7ème édition

Conférence Maghrébine sur

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

ASD 2013

ASD 2013

Actes de la 7ème édition

Conférence Maghrébine sur

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

Edités par

Azedine Boulmakoul, Omar Boussaid

25-27 mai 2013

Marrakech, 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 grandissant et dans le prolongement des éditions précédentes (Agadir–Maroc 2006, Sousse-Tunisie 2007, Mohammedia-Maroc 2008, Jijel-Algérie 2009, Sfax-Tunisie 2010 et Blida-Algérie 2012), ASD fait peau neuve et se mue en Conférence Maghrébine sur les Avancées des Systèmes Décisionnels. Comme lors de la première édition d'ASD, c'est le Maroc qui nous fait l'honneur d'accueillir ASD 2013 sous la nouvelle version de Conférence Maghrébine.

ASD 2013 ambitionne de consolider les expériences conduites par les chercheurs, industriels et utilisateurs issus de communautés travaillant sur les systèmes décisionnels. L'objectif de cette septième 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 2013 a reçu 75 soumissions d'articles en provenance de nombreux pays (Algérie, Canada, France, Maroc, Suisse, Tunisie). Après évaluation par les membres du comité scientifique, composé par 61 experts internationaux du domaine, 35 articles longs et 7 articles courts ont été retenus. Ces papiers couvrent différents thèmes de recherche et d'application sur les systèmes décisionnels.

ASD 2013 a reçu le soutien de différentes institutions publiques d'enseignement et de recherche que nous tenons à remercier : l'université HASSAN II Mohammedia Casablanca, la Faculté des Sciences et Techniques de Mohammedia, l'Ecole Marocaine des Sciences de l'Ingénieur, Laboratoire ERIC, université Lyon2, l'université de Sfax, l'association IOSIS, l'association Neoterizea 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 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
A. BOULMALKOUL, O. BOUSSAID

Présidents du comité d'organisation de la conférence

- Azedine BOULMAKOUL, FSTM, Maroc
- Kamal DAISSAOUI, EMSI, Maroc

Comité de pilotage

- BEN ABDALLAH Hanêne, MIRACL, Université de Sfax, Tunisie
- 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 Kamel, Université Oran, Algérie
- ADAM Frederic, BIS University College Cork, Ireland
- AHMED NACER Mohamed, USTHB Alger, Algérie
- AHMED OUAMER Rachid, LARI, Université Tizi Ouzou, Algérie
- ALIMAZIGHI Zahia, USTHB Alger, Algérie
- ASFARI Ounas, Université Lyon2, France
- ATMANI Baghdad, Université d'Oran, Algérie
- BADACHE Nadjib, CERIST Alger, Algérie
- BADARD Thierry, CRG Université Laval, Canada
- BADIR Hassan, ENSAT, Maroc
- BADRI Abdelmajid, FST Université Hassan II, Maroc
- BELLAFKIH Mostafa, INPT Rabat, Maroc
- BELLATRECHE Ladjel, ENSMA Poitiers, France
- BEN BLIDIA Nadjia, LRDSI, Blida Algérie
- BEN-ABDALLAH Hanen, Université de Sfax, Tunisie
- BENHARKAT Nabila, LIRIS Lyon, France
- BENSLIMANE Djamel, LIRIS Lyon, France
- BENTAYEB Fadila, Université Lumière Lyon 2, France
- BIMONTE Sandro, Cemagref, Clermond-Ferrand, France
- BOUAZIZ Rafik, MIRACL, Université de Sfax, Tunisie
- BOUCELMA Omar, Université d'Aix-Marseille, France
- BOUFAIDA Mahmoud, LIRE Constantine, Algérie
- BOUFAIDA Zizette, LIRE Constantine, Algérie

- BOUFARES Faouzi, LIPN Paris France
- BOUKHALFA Kamel, LSI, USTHB
- BOULMAKOUL azedine, Université Hassan II, Maroc
- BOURAMAOUL, Abdelkrim, Misc, Constantine, Algérie
- BOUSSAID Omar, Université Lumière Lyon 2, France
- DARMONT Jérôme, ERIC Lyon, France
- EL HEBIL Farid, INPT Rabat Maroc
- FAVRE Cécile, Université Lyon 2, France
- FEKI Jamel, Université de Sfax, Tunisie
- 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
- KABACHI Nadia, Université Lyon1, France
- KAZAR Okba, Université Biskra, Algérie
- KHOLLADI Med-khireddine, LIRE Constantine, Algérie
- KHROUF Kais, MIRACL, Université de Sfax, Tunisie
- LEMIRE Daniel, UQ Montréal, Canada
- LOUDCHER Sabine, ERIC, Lyon, France
- MAHDAOUI Latifa, LSI, USTHB
- MALKI Mimoune, USB Sidi Bel Abbes, Algérie
- MARGHOUBI Rabia, Université Hassan II, Maroc
- MELIT Ali, LAMEL Jijel, Algérie
- MEZIANE Abdelkrim, CERIST, Algérie
- MISSAOUI Rokia, LARIM U.Q. Outaouais, Canada
- MOUSSAOUI Abdelouaheb, Université de Sétif, Algérie
- NABLI Ahlem, MIRACL Sfax, Tunisie
- OUKID Saliha, LRDSI Blida, Algérie
- OULAD HAJ THAMI Rachid, ENSIAS Rabat, Maroc
- RAVAT Frank, IRIT, Toulouse, France
- REGUIEG F. Zohra, LRDSI Blida, Algérie
- SEKHRI Larbi, LIIR, Univ. Oran
- SIDHOM Sahbi, LOREA, Nancy, France
- TESTE Olivier, IRIT, Toulouse, France
- TOURNIER Ronan, IRIT, Toulouse, France
- ZAROUR Nasreddine, LIRE Constantine, Algérie
- ZEGOUR Djamel Eddine, ESI d'Alger, Algérie

Comité d'organisation :

- ALAMI Karim, EMSI, Maroc
- AZIZ Mabrouk, Université Abdelmalek Essaadi MAROC
- BESRI Zineb, FST Mohammedia, Maroc

- BOULMAKOUL Azedine , Université Hassan II, Maroc
- El BOUZIRI Adil , Université Hassan II, Maroc
- ELFILALI Sanaa , FSBM, Maroc
- IDRI Abdelfatah , FST Mohammedia, Maroc
- KARIM Lamia, Université Hassan II, Maroc
- MANDAR Meriem ,Université Hassan II, Maroc
- MARGHOUBI Rabia, Université Hassan II, Maroc
- MIKOU Hamza, Association Neoterizea
- MOALLA Mohamed Sahbi, ISET Sfax - Tunisie
- NAIM Abdelah, Association Neoterizea

Sommaire

Chapitre I : Data warehouse et entrepôt de données

Chapitre II : Aide à la décision spatiale

Modélisation de processus ETL dans un modèle MapReduce

Mahfoud Bala*, Zaia Alimazighi**
*LRDSI, Université Saad Dahleb, Blida, Algérie
mahfoud.bala@gmail.com,
**LSI, Université des Sciences et de la technologie Houari Boumediene, Alger, Algérie
zlimazighi@usthb.dz,

Résumé. Les processus ETL sont pris en charge par des moulinettes logicielles classées en trois catégories (1) L'extraction des données à partir des sources, (2) la transformation permettant de livrer des données de qualité ayant une valeur pour l'analyse (3) le chargement des données préparées dans l'entrepôt. En fait, les données sont considérées comme de la matière première qu'on traite à l'aide de systèmes d'information décisionnels pour produire des informations utiles pour l'aide à la décision. Ces donnés sont appelées à se métamorphoser et connaissent de nouvelles structures et des formats variés. De plus, de nouveaux environnements et paradigmes se développent. Le processus ETL n'est pas à l'abri de ces évolutions, vu qu'il est chargé de capturer toute donnée quelque soit sa nature, son format, du moment qu'elles soient pertinentes et peuvent ramener de la valeur au processus d'analyse. Les processus ETL deviennent de plus en plus complexes face à cette variété de formats de données et particulièrement aux données massives (*Big data*). Dans ce papier, nous proposons une approche de modélisation de processus ETL traitant des données de dimension big data selon le paradigme MapReduce.

1 Introduction

Grâce à leur vulgarisation et popularité, Internet, le web et Smartphones ont connu ces cinq dernières années une utilisation sans précédent. Entreprises et particuliers utilisent des applications web qui créent du contenu sur Internet à grande échelle où les données prennent des dimensions extraordinaires (mesurées en Péta-Bytes, voire en Zetta-Bytes). Les applications e-commerce, les réseaux sociaux, les sms échangés sur téléphones portables produisent des quantités gigantesques de données. L'analyse de ces données produit des informations synthétisées et plus intéressantes pour la prédiction et l'aide à la décision. Nous nous intéressons, particulièrement, à la capture de ces données et leur préparation dans des processus appelés ETL (Extracting, Transforming, Loading).

