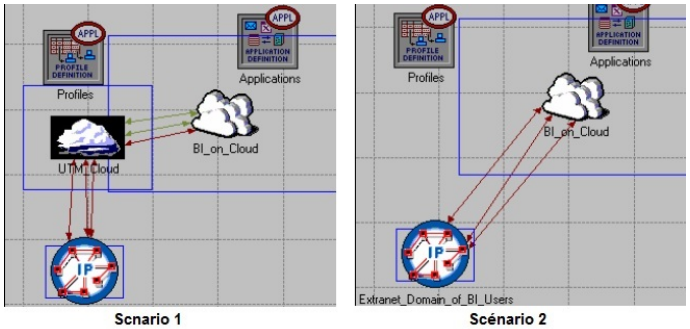


FIG. 2 – *comparaison entre les deux scénarios*

- UTM Unified Threat Management : ce modèle contient six réseaux locaux avec 500 postes de travail dans chaque réseau LAN, l'acheminement de tous les utilisateurs vers

le BI Cloud est assuré par le réseau UTM_Cloud qui regroupe les applications qui s'occupent de la connectivité de tous les utilisateurs à BI Cloud qui regroupe les entrepôts de données et les serveurs d'applications OLAP contenant les tableaux de bord OLAP et les vues temporaires.
– Modèle basé sur la distribution de contrôle de sécurité entre plusieurs serveurs dans le Cloud Computing. Dans ce modèle, UTM Cloud est éliminé et les utilisateurs sont directement liés aux commutateurs de Cloud BI. L'exécution de la simulation du deuxième modèle, les auteurs ont observé que le temps de simulation a été réduit de deux minutes à une minute. Cependant, ce modèle est difficile de gérer car le mécanisme de la sécurité est réparti dans des milliers serveurs.

Ray et Ray (2014) ont proposé un nouveau modèle de contrôle d'accès dans le Cloud Computing, basé sur la confiance. La décision d'autoriser une demande d'accès est faite si les trois conditions suivantes sont remplies : (1) le rôle est autorisé à la permission demandée ; (2) l'utilisateur peut activer le rôle ; et (3) l'utilisateur est autorisé à la permission demandée. Chaque utilisateur a un niveau de confiance qui est nombre réel entre 0 et 1, plus la valeur de niveau de confiance est moins, l'utilisateur risque de perdre l'accès.

3.4 Le contrôle d'accès aux entrepôts de données existants (niveau exploitation)

Online Analytical Processing (OLAP) est devenue de plus en plus une composante importante et répandue des systèmes d'aide à la décision. Le serveur OLAP est censé d'assurer des accès en fonction des habilitations de chaque utilisateur. Il peut refuser les accès aux données d'une mesure, d'une dimension, et/ou au delà d'un niveau dans une hiérarchie. Les droits d'accès peuvent être explicitement spécifiés sur les tables/colonnes des tables de l'entrepôt de données. Cependant, le serveur OLAP tout seul ne peut pas protéger l'accès aux données interdites. Des travaux ont été réalisés pour renforcer les droits d'accès/habilitations des utilisateurs, et pour interdire tout utilisateur malicieux d'inférer des données qui lui sont interdites à partir des données auxquelles il a accès.

Kirkgoze et al. (1997) Rosenthal et Sciore (2000)Saltor et al. (2002) ont défini un modèle sécurisé pour les entrepôts de données qui consiste à élaborer un cube personnalisé possédant ses propres dimensions et hiérarchies. Ce modèle repose sur la politique de gestion AMAC. Il s'agit d'une extension du modèle MAC qui permet de spécifier les tâches que l'utilisateur peut exécuter selon son rôle au sein de l'organisation. L'intérêt d'un modèle comme celui-ci, est la flexibilité de l'assignation des rôles aux différents cubes virtuels.

Priebe et Pernul (2001) poursuivent leurs recherches concernant la création des mécanismes de contrôle des accès afin d'assurer la confidentialité des données, et ils ont créé un mécanisme de contrôle d'accès sous forme d'un langage exprimant, au cours de la phase conceptuelle, les contraintes liées à la sécurité. Il s'agit d'un langage basé sur MDX «MulDimensionnale Xpression », celui-ci étant un langage de requête spécialisé dans l'interrogation et la manipulation des données multidimensionnelles. Il est comparable au langage SQL.

Triki et al. (2010) ont proposé une approche qui ne nécessite pas un traitement supplémentaire, après chaque phase d'alimentation de l'entrepôt de données. Elle est basée sur les réseaux Bayésiens afin de protéger un entrepôt de données contre les inférences, ils utilisent un module de contrôle, qui vise à interdire à un utilisateur d'inférer des données protégées à partir des données qui lui sont accessibles en utilisant les fonctions d'agrégations Min et Max.

Eavis et Althamimi (2012) ont présenté un cadre d'authentification qui s'appuie sur une algèbre spécialement conçue pour OLAP. Il est orienté objet et utilise des règles de réécriture de requêtes afin d'assurer l'accès aux données cohérentes à travers tous les niveaux du modèle conceptuel. Le processus est essentiellement transparent pour l'utilisateur, une notification est fournie dans le cas où un sous-ensemble de la demande initiale est renvoyé. Le résultat final est une approche intuitive et puissante pour l'authentification de base de données qui est uniquement adaptée au domaine OLAP.

4 Synthèse des travaux existants

Suite à l'étude des travaux existants, nous avons constaté les points suivants :

- Les auteurs Villarroel et al. (2006), Soler et al. (2008), Rodriguez et al. (2011), Priebe et Pernul (2001), Kirkgoze et al. (1997) ont réussi à bien définir les contraintes de la confidentialité des l'entrepôt de données, et proposer des approches intéressantes mais qui restent insuffisantes lors de l'hébergement de ces entrepôts de données dans le CC.
- La plupart de ces travaux qui traitent la confidentialité de l'entrepôt de données dans la phase conceptuelle sont appuyés sur le méta modèle CWM qui présente une norme pour l'échange et l'interopérabilité des métadonnées, cependant aucune proposition ne précise comment identifier le niveau de sensibilité des données, car la plupart des auteurs se basent sur le concepteur de l'entrepôt de données.
- Bien que les autorisations présentent l'axe principal pour garantir la confidentialité de l'accès à l'entrepôt, cependant l'absence d'une norme qui gère la précision de ces autorisations peut provoquer des incohérences et des inférences comme conséquences. Dans ce sens, on trouve le travail de Saltor et al. (2002) qui ont proposé l'utilisation du schéma des autorisations défini pour les bases de données fédérées sans aucune modification pour construire un entrepôt de données sécurisé, et Rosenthal et Sciore (2000) qui proposent la réécriture des requêtes afin de vérifier que ces dernières respectent les restrictions définit au niveau des sources.
- A noter également, que la notion d'inférence a été cité dans plusieurs travaux en tant qu'élément essentiel pour garantir la confidentialité, et dont la maîtrise est cruciale. Dans ce sens, on trouve le travail de Triki et al. (2011) qui ont proposé une approche qui permet de détecter les inférences partielles et précises, Blanco et al. (2010) qui proposent une approche une approche basée sur le diagramme d'état-transactions permettant de détecter les inférences dans la phase conceptuelle. Néanmoins, malgré les risques élevés d'inférences, il n'est pas suffisamment pris en compte dans la phase conceptuelle.

D'autre part, l'intégration d'un entrepôt de données dans un environnement Cloud Computing a pris beaucoup d'intérêt par les organisations et chercheurs, cela est dû aux avantages qu'elle offre, alors que la gestion d'accès reste un risque à traiter selon les spécificités de cet

	$Classification$	$Autorisation$	$Inference$
Kirkgoze et al. (1997)	Non	Non	Non
Rosenthal et Sciore (2000)	Non	Oui	Oui
Priebe et Pernul (2000)	Non	Non	Non
Saltor et al. (2002)	Non	Oui	Non
Sweeney (2002)	Non	Non	Oui
Fernandez-Medina et al. (2006)	Non	Non	Non
Villarroel et al. (2006)	Non	Non	Non
Soler et al. (2008)	Non	Non	Non
Blanco et al. (2010)	Non	Non	Oui
Triki et al. (2010)	Non	Non	Oui
Rodriguez et al. (2011)	Non	Non	Non
Triki et al. (2011)	Non	Non	Oui
Eavis et Althamimi (2012)	Non	Non	Non

TAB. 1 – *Comparaison des travaux sur la confidentialité au niveau modélisation et exploitation des entrepôts de données*

environnement. Dans ce sens on trouve le travail de Al-Aqrabi et al. (2013) ont proposé 2 modèles qui se focalisent sur la centralisation et la distribution des applications de la sécurité des systèmes décisionnels hébergés dans le Cloud, les auteurs ont observé que le temps de simulation a été réduit dans le modèle ou les applications de la sécurité sont distribuées mais il reste un modèle difficile à gérer, ainsi qu'ils n'ont pas précisé un mécanisme de contrôle d'accès. Bensaidi et al. (2012) ont proposé un modèle de contrôle d'accès hybride basé sur la confiance, et sur le modèle de recommandation d'accès dédiés aux systèmes d'information distribués. Le besoin d'un contrôle d'accès plus standardisé pour le Cloud devient plus qu'urgent, cependant l'inexistence à ce jour de normes reste une sérieuse limitation du Cloud.

5 Conclusion et perspectives

Dans cet article, nous avons défini la problématique de la confidentialité des entrepôts de données dans le CC. Ensuite, nous avons présenté un état de l'art sur la confidentialité des entrepôts de données en général, qui reposent sur le contrôle d'accès, et celle des entrepôts de données dans le CC. Nous avons aussi précisé les limites des travaux étudiés comme une base pour définir nos perspectives. Notre travail se compose de deux parties, La première consiste à traiter la confidentialité des données de l'entrepôt :

- Puisque la plupart des auteurs se basent sur le concepteur de l'entrepôt de données pour définir le niveau de sensibilité des données de l'entrepôt, ce qui peut provoquer des risques de divulgation des données sensibles en cas d'absence d'une approche. Nous avons l'intérêt de définir une approche qui consiste à classifier les données selon leur niveau de sensibilité(Sensible, Trop sensible, Confidentiel...).
- Les politiques de contrôle d'accès à partir des sources de données opérationnelles est un domaine attractif, qui doit être intégré dans la phase conceptuelle de l'entrepôt de données Fernandez-Medina et al. (2006). Ce domaine dans lequel certain chercheurs ont effectué des efforts tel que Saltor et al. (2002) et Rosenthal et Sciore (2000), mais

qui restent insuffisants. Dans ce sens , nous avons l'intention de définir une approche qui permet de coordonner les droits d'accès à l'entrepôt avec ceux des sources de données d'une façon automatique, pour bien définir les autorisations d'accès à l'entrepôt de données.

– Les inférences constituent une menace grave sur la vie privée des utilisateurs.Néanmoins sauf le travail de Triki et al. (2011), l'état de l'art ne traite pas suffisamment ce menace. Dans ce sens nous avons l'intérêt de continuer sur les perspectives de Blanco et al. (2010) en détectant les inférences possibles d'une façon automatique en analysant les sessions des utilisateurs.
– Parmi nos perspectives est de mettre en place un profil d'usage d'un utilisateur, ce profil aidera l'administrateur à bien gérer l'utilisation et l'accès à l'entrepôt de données. Parmi les privilèges qu'il traitera sont : la gestion des mots de passes , le nombre des tentatives de violation des droits d'autorisations en précisant des réactions automatiques, la localisation d'un utilisateur lors d'une tentative de violation, traçabilité des actions, nombre des sessions possibles pour un utilisateur et la durée d'une session.

La deuxième partie de notre travail consiste à traiter la gestion d'accès aux données de l'entrepôt hébergées dans le Cloud Computing et proposer une approche qui permet de gérer de manière fine et précise qui a accès à quoi, quand comment et selon quelles conditions , pour un entrepôt de données hébergé dans le Cloud Computing, en se basant sur le profil utilisateur et le profil d'usage en tenant compte des spécificités de l'environnement Cloud Computing.

Références

Al-Aqrabi, H., L. Liu, R. Hill, Z. Ding, et N. Antonopoulos (2013). Business intelligence security on the clouds: challenges, solutions and future directions. *Service Oriented System Engineering (SOSE), IEEE 7th International Symposium on (pp. 137-144). IEEE*.

Bensaidi, M., A. Aboulkalam, et A. Marzouk (2012). Politique de contrÃ´le d'accÃ¨s au cloud computing: Recommandation Ã base de confiance. *Network Security and Systems (JNS2), 2012 National Days of (pp. 90-96). IEEE*.

Blanco, C., E. FernÃ¡ndez-Medina, J. Trujillo, et J. Jurjens (2010). Towards the secure modelling of olap users behaviour.

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

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

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

Moussa, R. et H. Badir (2013). Data warehouse systems in the cloud: rise to the benchmarking challenge. *Journal International of Computers and Their Applications*, 245.

Priebe, T. et G. Pernul (2000). Towards olap security design - survey and research issues. *Proceedings of the 3rd ACM International Workshop on Data Warehousing and OLAP*

(DOLAP'00), 33–40.

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

Ray, I. et I. Ray (2014). Trust-based access control for secure cloud computing. *High Performance Cloud Auditing and Applications (pp. 189-213). Springer New York*.

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

Rosenthal, A. et S. Sciore (2000). View security as the basis for data warehouse security.

Saltor, F., M. Oliva, A. Abello, et J. Samos (2002). Building secure data warehouse schemas from federated information systems.

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

Sweeney, L. (2002). k-anonymity: A model for protecting privacy.

Triki, S., H. Ben-Abdallah, J. Feki, et N. Harbi (2010). SÃ©curisation des entrepÃ´ts de donnÃ©es contre les infÃ©rences en utilisant les rÃ©seaux bayÃ©siens. *6Ã¨mes JournÃ©es francophones sur les EntrepÃ´ts de DonnÃ©es et l'Analyse en ligne*, 35.

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

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

Villarroel, R., E. Fernandez-Medina, et M. Piattini (2006). A uml 2.0/ocl extension for designing secure data warehouses. *Journal of Research and Practice in Information Technology 38*, 31–43.

Summary

A data warehouse is a database of all functional data and criticism employed by management organizations for taking strategic decisions. The hosting of these data warehouses in the Cloud Computing (CC) overcomes the endless expansion of the data, because of its processing capacity and data storage. However, the confidentiality of these data warehouses in the CC needs many improvements and implementing specific standards which aims to adapt traditional methods of access control to this new paradigm CC. The objective of this article is to provide an overview aspects of access control to data warehouses. this aspect, which that presents one of the most important confidentiality mechanisms cloud computing because the cloud service can't apply classical access control model because of its access characteristics, and there is no standard protocol for managing the connectivity of users to hosted cloud resources.

An MDA Approach to Consider Dependability Requirements in Data Warehouse System Development

Imane Hilal, Nadia Afifi, Mohammed Ouzzif, Hicham Belhadaoui
RITM Lab. ESTC, CED ENSEM, Hassan II University, Casablanca, Morocco
imanehilal@gmail.com, {nafifi, belhadaoui, filali, ouzzif}@est-uh2c.ac.ma
http://www.est-uh2c.ac.ma

Abstract. Data Warehouse is one of the principal systems used by managers and decision makers to analyze the past, the present and the future of any organization. Since it handles crucial information, the Data warehouse system must ensure a high level of dependability. Traditionally, modeling approaches have focused on functional requirements but have rarely dealt with dependability aspects. Furthermore, most of the latter approaches offer partial solutions that only deal with isolated aspects of the Data Warehouse system's dependability, such as availability, reliability, maintainability and security. And consequently, they do not provide developers with an integrated and standard framework for modeling and integrating these requirements in the early development process. To overcome this problem, the current paper proposes an integrated approach based on the MDA (Model Driven Architecture) so as to consider dependability aspects from the first development's stages in order to address their interaction and ensure their traceability. For that purpose, we will describe how to develop the MDA models based on different UML profiles. Finally, a case study is provided to exemplify the benefits of our approach.

1 Introduction

Data Warehouse system provides a collection of technologies aimed at enabling managers and decision makers to analyze the past, the present and the forthcoming performance of their organization. Since it handles crucial information for strategic or operational decisions it must ensure a certain level of confidence to worthy its valuable missions. To reach that level many works addressed data quality as the most important issue. Others focused on the systems features such as: functionality, usability, performance, security, etc.

Traditionally, modeling approaches used to focus on functional requirements. But, even when the final system meets exactly the required functionality, it could be obsolete if its confidence level doesn't satisfy user's expectations. These latter are usually delicate to gather, consequently they are almost classified as non-functional requirements and are generally addressed at the final stages of the development cycle. The problem becomes serious when this type of requirement constrains the implementation of functional features. In this case the non-functional requirement must be elicited and treated at the earlier levels of system's development.

To deliver a Data warehouse system satisfying the trust level expected by its users, our work focuses on its dependability as a part of non-functional requirement. The raison behind choosing dependability is its definition which is "the system ability to deliver service that can justifiably be trusted" (Avižienis et al. 2004). To attain the trusted service announced, dependability encompasses different attributes: reliability, availability, security and main-

tainability. Such attributes are crucial requirements for data warehouse systems and can directly and severely influence its trust level. Moreover, such attributes are not usually homogenous since they could have mutual or conflicting interactions.

In this paper we present a comprehensive approach to integrate dependability aspects at early stages of data warehouse systems. Our approach is based on the OMG's (Object Management Group) standard MDA which defines models at different abstraction levels, and thus addresses the complete development lifecycle, starting from requirement analysis and design, programming, testing, and ending with deployment and maintenance (Bezivin & Gerbe 2001).

To align our approach to MDA, we propose the use of suitable UML profiles to refine and model the different dependability attributes in each of the abstraction's level advocated by the standard. During this phase, we confront the dependability aspects models in order to highlight their interactions and iteratively solve their conflicts. Then, at the last modeling level we recommend the use of Aspect Oriented software development (AOSD) (Schauerhuber et al. 2007) to integrate and "weave" the dependability attributes to functional models in order to obtain a final model gathering both functional and dependability aspects.

This paper is organized as follows: First, we discuss in Section 2 the approaches related to handling the dependability attributes in DWS. Section 3 presents general concepts of our approach as well as the UML profile suggested for each stage. Section 4 precise a roll-out scenario of our method. Section 5 describes the contributions of our approach to address dependability aspects of Data warehouse dedicated to Urgent Care Clinical Dashboard, a project led by National Health Services of United Kingdom. Finally, Section 6 concludes the paper by providing a summary and considerations for future research.

2 Related work

In this Section, we provide a brief overview on techniques addressing dependability aspects in the context of DWS, and then we present approaches focusing on requirement analysis in DWS especially non-functional ones that integrate dependability requirements.

As we mentioned previously, the original definition of dependability for a computing system is the ability to deliver service that can justifiably be trusted. This definition highlights the importance of trust justification. Another alternate definition specifies that dependability as the "ability to avoid service failures that are more frequent and more severe than is acceptable"(Laprie & Roche 1995).

Those abstract definitions integrate many attributes, encompassed in the large concept of dependability, and which are summarized in:

- Availability: ability of the system to be operational at the requested time,
- Reliability: ability of the system to provide continuity of service,
- Confidentiality: ability of the system to prohibit unauthorized disclosure of information,
- Integrity: ability of the system to prevent non-occurrence of improper alteration of information,
- Maintainability: the ability of the system to handle repairs and updates.

Considering the previous definitions both availability and reliability emphasize the avoidance of failures, and can be grouped together, and collectively defined as the avoidance

or minimization of service outages. Moreover, Associating integrity and availability together, as well as confidentiality, lead to security (Aviz & Randell 2001).

In the context of data warehouse systems, each of dependability attributes has been widely discussed in various works (Bellatreche et al. 2007; Bellatreche & Boukhalfa 2005).Thus, several proposals have focused on high availability data warehouse(Lau & Madden 2006) as well as the implementation of real-time ETL (Golab et al. 2009) and evaluation of the data reliability (Destercke et al. 2013). On the other hand, the dynamic nature of the system through its growing requirements, imposed to consider seriously its maintenance (Golfarelli & Rizzi 2009). Finally the criticality of the manipulated information gave rise to a significant number of works dealing with security issues(Blanco et al. 2009; Singh & Kumar 2012; Ahmad & Ahmad 2010). Despite the abundance of works dealing with dependability attributes, most have addressed them as separated and isolated aspects neglecting their potential interactions and their complementarity to ensure a dependable system. Moreover, the threats and means are not evoked except for (Lau & Madden 2006) who focused on fault tolerance to attain high availability of DWS.

Requirement analysis is the first phase of DWS that identifies which information is pertinent considering the user needs or/and the available data. The most cited works treating this phase can be subdivided under the following categories: supply-driven, goal-driven, user-driven (demand-driven) and model Driven (Golfarelli 2009).

All the previous cited works focused on the functional requirements and totally mistreated the non-functional ones which integrates dependability requirements (Golfarelli 2009). Only few works gave a special attention to such particular requirements.

To summarize, the consideration of dependability attributes from the expression of requirements is necessary to develop a trustworthy data warehouse. However, several challenges persist since these attributes are implicit and difficult to obtain from the first levels; in particular they are not objective and can be conflicted.

3 MDA approach for dependable Data warehouse Systems

MDA is an OMG's standard. it provides model-driven software development based on the promotion of machine-readable models in order to (i) overcome technology obsolescence, (ii) insure portability and integration, (iii) enhance productivity by automating many development tasks, thus focusing more on core logic business, and finally (iv) improve quality, maintenance and reusability by the concept of concerns separation. Basing on those specifications, It proposes the use of three kinds of models at different abstraction levels, to address the complete development lifecycle(Blanc & Salvatori 2011):

CIM: Computation Independent Model which represents the Business model based on system requirements.

PIM: Platform Independent Model which represents the Conceptual models without including information about specific platforms and technologies.

PSM: Platform Specific Model which represents the Logical models for a target platform or technology.

The system requirements are represented in a CIM. Then, this latter is refined into a PIM which can be automatically transformed into PSM for the target implementation technology. Finally, code can be generated from PSM basing on transformation Rules.

3.1 Proposed UML profiles for each MDA abstraction level

According to MDA, used models must be in conformity with their MetaModel or profiles which must be also conform to the MOF. In the following sections we will detail the different profiles proposed in each of the abstraction levels preconized by MDA.

3.1.1 Computation Independent Level

At this early level of abstraction, modeling dependability requirements reveals several challenges. These difficulties are usually caused by a strong focus on the functional requirements, and the user's profile, typically as decision-makers; therefore they are not familiar with such requirements. After a deep search for the existing profiles used at this level (Supakkul & Chung 2012)we find out that the profile NFR Framework(Chung & Leite 2009) is the most suitable. Because of the limited length of this paper, we have not conducted a comparative study to justify our choices, but the following paragraphs detail its foundations.

The NFR (Non Functional Requirement) Framework (Chung & Leite 2009)is a goal-oriented approach for addressing NFRs which include the dependability ones. It considers NFRs as softgoals to be addressed by applying an iterative process based on many concepts such as decomposition (AND/OR), operationalization and argumentation.

The reason behind choosing this framework is its implicit and realistic recognition of the difficulties that are associated with NFRs: (i) Difficulty to define an NFR term completely unambiguously without using any other NFR term, which in turn will have to be defined. (ii) Difficulty to explore a complete list of possible solutions and choose the best, or optimal, solution.

3.1.2 Platform Independent Level

This level is characterized by the development of PIM models corresponding to the analysis and design. Thus, at this stage the NFR Framework is more appropriate because it is centered on the requirements. Among the profiles provided by the OMG, QoS-FT profile (Solberg et al. 2004) provides a set of concepts for the integration of modeling aspects of service quality and fault tolerance, particularly in the context of real-time systems for analysis and design. This profile is based on the construction of models that can be used to make predictions about the characteristics of QoS. These models facilitate the communication between developers in standardized approach. They also allow interoperability between the various design tools.

3.1.3 Platform Specific level.

At this stage, the refinement of dependability aspects, was carried out, and their optimal combination is found and justified. We simply need to specify them according to the target platform as specified by the MDA standard. However, we encountered two difficulties mainly because this is the final phase that represents the modeling level and it should consider the possibility of code generation.

The first problem is due to the fact that the dependability aspects do not make sense if they are not related to system features which will be improved. In addition, we note that experience shows that capturing non-functional requirements without mapping them into the conceptual model may provoke loss of information (Golfarelli & Rizzi 2009). To solve this problem we propose to consider aspects resulting from the previous steps as "crosscutting concerns" so that they can benefit from the Aspect Oriented Software Development (AOSD) (Krechetov et al. 2006) paradigm based on the separation of concerned. Thus aspects can be woven with functional component and finally form a complete model.

However, the second problem is related to the Multi-Layer nature of the data warehouse system. The latter included several type of platform used for each layer (heterogeneous data sources, tools for the extraction, transformation and data loading, warehouse databases, restitution tools and data exploration). Thus, to achieve modeling the dependability aspects for specific platform we must first specify the concerned layer.

In this level we can distinguish two categories of dependability aspects: those related to software tools and those concerning physical infrastructure. The modeling of the two categories will fulfill the standard required system dependability. To model this latter, while specifying the destination platform, we propose to use the component diagrams and deployment.

The component diagram is used to represent the software aspects in the form of component that can be woven with the functional components of DWS according to the AOSD mechanisms which are out of scoop. Whereas the deployment diagram is used to represent the hardware that will be used to provide the required level of dependability on which the components (dependability and functionality) will be deployed.

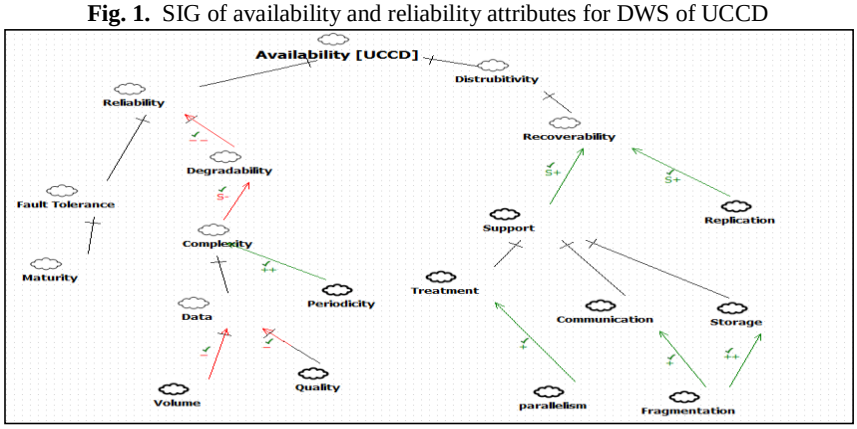
4 Case study: Data warehouse system for Urgent Care Clinical Dashboard

This case study is extracted from the system specifications for Urgent Care Clinical Dashboard (UCCD) elaborated by the National Health Service (NHS) of UK[1]. NHS is responsible for all healthcare services of the Kingdom. In recent years, it has initiated UCCD project whose main objectives are:

- Implement Urgent Care Clinical Dashboard to all health community across England starting with 12 pilot sites;
- Create a toolkit to standardize Urgent Care Clinical Dashboard including standards, logical architecture, library data flow, detailed design documentation, project management objects.

We start by developing CIMs. In this sense, Figures 1, 2 and 3 illustrate the SIG of the NFR Framework profile developed for each of dependability attributes. Then, the integration of the different CIMs, allows confronting dependability aspects to determine aspects to retain or to decline for the next level as shown in Figure 4.

Fig. 1. SIG of availability and reliability attributes for DWS of UCCD



[1] https://www.networks.nhs.uk/nhs-networks/qipp-urgent-care-gp-dashboard/documents/test-1/clinical-metrics/?order=date

Fig. 2. SIG Maintainability attribute for DWS of UCCD

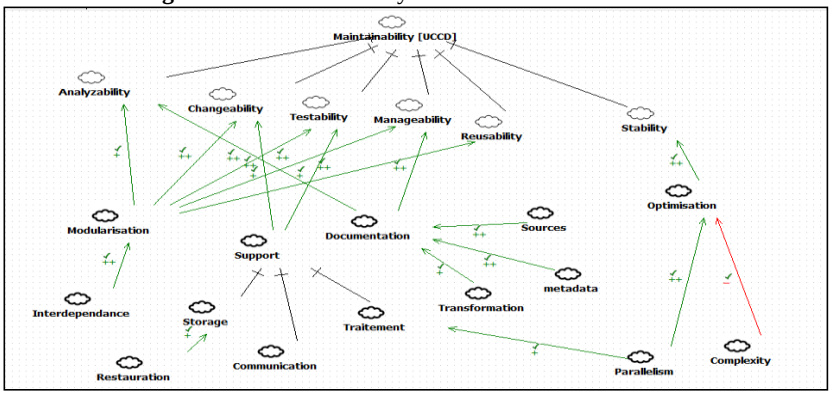


Fig. 3. SIG of security attribute for DWS of UCCD

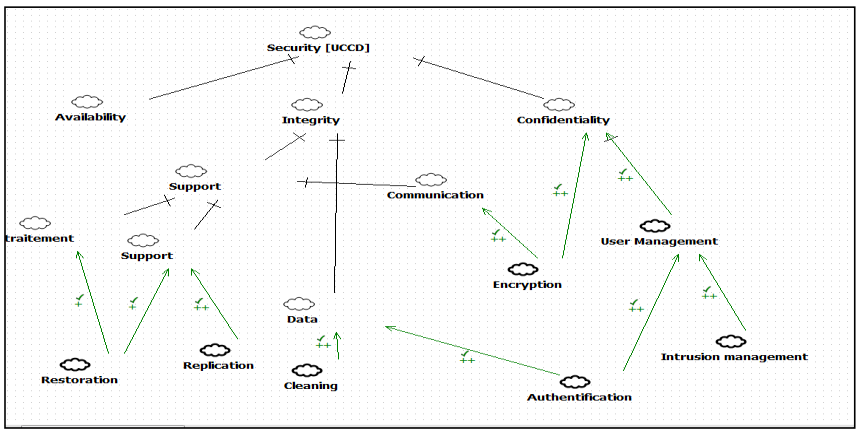


Fig. 4. A part of the Integration of SIG's dependability attributes

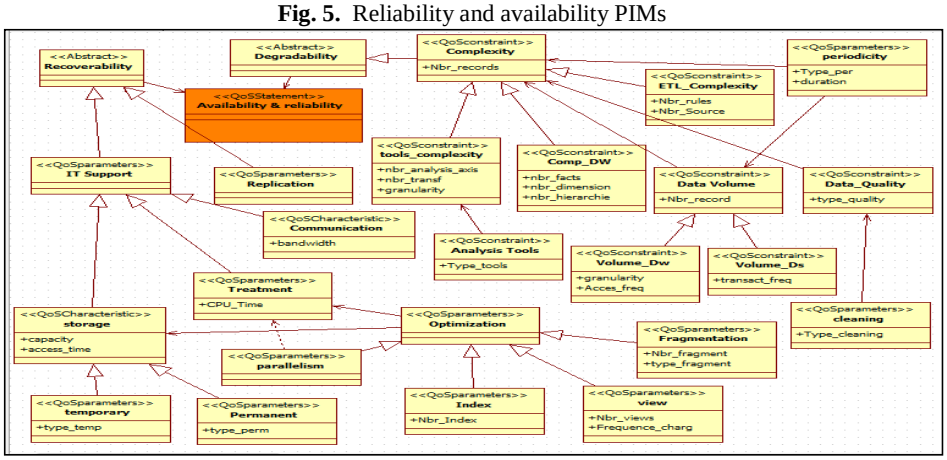
The integration in figure 4 shows that the data encryption improves the integrity and Privacy, which are essential aspect to ensure security. However encryption techniques increase the complexity (Waters 2011). It directly affects the modularity and thus affects the maintainability. On the other hand, the complexity makes it difficult to achieve the distributivity and thus decreases the reliability of the system, which automatically assigns its availability.

If we take the hypothesis that the UCCD system will be a system that will be applied in 90% internally, the encryption implementation as required by NHS in (Wood & Penny 2012) may affect the availability and reliability of the system (Waters 2011). Thus, through the interactions study, we can justify and propose to apply encryption on the data which will be transmitted outside the clinic system. This choice will not penalize maintainability, availability and reliability of UCCD system all by ensuring an adequate level of security.

4.1 Platform Independent Model

Development of PIM based on the QOS-FT profile, will allow detailing the dependability aspects retained from the previous level. For this, we will enrich these aspects without involving the characteristics of any platform. Thus, the models will be portable and achievable regardless of the technical details of implementation. The following figures 5,6,7,8 show the proposed PIMs.