Un grand intérêt a été réservé au domaine de l'ETL par la communauté. Pour l'aspect modélisation, nous citons les propositions de Stöhr et al. (1999) ; Vassiliadis et al. (DOLAP 2002) ; Trujillo et Luján-Mora (ER 2003) ; Luján-Mora et al. (2004) ; Davidson et Kosky (1999) ; Vassiliadis et al. (2001) ; Vassiliadis et al. (CAiSE 2003, IS 2005, DaWaK 2005, ER 2005). En ce qui concerne la sémantique de l'ETL, nous citerons les travaux de Simitsis (2005) ; Simitsis & Vassiliadis (DSS 2008) ; Skoutas et Simitsis (DOLAP 2006, IJSWIS 2007) ; Skoutas et Simitsis (NLDB 2007). Darmont et al. (2005) et Vassiliadis et al.(2007)

ont travaillé sur l'aspect Benchmarking. Un panorama complet sur les différentes contributions dans le domaine de l'ETL a été présenté par Vassiliadis (2009). Récemment, l'avènement du *Big data* et du *Cloud computing* a attiré l'intérêt de la communauté. Les travaux ont porté essentiellement sur des aspects très techniques liés à l'amélioration des performances : Liu et *al.* (2011), implémentation de techniques OLAP sous ces nouveaux environnements : Laurent d'Orazio et Sandro Bimonte (2010), etc.

Etants convaincus de l'importance de la modélisation qui permet de dessiner un processus ETL avant même d'aborder son implémentation, nous avons proposé une approche de modélisation pour le *Big data* à l'aide du paradigme *MapReduce*. La section 2 présentera l'approche de Vassiliadis pour la modélisation ETL. Nous présentons, dans la section 3, notre contribution qui est l'approche de modélisation ETL pour le Big data que nous illustrons par un exemple, celui d'un établissement universitaire. Notre prototype est présenté dans la section 4 et nous terminerons par la section 5 réservée pour une conclusion et des perspectives.

2 Approche de Vassiliadis

Le formalisme de Vassiliadis se distingue principalement des autres propositions par l'importance donnée à l'attribut (élément atomique d'une donnée) dans le processus ETL. Aussi, l'approche de Vassiliadis ne se contente pas à des schémas classiques consistant à ne modéliser que les données manipulées. L'aspect dynamique est aussi modélisé : les tâches d'extraction, de transformations, de chargement, de mappage ainsi que de synchronisation (flux de travail). La sémantique des tâches est largement décrite pour comprendre la logique du processus ETL avec tous ses détails. Des extensions ont été faites et ont été implémentées dans un prototype qui porte le nom d'ETL-XDesign, M.Bala et Z.Alimazighi (2012). La figure 3 présente un exemple de processus ETL relatif à la gestion d'un établissement universitaire.

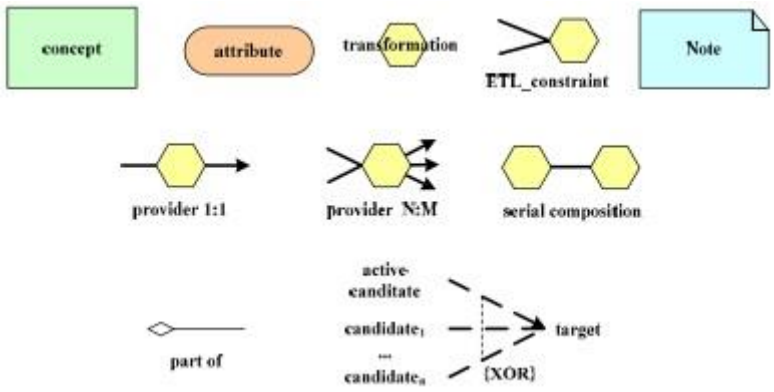


FIG. 2 – *Notations utilisées pour la modélisation conceptuelle d'un processus ETL*

L'exemple de la figure 3 montre comment à partir d'un fichier source (*scolarité en cours*), transféré vers le DSA (*S1.Etudiant*), un ensemble de transformations sont exécutées

pour préparer les données et calculer les effectifs par année, spécialité, cycle avant leur chargement dans l'entrepôt de données (*DW.Etudiant*). quatre transformations sont prévues dans cet exemple : NOT NULL (*NN*) qui contrôle l'existence de valeurs NULL, conversion d'une date Américaine en format Européen (*A2EDate*), fonction (*f*) d'extraction de l'année à partir d'une date (*year()*) et l'agrégation (γ) qui consiste à calculer les effectifs des étudiants par cycle, spécialité et année. L'exemple montre aussi que certains attributs (*bourse* et *sport*) sources ne sont pas utilisés dans le processus ETL.

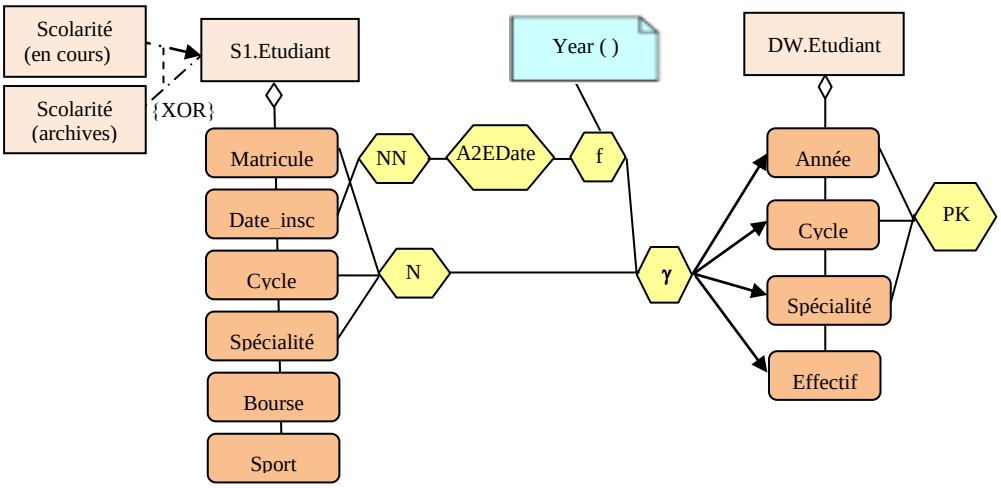


FIG. 3 – *Exemple d'un processus ETL relatif à un établissement universitaire*

3 Approche de modélisation d'un processus ETL dans un modèle MapReduce

Face au phénomène des données massives, plusieurs aspects dans les systèmes décisionnels méritent d'être réétudiés. Le processus ETL, en particulier, est affecté puisque toutes les données y transitent avant d'atteindre l'entrepôt de données. Des données avec une dimension Péta-Bytes et Zetta-Bytes affaiblissent les performances de l'ETL. Pour surmonter cette difficulté, nous adoptons le paradigme MapReduce, un standard à grande échelle pour le traitement parallèle sur des clusters d'ordinateurs destiné pour le traitement de données à forte intensité connues aujourd'hui sous le nom de Big Data. L'approche proposée est décrite dans la *section 3.3*.

3.1 Big data

Le Big data s'attaque au traitement spécifique de données massives. En fait, la volumétrie n'est pas la seule caractéristique, ce type de données est connu aussi pour sa variété en termes de formats et de nouvelles structures ainsi qu'une exigence en termes de rapidité dans le traitement. Selon IBM, chaque jour, nous générons 2,5 trillions d'octets de données. A tel

point que 90% des données dans le monde ont été créées au cours des deux dernières années seulement. Ces données proviennent de partout : de capteurs utilisés pour collecter les informations climatiques, de messages sur les sites de médias sociaux, d'images numériques et de vidéos publiées en ligne, d'enregistrements transactionnels d'achats en ligne et de signaux GPS de téléphones mobiles, pour ne citer que quelques sources. Ces données sont appelées Big Data ou volumes massifs de données.

3.2 Paradigme MapReduce (MR)

MapReduce est un modèle de programmation ayant connu plusieurs implémentations sous forme de Frameworks destiné pour le traitement de données massives. Dans ce modèle, le programmeur spécifie le traitement en deux étapes en utilisant une fonction Map() et une fonction Reduce(). Le système MapReduce, fonctionnant sur une plateforme de type cluster, parallélise alors automatiquement le traitement en découpant le processus en sous processus où chacun sera confié à un nœud (fonction Map exécutée sur une machine du cluster) dont les résultats partiels seront soumis aux reducers (fonctions Reduce exécutées sur une machine du cluster) afin de restituer le résultat final.

Google étant le pionnier dans ce domaine, Jeffrey Dean et Sanjay Ghemawat (2004) ; il a introduit le paradigme en 2004 avec leur framework MapReduce utilisé dans plusieurs de ses applications. D'autres Frameworks ont été développés, on citera Apache Hadoop et Disco. Hadoop est devenu une référence dans les plateformes MapReduce open source.