Fig. 5. Reliability and availability PIMs



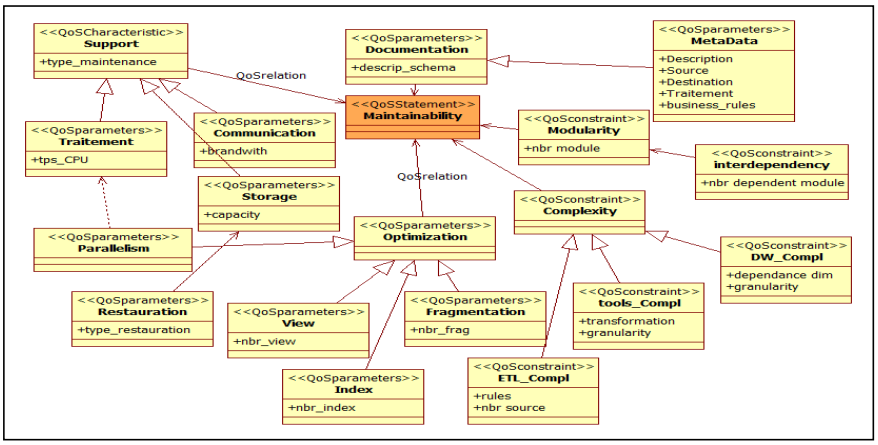
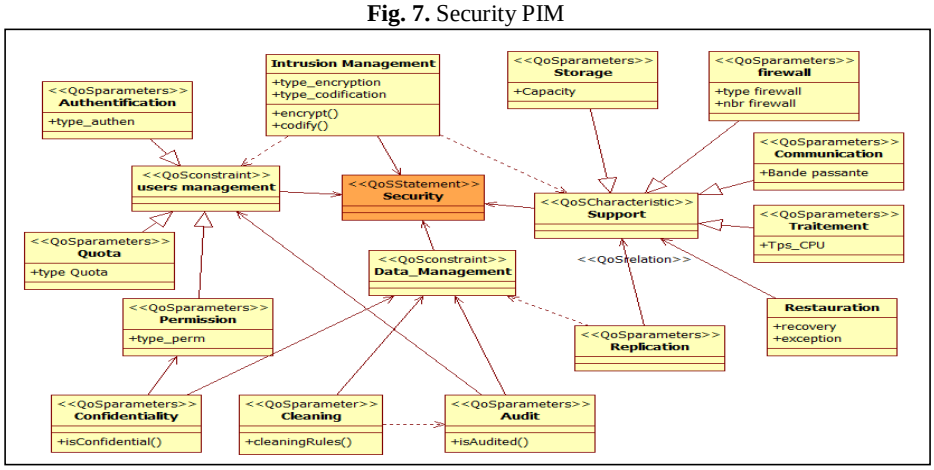


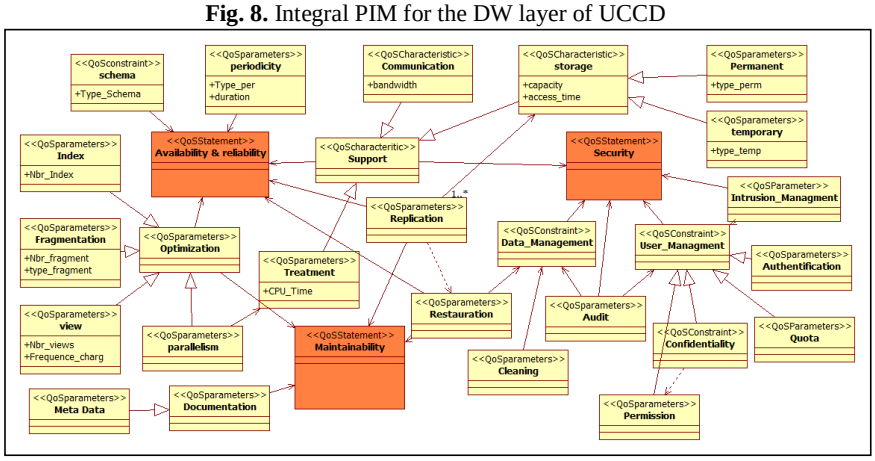
Fig. 6. Maintainability PIM

Fig. 7. Security PIM



At this level we will develop the component and deployment diagrams. Thus, we'll highlight the weaving of aspects with the functional part in modeling, just before their implementation. We will apply these concepts to DW layer to highlight the integration of aspects of dependability refined from the first step of our approach.

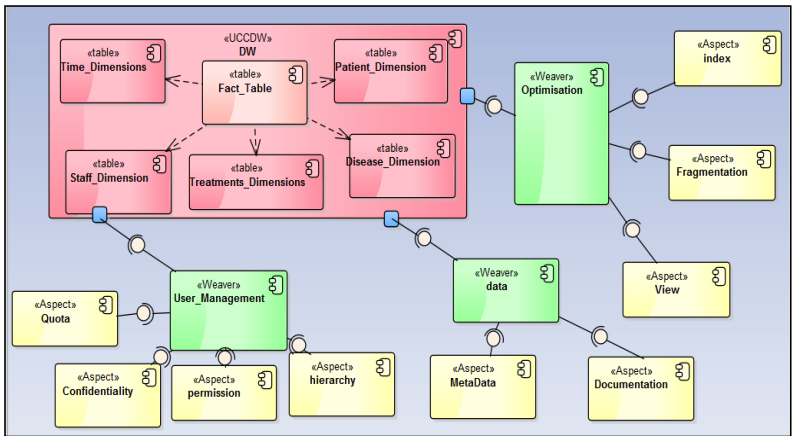
Fig. 8. Integral PIM for the DW layer of UCCD



4.2 Platform Specific Level

The component diagram represents the first part of the modeling of PSM as shown at figure 9. Thus all aspects will be woven with the functional DW following an approach Aspect Oriented. DW schema was realized based on the examples dashboard provided by NHS.

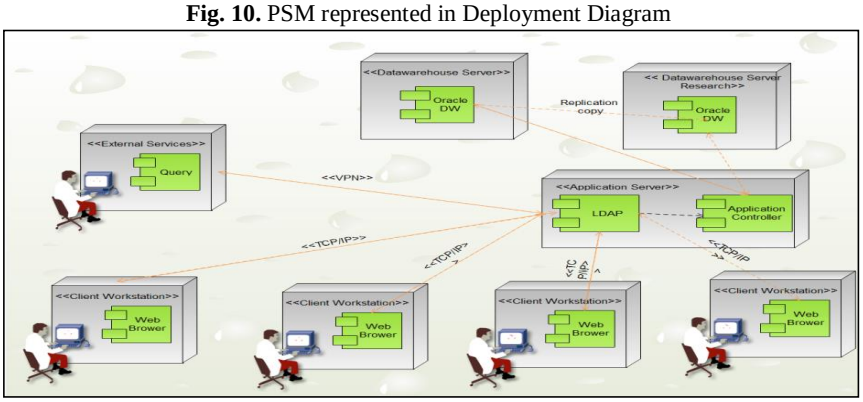
Fig. 9. Composite Diagram representing the weaving with functionality system



The deployment diagram in Figure 10 explains the technical architecture choices that resulted compromises made from the previous phases. This diagram is complementary to the component diagram to developing an integral PSM of the data warehouse layer.

We specify that the NHS has already proposed a set of technical architecture scenarios in their specifications. Based on these, we developed the model above, taking into account the constraints resulting from the interaction of the aspects of dependability handle by our methodology.

Fig. 10. PSM represented in Deployment Diagram



Concerning availability and reliability attributes, we propose to achieve DW server replication using two independent servers. The latter will ensure the continuity of the UCCD services in the event of a failure of the primary server DW. The application server will switch client requests to the Replica server after the passivity detection of the primary server, which guarantees the system reliability. In the case of the primary server overload periods, when the response time begins to approach an unacceptable threshold (to be specified), the application server can load balance requests based on their urgency and criticality (eg: a query that is sent from an operations room or ambulance is more critical than the one that comes from a consultation room).

Regarding maintainability attribute, the low coupling between software components thanks to Object Oriented Design increases significantly the system maintainability. In addition the physical duplication mechanism and its separation from the application server can execute maintenance operations without assigning system response.

From the security point of view, although the NHS insists in a special document (Wood & Penny 2012) on the data encryption to ensure their confidentiality , the confrontation aspect's model to find their optimal combination has result to not recommend encrypting data in the clinical local network, because it interferes directly to data availability especially during overload periods. But to ensure an adequate level of security, we propose the use of LDAP (Active Directory for Windows, Linux Open LDAP, etc.) to provide centralized management of users, supported by a connection policy based on security certificates. Thus the users cannot access the data if they are identified by the LDAP server and on machines that contain security certificates authorized by the system administrator.

5 Discussion and open challenges

In this paper, we introduced an MDA approach to deal with dependability aspects in Data warehouse context. Our focus lies to the fault prevention and fault tolerance. Our aim, in one hand, is to prevent faults that can affect dependability aspects at the earlier stages of data warehouse development. Moreover, Due to the unavoidable occurrence of faults, the degree to which a system holds the attributes of dependability is related to their priority and criticism. For example the data warehouse system for healthcare requires dependability attributes levels different from the military or banking one. Therefore, we propose to address the fault tolerance by confronting the dependability attributes in order to find out their optimal combination for each abstraction level in the other hand. To attain our objectives, we propose the use of different UML profiles for each MDA model: NFR Framework for CIM, QOS-FT for PIM and deployment and composite UML diagrams for PSM. Finally, we conclude the paper with a case study of the Data warehouse system for Urgent Care Clinical Dashboard. In our future work, we plan to define the transformation rules to enable automation of our approach until the code generation and to generate a trace model to justify the traceability of dependability requirements. Furthermore a case study, with dependability quantitative metrics, has to be done. However, we think this is the first step to a broad area to be explored.

References

Ahmad, S. & Ahmad, R., 2010. An improved security framework for data warehouse: a hybrid approach. In *Information Technology (ITSim), 2010 International Symposium in*. IEEE, pp. 1586–1590.

Aviz, A. & Randell, B., 2001. Fundamental Concepts of Computer System Dependability. , pp.1–16.

Avižienis, A. et al., 2004. Basic concepts and taxonomy of dependable and secure computing. *IEEE Transactions on Dependable and Secure Computing*, 1(1), pp.11–33. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1335465 [Accessed August 6, 2014].

Bellatreche, L. et al., 2007. Selection and pruning algorithms for bitmap index selection problem using data mining. In *Data Warehousing and Knowledge Discovery*. Springer, pp. 221–230.

Bellatreche, L. & Boukhalfa, K., 2005. An evolutionary approach to schema partitioning selection in a data warehouse. In *Data Warehousing and Knowledge Discovery*. Springer, pp. 115–125.

Bezivin, J. & Gerbe, O., 2001. Towards a precise definition of the OMG/MDA framework. In *Proc. 16th Annu. Int. Conf. on Automated Software Engineering (ASE 2001)*. IEEE Comput. Soc, pp. 273–280. Available at: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=989813&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D989813 [Accessed August 26, 2014].

Blanc, X. & Salvatori, O., 2011. *MDA en action: Ingénierie logicielle guidée par les modèles* EYROLLES, ed., Available at: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:MDA+en+action#0 [Accessed August 26, 2014].

Blanco, C. et al., 2009. Data Warehouse Security. In *Encyclopedia of Database Systems*. Springer, pp. 675–679.

Blanco, C. et al., 2010. Defining and transforming security rules in an MDA approach for DWs. *International Journal of Business Intelligence and Data Mining*, 5(2), p.116. Available at: http://www.inderscience.com/link.php?id=31283 [Accessed August 26, 2014].

Bondavalli, A. et al., 2001. Dependability analysis in the early phases of UML-based system design. , pp.265–275.

Chung, L. & Leite, J.C.S. do P., 2009. On non-functional requirements in software engineering. In A. T. Borgida et al., eds. *Conceptual Modeling: Foundations and Applications*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 363–379. Available at: http://www.springerlink.com/index/10.1007/978-3-642-02463-4 [Accessed August 26, 2014].

Destercke, S., Buche, P. & Charnomordic, B., 2013. Evaluating data reliability: an evidential answer with application to a Web-enabled data warehouse. *Knowledge and Data Engineering, IEEE Transactions on*, 25(1), pp.92–105.

Gerhard, G., Juris, H. & Jan, van L., 2012. Measurement, Modelling, and Evaluation of Computing Systems and Dependability and Fault Tolerance. In J. B.Schmitt, ed. *16th International GI/ITG Conference MMB & DFT 2012 Kaiserslautern, Germany,*. Springer. Available at: http://www.ulb.tu-darmstadt.de/tocs/59142804.pdf [Accessed February 26, 2015].

Golab, L., Johnson, T. & Shkapenyuk, V., 2009. Scheduling updates in a real-time stream warehouse. In *Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on*. IEEE, pp. 1207–1210.

Golfarelli, M., 2009. From User Requirements to Conceptual Design in Data Warehouse Design. In L. Bellatreche, ed. *Data Warehousing Design and Advanced Engineering Applications : Methods for Complex Construction*. IGI Global, pp. 1–18. Available at: http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-60566-756-0 [Accessed August 26, 2014].

Golfarelli, M. & Rizzi, S., 2009. *Data warehouse design: Modern principles and methodologies*, McGraw-Hill, Inc.

Krechetov, I. et al., 2006. Towards an integrated aspect-oriented modeling approach for software architecture design. In *8th Workshop on Aspect-Oriented Modelling (AOM. 06), AOSD*. Citeseer.

Laprie, J. & Roche, C., 1995. Dependable Computing : Concepts , Limits , Challenges. In *Proc. 25th IEEE Int. conf Fault-tolerant computing*. WA: IEEE Comput. Soc, pp. 42–54. Available at: http://dl.acm.org/citation.cfm?id=1899261.

Lau, E. & Madden, S., 2006. An integrated approach to recovery and high availability in an updatable, distributed data warehouse. In *Proceedings of the 32nd international conference on Very large data bases*. VLDB Endowment, pp. 703–714.

Mazón, J.-N. & Trujillo, J., 2008. An MDA approach for the development of data warehouses. *Decision Support Systems*, 45(1), pp.41–58. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0167923606002077 [Accessed August 26, 2014].

Paim, F.R.S. & de Castro, J.F.B., 2003. DWARF: an approach for requirements definition and management of data warehouse systems. In *Proc. 11th IEEE Int. Conf. Requirements Engineering*. CA: IEEE Comput. Soc, pp. 75–84. Available at:

http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1232739 [Accessed August 26, 2014].
Rilston, F., Paim, S. & Castro, J.F.B., 2002. Enhancing Data Warehouse Design with the NFR Framework. In *Proc. 5th Workshop on Requirements Engineering (WER2002)*. Valencia, Spain.
Schauerhuber, A. et al., 2007. *A survey on aspect-oriented modeling approaches*, Available at: http://publik.tuwien.ac.at/files/pub-inf_4920.pdf.
Singh, I. & Kumar, M., 2012. Evaluation of approaches for designing secure data warehouse. In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics*. ACM, pp. 69–73.
Solberg, A., Oldevik, J. & Aagedal, J.Ø., 2004. A framework for QoS-aware model transformation, using a pattern-based approach. In *On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE*. Springer, pp. 1190–1207.
Soler, E., Stefanov, V. & Mazón, J., 2008. Towards comprehensive requirement analysis for data warehouses: Considering security requirements. *Proc. 3rd Int. Conf. Availability, Reliability and Security*, pp.104–111. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4529327 [Accessed August 26, 2014].
Supakkul, S. & Chung, L., 2012. The RE-Tools: A multi-notational requirements modeling toolkit. In *Requirements Engineering Conference (RE), 2012 20th IEEE International*. IEEE, pp. 333–334.
Trivedi, K.S. et al., 2009. Dependability and security models. In *2009 7th International Workshop on Design of Reliable Communication Networks*. Ieee, pp. 11–20. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5340029.
Waters, B., 2011. Ciphertext-policy attribute-based encryption: An expressive, efficient, and provably secure realization. In *Public Key Cryptography–PKC 2011*. Springer, pp. 53–70.
Wood, J. & Penny, M., 2012. *Approved Cryptographic Algorithms Good Practice Guideline*, Available at: http://systems.hscic.gov.uk/infogov/security/infrasec/gpg/acs.pdf.

Résumé

Les systèmes décisionnels sont l'un des principaux systèmes utilisés par les gestionnaires et les décideurs pour analyser le passé, le présent et l'avenir de toute organisation. Ainsi, ils gèrent des informations cruciales, et donc ils doivent assurer un haut niveau de sûreté de fonctionnement. Traditionnellement, les approches de modélisation ont focalisé sur les exigences fonctionnelles, mais ont rarement traité les aspects de sûreté de fonctionnement dans leurs globalités. En outre, la plupart de ces dernières approches offrent des solutions partielles qui ne traitent que des aspects isolés de la sûreté de fonctionnement des systèmes décisionnels, tels que la disponibilité, la fiabilité, la maintenabilité et la sécurité. Et par conséquent, ils ne fournissent pas un cadre intégré et standard pour la modélisation et l'intégration de ces exigences dans le processus de développement précoce. Pour répondre à cette problématique, notre papier propose une approche intégrée basée sur la MDA (Model Driven Architecture) afin de tenir en compte les aspects de sûreté de fonctionnement dès les premiers stades de développement afin de résoudre leur interaction et d'assurer leur traçabilité. Pour ce faire, nous allons décrire les différents profils UML pour concevoir les modèles MDA. Enfin, une étude de cas est fournie pour illustrer et démontrer l'impact de notre approche.

Features Extraction from Mammographic Masses Using Genetic Active Contours

Ferkous Chokri[*]. Merouani Hayet Farida [**]
[*]Laboratoire des Sciences et Technologie de l'Information et de la Communication (LabS-TIC), Université 08 mai 1945 Guelma, Algérie, Email: ferkous.chokri@labstic.net
[**]Laboratoire de Recherche en Informatique (LRI),
Université Badji Mokhtar Annaba, Algérie, Email : hayet_merouani@lri-annaba.net

Abstract. In this paper we propose a system for extracting features from mammographic masses, this system is inspired overall by the approach of the doctor during the radiologic examination as it was agreed in BI-RADS (Breast Imaging reporting System and Data System). The segmentation of the masses in our approach is manual because we assume that the detection is already made. Thereafter it is sought to detect the contours of the mass using a genetic active contours, the best snake of the population at a given instant is considered as approximating contour of the abnormality. Energies of continuity, curvature and the external energy of the final snake, in addition to the size and intensity of the abnormality form the features vector.

1 Introduction

The cancer of the center is one of the great concerns of the health which started to claim importance in the medical research because of its high prevalence and the rates of detection.

According to the statistics of the national institute of the Algerian public health (INSP) [6], the number of deaths annual due to this disease east estimates at 3500. Moreover, we count 7000 new cases of disease per year; the incidence was multiplied by 5 in 20 years.

During an examination mammographic, the radiologist does nothing but solve one problem of vision. It is starting from this point and of recent techniques of the analysis and pretreatment of image, that the idea to use the machine to facilitate the work of the radiologist.

An assistance system with the typical decision must include four big steps which are the segmentation, extraction of the characteristics and classification.

The phase of extraction of the characteristics particularly interests us in our research. In this paper we propose a system which makes it possible to extract starting from a mammographic mass a vector from characteristics which will be used thereafter in the classification.

2 Bi-Rads features

BI-RADS (Breast Imaging System And Dated-System) is a system of assistance to the drafting of the account returned used more and more in the world.

According to BI-RADS [9], a mass is characterized by its form (round, oval, lobule, irregular), its contour (circumscribed, micro lobule, masked, indistinct, speculated), its density (high, average, weak, lubricating and mixed).

3 Active contours theory

Active contours, or snakes, are computer-generated curves that move within images to find object boundaries, the goal of this method is to find a contour that best approximates the perimeter of an object, an active contour is a curve that moves towards the sought-for shape in a way controlled by internal forces - such as rigidity and elasticity - and an image force. The image force should attract the contour to certain features, such as edges, in the image. This is done by creating an attractor image, which defines how strongly each point in the image should attract the contour.

The general form of Snake, described by Kass et al [1]. represents contour by a vector having a length of arc *s*, the equation making it possible to measure energy is given by:

$$E_{Snakes} = \int (E_{\text{int}}(v(s)) + E_{ext}(v(s)) + E_{image}(v(s)))ds \quad (1)$$

Where E_{int} is the internal energy which manages the coherence of the curve, It maintains the cohesion of the points and the stiffness of the curve, the term of regularization is:

$$E_{\text{int}}(v) = \int_0^1 (\frac{\alpha}{2}(s)\|v'(s)\|^2 + \frac{\beta}{2}(s)\|v''(s)\|^2)ds \quad (2)$$

The *v'* terms and *v''* are the derivative first and second of *v* compared to *s*.

E_{ext} is external energy, makes it possible to manage the regularization of active contour, it corresponds to the adequacy with the data. This external energy takes into account the characteristics of the image; The last term, E_{image}, measurements the quality of the image, such as the intensity or the force of the edges.

The parameters α, β and γ are used to balance the influence of these three terms.

4 Genetic Active Contours

It acts, to put the exploratory capacity of the genetic algorithms at the research center of a contour which we define starting from the method of active contours. Thus, the required solution is a contour, i.e. a whole of points and the function of evaluation is related to the total energy of active contour. It is thus a question of minimizing this energy. The difference with the majority of the algorithms in active contours lies in the fact that one does not make evolve a snake, but a population of snakes. Only best the snake of the population at a given moment is regarded as the approximation of contour.

4.1 Coding of the chromosomes

We used polar codification, this method decode a chromosome in coordinate polar (θ, ρ), having for origin the center (x_c, y_c) of the image.

$$\rho = \sqrt{(x - x_c)^2 + (y - y_c)^2} \quad (3)$$

In this case, the angle θ will vary from *0* with 2π and for an image of width *W* and height *H* the space of definition becomes:

$$\rho_{\max} = \left\{ \begin{array}{ll} \dfrac{W}{2Cos_\theta} & if\ \ \theta \in [0.\dfrac{\pi}{4}] \cup [\dfrac{7\pi}{4}.2\pi] \\ \dfrac{W}{2Sin_\theta} & Si\ \ \theta \in [\dfrac{\pi}{4}.\dfrac{3\pi}{4}] \\ \dfrac{-W}{2Cos_\theta} & Si\ \ \theta \in [\dfrac{3\pi}{4}.\dfrac{5\pi}{4}\ ] \\ \dfrac{-W}{2Sin_\theta} & Si\ \ \theta \in [\dfrac{5\pi}{4}.\dfrac{7\pi}{4}] \end{array} \right\} \qquad (4)$$

In order to reduce the space of research, knowing that we seek a contour closed more or less centered on the image, we make implicit the value of θ. The points are distributed in all the directions starting from the origin, i.e. for a snake of 40 points, we takes

$$v_i = (\rho_i, \theta_i),\ \ \theta_i = \frac{i}{20}\pi$$

In this coding, only ρ is explicitly coded; that reduces half the length of the chromosomes.

4.2 Creation of the initial population

The first stage of the genetic algorithm is the construction of an initial population, a whole of chromosomes which represent contours, the chromosomes must be by chance given, i.e. for each contour, a succession of points generated by chance ordered by the angle θ.

4.3 Selection

We used the method of selection in K-tournaments, for a population of size N; the method can be described by three stages.

1. Choose K individuals among the N present in the population.
2. Recopy best these K individuals in the new population.
3. Repeat the process N 'times.time.

4.4 Crossing and Mutation

The crossings with one and several points of cut, as well as the uniform crossing, the experimentation of Rossel [3] shows that the choice of a method of crossing affects little the quality of the final solution obtained. On the other hand, that has a strong influence on the speed of convergence of the algorithm. It also noted that beyond ten points of cut, the speed of convergence increases little. We thus adopted this method which has the advantage of providing very correct results without to be too expensive in computing times.

The choice of the rate of mutation has a great influence on the effectiveness of the algorithm. Traditionally the rate of mutation μ is fixed by the formula $\mu=l/L$, it could be interesting to increase the rate of mutation only on the points considered as “bad” compared to the others (in this case, the change is known as adaptive).

5 Process of Features Extraction

5.1 Over-densities Segmentation

The goal of this stage is to extract the areas from interest in the gland mammary which are the abnormal over-density;

The Over-densities of interest can have on the stereotype mammographic of well-defined contours or badly definite according to their nature.

In our system the stage of segmentation is done manually, because we supposes that the detection of the mass is already made, therefore the entry of our system is a binary image which is composed of a white mass and a black background.

5.2 The Setting in scale

The setting in scale plays a big role during the calculation of the characteristics of form, by respecting a scale of reference makes it possible the mass identical to have same "shape features" whatever their sizes.

5.3 Edge detection

To detect contours from an abnormality. two phases are necessary; in the first phase we seek the center $C\ (x_c, y_c)$ of the mass, and in the second phase, starting from the center $C\ (x_c, y_c)$ we seek to detect contours of the anomaly by exploiting the algorithm of genetic active contours (GAC).

A population of 200 contours is generated by chance, and each individual (contour) contains 40 chromosomes (points), each chromosome represents a point in the space of research. We code the points of the contours generated in the chromosome into polar, the zone of research is defined by a circle of a diameter of 500Pixels, around the center of the anomaly $C\ (x_c, y_c)$, the space of definition becomes:

$$\rho_{\max} = \frac{D}{2} = \frac{500}{2} = 250 \qquad \forall\ \theta \in [0,\ 2\pi]$$

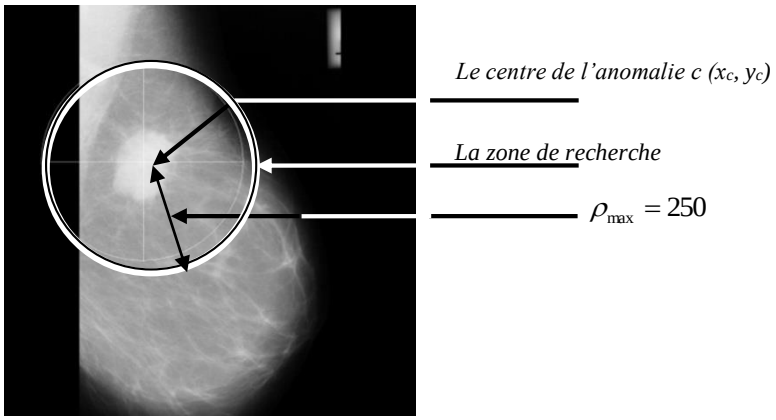


FIG. 1– *representation of the search area in GAC*

Using polar coding; only ρ is coded, and each chromosome is encoded by the union of 40 axes generated ($\rho_1, \rho_2, \ldots \rho_{40}$).

Fitness criterion

The fitness of genetic active contours (CAG) is calculated by the following function:

$$F_{CAG} = \sum_{i=1}^{40} f(i) \qquad (5)$$

Where $f(i)$ is the fitness of the chromosome i.

$$f(i) = E_{\text{int}}(i) + E_{ext}(i) \qquad (6)$$

Stop Criterion

Once the maximum number of generations reached, fixed at 200 generations, we choose the best final contour of the population like the approximation of the contour of the anomaly; the following figure illustrates the results of detection of contours of anomaly of some withdrawn mammography of MIAS database [8].

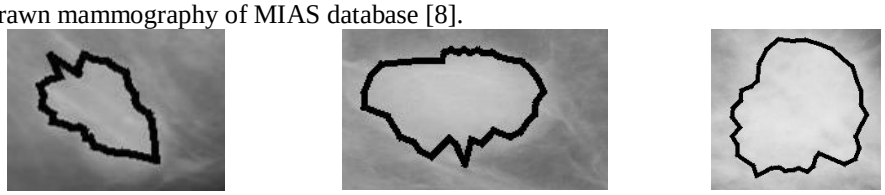


FIG. 2 – results of the detection of contours of the masses of the mammographies mdb025, mdb005, mdb081 of MIAS database with active contours genetics.

6 Features extraction

The goal of our work is to extract the characteristics relating to the zone included by the contour elected like better in the preceding stage, for given the characteristics of this mass it is enough to extract information according to:

The form of the mass: the form of the mass is strongly related to the internal energy of elected contour, i.e. energies of curve and continuity

The curve which has a behavior like a thin section. It takes an important value on the other hand when the curve curves quickly i.e. for obtaining corners and the forms complex, when this energy is weak the curve in the case of tightens towards a circle a contour closed.

Type of contours of the mass: the type of contours so related to the external energy of elected contour. Because external energy takes an important value on the other hand when the mass has well defined and circumscribed contours, one obtains a weak energy in the case of badly definite or unobtrusive contours.

Size of the mass: it is easily to calculate the size "T" of the mass which is represented by the zone included in the elected contour.

Intensity of the mass: the intensity "I" of the mass equal to the average of intensities of the pixels of the zone is included by elected contour.

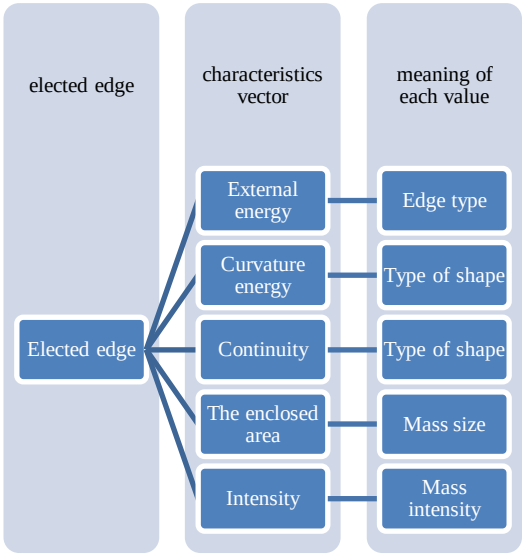


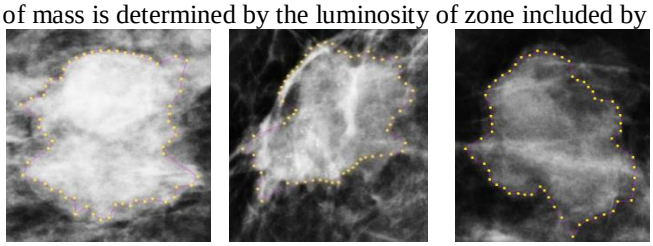
FIG. 3 – *characteristics vector of elected edge*

7 Tests

These tests were made, by fixing the coefficients : $\alpha = 0,5$; $\beta = 0,5$; $\gamma = 5$.

We balanced internal energy with a small value, since we do not have information a priori on the degree of curve or mass continuity, and since the lesions mammary do not have a form not defined well, only the gradient has an important significance on contours of the mass, for that, we balanced external energy with a great value.

The density of mass is determined by the luminosity of zone included by contour.



Weak Strong Average

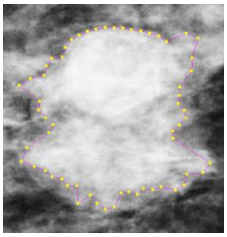
FIG. 4 – *detection from various densities.*

Table (1) represents the various values obtained for the three types of density:

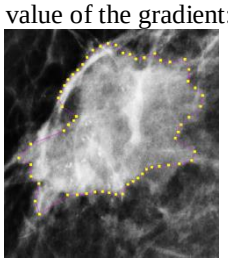
	Mass density		
	Fort	Medium	Low
Area density	3425,4	2711 ,96	1810,1

TAB. 1 – *Results of densities of the 3 masses of figure 4.*

The type of contour is connected directly with the value of the gradient:



a- undefined



b- well-defined

FIG. 5 – masses (a) with an undefined contour (b) with a well-defined contour.

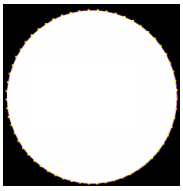
With the same parameters:

	Type du contour	
	undefined	well-defined
external Energy	684,4	865,5

TAB. 2 – Results of the parameters of contours unobtrusive and clear of figure 5.

A clear contour is characterized by a higher gradient; on the other hand an unobtrusive contour is characterized by a weak gradient.

Energies of continuity and curve describe the form of the mass, a circular mass of form is represented by a contour having minimal energies, and on the other hand an irregular form is represented by a contour having maximum energies.



a- Circle

b- Lobulated

c- Irregular

FIG. 6 – various shapes (A) circle (b) Lobulated (c) irregular.

The following table shows the results of continuity and curve obtained on form circular, lobule and irregular.

	Mass shapes		
	Circle	Lobulated	Irregular
Curvature	123	1543	2140
Continuity	3034,84	3262,60	3232,12

TAB. 3 – Results of the parameters of form of figure 6.

8 Conclusion

In this paper we presented a system of extraction of the characteristics of the mammographic masses, the vector of the characteristics is composed of four principal characteristics, the form which results in energies of continuity and curvature, the type of contour which is determined by external energy, the size expressed by the number of pixels of the mass and the density by the average intensity of the pixels of the abnormality.