3.3 Approche de modélisation dans un environnement MapReduce

Il s'agit de reprendre l'approche de Vassiliadis très connue dans ce domaine et l'étendre pour prendre en charge des aspects spécifiques à un environnement MapReduce (MR). Le formalisme proposé par Vassiliadis ne modélise pas la distribution des données et le découpage du processus pour une exécution parallèle. Nous proposons de nouvelles notations (graphiques) pour modéliser le partitionnement des données et les phases Map et Reduce. Dans cette contribution, nous avons veillé à ne pas perturber le formalisme initial et que les nouvelles notations proposées s'intègrent en symbiose avec le formalisme de Vassiliadis. Dans l'approche proposée, les transformations permettant la normalisation des données (nettoyage, filtrage, conversions, projection, ...) sont prises en charge dans la phase Map. Ce type de transformations s'opère sur les tuples un à un indépendamment des autres. Par contre, les transformations de type fusion et agrégation s'opèrent sur une population de tuples, c'est la vocation de la phase Reduce.

Afin de préserver la trajectoire des données le long du processus, le mappage des données est représenté entre les données sources, la tâche de partitionnement des données et la phase Map, entre la phase Map et la phase Reduce et bien sûr entre la phase Reduce et le Datawarehouse. Le partitionnement est représenté afin de montrer comment les données sources sont partitionnées pour soumettre à chaque Mapper sa partition sur laquelle il opère les transformations nécessaires. Entre la phase Map et la phase reduce, le partitionnement montre comment les résultats partiels des mappers sont soumis aux reducers. La sémantique des Partitionneurs, des Mappers et des Reducers est explicitée par des Notes.

Pour de représenter un processus ETL avec le formalisme de Vassiliadis dans un modèle MapReduce, nous intégrons trois nouvelles notations comme décrites dans la figure 4.

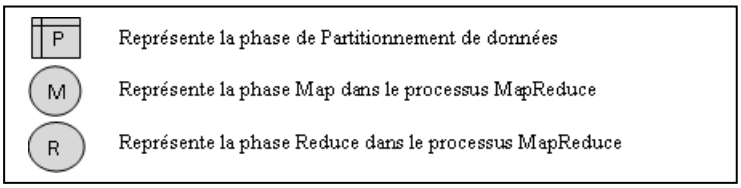


FIG. 4 – *Notations proposées pour la modélisation du processus MapReduce*

Afin de comprendre le principe de l'approche proposée, l'exemple sur l'établissement universitaire est représenté dans le modèle MapReduce (figure 5).

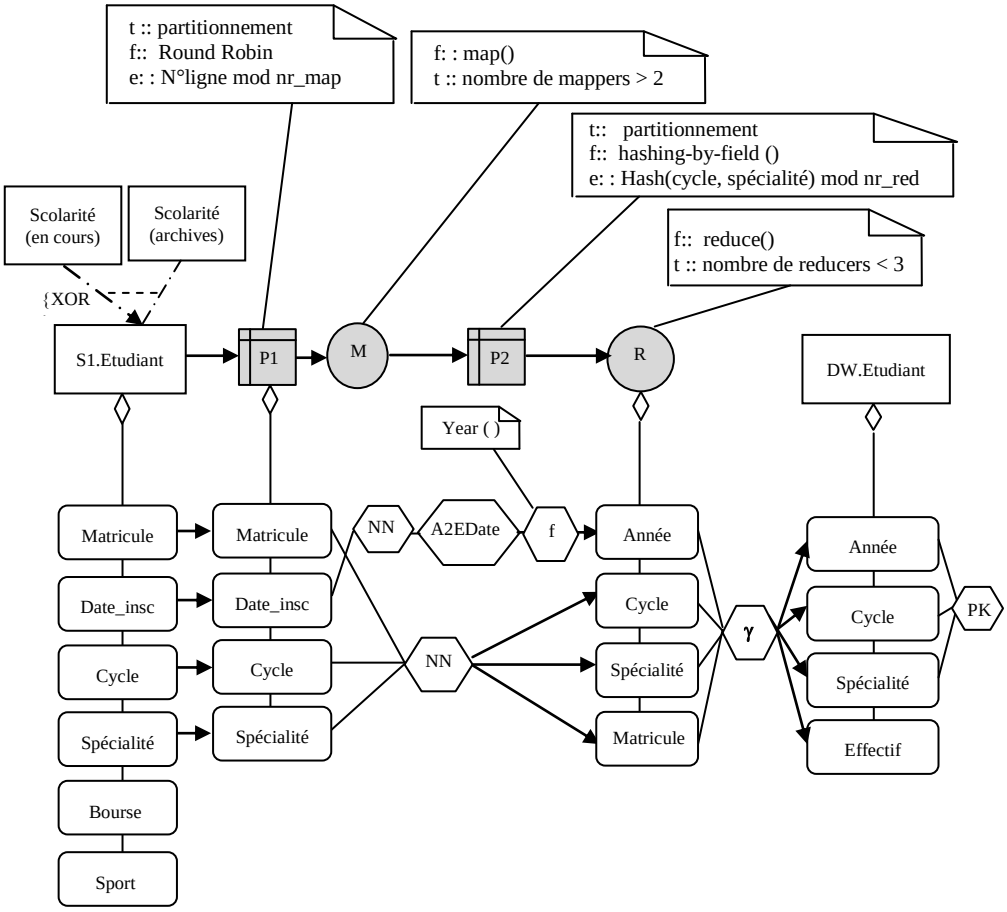


FIG. 5 – *Exemple de l'ETL relatif à un établissement universitaire dans un modèle MapReduce*

4 Implémentation

Nous avons mis en œuvre un prototype qui permet, grâce à une interface très conviviale (figure 6), d'exprimer tous les détails d'un ETL. Dès que le paramétrage de l'ETL terminé, un bouton « proceed » permet de lancer, en interactif, l'exécution du processus ETL. Le paramétrage peut être sauvegardé dans un fichier de configuration pour une exécution en batch. Le prototype a été développé sous l'environnement suivant :

- OS : Linux Ubuntu 11.10 édition desktop
- Framework MapReduce : Disco
- Langage : Python 2.6
- Sources supportées : Base PostgreSQL, documents XML, fichiers csv

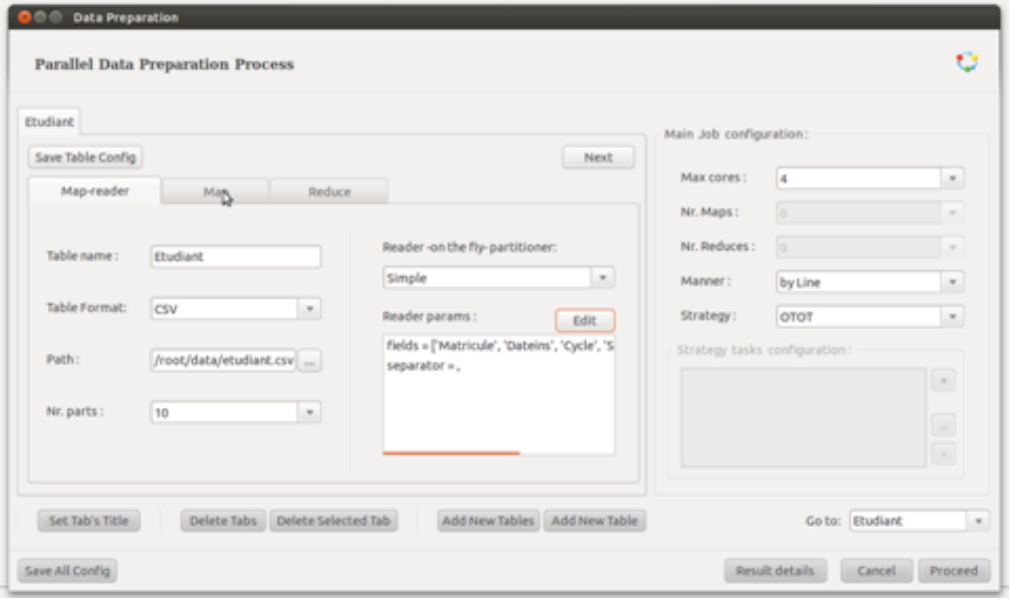


FIG. 6 – *Onglet Paramétrage des sources de données et partitionnement à la volée*

4.1 Partitionnement

Nous avons mis en œuvre deux catégories de partitionneurs : partitionneurs physiques et partitionneurs à la volée.

4.1.1 Partitionneurs physiques

Il s'agit de générer à partir d'une source de données (table relationnelle, document XML ou fichier plat) un ensemble de partitions physiques (fichiers) dont la fusion donnera la source initiale. Cette tâche ne peut se faire en parallèle.

a. ***Partitionneur simple*** : générer un nombre de partitions physiques (fichiers) dont la taille est approximativement égale à **Taille (fichier_source) / nb_part.**

Algorithm 1 Partitionneur simple (nb_part)

```
Début
Taille_partition = taille(fichier sources)/nb_part;
i=0 ;
T : enregistrement ;
Tant que non fin (fichier source)
Faire
        Tant que taille(Pi) <taille_partition
        Faire
                Lire(fichier source, T) ;
                Ecrire (Pi, T) ;
        Fait ;
        i=i+1 ;
Fait ;
Fin.
```

b. ***Partitionneur Hashing by fields***: un ou plusieurs attributs du fichier source sont retenus pour le partitionnement en utilisant la fonction Hash. Soit nb_part le nombre de partitions à générer. Un tuple du fichier source sera inséré dans la partition Hash(att1, att2, ..) mod nb_part.

Algorithm 2 Partitionneur Hashing by fields (Att1, Att2, …, nb_part)

```
Début
Entier i=0;
T : enregistrement ;
Tant que non fin (fichier source)
Faire
        Lire (fichier source, T) ;
        i=Hash (T.att1, T.att2, …) mod nb_part ;
        Ecrire (Pi, T) ;
Fait ;
Fin.
```

c. ***Partitionneur Round Robin simple*** : le tuple (ligne) est inséré dans une partition selon son rang dans le fichier source. N°partition = N°ligne mod nb_part

Algorithm 3 Round Robin simple (nb_part)

```
Début
Entier i=1 ;
T : enregistrement ;
Tant que non fin (fichier source) ;
Faire
        Lire (fichier source, T) ;
        i= i mod nb_part ;
        Ecrire (Pi, T) ;
        I=i+1 ;
 Fait ;
Fin.
```

d. ***Partitionnement Round Robin par groupe*** : Même principe que RR simple, seulement il s'agit d'affecter une partition à un groupe de tuples et non pas pour un tuple. En plus du paramètre nb_part, cet algorithme aura comme deuxième paramètre le nombre de tuples par groupe (nb_tuples_groupe).

e. ***Partitionneur Round Robin Multiple*** : combinaison des deux techniques précédentes en 2 phases. La phase 1 consiste à constituer des partitions provisoires avec RR simple et dans la phase 2, il s'agit d'affecter pour chacune des partitions provisoires la partition finale en utilisant RR par groupe.

f. ***Partitionneur Hybride*** : combinaison du Hashing by fields avec RR. La phase 1 consiste à constituer les partitions provisoires par Hashage et la phase 2 consiste à affecter pour chacune de celles-ci la partition finale en utilisant RR par groupe.

4.1.2 Partitionneurs à la volée

Le partitionnement à la volée (ou lecture sélective) est une technique de partitionnement appliquée au moment de la lecture des tuples par les MapReaders. MapReader() est une interface permettant de lire des tuples à partir de la partition (logique) appropriée pour les soumettre au mapper correspondant. Le partitionnement à la volée peut se faire en parallèle puisque plusieurs Mapreaders accèdent en même temps à la source mais chacun ne prendra que les tuples appropriés selon la technique de partitionnement. Nous avons mis en œuvre six (06) mapreaders() correspondant aux techniques de partitionnement présentées précédemment.

4.2 Processus MapReduce

Un ensemble d'UDF ont été implémentées pour traiter différentes catégories de transformations : conversion de type de données, fonctions de type date, fonctions de type chaines, traitement des valeurs creuses, … Le paramétrage de la phase Map (figure 7) consiste à énumérer les transformations à opérer sur les tuples. Ces transformations sont exécutées, en parallèle, par chaque mapper sur sa partition.

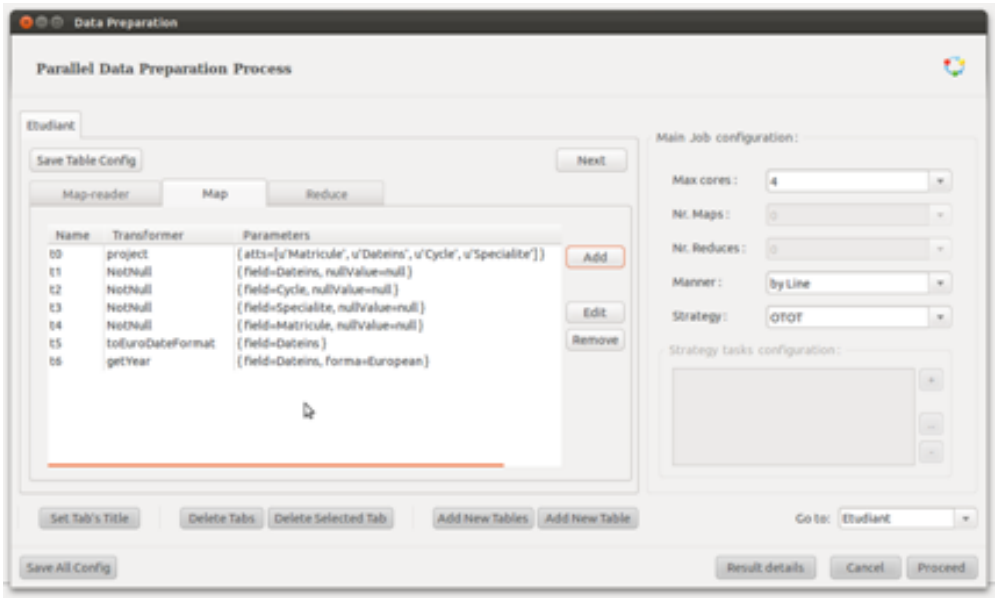


FIG. 7 – *Onglet Paramétrage de la phase Map*

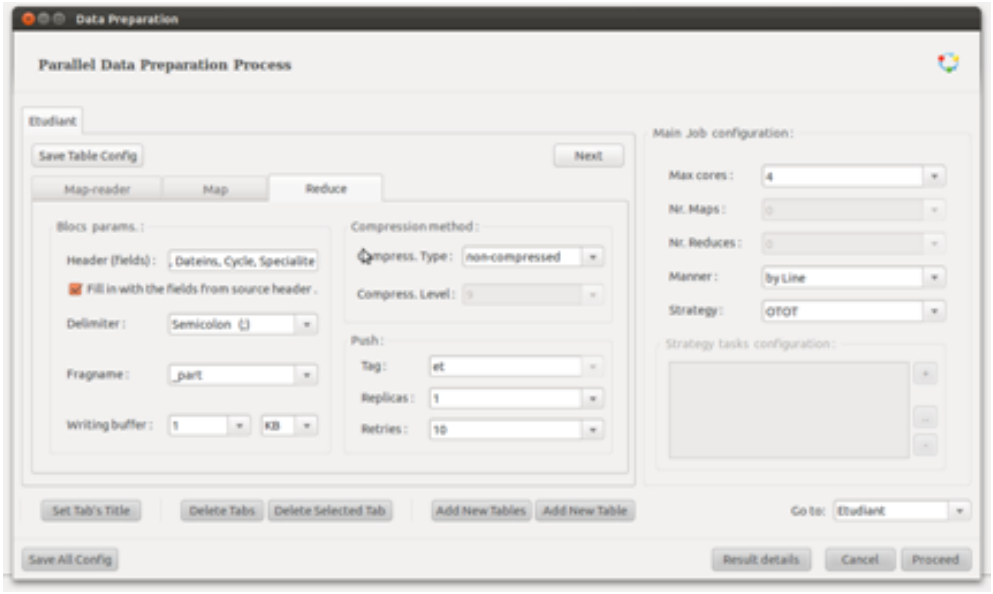


FIG. 8 – Onglet *Paramétrage de la phase Reduce*

Pour la phase reduce (figure 8), nous n'avons pas encore implémenté les UDF d'agrégation et de fusion. L'onglet Reduce permet de choisir :

- Stratégies de chargement : OTOT (One Table One Task), OTAT (One Table All Tasks), OTnT (One Table n Tasks) et nTOT (n Tables One Tasks).

- Stratégies de lecture/écriture : par ligne ou par bloc (buffering)
- Paramètres de génération des partitions physiques (résultats de la phase Map)

5 Conclusion et perspectives

La modélisation est un moyen qui facilite la compréhension des détails de fonctionnement de l'ETL et permet alors de maîtriser sa complexité et anticiper sur les éventuels problèmes et risques avant l'implémentation ou le paramétrage de l'outil ETL. L'approche de Vassiliadis étant une des plus intéressantes dans ce domaine.

Face à l'avènement du Big data, les systèmes décisionnels connus comme de grands consommateurs de données doivent être réétudiés pour évaluer l'impact du big data sur les aspects modélisation, performances, etc. Dans ce contexte, nous avons proposé une approche de modélisation ETL basée sur le modèle MapReduce. Cette approche fait apparaître trois nouveaux concepts dans le modèle ETL à savoir le partitionnement des données sources, transformations dans la phase Map et enfin la fusion et agrégation des données dans la phase Reduce.

Le papier a abordé la question du Big data dans la modélisation ETL mais celle-ci mérite d'être expérimentée sur des données massives pour évaluer l'impact du Big data ainsi que l'efficacité du modèle MapReduce dans l'amélioration des performances. Pour l'amélioration des performances de l'ETL face au big data, il ne suffit pas de rester sur le traitement parallèle sur un cluster. Il est intéressant aussi d'évaluer ce que peuvent apporter les structures NoSQL à l'ETL.