We must finish the decisional part of the algorithm with regard to the classification of the anomalies detected according to Bi-RADS from 1 to 5.

References

[1] Kass M, Witkin A, Terzopoulos D., Snakes: Active Contour Models 1(4). pp 321-31, 1987

[2] Roshan D, Koichi H, Breast Skin-Line Estimation and Breast Segmentation in Mammograms using Fast-Marching Method. Information technology journal 7(3): pp 490-496. 2008

[3] Jean-Jacques R, Les contours actifs, une méthode de segmentation application à l'imagerie médicale. Thèse de Doctorat en informatique de l'Université tours, pp 10-35, 2003

[4] Montanari U. A method for obtaining skeletons using a quasi-euclidean distance. Journal of the Association for Computing Machinery, 15 , pp600–624, 1968. 1968.

[5] Borgefors G. Distance transformations in arbitrary dimensions. Computer Vision, Graphics, and Image Processing 27, pp 321–345,1984.

[6] Abid L, cancer epidemiology in algeria: best use of cancer registers ; african journal of cancer, Vol 1, Springer, pp 98-103, 2009.

[7] Suckling J. The Mammographic Image Analysis Society Digital Mammogram Database Exerpta Medic. Exerpta Medica, International Congress Series 1069, pp.375-378. 1994.

[8] Vachier C, « Extraction De Caracteristiques,Segmentation D'image Et Morphologie Mathématique », Thèse De Doctorat , 1995.

[9] L. Lévy, M. Suissa, J. Bokobsa, H. Tristant, J.-F. Chiche, B. Martin, G. Teman, «Breast Imaging Reporting System and Data System », Gynécologie Obstétrique & Fertilité 33, 338–347, Elsevier, 2005.

Resumé

Dans ce papier nous proposons un système pour l'extraction des caractéristiques des masses mammographiques, ce système s'inspire globalement de l'approche du médecin lors de l'examen radiologique en se basant sur le système d'aide à la rédaction des comptes rendu s(Breast Imaging Reporting System and Data System) BI-RADS qui permet de décrire les anomalies rencontrées en mammographie ; La segmentation des masses dans notre approche est manuelle car on suppose que la détection est déjà faite. Par la suite on cherche à détecter les contours de l'anomalie en utilisant les contours actifs génétiques, le meilleur snake de la population à un instant donné est considéré comme l'approximation du contour de la masse. Les énergies de continuité, de courbure et l'énergie externe du snake final ainsi la taille et l'intensité de l'anomalie forment le vecteur des caractéristiques.

Graph topology for Protein-Protein Interaction Networks Clustering

Smail BOUZERGANE*, Mustapha DERRHI *, Ahmed MOUSSA *

LabTIC, ENSAT, Abdelmalek Essaadi University, BP 1818, Tangier Morocco *
*bouzergane.smail@gmail.com
*mderrhi@yahoo.fr
*amoussa@uae.ac.ma

Abstract. Protein-protein interaction (PPI) networks play important biological roles in cells. Their study allows understanding the behavior of proteins inside the cell. Many computational methods have been proposed to detect protein complexes, to isolate groups of interacting proteins involved in the same biological processes or to perform together specific biological functions. In this paper, we first review the main graph-based clustering methods that have been applied to PPI networks. Then, we present an experimental comparison between these graph-based approaches and similarity-based approaches, using two types of data. We will also compare efficiencies and relevance of these methods.

1 Introduction

The increasing amount of PPI data has enabled us to detect protein complexes from the PPI networks. A PPI network is mostly represented as an interaction graph with nodes represents involved proteins and edges encode their interactions. Various topological properties of protein interaction networks have been studied, such as the network diameter, the distribution of vertex degree, the clustering coefficient and so on (Bader and Hogue, 2003). These network analysis have shown that protein interaction networks have the features of a scale-free network (Li et al., 2004) and "small-world effect" (Del Sol et al., 2005). Beyond the discussions of the scale-free and small-world properties, an important challenge for system biology is to understand the relationship between the organization of a network and its function. It has been shown that clustering protein interaction networks is an effective approach to achieve this goal (Brohée and van Helden, 2006).

The problem of detecting protein complexes using PPI networks can be computationally addressed by using clustering techniques. Clustering in PPI networks means grouping together proteins which share a larger number of interactions, which are considered to represent functional modules. Possible uncharacterized proteins in a cluster may be assigned to the biological function recognized for that module. PPI networks have various characteristics which have to be taken into account when developing clustering algorithms for detecting functional complexes. Therefore a number of clustering approaches have been proposed to extract relevant modules from PPI networks.

In this work, we mainly focus on methods that only use graph topology for detecting clusters in PPI networks, and employ similarity measures between proteins. Firstly, we provide a review of the main clustering techniques proposed for detecting protein complex from PPI networks in order to guide researchers to develop new methods and practitioners to apply these methods by providing information about their availability. Then, we compare the results obtained using graph-based approaches with those returned by the recent similarity-based approaches and we discuss open challenges that can stimulate further research. Finally, to guide the reader in the choice and application of existing PPI network clustering tools, we also provide information about their availability.

2 Review of Clustering methods

Clustering approaches for PPI networks can be broadly characterized as distance-based or graph-based (Lin et al., 2006). Distance-based clustering approaches focuses on the definition of the distance between proteins (Pein and Zhang, 2005). In PPI Network a distance can be easily obtained from a simple matching coefficient that calculates the similarity between two elements. The similarity value can be normalized between 0 and 1, and the distance can be derived from 1− (0|1). If the similarity value of two elements is high, the spatial distance between them should be short. Graph-based clustering techniques consider the topology of the PPI network. These latter techniques are deeply studies in other research fields, such as physics and data mining, and are known as community detection methods (Girvan and Newman, 2002).

Based on the structure of the network, the density of each subgraph is maximized or the cost of cutoff minimized while separating the graph. The following sections will discuss in summary these clustering approaches. We distinguish six main categories of algorithmic approaches employed in methods for complex detection in PPI networks: Local neighborhood Density search, Cost-based Local search, Flow Simulation, Link Clustering, Population-based Stochastic search and the recent Similarity-based Approaches.

In the following, we first introduce a description of the categories listed above and also of the recent similarity-based methods. Figure 1, summarizes the main characteristics of each method and highlights which software is publicly available.

2.1 Local neighborhood Density search

Many methods, including the most popular ones, are based on local neighborhood density search. Their objective is to find dense subgraphs within the input PPI network. They aim at maximizing the density of each found subgraph. The representative methods of this approach are: MCODE (Bader and Hogue, 2003), DPClus (Altaf-Ul-Amin, M. et al., 2006), SWEMODE (Lubovac et al., 2006), DECAFF(Li et al., 2007), CFinder (Adamcsek et al., 2006), RANCoC (Pizzuti and Rombo, 2012a), MF-PINCoC(Pizzuti, and Rombo, 2008), PINCoC(Pizzuti, and Rombo, 2007), PCP (Chua et al., 2007), DME (Georgii et al., 2009) and CMC (Clustering by Maximal Cliques) (Guimei et al., 2009).

MCODE, DPClus and SWEMODE adopt a rather similar search strategy. First, they define the weight of each node, then they choose the node with highest weight as seed cluster and finally they add neighboring nodes to the current cluster if some threshold parameters are satisfied. The main difference concerns the definition of the weight. In contrast to

SWEMODE, which combines also semantic information coming from the Gene Ontology database, MCODE and DPClus use only the network topology. Thus, the former approach should give increased confidence in the predicted function, although it is worth pointing out that it does not allow the participation of a protein to more than one cluster. In contrast to CFinder, DECAFF and PCP, which use the concepts of maximal k-clique or local cliques to grow clusters, PINCoC, MF-PINCoC and RAN-CoC rely on co-clustering to find dense subgraphs. All such methods need some parameters that biases the number and kind of output clusters.

2.2 Cost-based Local search

Methods based on cost-based local search extract modules from the interaction graph by partitioning the graph into connected subgraphs using a cost function for guiding the search towards a best partition. The most popular methods in this category are: SL(Samantha and Liang, 2003), RNSC (King et al., 2004; Pizzuti and Rombo, 2014), MODULAND (Kovacs et al., 2010) and OCG (Becker et al., 2012).

SL is greedy approach, where the former optimizes the concept of P-value to build clusters recursively merging protein pairs having the smallest P-value, and the latter optimizes the community strength of a module. In contrast to RNSC, that moves the nodes among the clusters to improve its cost function, OCG merge the clusters for optimizing the modularity, and it use the strategy recently proved in Farutin et al. (2006) for overcoming the resolution limit problem. This strategy relies on the fact that methods maximizing modularity could not discover structures at small scales, hidden within large clusters. ModuLand is based on a different approach, as it uses different influence functions of nodes to find regions where nodes influence each other. These regions are then explored to obtain local maxima corresponding to communities.

2.3 Flow Simulation

Methods based on the FS approach mimic the spread of information on a network, using random walk, or biological knowledge for passing information between proteins in the network to cluster proteins. Two main methods are used: MCL(Enright et al., 2002; Van Dongen, 2008) and RRW(Macropol et al., 2009). MCL simulates the behavior of many walkers starting from the same point and moving within a graph in a random way. RRW uses random walks to compute the nearest proteins of a cluster.

2.4 Link Clustering

LC methods group the set of edges rather than the set of nodes of the input network, often exploiting suitable techniques to compute edge similarity. LC is used to discover overlapping communities in complex networks different than PPI networks. Two LC techniques are applied to PPI networks: PEREIRA (Pereira et al., 2004) and AHN (Ahn et al., 2010).

LC approaches have the main advantage that nodes are automatically allowed to be present in multiple communities, without the necessity of performing multiple clustering on the set of edges. As a negative point, if the input network is dense then LC may become compu-

tationally expensive. We also observe that the performances of these techniques may depend on the link similarity measure they adopt.

2.5 Population-based Stochastic search

PS has been used to develop algorithms for network community detection, although only the works summarized later in the text have been applied to PPI networks. Among these methods we find: IGA (Ravaee et al., 2010) and GA-PPI (Pizzuti and Rombo (2012b, 2013). The methods described earlier in the text use different representations of candidate solutions and different fitness functions. Individuals are represented by bit strings associated with the presence of edges in CGA, nodes in IGA and connections between pairs of nodes in GA-PPI. As for LC methods, CGA performances may become worse when the input networks have a large number of edges.

2.6 Similarity-based Approaches

Clustering approaches for PPI networks can use the similarity-based values to create overlapping neighborhood clusters. The recent methods of this approach are: ProRank (Zaki et al., 2012), PEWCC (Zaki et al. 2013) and ClusterONE (Nepusz et al. 2012). These recent methods are compared with SVM-Net method for predicting multi-protein complexes introduced in (Nazar 2014).

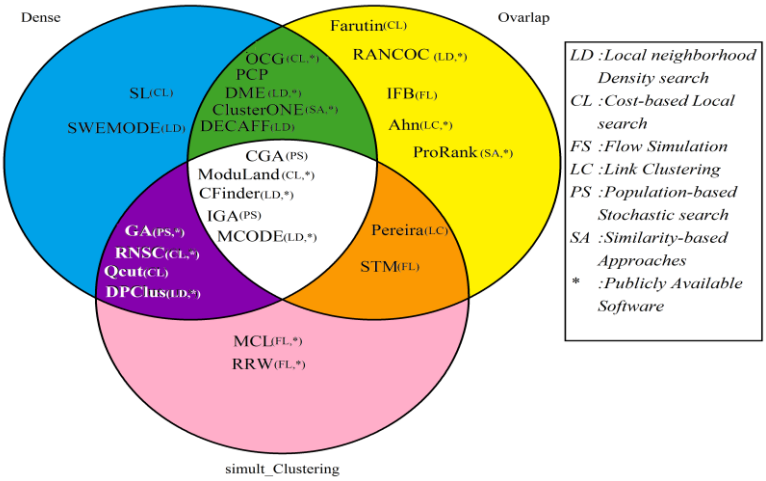


FIG. 1 – Summary of the main characteristics of PPI network Clustering methods.

3 Experimental results and evaluation

In this section we present the results of the popular graph-based approaches. These results are compared with those returned by the recent similarity-based approaches. We evaluated the effectiveness of different methods using two different yeast PPI datasets. The first dataset (Yeast PPI-Data1) is prepared by Gavin et al. (2006) and the second dataset (Yeast PPI-Data2) is a combined PPI dataset containing yeast protein interactions generated by six indi-

vidual experiments, including interactions characterized by mass spectrometry technique, and interactions produced using two-hybrid techniques.

3.1 Validation measures

To estimate the cumulative quality of the prediction, we compare the number of matching complexes with the number of known complexes using recall (TP/TP+FN), precision (TP/TP+F) and F-measure (2*(Recall* Precision)/(Recall+Precision)), where: TP: related protein classified as "related", TN: unrelated protein classified as "related", FP: related protein classified as "unrelated", FN: unrelated protein classified as "unrelated".

The values of performance measures are computed for each method employed. Figure 2, Figure 3 and Figure 4, show the Recall, Precision, and F- measure values for graph-based approaches and similarity-based approaches.

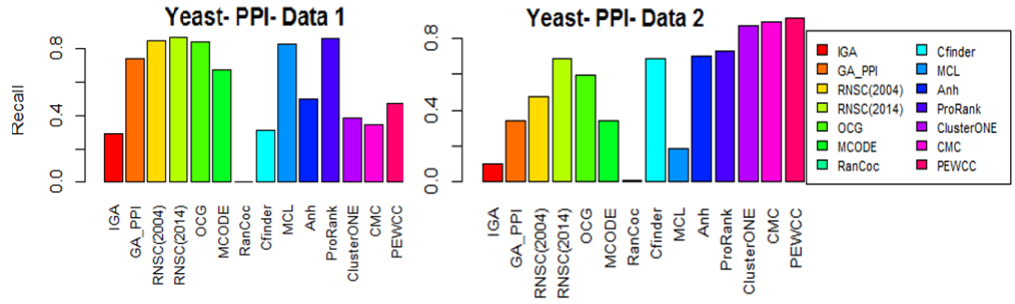


FIG. 2 – *Recall values for Yeast PPI Data 1 and Yeast PPI Data 2.*

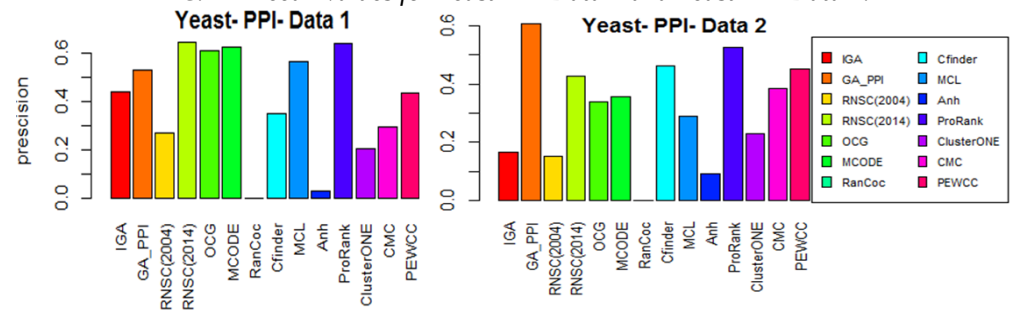


FIG. 3 – *Precision values for Yeast PPI Data 1 and Yeast PPI Data 2.*

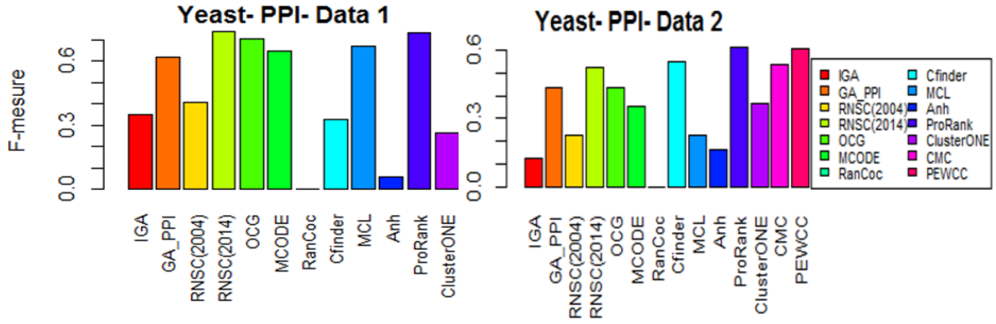


FIG. 4– *F-measure values for Yeast PPI Data 1 and Yeast PPI Data 2*

3.2 Discussions

The first observation is that the similarity-based methods and evolutionary RNSC obtain higher Recall values with respect to the other graph-based approaches for all the fitness funtions, on all the considered data sets. However, high values of precision indicate a more accurate prediction, since the predicted complexes are composed by a high percentage of proteins belonging to the true complex, thus the fraction of false positive is low. In this case, evolutionary RNSC, PRORANK, MCODE and OCG are superior to all the other approaches on Yeast PPI-Data1, while the GA-PPI approaches overcome RNSC, AHN and RANCOC. IGA-PPI also performs better than evolutionary RNSC, and obtains higher Precision values than those obtained by GA-PPI. Regarding Yeast PPI-Data2, the results obtained by RNSC, AHN, RANCOC and IGA are the worst performing, while GA-PPI obtains the best value of precision.