Références

Bala M., Alimazighi Z. (2012), ETL-XDesign : Outil d'aide à la modélisation de processus ETL. 6ème édition des Avancées sur les Systèmes Décisionnels, pp 155-166, Blida, Algérie, 01-03 Avril 2012

Darmont, J., Bentayeb, F., & Boussaïd, O. (2005). DWEB: A Data Warehouse Engineering Benchmark. Proceedings 7th International Conference Data Warehousing and Knowledge Discovery (DaWaK 2005), pp. 85–94, Copenhagen, Denmark, August 22-26 2005

Davidson, S., and Kosky, A. (1999). Specifying Database Transformations in WOL. Bulletin of the Technical Committee on Data Engineering, 22, 1, 25-30.

d'Orazio L., Bimonte S. (2010) : Multidimensional Arrays for Warehousing Data on Clouds. Globe 2010: 26-37

Kimball, R., and Caserta, J. (2004). The Data Warehouse ETL Toolkit. Wiley Publishing, 2004

Jeffrey Dean, Sanjay Ghemawat (2004): MapReduce: Simplified Data Processing on Large Clusters. OSDI 2004: 137-150

Liu X., Thomsen C., and Pedersen T.B. (2011), ETLMR: A Highly Scalable Dimensional ETL Framework Based on MapReduce. DaWaK 2011, LNCS 6862, pp. 96–111, 2011. © Springer-Verlag Berlin Heidelberg 2011.

Lujàn-Mora S (2005), PhD thesis Department of Software and computing systems, University of Alicante, 2005

Luján-Mora, S., Vassiliadis, P., & Trujillo, J. (2004). Data Mapping Diagrams for Data Warehouse Design with UML. In Proc. 23rd International Conference on Conceptual Modeling (ER 2004), pp. 191-204, Shanghai, China, 8-12 November 2004.

Shilakes, J. Tylman. Enterprise Information Portals. Enterprise Software Team, Merrill Lynch, 1998

Skoutas, D., & Simitsis, A., (2006). Designing ETL processes using semantic web technologies. In Proceedings ACM 9th International Workshop on Data Warehousing and OLAP (DOLAP 2006), pp.:67-74, Arlington, Virginia, USA, November 10, 2006

Skoutas, D., & Simitsis, A., (2007). Ontology-Based Conceptual Design of ETL Processes for Both Structured and Semi-Structured Data. Int. Journal of Semantic Web Information Systems (IJSWIS) 3, 4, 1-24

Skoutas, D., & Simitsis, A., (2007). Flexible and Customizable NL Representation of Requirements for ETL processes. In Proceedings 12th International Conference on Applications of Natural Language to Information Systems (NLDB 2007), pp.: 433-439, Paris, France, June 27-29, 2007

Simitsis A, Modeling and Optimization of Extraction-Transformation-Loading (ETL) Processes in Data Warehouse Environments. PhD thesis, National Technical University of Athens School of electrical and computer engineering, Division of computer science, Athens, 2004

Simitsis, A. (2005). Mapping conceptual to logical models for ETL processes. In Proceedings of ACM 8th International Workshop on Data Warehousing and OLAP (DOLAP 2005), pp.: 67-76 Bremen, Germany, November 4-5, 2005

Simitsis, A., & Vassiliadis, P. (2008). A Method for the Mapping of Conceptual Designs to Logical Blueprints for ETL Processes. Decision Support Systems, 45, 1, 22-40.

Simitsis, A., Vassiliadis, P., Terrovitis, M., & Skiadopoulos, S. (2005). Graph-Based Modeling of ETL activities with Multi-level Transformations and Updates. In Proc. 7th International Conference on Data Warehousing and Knowledge Discovery 2005 (DaWaK 2005), pp. 43-52, 22-26 August 2005, Copenhagen, Denmark.

Stöhr, T., Müller, R., & Rahm, E. (1999). An integrative and Uniform Model for Metadata Management in Data Warehousing Environments. In Proc. Intl. Workshop on Design and Management of Data Warehouses (DMDW 1999), pp. 12.1 – 12.16, Heidelberg, Germany, (1999)

Trujillo, J., & Luján-Mora, S. (2003). A UML Based Approach for Modeling ETL Processes in Data Warehouses. In Proceedings of 22nd International Conference on Conceptual Modeling (ER 2003), pp. 307-320, Chicago, IL, USA, October 13-16, 2003

Vassiliadis, P., Karagiannis, A., Tziovara, V., & Simitsis, A. (2007). Towards a Benchmark for ETL Workflows. 5th International Workshop on Quality in Databases (QDB 2007), held in conjunction with VLDB 2007, Vienna, Austria, 23 September 2007

Vassiliadis, P. (2009). A Survey of Extract–Transform–Load Technology. International Journal of Data Warehousing and Mining, Volume 5, Issue 3. edited by David Taniar © 2009, IGI Global

Vassiliadis, P., Quix, C., Vassiliou, Y., & Jarke, M. (2001). Data Warehouse Process Management. Information Systems, 26, 3, pp. 205-236

Vassiliadis, P., Simitsis, A., Georgantas, P., & Terrovitis, M. (2003). A Framework for the Design of ETL Scenarios. In Proc. 15th Conference on Advanced Information Systems Engineering (CAiSE 2003), pp. 520- 535, Klagenfurt/Velden, Austria, 16 – 20 June, 2003

Vassiliadis, P., Simitsis, A., Georgantas, P., Terrovitis, M., & Skiadopoulos, S. (2005). A generic and customizable framework for the design of ETL scenarios. Information Systems, 30, 7, 492-525

Vassiliadis, P., Simitsis, A., Terrovitis, M., & Skiadopoulos, S. (2005). Blueprints for ETL workflows. In Proc. 24th International Conference on Conceptual Modeling (ER 2005), pp. 385-400, 24-28 October 2005, Klagenfurt, Austria

Vassiliadis, P., Simitsis, A., & Skiadopoulos, S. (2002). Conceptual Modeling for ETL Processes.In Proc. ACM 5th International Workshop on Data Warehousing and OLAP (DOLAP 2002), McLean, VA, USA November 8, 2002.

Summary

ETL processes are supported by software pieces classified in 03 categories (1) Extracting of data from sources, (2) Transforming to deliver data quality with a value for the analysis (3) Loading the prepared data in the datawarehouse. In fact, the data are considered raw material used by decision-making information systems to produce useful information for decision support. Today, data are very complex with excessive sizes and present new structures and a variety of formats. In addition, new environments and new paradigms are developed. The ETL process is not immune to these changes, as it is responsible for capturing any data whatever its nature, size, provided that they are relevant and can bring value to the analysis process. The ETL processes are becoming increasingly complex facing the variety of these data formats and especially for Big Data. Facing to the Big Data, we propose, in this paper, a modeling approach for ETL processes according the MapReduce paradigm.

Construction de cube OLAP avec MapReduce dans un environnement Cloud Computing

Billel ARRES*, Nadia KABACHI**

* Université Lumière – Lyon 2
5 avenue Pierre Mandès-France, 69676 Bron, France
Billel.Arres@univ-lyon2.fr
** Université Claude Bernard – Lyon 1
43 Boulevard du 11 Novembre 1918, 69100 Villeurbanne, France
nadia.kabachi@univ-lyon1.fr

Résumé. L'informatique dans les nuages, sous l'impulsion des grandes compagnies telles que Google, Microsoft ou encore Amazon, a récemment suscité une attention particulière que ce soit dans le monde industriel ou académique. L'arrivée d'*Hadoop*, basé sur le paradigme *MapReduce*, pour le traitement parallèle des grandes quantités de données, a permis de faciliter l'accès à un tel environnement. L'objectif de cette étude est de mettre en place un environnement informatique dans les nuages permettant de créer des entrepôts de données et de réaliser des analyses en ligne. Il s'agit notamment de manipuler de larges bases de données non relationnelles, support des entrepôts de données, avec une nouvelle génération de systèmes de gestion de bases de données (SGBD) tels que *Hive*. Pour cela, nous avons implémenté un entrepôt de donnés issu de SSB[1] sous *Hive*. Nous avons ensuite comparé les résultats des temps de chargement de l'entrepôt et de construction de cube OLAP, ainsi que la montée en charge avec différentes variantes d'architectures. Les résultats obtenus peuvent servir de base pour de futurs travaux dans un but de comparaison des performances, d'aide dans le choix de plateforme, où des applications client peuvent être développées pour traduire des requêtes SQL vers des requêtes *Hive-QL*.

1 Introduction

Les entrepôts de données et les systèmes OLAP (*On-Line Analytical Processing*) représentent des technologies d'aide à la décision qui permettent l'analyse en ligne de gros volumes de données, W.Inmon (1996). Si leurs usages étaient initialement dédiés pour organiser, stocker et exploiter de manière optimale des données simples, ils n'avaient pas été conçus pour les nouvelles unités de mesures des données accumulées jusqu'à aujourd'hui. S.Chen (2010).

Popularisé par *Google*, un modèle novateur de traitement parallèle de données, nommé *MapReduce*, est présenté en 2004 dans un article désormais célèbre, J. Dean and S. Ghemawat (2004). La percée de ce modèle a rendu possible la réalisation, avec un matériel ordinaire, le traitement en une minute d'un problème qui auparavant exigeait une heure, à condition de multiplier par 60 le nombre de machines. S.Genaud (2011).

1- Star Schéma Benchmark Web Page. http://www.cs.umb.edu/~poneil/StarSchemaB.PDF1

Les architectures hautes performances visent à faire face aux besoins croissants en termes de calcul et de stockage des applications scientifiques et industrielles, M.Armbrust, et al (2009). Parmi ces architectures, le nuage informatique (*Cloud Computing*), qui consiste à externaliser les traitements et les données stockées. Par ailleurs, l'arrivée du projet *Hadoop*, T.White (2009), basé sur *MapReduce*, a permis d'exploiter facilement de telles architectures. Parmi les projets qui ont accompagné *Hadoop*, on retrouve *Hive,* qui est un SGBD pour la gestion de larges quantités de données non relationnelles et des grands entrepôts.

Il faut savoir que la construction des entrepôts de données et l'analyse en ligne dans les nuages est un domaine de recherche vraiment nouveau. En effet, si d'un côté M.Armbrust, et al (2009), de l'Université de Californie ont identifié des axes de recherches pour les nuages (disponibilité, qualité de service, sécurité, etc.), de l'autre côté, peu de travaux se sont intéressés à ce qui se fait techniquement par les grands acteurs du web dans ce domaine. D'où, la nécessité d'aller dans ce sens. Pour cela, nous nous sommes fixés comme objectif de monter un environnement de développement pour l'informatique décisionnelle basé sur *Hadoop* et *Hive* afin de mettre à l'épreuve cette nouvelle génération d'SGBD et de nous familiariser avec l'informatique dans les nuages et des larges bases de données. Nous proposons d'abord, de construire un entrepôt de donnée dans le *Cloud*, et ensuite d'évaluer les performances de différentes variantes d'architectures de plateformes mise en place avec *Hadoop*. Nous avons opté d'utiliser un échantillon de données SSB, qui est un banc d'essai d'aide à la décision, conçu pour mesurer les performances d'un entrepôt de données en étoile. Nous proposons de réaliser un tel environnement avec l'implémentation d'un entrepôt de donnés sous *Hive*. D'exploiter les fonctions *Map* et *Reduce* de cet environnement afin de comparer les résultats des temps de chargement de l'entrepôt de données et de construction d'un cube OLAP. Ceci se fera, d'une part, sur un cluster virtuel, sous forme d'une machine virtuelle (VM), puis sur un cluster physique à un seul nœud. D'autre part, nous envisageons d'évaluer les performances de *Hive*, et ce, en montée en charge (c'est-à-dire pour différentes tailles de l'entrepôt de données) sur un cluster physique à quatre nœuds puis à six nœuds. Les résultats obtenus pourront nous servir de base pour de futurs travaux dans un but de comparaison des performances, d'aide dans le choix de plateforme, où des applications client peuvent être développées pour traduire des requêtes SQL vers des requêtes *Hive-QL*, et enfin de pouvoir valider l'adéquation de modèles non-relationnels avec l'entreposage des données complexes.

Ce papier est organisé comme suit : la section 2 dresse un état de l'art pour présenter les différentes notions liées à notre travail de recherche, à savoir, le *Cloud Computing*, le paradigme *MapReduce* et le projet *Hadoop* d'*Apache*. La section 3 est consacrée à l'approche proposée pour l'étude de tels environnements. Nous avons effectué plusieurs expérimentations pour tester ces environnements que nous développons dans la section 4. Les résultats seront exposés et commentés dans la section 5. Et enfin nous terminerons par une conclusion et les perspectives de notre travail.

2 Etat de l'art et travaux connexes

2.1 Cloud Computing

Le *Cloud Computing* ou l'informatique dans les nuages est un ensemble de services déployés sur un réseau. C'est un concept qui consiste à déporter sur des serveurs distants des stockages et des traitements informatiques habituellement localisés sur des serveurs locaux.

L'architecture des nuages informatique se base généralement sur une organisation en couches. L. D'Orazio et al (2011). Le premier niveau correspond à l'infrastructure (*IaaS*). En général, il est composé de data center(s), mis à disposition par les fournisseurs de nuages. *Microsoft Azure* et *Amazon EC2* sont des exemples de telles infrastructures. Le deuxième niveau est dédié aux plateformes (*PaaS*). Il permet de mettre à disposition un environnement d'exécution rapidement disponible. L'exemple le plus connu à ce niveau est *MapReduce*, J. Dean and S.Ghemawat (2004), et son implémentation open source *Hadoop*, T.White (2009). Le troisième niveau représente l'environnement d'exécution (*SaaS*). Ce dernier propose des solutions logicielles sous forme de services hébergés et a pour objectif de permettre une facilitée d'utilisation et une totale transparence des autres couches de l'architecture. *Hive* de Facebook, A.Thusoo et al (2009), ou *Scoope* de Microsoft, sont des exemples de cette dernière couche basés sur des modèles de données particuliers, comme le modèle orienté colonne ou des modèles étendant le modèle relationnel tel que le *NoSQL* (*Not Only SQL*), Q.Wang et al (2007).

Le nuage informatique fournit une propriété d'élasticité permettant d'ajuster les ressources en fonction des applications, en augmentant les capacités de calcul et de stockage en cas de pics d'utilisation et les diminuant lors de périodes creuses, tout en autorisant la parallèlisation du stockage et des traitements des données. Selon les approches, on distingue deux modèles dominants de déploiement des services du *Cloud Computing* : Le *Cloud Publique*, accessible à un large public et appartient à un fournisseur de services, et le *Cloud Privé*, où l'infrastructure est dédiée complètement à une organisation unique. Dans notre travail, une architecture de *Cloud privé*, basée sur *Hadoop* et *Hive* est implémentée. Elle permet la compilation d'instructions *Hive-QL* et leurs exécution dans des environnements parallèles (section 3).

2.2 MapReduce et Hadoop

2.2.1 Le paradigme MapReduce

MapReduce est un modèle de programmation massivement parallèle adapté au traitement de très grandes quantités de données. J. Dean and S. Ghemawat (2004). Ce modèle de programmation s'articule autour de deux étapes principales *Map* et *Reduce* (cf. figure 1). Dans l'étape *Map* le nœud (machine) à qui est soumis un problème, le découpe en sous-problèmes, et les délègue à d'autres nœuds (qui peuvent en faire de même récursivement). Dans l'étape *Reduce* les nœuds les plus bas font remonter leurs résultats aux nœuds parents qui les avaient sollicités. À la fin du processus, le nœud d'origine peut recomposer une réponse au problème qui lui avait été soumis. Introduit par *Google*, *MapReduce* a été utilisé, pour traiter plus de 400 TB de données en 6 minutes seulement, avec un nombre de machines égal à 436, S.Genaud (2011).

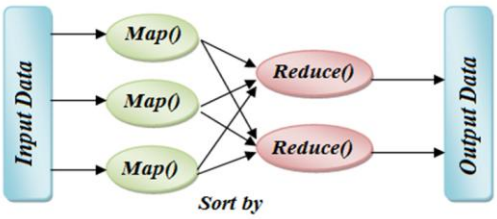


Fig. 1 – *Schéma global du principe MapReduce.*

L'implémentation la plus populaire du paradigme *MapReduce* est le framework *Hadoop*, qui permet aux applications d'être exécutées sur de larges clusters déployés sur des machines à faible coût. D'autres implémentations du paradigme *MapReduce* sont disponibles pour différentes architectures, telles que les architectures multicore, les architectures à multiples machines virtuelles, les environnements de *Grid Computing* ou encore les environnements mobiles.

2.2.2 Hadoop

Hadoop est un projet Open Source basé sur le paradigme *MapReduce* et *Google File System*. Il peut être considéré comme un système de traitement de données évolutif pour le stockage et le traitement par lot de très grandes quantités de données. Il est tout à fait adapté aux stockages et aux analyses de type "ad hoc" sur de très grandes quantités de données. T.White (2009).

Dans *Hadoop*, puisqu'il s'agit de la gestion de très gros volumes de données, il faut aussi optimiser l'utilisation de la bande passante sur le réseau. C'est pourquoi, *MapReduce* est généralement utilisé en combinaison avec un système de gestion de fichiers distribués, dans le cas d'*Hadoop* il s'agit de l'*HDFS* (*Hadoop Data File System*). *HDFS* a une architecture de type maître/esclave. V.Guana et J.Davidson (2012). Dans cette logique, un cluster *Hadoop* est constitué d'un unique serveur maître, nommé *NameNode*, qui gère le système de fichiers et les droits d'accès; mais aussi des serveurs qui sont à la fois un outil de calcul et un outil de stockage, nommés *DataNodes*, en général un par nœud. *Hadoop* a été largement adopté par la communauté du décisionnel et a rendu le domaine de l'entreposage de données sur le *Cloud* plus accessible.