As regards F-measure, Evolutionary RNSC and OCG obtain the bests results on Yeast PPI-Data1 and PEWCC obtains the bests results on Yeast PPI-Data 2. However, ProRank obtain the best results on all the considered datasets.

ProRank and Evolutionary RNSC are the best performing on Yeast PPI-Data1 and also on Yeast PPI-Data2. As regards Yeast PPI-Data2, in addition to methods ProRank and Evolutionary RNSC, the best other methods are PEWCC and CFinder.

MCL predicts protein complexes with high precision and recall values on Yeast PPI-Data1, while, Recall, precision and F-measure values are very low values when using Yeast PPI-Data2.

ClusterONE also achieves high recall values than ProRank, GA-PPI and evolutionary RNSC when using Yeast PPI-Data1, while, ClusterONE has a lower performance when using Yeast PPI-Data2.

PEWCC like CMC and CFinder predicts complexes with high quality and gains a high precision, Recall and F-measure on Yeast PPI-Data2 and returns to low sensitivity, precision and f-measure values on Yeast PPI-Data2.

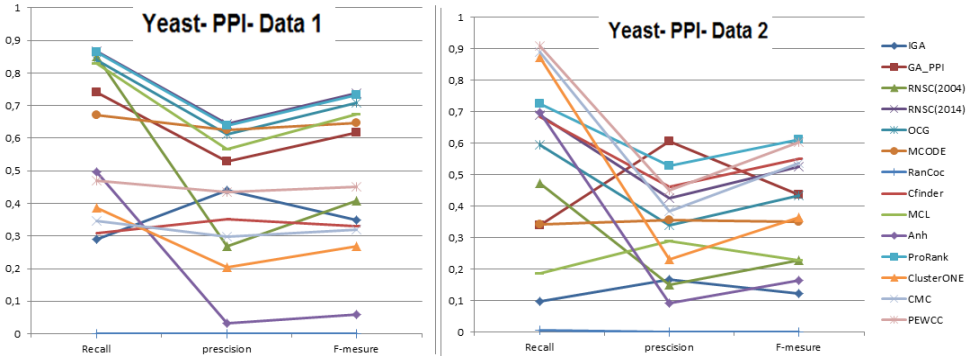


FIG. 5 – *The performance comparison for various algorithms for Yeast-PPI-Data1 and Yeast-PPI-Data 2.*

4 Conclusion

The goal of this review is firstly providing a compact overview of the main techniques presented in the literature for PPI networks clustering. Secondly, the review present an ex-

perimental comparison to show the capability between Graph-based approaches and similarity-based approaches in extracting clusters from PPI networks, according to different topology-based fitness functions, and also with respect to the main other approaches. Both aspects allow us to draw interesting considerations and conclusive remarks, as a multiple algorithms have been proposed in recent years have their own strengths and disadvantages. In this study, we first reviewed the recent and popular algorithms and then performed a comprehensive comparison among them on two different PPI data by using various evaluation criteria.

Generally, proteins may participate in multiple biological processes. Thus, methods that assign a protein to only one cluster, such as non-overlapping approaches tested here, are limited in potentiality for describing the complexity of biological systems, because the proteins may participate in multiple biological processes. This limitation is also proved by the experimental comparison provided here. Generalization to allow overlapping clusters would be desirable to enhance the prediction capability of such approaches.

The experimental study detailed in Section 3 show that the performances of PPI network clustering techniques may become different when they are applied on different interaction datasets, and only a few of recent are able to keep a good accuracy.

ProRank, PEWCC, ClusterONE and some Cost-based local search methods like Evolutionary RNSC are relatively recent and not yet enough explored, presenting interesting potentialities. Thus, additional research is necessary and desirable to improve the predictive power of these approaches.

References

Adamcsek,B. et al. (2006*). CFinder: locating cliques and overlapping modules in biological networks.* Bioinformatics, 22, 1021–1023.

Ahn,Y.-Y. et al. (2010). *Link communities reveal multiscale complexity in networks*. Nature, 466, 761–764.

Altaf-Ul-Amin, M. et al. (2006) *Development and implementation of an algorithm for detection of protein complexes in large interaction networks.* BMC Bioinformatics, 7, 207.

Bader, G. and Hogue, H. (2003). *An automated method for finding molecular complexes in large protein–protein interaction networks*. BMCBioinform.4(2).

Becker,E. et al. (2012). *Multifunctional proteins revealed by overlapping clustering in protein interaction network.* Bioinformatics, 28, 84–90.

Brohée, S. and van Helden, J. (2006). *Evaluation of clustering algorithms for protein-protein interaction networks.* BMC Bioinformatics, 7:48.

Chua,H. et al. (2007). *Using indirect protein-protein interactions for protein complex prediction.* In: Proceedings of Computational Systems Bioinformatics Conference (CSB07). pp. 97–109.

Del Sol, A., Fujihashi, H. and O'Meara, P. (2005*). Topology of small-world networks of protein-protein complex structures.* Bioinformatics, 21(8):1311-131.

Enright,A. et al. (2002). *An efficient algorithm for large-scale detection of protein families.* Nucleic Acids Res., 30, 1575–1584.

Farutin,V. et al. (2006) .*Edge-count probabilities for the identification of local protein communities and their organization*. Proteins, 62, 800–818.

Gavin,A.C. et al. (2006). *Proteome survey reveals modularity of the yeast cell machinery.* Nature, 440, 631–636.

Georgii,E. et al. (2009). *Enumeration of condition-dependent dense modules in protein interaction networks.* Bioinformatics, 25, 933–940.

Girvan, M. and Newman, M. E. J. (2002). *Community structure in social and biological networks*. In Proc. National. Academy of Science. USA 99, pages 7821–7826.

Guimei L., Wong L. and Chua H. N. (2009). *Complex discovery from weighted PPI networks*. Bioinformatics, 25(15), 1891–1897.

King,A.D. et al. (2004) *Protein complex prediction via cost-based clustering.* Bioinformatics, 20, 3013–3020.

Kovacs,A.I. et al. (2010). *Community landscapes: an integrative approach to determine overlapping network module hierarchy, identify key nodes and predict network dynamics*. PLoS One, 5, e12528.

Li S., Armstrong, C., Bertin, N. (2004). *A map of the interactome network of the metazoan.* Science, 303(5657):540-543. 7.

Li,X.-L. et al. (2007). Discoverin*g protein complexes in dense reliable neighborhoods of protein interaction networks.* In: Proccedings of Computer System Bioinformatics Conference. (CSB07). pp. 157–168.

Lin, C., Cho, Y., Hwang, W., Pei, P. and Zhang, A. (2006). *Clustering methods in proteinprotein interaction network. in Knowledge Discovery in Bioinformatics: Techniques, Methods and Application*. John Wiley & Sons,Inc.

Lubovac,Z. et al. (2006). *Combining functional and topological properties to identify core modules in protein interaction networks*. Proteins, 64, 948–959.

Macropol,K. et al. (2009). *RRW: repeated random walks on genome-scale protein networks for local cluster discovery.* BMC Bioinformatics, 10, 283.

Nepusz, T., Yu, H. and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. Nature Methods, vol. 9, pp. 471-472.

Pein, P. and Zhang, A. (2005). *A two-step approach for clustering proteins based on protein interaction profiles.* In IEEE Int. Symposium on Bioinformatics and Bioengeneering (BIBE'2005), pages 201–209.

Pereira,J.B. et al. (2004). *Detection of functional modules from protein interaction networks.* Proteins, 54, 49–5.

Pizzuti, C. and Rombo, S.E. (2012a). *A coclustering approach for mining large protein-protein interaction networks*. IEEE/ACM Trans. Comput. Biol. Bioinform., 9, 717–730.

Pizzuti,C. (2009). *Overlapped community detection in complex networks*. In: Proceedings of the 11th Annual conference on Genetic and Evolutionary computation, GECCO'09. pp. 859–866.

Pizzuti,C. and Rombo,S.E. (2007). *Pincoc: a co-clustering based approach to analyze protein-protein interaction networks*. In: Procedings of the 8th International Conference on Intelligent Data Engineering and Automated Learning. pp. 821–830.

Pizzuti,C. and Rombo,S.E. (2008). *Multi-functional protein clustering in ppi networks.* In: Proccedings of the 2nd International Conference on Bioinformatics Research and Development (BIRD). pp. 318–330.

Pizzuti,C. and Rombo,S.E. (2012b). *Experimental evaluation of topological-based fitness functions to detect complexes in ppi networks*. In: Genetic and Evolutionary Computation Conference (GECCO). pp. 193–200,

Pizzuti,C. and Rombo,S.E. (2013). *Restricted neighborhood search clustering revisited: an evolutionary computation perspective*. In: Proccedings of the 8th IAPR International Conference on Pattern Recognition in Bioinformatics (PRIB). pp. 59–68.

Pizzuti,C. and Rombo,S.E. (2014). An evolutionary restricted neighborhood search clustering approach for PPI networks. Neurocomputing145, 53–61.

Ravaee,H. et al. (2010). *Improved immune genetic algorithm for clustering proteinprotein interaction network.* In: Proccedings of the 2010 IEEE International Conference on Bioinformatics and Bioengineering. pp. 174–179.

Samantha,M. and Liang,S. (2003). *Predicting protein functions from redundancies in large-scale protein interaction networks*. Proc. Natl Acad. Sci. USA, 100, 12579–12583.

Van Dongen,S. (2008). *Graph clustering via a discrete uncoupling process.* SIAM J. Math. Anal. Appl., 30, 121–141.

Zaki, N., Efimov, D. and Berengueres, J. (2013). *Protein complex detection using interaction reliability assessment and weighted clustering coefficient.* BMC Bioinformatics, 14:163.

Zaki,N. et al. (2012). *Prorank: a method for detecting protein complexes.* In: Proccedings of the Genetic and Evolutionary Computation Conference. pp. 209–216.

Résumé

Les réseaux d'interactions protéine-protéine (PPI) jouent des rôles importants dans les cellules. Leur étude permet de comprendre le comportement des protéines à l'intérieur de la cellule. De nombreuses méthodes de classification ont été proposées pour détecter des complexes protéiques ou d'isoler les groupes de protéines en interaction qui participent aux mêmes processus biologiques ou qui effectuent un ensemble des fonctions biologiques spécifiques. Dans ce papier, nous passons tout d'abord en revue les principales méthodes de classification fondées sur les propriétés topologiques des graphes qui ont été appliquées aux réseaux PPI. Ensuite, nous présentons une comparaison expérimentale entre les approches fondées sur les propriétés topologiques des graphes et ceux basées sur la similarité entre les protéines en utilisant deux types de données. Nous allons également comparer l'efficacité et la pertinence de ces méthodes.

The Power of Contingency Tables in Prediction

Faraj A. El-Mouadib
University of Benghazi
Faculty of Information Technology
Department of Computer Science
elmouadib@yahoo.com

Abstract. Real world repository system such as, Data warehouses, Databases, etc are very useful source of different forms of knowledge. Real world data are highly susceptible to noise, missing, and inconsistent values due to their typically huge size. Quality knowledge is based on quality data. To accomplish the goal of getting quality data, preprocessing the data is a mandatory step to get the data in a position proper for mining, and, ultimately, the mining step towards achieving the expected results. Our focus in this paper is to demonstrate the power of contingency tables in prediction by the of the set of lambda (λ) measures of association introduced by Goodman and Kruskal (1954).

1. Introduction

Contingency tables have been used as tools in statistics for many statistical measures since their introduction by Karl Pearson in 1904 in his paper in title of: On the Theory of Contingency and Its Relation to Association and Normal Correlation, Pearson K. (1904). Contingency tables are used as a base for the well known chi-square (χ^2) statistic, which is test of independence and the set of lambda (λ) measure of association introduced by Goodman and Kruskal in 1954.

Nowadays, the Information Industry is being flooded with the huge amounts of data by the grace of day to day technological advancements in gathering and storage capacity. But such data is highly susceptible to noise, missing, and inconsistent values. This data is very important to the fast evolving world to discover the buried implicit knowledge from the data tombs upon which decisions are made. So the data must be preprocessed in order to acquire high quality knowledge that can be used by decision makers. The purpose of the preprocessing phase is to improve the data quality and, consequently, improve the accuracy and efficiency of the drawn knowledge.

This paper focuses on the power of contingency tables in prediction. Contingency tables are used in conjunction with the set of lambda (λ) measure of association to predict the value of one attribute with or without the knowledge of another attribute. Such an idea can be used to deal with the problem of the missing values that accompany the real world data.

2. Contingency tables

In statistics, a contingency table also referred to as cross tabulation or crosstab is a type of table in a matrix format that displays the multivariate actual frequency distribution of the variables (attributes). A contingency table is a two-way table with r rows and c columns, which summarizes the cross classification of sample elements according to two

characteristics. The rows correspond to one classification (or category) and the columns correspond to the other. Entries (cells) in the table represent the actual number of cases that have the corresponding combination of values of both characteristics. Assume that we have two variables A and B, variable A has a domain or level values of: A_1, A_2, ..., A_α and variable B has the values of: B_1, B_2, ... , B_β may be represented by a 2D-table of the following form:

		B				*Total*
		B_1	B_2	. . .	B_β	
A	A_1	n_{11}	n_{12}	. . .	$n_{1\beta}$	$n_{1.}$
	A_2	n_{21}	n_{22}	. . .	$n_{2\beta}$	$n_{2.}$
	.	.	.	. . .	.	.
	.	.	.	. . .	.	.
	A_α	$n_{\alpha 1}$	$n_{\alpha 2}$	. . .	$n_{\alpha\beta}$	$n_{\alpha.}$
Total		$n_{.1}$	$n_{.2}$	. . .	$n_{.\beta}$	*N*

TAB. 1– *Contingency table.*

Total number of cases are represented by the following formula:

$$N = \sum_{i,j} n_{ij}$$

Where n_{ij} represents the actual counts of objects classified as belong to the cell (A_i, B_j). The marginal frequencies $n_{\cdot 1}$, ... , $n_{\cdot\beta}$ and $n_{1\cdot}$, ... , $n_{\alpha\cdot}$ denote the marginal counts (the column totals and row totals respectively). The population can be represented in terms of proportions, where $p_{ij}=n_{ij}/N$.

3. Chi Square χ^2 as a Test of Independence

We use Chi Square Test as a test of independency for a contingency table that has *r* rows and c columns to determine the independency of categorical variables. This test contains null and alternatives hypothesis such as:

h_o: it is used to determine that the two given categorical variables are independent.

h_a: it is used to show the relationship between two categorical variables.

The Chi-square (χ^2) statistics is based on the difference between the actual distribution Aij and the expected distribution Eij based on the null hypothesis of independence. The Chi-square (χ^2) statistics is computed by:

$$\chi^2 = \sum_{i,j} \frac{(A_{ij}-E_{ij})}{E_{ij}}$$

Where:

A_{ij} denotes the actual count that falls in row *i* and column *j*.

E_{ij} denotes the expected count:

$$E_{ij} = \frac{n_{i.}.n_{.j}}{N} \quad \text{for } i=1,\ldots,r \text{ and } j=1,\ldots,c$$

$n_{i.}$ and $n_{.j}$ are row totals and column totals, respectively.

An excellent test of independence may be based on χ^2 which, doesn't mean at all that χ^2, or some function of it, is an appropriate *measure* of degree of association, Bartoszynski R.

and Niewiadomska-Bugaj M. (1996). The association that we are concerned with is the association between two variables (e.g. *A* and *B*) which are represented by a 2D- contingency table.

4. Measures of association

A measure of association is a quantity calculated from frequency distribution intended to describe the relationship between two cross-classified variables or attributes. Several measures of association stem from the standard Chi-square (χ^2) statistics upon which a test of independence is usually based. The ease of interpretation of measures of association varies considerably. The measures of association based on χ^2, are little more than index numbers to indicate when one table shows a greater association than another, Bishop Y., Fienberg S. and Holland P. (1975). As stated in Bartoszynski R. and Niewiadomska-Bugaj M. (1996), an excellent test of independence may be based on χ^2 statistics or, some function of it, is not an appropriate measure of the degree of association. Usually the relationship between test of independence and measure of association is somehow confusing. Test of independence determines whether a relationship exists between two cross-classification variables, while measure of association helps us understand the particular type and the extent of the relationship between the two cross-classified variables. It is obvious that measures of association that are based on χ^2 statistics don't have clear interpretation and there are some other measures of association that were designed with an ease and clear interpretation, among them the lambda (λ) measure that was proposed by Goodman and Kruskal in (1954).

According to Goodman L. and Kruskal W. (1979), measures of association may be classified into the following categories:

1. Measures based on the ordinary chi-square statistics used to test independence in a complete two-dimensional table.
2. Measures based on the cross-product ratio for 2×2 tables.
3. A "Proportional Reduction of Error (PRE)" measure that indicates the relative value of using the row (column) categories to predict the column (row) categories.
4. A "proportion of explained variance" measure which provides a close analogy for categorical data to the squared correlation coefficient for continuous data.

Definition: An Equivalence relation between two attributes is expressed by 2-D contingency table in which the population is concentrated in cells, no two of which are in the same row or column of the table.

Whether a contingency table represents equivalence or approximate equivalence relation, can carried out by the use a set lambda λ measures of association for unordered (nominal) data values. The set of lambda measures is based on the concept of PRE.

5. The set of lambda λ measures

The set of lambda (λ) measures belongs to the third type. The measure is based on the optimal prediction and is useful, in particular, in a situation of the following nature:

i. There are two cross-classified variables *A* and *B*.

ii. There are no relevant underlying continua.
iii. There is no natural ordering of interest.
iv. The *A* classification precedes the *B* classification chronologically, causally, or otherwise.

The set of measures enable us to predict optimally the category for the column variable from the category of the row variable and vice versa. We can predict the category of variable *B* from the category of variable *A* by:

1. Given no information (assuming *B* is statistically independent of *A*), or
2. given the *A* category (assuming *B* is a function of *A*).

Let $\rho_{.m}$ represents the largest marginal proportion among *B* classes (the maximum of columns totals) and let ρ_{am} represents the largest proportion in the a^{th} row of the cross-classification table (the maximum of rows totals):

$$\rho_{.m} = \max_{b} \rho_{.b}$$

$$\rho_{am} = \max_{b} \rho_{ab}.$$

In the first case it is best to predict that B_b for which $\rho_{.b} = \rho_{.m}$ that is to predict the *B* category that has the largest marginal proportion and the probability of error would be: $1 - \rho_{.m.}$

In the second case it is best to predict that B_b for which $\rho_{ab} = \rho_{am}$ that is to predict the *B* category that has the largest proportion in the observed *A* category and the probability of error would be:

$$1 - \sum_{a} \rho_{am}$$

A Proportional Reduction in Error (PRE) strategy relates these two cases by:

$$PRE = \lambda_b = \frac{\text{probability of error in case(1)} - \text{probability of error in case(2)}}{\text{probability of error in case(1)}}$$

A simple replacement of the probability of errors in the two cases by their correspondence formula, the measure λ_b is then:

$$\lambda_b = \frac{1 - \rho_{.m} - (1 - \sum_{a} \rho_{am})}{1 - \rho_{.m}}$$

$$\lambda_b = \frac{\sum_{a} \rho_{am} - \rho_{.m}}{1 - \rho_{.m}}$$

The measure λ_b represents the relative improvement in predicting the *B* category when the *A* category is known, as opposed to when the *A* category is not known. In this sense, it is an asymmetric or directional measure of predictability.

6. More on asymmetric measures

Several important properties of λ_b are given and can be easily seen:

i. λ_b is indeterminate if and only if the population lies in one column (lies in one *B* class), otherwise $0 \leq \lambda_b \leq 1$ inclusive.
ii. λ_b is 0 if and only if knowledge of the *A* classification is of no help in predicting the *B* classification.

iii. λ_b is 1 if and only if knowledge of an individual's *A* class completely specifies its *B* class, if each row of the table contains at most one non-zero probability.
iv. In the case of statistical independence λ_b, when determinate, is zero. The converse need not hold: λ_b may be zero without statistical independence holding.
v. λ_b is unchanged by permutation of rows or columns.

The λ_a measure can be defined analogously to λ_b. The measure λ_a represents the relative improvement in predicting the *A* category when the *B* category is known, as opposed to when the *B* category is not known. The λ_a measure is the then as follows:

$$\lambda_a = \frac{\sum_b \rho_{mb} - \rho_{m.}}{1 - \rho_{m.}}$$

where

$$\rho_{m.} = \underset{a}{MAX}\ \rho_{b.},$$

$$\rho_{mb} = \underset{a}{MAX}\ \rho_{ab}.$$

Property (iv), that says λ_b can be zero even in the absence of statistical independence, does not imply that it is unsatisfactory measure. The rationale underlying λ_b is a predictive interpretation of association. Moreover when other measures of association, such as those based on chi square, indicate that a particular table shows a substantial departure from independence, a λ_b value of 0 for the same table can occur, and this indicates the absence of predictive association when predicting column categories from row categories. Rather than being a drawback, this predictive interpretation is what differentiates λ_b and λ_a from other measures.

The main attraction of both λ statistics is their interpretability as PRE. An asymmetric or directional prediction may be particularly useful when causal or chronological direction between variables is suspected.

7. Symmetric measures of association

In many cases the situation is symmetrical and we want to predict the *A* class half of the time (at random) and the *B* class half of the time (at random) either:

1. Given no further information, or
2. Given the A_a class when the B_b class is to be predicted and vice versa.

Clearly the measure λ can now be defined in the same way as λ_b. In the first case, where we have no a priori information, the probability of error in prediction is 1- (1/2) ($\rho_{.m} + \rho_{m.}$) and in the second case the probability of error in prediction is $1 - \frac{1}{2}\left(\sum_a \rho_{am} + \sum_b \rho_{mb}\right)$. The relative decrease in probability of error is defined by the coefficient:

$$\lambda = \frac{\frac{1}{2}[\sum_a \rho_{am} + \sum_b \rho_{mb} - \rho_{.m} - \rho_{m.}]}{1 - \frac{1}{2}(\rho_{.m} + \rho_{m.})}$$

Both λ and λ_b measures have analogous properties due to the fact that they are based on the same principle PRE.

The lambda measure has the following properties (as given by Goodman and Kruskal):

i. λ is indeterminate if and only if when the entire population lies in a single cell of the table, otherwise $0 \leq \lambda \leq 1$ inclusive.
ii. λ is 1 if and only if all population is concentrated in cells no two of which are in the same row or column; there is only one non-zero probability cell in each row.
iii. λ is 0 in the case of statistical independence, but the converse need not hold.
iv. λ is unchanged by permutation of rows or columns.
v. λ lies between λ_a and λ_b inclusive.

8. Additional comments

When the population is given in terms of cells counts instead of proportions (ρ_{ab}'s) then the set of λ measures can be computed very easily. Let N be the size of the population (total counts), $n_{ab} = N*\rho_{ab}$, $n_{am} = N*\rho_{am}$, $n_{mb} = N*\rho_{mb}$, and so on; then:

$$\lambda = \frac{\sum_a n_{am} + \sum_b n_{mb} - n_{.m} - n_{m.}}{2N - (n_{.m} + n_{m.})}$$

$$\lambda_b = \frac{\sum_a n_{am} - n_{.m}}{N - n_{.m}}$$

$$\lambda_a = \frac{\sum_b n_{mb} - n_{m.}}{N - n_{m.}}$$

9. Preprocessing

Real world data repositories (databases and data warehouses) are highly susceptible to all kind of errors such as noise, missing, and inconsistency. Such massive amounts of data, several gigabytes or more, contains hidden knowledge. Erroneous data can cause confusion to the mining procedure, resulting in unreliable results.

The Incompleteness of data can be related to number of reasons, such as; the unavailability of values for some attributes or the values had been overlooked during the data entry or other reasons, Han J. et al, (2012). The knowledge can be discovered or minded via a variety of mining functionalities depending on a number of criteria such as: type of database, kind of knowledge sought, type of utilized techniques and type of applications. In order to acquire high quality knowledge that can to be used as the corner stone of decision making, the data must be preprocessed as to elevate its' quality and, consequently, improve the accuracy and efficiency of the subsequent mining process. Data preprocessing is an important step in the process of knowledge discovery, since quality decisions must be based on quality data, Wang L. and Fu X., (2005).

Detecting and rectifying data anomalies can lead to huge payoffs for decision-making. Here, our main concern is on missing data values.

Missing data values lead to incomplete or erroneous knowledge. According to Han J. et al, (2012), there are number of methods to rectify the data and these methods are:

1. Ignore or delete the tuple that contains missing attribute values.

2. Fill in the missing values manually.
3. Use a global constant to replace all the missing values with a designated value such as; *"Unknown"* or∞.
4. Use the mean or the median of the available values to fill in the missing values.
5. Use the attribute mean or median of all instances that belong to the same class as the tuple with missing value.
6. Use the most probable value to fill in the missing value.

According to Han J. et al, (2012), the first five methods are not robust and some of them might introduce more incorrect values. Method 6 is more popular strategy because it considers the information from other attributes' values in the prediction estimation process. Our proposed method of prediction uses contingency tables and lambda measures of association in the prediction estimation process.

10.Statistical measures computation

The measures λ_b and λ_a are asymmetric measures of the predictive power of a contingency table, while the measure of λ is a symmetric measure which is considered as a measure of equivalence (representing the accuracy of the prediction).

Examples: if we have the following contingency table, which has the value of λ= 1.00, we can predict the value of the *A* attribute given the value of the *B* attribute.

		A		
		$\mathbf{a_1}$	$\mathbf{a_2}$	$\mathbf{a_3}$
B	$\mathbf{b_1}$	37	0	0
	$\mathbf{b_2}$	0	0	26
	$\mathbf{b_3}$	0	15	0

TAB. 2– *Contingency table that represents an equivalence relation (λ=1.00).*

From the above table-2, we can draw the following prediction rules:

If *A*=a_1 then *B*=b_1 with accuracy of *100%*.
If *A*=a_2 then *B*=b_3 with accuracy of *100%*.
If *A*=a_3 then *B*=b_2 with accuracy of *100%*.

		A	
		1	**2**
B	**0**	9	0
	1	1	17

TAB. 3– *Contingency table that represents an approximate equivalence relation (λ=.90).*

From the above table-3, we can draw the following prediction rules:

If *A*=1 then *B=0* with accuracy of *90%*.
If *A*=2 then *B*=1 with accuracy of *90%*.

11.Conclusion

We have demonstrated that some special type of contingency tables, which represent equivalence relation, with the use of the set of lambda measures of association can be used in prediction. The obtained results of the two simple examples had proved the validity of this work argument that can be used in big data for filling in the missing values.

12.Further directions

The authors would like to make some suggestions for future work by making more experiments with real or synthetic data sets and to try other types of data values such as continuous, ordinal and ratio.

References

1. Bartoszynski R. and Niewiadomska-Bugaj M. 1996. Probability and Statistical Inference", Wiley series in probability and statistics, Wiley inter science, USA, pp. 755–798.
2. Bishop Y., Fienberg S. and Holland P. 1975. Discrete Multivariate Analysis Theory and Practice, The MIT Press Cambridge, USA, pp.373– 400.
3. Goodman L. and Kruskal W. 1979. Measure of association for Cross Classification, Springer Series in statistics, Springer-Verlag, 1979, New York, USA. Reprinted from the Journal of The American Statistical Association, 1954, 49, pp. 732–764.
4. Han J. et al, 2012. Data Mining: Concepts and Techniques. 3rd Edition. U.S.A: Morgan Kaufman Publishers.
5. Pearson K., 1904. Mathematical contribution to the theory of evolution, On the Theory of Contingency and Its Relation to Association and Normal Correlation, London: Dulau and Co.
6. Wang L. and Fu X., 2005. Data Mining with Computational Intelligence. Germany: Springer-Verlag Berlin Heidelberg.

Résumé

Les données du monde réel sont très sensibles au bruit, disparues, et des valeurs incompatibles en raison de leur taille généralement énorme. La connaissance de la qualité est basée sur des données de qualité. Pour atteindre l'objectif d'obtenir des données de qualité, prétraitement des données est une étape obligatoire pour obtenir les données dans une bonne position pour l'exploitation minière, et, finalement, l'étape de l'exploitation minière en vue d'atteindre les résultats escomptés. Notre objectif dans cet article est de démontrer la puissance de tableaux de contingence en prévision par le de l'ensemble de lambda (λ) des mesures d'association mis en place par Goodman et Kruskal (1954). Cette idée peut être utilisé à remplir les données manquantes comme une étape de prétraitement.

Aide à la décision de groupe dans le processus de maintenance logicielle

Mohammed Zoheir Dinedane[1], Mustapha Kamel Abdi[2]

[1,2] Département d'informatique, Faculté des sciences exactes et appliquées, Université d'Oran 1 Ahmed BenBella , 31000, Oran, Algérie

din_danos@hotmail.fr[1]
abdi.mustapha@univ-oran.dz[2]

Résumé. -Il est important de gérer les changements dans la maintenance logicielle pour répondre aux besoins changeants du client afin de toujours le satisfaire. Accepter trop de changements entraîne des retards dans l'achèvement du projet et par conséquent des frais supplémentaires. Rejeter des changements peut causer l'insatisfaction du client. Ainsi, il est important pour le gestionnaire de la maintenance logicielle de prendre des décisions efficaces en matière de gestion des changements proposés. Un type d'information qui contribue à la prise de décision est la prédiction du nombre de classes affectées suite aux modifications. Cette prédiction peut être réalisée en effectuant une analyse d'impact de changement. Nous proposons dans ce papier une approche d'aide à la décision de groupe qui permet de choisir une solution parmi plusieurs alternatives, où différents points de vue sont pris en compte. Cette approche est basée sur un protocole de négociation couplé avec la méthode multicritère Electre-III. Ce protocole dispose d'un agent coordinateur et un ensemble d'agents participants, qui essayent de trouver un compromis répondant le mieux à tous les décideurs.
Mot clé : Analyse d'impact, maintenance, aide à la décision, ELECTRE III, protocole de négociation.

1 Introduction

Le coût de la maintenance dépend du degré de dépendance entre les entités d'une architecture logicielle. Un changement peut avoir des effets considérables et inattendus sur le reste du système. Le danger encouru lors d'une modification est la propagation du changement. De ce fait, il est préférable d'avoir une idée sur l'architecture du logiciel pour estimer l'impact de changement et ainsi réduire le coût de la maintenance. Un produit logiciel est dit modulaire si ses composants présentent un faible degré de couplage. Dans le cadre des applications orientées objets (OO), il existe différents types de couplage entre classes.

Mesurer ces types de relations permet de mieux comprendre le lien qui existe entre le couplage des classes et les attributs de qualité. Nous définissons un changement dans un

programme comme une modification apportée à un de ses éléments (classe, méthode ou variable). L'analyse de l'impact est une activité dont l'objectif est de déterminer l'étendue d'une requête de changement. Elle estime les éléments affectés au niveau du code source. Plus une classe est couplée avec d'autres classes, plus elle est sensible aux changements effectués dans ces classes et plus elle est susceptible de subir des erreurs. D'autres facteurs influencent la maintenance logicielle, nous citons notamment la nature de changement proposé et les critères de l'ingénieur de la maintenance (expérience dans le domaine, familiarisation avec le langage de programmation,...etc.).

Dans la maintenance des systèmes logiciels, les gestionnaires sont confrontés au problème de décision de groupe, où plusieurs points de vue souvent conflictuels rentrent en jeux. Dans le contexte de ce papier, l'aide à la décision pour la maintenance des systèmes à objet est motivée par le manque d'outils.