2.2.3 Hive

Hive est un logiciel open-source qui permet aux programmeurs d'analyser de grandes quantités de données sur *Hadoop*, E.Capriolo et al (2012). Il implémente un langage de requête orienté SQL, nommé *HiveQL*, dont la mise en œuvre se traduit par l'exécution de *jobs* (tâches) *Map/Reduce* orchestrés par *Hadoop*. Il permet aussi le calcul à grande échelle ainsi que la tolérance aux pannes pour le stockage et le traitement sur du matériel de base.

Hive opère directement sur des fichiers bruts et hétérogènes, lesquels sont chargés et distribués sur le système de gestion de fichier *HDFS*. Il peut s'agir du résultat de requêtes en base de données, d'enregistrements au format csv ou de fichiers logs. *Hive* fournit donc un langage de requête simple basé sur SQL, et permet, aux utilisateurs qui y sont familiers, d'interroger, de résumer et d'analyser facilement les données. En même temps, *Hive* permet, aux programmeurs *MapReduce* traditionnels de charger leurs fonctions *Map* et *Reduce* pour faire des analyses plus sophistiquées.

3 Approche proposée

Nous avons mis en place une architecture de *Cloud Computing privé*, limitée en termes d'infrastructure. L'objectif est de tester la faisabilité de notre approche de construction d'un entrepôt de données dans le *Cloud*. Elle ne nous permet pas d'étudier le passage à très grande échelle. Cependant, elle nous aidera à déployer la parallèlisation du stockage et des traitements des données de l'entrepôt. Cela nous permet d'observer la réaction de cette approche

en effectuant une montée en charge, bien que celle-ci reste relativement limitée. Le but est de constater que les performances ne se dégradent pas lorsque nous augmentons la taille de l'entrepôt de données. Notre étude consiste donc, dans un premier temps, à la mise en œuvre du paradigme de traitement parallèle des données *MapReduce*, à travers de multiples clusters distribués. Ensuite, nous proposons une évaluation des performances de *Hive* en temps de chargement d'entrepôts de données et de construction de cubes OLAP via des requêtes *HiveQL* et cela selon différentes variantes d'architecture.

L'architecture proposée (cf. Figure 2) permet le partitionnement des données de l'entrepôt sur les différents clusters (nœuds), la construction du cube OLAP et son interrogation par l'utilisateur. Nous avons opté d'utiliser le modèle de données SSB, P.O'Neil et al (2009). C'est un banc d'essai, facilement exploitable, conçu pour mesurer les performances d'un entrepôt de données en étoile (cf. Figure 3).

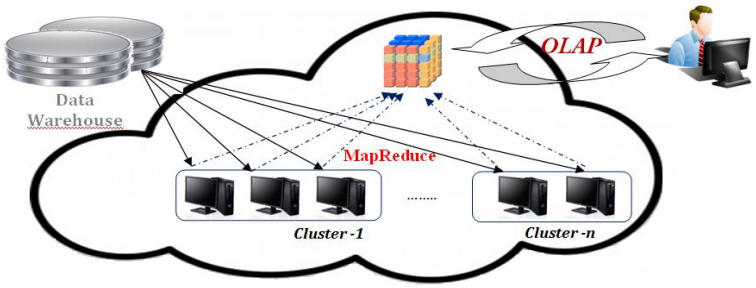


Fig. 2 – *Architecture proposée.*

3.1 Phase de construction de l'entrepôt de données

En se basant sur un modèle en étoile classique de ventes, l'entrepôt de données mis en place est constitué d'une table de FAIT appelée LINEORDER. Cette dernière dispose de dix-sept attributs pour renseigner une commande dont une clé primaire, composée de ORDERKEY et de LINENUMBER, et des clés étrangères des tables de dimension CUSTOMER, PART, DATE et SUPPLIER.

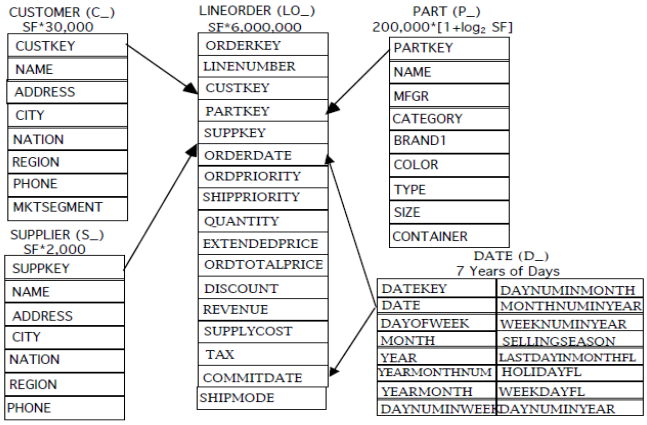


Fig. 3 – *Schéma de l'entrepôt de données.*

3.2 **Phase de chargement et de construction du cube OLAP**

La majorité des opérations supportées par le *HiveQL* sont très similaires au *SQL*. Cependant, la différence architecturale des deux systèmes sur lesquels se base ces langages, et particulièrement, l'utilisation de l'*HDFS* comme système de gestion de fichiers par *Hive*, impose d'autres opérations supplémentaires qui nécessitent l'adaptation de l'utilisateur.

Ainsi, pour mesurer les performances de *Hive* en temps de chargement de l'entrepôt de données et de construction du cube OLAP, nous avons élaboré des requêtes en *HiveQL*. Nous avons choisi de construire un cube de données répondant à la requête décisionnelle suivante : *« Quelle est la somme des revenues des ventes par année et par marque pour la région Asie depuis 1992 ? »*. On cherche ici la somme des revenus (REVENUE dans LINEORDER) comme mesure selon les dimensions PRODUIT (PART) et PERIODE (DATE). Fig. 3.

// Exemple de création de la table « SUPPLIER » avec *Hive* :

```
CREATE TABLE supplier (ssuppkey INT, sname STRING, saddress STRING, scity
STRING, snation STRING, sregion STRING, sphone INT) ROW FORMAT DELIMITED
FIELDS TERMINATED BY '|' LINES TERMINATED BY '\n';
```

// Chargement des données dans la table «SUPPLIER» à partir du fichier supplier.tbl :

```
LOAD DATA LOCAL INPATH'/home/Public/Stage/SSB/1Go/supplier.tbl' INTO TABLE
supplier;
```

// Groupe de requêtes de construction du cube OLAP en *HiveQL* :

```
INSERT OVERWRITE TABLE store select * from (
    select sum(l.lorevenue) as revenu, d.dyear as ANNEE, p.pbrand1 as MARQUE
    from lineorder l JOIN ddate d ON (l.loorderdate = d.ddatekey) JOIN part p On
    (l.lopartkey = p.ppartkey)JOIN supplier s ON (l.losuppkey = s.ssuppkey)
    where d.dyear >= 1992 and s.sregion = 'ASIA'
    group by d.dyear, p.pbrand1
UNION ALL
    select sum(l.lorevenue) as revenu, d.dyear as ANNEE, 'ALL' as MARQUE
    from lineorder l JOIN ddate d ON (l.loorderdate = d.ddatekey) JOIN part p ON
    (l.lopartkey = p.ppartkey) JOIN supplier s ON (l.losuppkey = s.ssuppkey)
    where d.dyear >= 1992 and s.sregion = 'ASIA'
    group by d.dyear
UNION ALL
    select sum(l.lorevenue) as revenu, 'ALL' as ANNEE, p.pbrand1 as MARQUE
    from lineorder l JOIN ddate d ON (l.loorderdate = d.ddatekey) JOIN part p ON
    (l.lopartkey = p.ppartkey) JOIN supplier s ON (l.losuppkey = s.ssuppkey)
    where d.dyear >= 1992 and s.sregion = 'ASIA'
    group by p.pbrand1
UNION ALL
    select sum(l.lorevenue) as revenu, 'ALL' as ANNEE, 'ALL' as MARQUE
     from lineorder l JOIN ddate d ON (l.loorderdate = d.ddatekey) JOIN part p ON
    (l.lopartkey = p.ppartkey) JOIN supplier s ON (l.losuppkey = s.ssuppkey)
    where d.dyear >= 1992 and s.sregion = 'ASIA'
    ) lineorder
```

4 Expérimentation

Dans cette section, nous allons évaluer les performances de *Hive* en temps de chargement de l'entrepôt et en temps de construction du cube OLAP, selon différentes variantes de l'architecture définie. Les deux premières expérimentations permettent de comparer les performances de *Hive* entre un environnement virtuel et un environnement physique, et cela, en fixant le nombre de nœuds (1 nœud) et la taille de l'entrepôt de données mis en place (1Go). La troisième expérimentation permet le passage à l'échelle dans un environnement physique, avec la variation du nombre de nœuds (de 4 à 6 nœuds) ainsi que la taille de l'entrepôt de données (1Go à 1To).