1.1 Problématique et contribution

Dans la gestion des projets logiciels, plusieurs changements sont proposés pour résoudre le même problème au niveau d'un code et satisfaire le même besoin de l'utilisateur d'un système logiciel. Cependant, différents buts et objectifs doivent être pris en considération lorsqu'il s'agit d'un groupe de décideurs. A cet effet, nous avons proposé une approche d'aide à la décision de groupe qui permet d'apporter une aide dans le processus de maintenance afin de minimiser sont coût. Minimiser le cout c'est minimiser le temps entre la proposition du changement, son implantation, et sa réalisation. Cette approche est basée sur le couplage entre l'analyse multicritère (AMC) et les systèmes multi agent (SMA). On munit le module SMA par un protocole de négociation sur la base de la médiation. Ce protocole comporte un agent coordinateur (initiateur) qui est responsable de la bonne conduite de la négociation et un ensemble d'agents participants. Les agents représentent les différentes entités qui regroupent une équipe de gestion de la maintenance.

L'AMC permet de classer les changements proposés du meilleur au moins bon en respectant les différents points de vue, souvent conflictuels des différents décideurs concernés. Le choix final dans cette situation intervient après un processus de négociation.

La section 2 fait un tour d'horizon sur les travaux concernant l'analyse d'impact, et les systèmes d'aide à la décision. La troisième section est réservée à notre approche. La section 4 explique l'expérimentation ainsi la mise en œuvre de l'outil. Les perspectives de notre travail sont discutées en conclusion.

2 Travaux de recherche connexes

2.1 Couplage vs analyse d'impact

Le couplage mesure la force de l'interconnexion entre les modules d'un système. Durant le processus de maintenance de logiciel, le couplage prédit la difficulté de changement des modules du programme et les implications pour les programmes dans les autres modules. Le couplage réfère au degré d'interdépendance entre les parties d'un programme. Plus une classe est couplée à d'autres classes, plus importante est sa sensibilité aux changements dans ces classes.

Un logiciel de bonne qualité doit obéir au principe de faible couplage Chidamber et Kimerer(1994). Un couplage faible facilite la maintenance vu que les dépendances entre les classes sont minimes.

2.2 Systèmes d'aide à la décision

Beaucoup de contributions en matière d'aide à la décision ont été publiées. On peut citer de manière non exhaustive les travaux de (Juan, 2013) qui a utilisé des facteurs sociaux dans l'argumentation et la négociation pour les SAD de groupe ; Shichao et Qiping (2010) qui a montré l'effet de l'utilisation d'aide à la décision de groupe dans les études de gestion de la valeur ;Santos et Hipu (2012) présente une approche fondée sur le consensus pour résoudre efficacement le problème de prise de décision de groupe multicritère ; Weiming et Qi Hao (2012) propose une approche d'intégration de systèmes à faible couplage pour l'aide à la décision en matière de gestion et de maintenance ; et enfin Hamdadou et Thérèse (2011) propose un système d'aide multicritère à la décision de groupe pour le diagnostic industriel en utilisant les SMA et l'AMC.

3 Approche proposée

Notre approche est inspirée du travail réalisé par Hamdadou et thérèse (2011), l'idée de cette approche est l'utilisation simultanée d'un SMA et AMC, en offrant une souplesse et une facilité pour prendre la bonne décision qui satisfait le groupe. Cette approche est composée de deux modules principaux : le module extraction de métriques de couplage, et le module SMA.

L'identification et la compréhension des changements qui peuvent être apportés aux applications OO s'avèrent importantes et fructueuses. Pour cela, on s'est inspiré du travail réalisé dans Cheikhi (2004). Nous avons repris un questionnaire visant à rassembler les perceptions des gens qui assurent la maintenance des logiciels à objets. Les changements sont groupés par classe, méthode, et attribut, et le résultat de ce questionnaire est soit le changement est souvent, rare, ou jamais.

Nous allons calculer les métriques de couplage concernées de toutes les classes à partir d'un code source Java, puis nous proposons quelques changements dont l'intérêt de résoudre un problème au niveau du code. L'intersection entre les changements et les métriques calculées des classes portant le changement nous donne une matrice. Cette matrice va être utilisée comme entrée dans la méthode ELECTREIII. Le résultat final sera un rangement de ces changements du meilleur au moins bon selon les préférences des décideurs.

La matrice de performance (figure 3) dans notre cas comprend dans les lignes les changements proposés pour la résolution d'un problème (les actions), et dans les colonnes les différentes métriques des classes portant le changement, la nature de changement, plus les critères de chaque ingénieur qui a proposé le changement (les critères).

Les décideurs sont ensuite invités à saisir leurs préférences (les poids des critères et les différents seuils). Ces données vont servir aux calculs des deux phases d'ELECTRIII qui affichent à la fin un rangement de changements pour chaque décideur selon ses préférences, puis le SMA fournit le processus de négociation afin de parvenir à un accord final qui satisfait le plus grand nombre de groupe.

3.1 Module extraction de métriques

Notre objectif consiste à extraire des métriques capturant les caractéristiques importantes comme le couplage. Nous avons choisi la propriété de couplage pour deux raisons :

- Avoir une idée sur la qualité du système en termes de couplage, et cela permet d'estimer la propagation de changement dans le système.
- Exploiter ces métriques comme critères dans la matrice de performance.

Nous avons développé un outil sous Eclipse qui permet d'analyser un code source java et d'extraire toutes les données importantes pour le calcul des métriques de couplage considérées pour chaque classe du système.

Dans notre travail, nous attribuons des indices à chaque changement, pour le changement jamais (indice 0), le changement rare (indice 1), et changement souvent (indice 2).

3.2 Module SMA

Ce module vise à représenter les différents acteurs qui ont leurs propres objectifs, et leurs préférences de décision. La technologie multi-agent a déjà, fait ses preuves dans de nombreux domaines, spécialement dans les applications d'aide à la décision collective (groupe de décideurs) grâce à la facilité qu'elle fournit.

Nous déléguons au SMA la sélection du changement élu selon un processus de négociation. Pour faire face à cette décision de groupe, il est nécessaire de passer par une procédure de négociation pour parvenir à un consensus bénéfique. À cette fin, nous dotons le SMA avec un protocole de négociation sur la base de la médiation impliquant deux types d'agents:

- L'agent coordinateur (gestionnaire): est l'agent responsable de la gestion de la négociation, la modification du contrat et le choix du changement élu final.
- Les agents participants (groupe): ce sont les agents impliqués dans la décision; l'objectif de chaque agent est que son changement préféré est choisi.

La négociation se fait entre l'agent coordinateur et tous les agents participants.

Le protocole de négociation en cours est caractérisé par une série de messages échangés entre l'agent coordinateur et les agents participants. Il procède en cinq phases.

1. La phase d'initialisation : Les participants sont invités à exprimer leurs préférences concernant les différents changements.
   Chaque agent établit une classification des changements du meilleur au pire, selon un ensemble de critères en utilisant la méthode multicritère ELECTRE III.
2. La phase de proposition: pendant cette phase, l'agent coordinateur propose un marché à tous les participants sur un changement donné. Ils acceptent ou rejettent le contrat selon leur vecteur de préférences, préalablement construit dans la phase d'initialisation.
3. La phase d'évaluation: lorsque le coordinateur reçoit toutes les réponses des participants concernant la proposition du contrat, il calcul le nombre des agents participants ayant accepté sa proposition. Si ce nombre est supérieur ou égal à un seuil donné, alors la négociation est réussie. Sinon, il doit effectuer une modification de l'accord.

4. La phase modification: pendant cette phase, le coordinateur est amené à faire une modification du contrat en prenant comme point de départ les propositions des agents. Il faut établir une synthèse de ce qu'il a reçu lors de la phase d'évaluation et retourne ensuite à la phase de proposition.
5. La phase de décision: celle-ci est la dernière phase du protocole proposé. Elle signifie la fin du processus de négociation. Une décision est prise par le coordinateur selon les réponses des participants concernant la proposition qu'il a faite.

4 Expérimentation

Pour l'expérimentation de notre approche, nous avons utilisé plusieurs systèmes de tailles petites et moyennes. Mais pour valider l'expérimentation nous avons utilisé un grand système BOAP El (Hamchemi et Snoussi, 2002) : il s'agit d'une application développée avec le langage Java, fournie par le CRIM (Centre de Recherche en Informatique à Montréal). BOAP (Boite à Outils pour l'Analyse des Programmes), est un ensemble d'outils logiciels intégrés qui permet à un expert d'évaluer rapidement le niveau de qualité d'un logiciel. Nous avons choisi BOAP parce nous l'avons déjà utilisé comme système test dans des travaux antérieurs (Abdi, 2007) et (Abdi et al, 2009) pour vérifier certaines hypothèses. Et d'autre part, parce qu'il a été utilisé aussi dans Chikhi (2004) pour estimer l'impact de changement par le biais d'algorithmes d'apprentissage automatique.

4.1 Mise en œuvre de l'outil

L'option extraction permet de parcourir le code source et donner son AST(Arbre Syntaxique).

Après l'obtention de l'AST, on va procéder à la structuration en classes, méthodes, attributs, héritage, invocation de méthode,… etc.

L'option structure (figure1) permet d'afficher cette structuration du code source sous une forme exploitable afin de faciliter la manipulation.

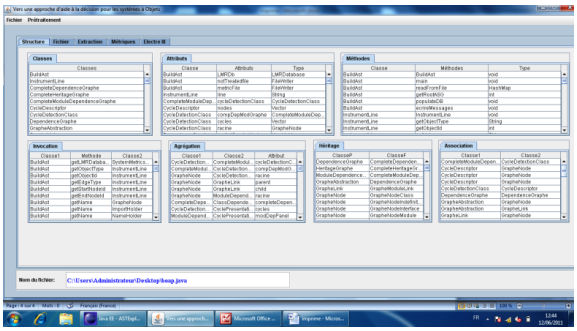


Figure1- Structure du code source

Le calcul des métriques est fait sur la base des informations extraites dans la figure 2.

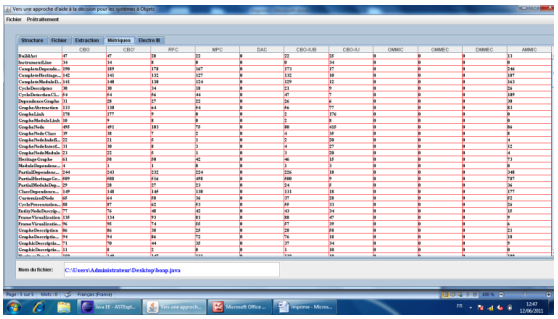


Figure2- Calcul des différentes métriques

Les changements proposés dans notre expérimentation pour un besoin de maintenance du système sont :

- Ajout d'une méthode à la classe BuildAst.
- Ajout d'une superclasse à la classe DependancePanel.
- Changement de type d'un attribut de la classe GraphModuleLink.
- Changement de la portée d'une méthode de la classe – CompleteModule
- Changement d'implémentation d'une méthode de la classe laModuleDependance.
- Suppression de la classe partielHeritageGraph.

Nous avons supposé que ces six changements résoudront le même problème. Donc quel est le meilleur en tenant en compte les différentes préférences des décideurs ?

Dans notre cas, on a pris un groupe de quatre décideurs (Analyste, Programmeur, Agent Testeur, et Ingénieur maintenance). On a attribué à chacun un poids exprimant son poids dans la négociation.

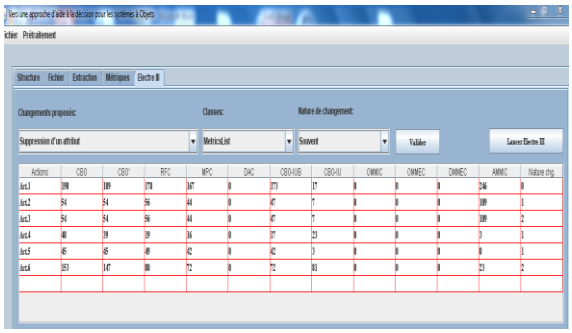


Figure3- Matrice de performance

Et après l'application d'ELECTREIII, on obtient des vecteurs de changement pour chaque décideur selon ses propres préférences (paramètres subjectives).

Avant de commencer le processus de négociation, il est nécessaire de fixer un seuil d'acceptation. Dans notre étude, il est fixé à (70%).

Après plusieurs modifications du contrat et au quatrième tour, les décideurs arrivent à un consensus, le changement sélectionné est le changement 3 avec un taux de 80% d'acceptation.

5 Conclusion

L'analyse d'impact est une technique qui permet d'estimer les changements afin de diminuer le coût croissant de la maintenance. L'utilisation des métriques de couplage comme indicateur d'impact de changement a été validé dans plusieurs travaux (Chaumun et al., 2000), Abdi (2007), Cheikhi (2004), et (Abdi et al., 2009). Le problème auquel est confronté le chef de projet de maintenance logicielle réside dans le choix de la solution optimale parmi plusieurs solutions proposées dans un groupe de décideurs. Ces derniers sont confrontés à leurs tours à des choix difficiles. Leurs décisions doivent assurer le meilleur changement selon leurs préférences et objectifs. Dans ce travail, nous avons développé un outil qui permet de calculer certaines métriques de couplage à partir d'un système en entrée et d'aider les décideurs dans leur choix du changement optimum concernant ce système. L'objectif de ce papier est de proposer un processus de décision de groupe avec l'utilisation du protocole de négociation implémenté dans un SMA. Dans nos travaux futurs, nous prévoyons faire d'autres expérimentations sur d'autres systèmes de grande tailles en considérant d'autres métriques de cohésion et d'héritage ; et perfectionner l'approche avec un protocole d'argumentation.

Références

Abdi, M. K (2007). *Analyse et Prédiction d'impact de Changement dans un système à objets*. Thèse de Doctorat d'Etat en Informatique, Université d'Oran, Es-Sénia, Avril 2007.

Abdi, M.K. H. Lounis, H. Sahraoui (2009). *Predicting Change Impact in Object-Oriented Applications with Bayesian Networks*. In Proceedings of the 33rd Annual IEEE International Computer Software and Applications Conference (COMPSAC2009), Seattle, Washington USA.

Chaumun, M. A. H. Kabaili, R. Keller, F. Lustman, et G. Denis (2000). *Design Properties and Object-Oriented Software Changeability*. In Fourth European Conference on Software Maintenance and Reengineering, Zurich, Switzerland, Pages 45-54.

Cheikhi, L (2004). *Estimation de l'impact de changement dans les programmes à objet*. Thèse de Master, Département d'Informatique et de Recherche Opérationnelle, Université de Montréal.

Chidamber, S. et C. Kemerer (1994). *A Metrics Suite for Object- Oriented Design*. In IEEE Transaction on Software Engineering", Vol. 20, No. 6, Pages 476-493.

EL Hachemi et Snoussi (2002). Alikacem EL Hachemi, Hicham Snoussi. *BOAP 1.1.0 : Manuel d'utilisation*, CRIM, Janvier 2002.

Hamdadou, D. Thérèse, L (2011). *A MultiCriteria Group Decision Support System for Industrial Diagnosis:* INFOCOMP, v. 10, no. 3, p. 12-24, September of 2011.

Juan A. Recio-García , Lara Quijano, Belen Díaz-Agudo (2013). *Including social factors In an argumentative model for group decision support system*. In decision support system No56 pages 48-55.

Maystre, J.Pictet, J.Simos. *Méthodes multicritères Electre*. Presses Polytechniques et universitaires Romandes, Lausanne, Suisse, 1994.

Santoso Wibowo, Hepu Deng (2012). *Consensus-based decision support for multicriteri a Group decision making*. In Computers & Industrial Engineering 66 (2013) 625-633.

Shichao Fan, Qiping Shen (2010). *The effect of using group decision support systems in Value management studies*. In International Journal of Project Management 29 (2011) 13-25.

Weiming Shen, Qi Hao, Yunjiao Xue (2012). *A loosely coupled system integration approach For decision support in facility management and maintenance*. In Automation in Construction 25 41–48.

Summary

It is important to manage the changes in the software to meet the evolving needs of the customer and hence, satisfy them. Accepting too many changes causes delay in the completion and it incurs additional cost. Rejecting the changes may cause dissatisfaction to the customers. Thus, it is important for the software project manager to make effective decisions when managing the changes during software development. One type of information that helps to make the decision is the prediction of the number of classes affected by the changes. This prediction can be done by performing change impact analysis.

We propose in this paper, a group decision support system that allows to choose a solution among several proposed taking into account the different preferences of decision makers. We use the benefits of Multi agent system to represent the diversity of decision makers.

This approach is based on a negotiation protocol coupled with the multi-criteria method Electra-III. This protocol has a coordinator agent and a set of participating agents, who try to find a compromise that satisfies all makers.

Keywords: Impact analysis, Maintenance, Decision support, ELECTREIII, Negotiation protocol.

Index des auteurs