4.1 Expérimentation sur un cluster virtuel (1 nœud)

Dans un but d'évaluation et de familiarisation avec le système *Hadoop*, la machine virtuelle est un bon choix de départ, du moment qu'elle permet le déploiement du système *Hadoop* avec un minimum de temps et de ressources matérielles. Nous avons choisi d'installer le package de développement *Cloudera-CDH4*[2], pour machine virtuelle (VM), qui offre une pré-installation d'*Hadoop* et ses composants dont *Hive*, sous *Ubuntu*. Le package a été installé sur une machine physique présentant des caractéristiques standards à savoir : mémoire installé de 4Go et un processeur Intel 3.10GHz x4. (Mêmes caractéristiques pour la VM).

La configuration est équivalente à un cluster virtuel à un seul nœud, où la machine fonctionne en tant que *NameNode* et *DataNode* en même temps. Dans cette première partie de l'expérimentation, nous avons utilisé un entrepôt de données de taille égale à 1 Go suffisant pour tester l'architecture proposée dans un environnement virtuel.

4.2 Expérimentation sur un cluster physique (1 nœud)

La deuxième partie de l'expérimentation a pour but de comparer les performances de *Hive* entre un cluster physique et le cluster virtuel déjà mis en place. Ainsi, la même plateforme a été déployée, mais cette fois, sur une machine physique, avec une mémoire installée de 4Go (RAM), d'un processeur Intel 3.10GHz x 4 et *Ubuntu* 12.10 type 64bits comme système d'exploitation. Nous avons utilisé la version d'*Hadoop.0.20.0* mis à disposition par *Apache*. L'entrepôt de données utilisé est le même que celui de la première partie de l'expérimentation (1Go). La configuration mise en place est équivalente à un cluster avec un seul nœud. La machine travaille en tant que *NameNode* et *DataNode* en même temps. *Hive* a été installé à partir d'une version stable 0.9.0, disponible sur le site d'*Apache*.

4.3 Montée en charge sur le cluster physique (4 – 6 nœuds)

Cette partie de l'expérimentation a pour objectif de mesurer les temps de réponse de la plateforme *Hadoop/Hive* en fonction de sa sollicitation. Ceci nous permettra de toucher au cœur du système *Hadoop* qui est l'utilisation avancée de l'*HDFS* et la répartition des données et des charges sur le réseau. Tout cela, avec la mise en place d'un cluster physique selon l'architecture proposée, et tester ainsi, modestement, ce qui se fait actuellement par les

2- Cloudera Entreprise – CDH4. Web Page. http://www.cloudera.com/

grands acteurs du web avec beaucoup plus de moyens, tout en essayant d'évaluer, à notre échelle, les performances de telles infrastructures. La première évaluation de *Hive* sur la scalabilité a été effectuée sur un cluster de machines à quatre nœuds identiques, y compris la machine maître. La deuxième évaluation de *Hive* sur la scalabilité a été effectuée, cette fois, sur une configuration à six nœuds identiques, y compris la machine maître. Les machines utilisées présentent toutes les mêmes caractéristiques à savoir une mémoire de 4Go (RAM) et d'un processeur Intel 3.10GHz x 4. Cela, permettra d'évaluer les performances de la plate forme *Hadoop* avec l'augmentation de la taille de l'entrepôt et du nombre de nœuds.

5 Présentation et analyse des résultats

Dans cette partie nous allons présenter les résultats des expérimentations effectuées selon deux volets. Le premier volet consiste à fixer la taille de l'entrepôt de données (1Go) et le nombre de nœuds (1 nœud) et comparer les résultats de performances entre un cluster physique et un cluster virtuel. Le deuxième volet consiste à comparer les résultats de performances d'un cluster physique, mais cette fois ci, en variant la taille de l'entrepôt de données (de 1Go à 1To) et le nombre de nœuds (4 puis 6 nœuds).

5.1 1er Cas : Configuration avec un cluster virtuel et un cluster physique

Les résultats des tests effectués sur les deux plateformes virtuelle et physique sont résumés dans le tableau ci-dessous. Le chargement de chaque table est mesuré en seconde. La somme du temps des chargements de toutes les tables équivaut au temps total de chargement de l'entrepôt de données par *Hive*. Le temps de construction du cube OLAP, est le temps nécessaire à *Hive* pour l'exécution de tous les jobs (tâches) *MapReduce* créés en compilant le groupe de requêtes *HiveQL* de construction du cube.

Temps de chargement des tables	Cluster virtuel (1Go)	Cluster physique (1Go)
SUPPLIER	1.337	0.763
CUSTOMER	0.435	0.418
PART	1.577	0.922
DDATE	0.317	0.312
LINEORDER	139.584	29.675
Totale temps de chargement (s)	143.243	32.090
Temps de construction du cube (s)	2148.157	1074.619

TAB. 1 – *Résultats de l'expérimentation des deux plateformes virtuelle et physique.*

Les graphiques suivant illustrent de manière plus visuelle les résultats des temps de chargement des tables et des temps de construction du cube OLAP entre le cluster physique à un nœud et le cluster virtuel à un nœud :

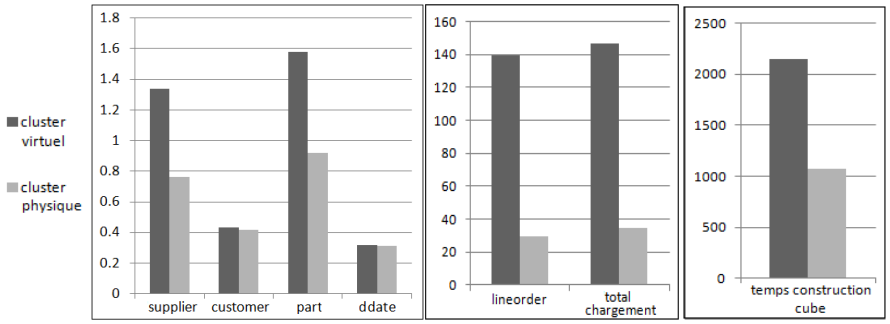


Fig. 4 – *Résultats des temps de chargement de l'entrepôt de données et de construction du cube OLAP entre le cluster virtuel à un nœud et le cluster physique à un nœud.*

Les résultats obtenus à partir des tests effectués montrent bien la supériorité de l'environnement physique en temps de chargement de l'entrepôt (32.09 s) et de construction du cube OLAP (17.91 mn), et ce en consommant la moitié du temps nécessaire au cluster virtuel pour le chargement (2.38 mn) et la construction du cube OLAP (35.8 mn). Le cluster virtuel permet une installation rapide et facile surtout en exploitant le package *Cloudera*. Il facilite la prise en main du système *Hadoop*, avec tous ses composants. Il permet également de créer un environnement facile à utiliser dans le but de se familiariser avec l'informatique de nuage et de l'entreposage de données tel que *Hive*. Cependant l'utilisation à grande échelle avec des données réelles montre les limites de l'environnement virtuel et avantage l'environnement physique.

5.2 2ème Cas : Configuration avec un cluster physique à 4 puis à 6 nœuds

Les résultats de l'expérimentation de la montée en charge sous *Hive*, en temps de chargement d'entrepôt de données et en temps de construction du cube OLAP, sur un cluster à quatre nœuds puis à six nœuds sont présentés dans le tableau ci-dessous. La taille de l'entrepôt de données varie de 1Go à 1To. Les temps sont mesurés en secondes :

	Temps de chargement de l'entrepôt de données (s)		Temps de construction du cube OLAP (s)	
Taille (Go)	4 nœuds	6 nœuds	4 nœuds	6 nœuds
1	14.815	13.445	740.01	693.39
10	128.712	118.079	2085.523	1737.121
30	381.307	375.403	6014.520	3987.577
60	810.339	671.207	10978.90	8115.648
80	968.586	963.814	15806.711	9512.574
100	1220.783	1161.447	17622.847	11247.618
250	1512.735	1323.995	19441.608	11561.249
500	1815.883	1484.313	22683.214	12101.649
750	2103.946	1615.055	23473.916	12842.988
1000	2406.340	1762.792	23967.860	13377.016

TAB. 2 – *Résultats de la montée en charge sur les clusters physiques à 4 et à 6 nœuds.*

On remarque que le temps de chargement de l'entrepôt de données évolue avec l'augmentation de sa taille. Ainsi, si on prend les résultats des temps de chargement et de construction du cube OLAP pour un cluster à 6 nœuds, ils augmentent clairement avec l'augmentation de la taille de l'entrepôt, allant de 13.44 secondes pour le chargement d'un entrepôt de 1 Go à plus de 29 minutes pour l'entrepôt de 1To ; et d'un peu moins de 12 minutes pour la construction du cube OLAP pour un entrepôt de 1 Go à moins de 4 heurs pour un entrepôt de 1To de données. Ce qui est à première vue énorme. Du fait que les machines utilisées pour les tests sont des machines ordinaires à faible capacités, par rapport à ce qui est utilisé généralement. Cependant, ces résultats restent encourageants, car il faut souligner qu'il y'a une différence importante lors du passage de quatre à six nœuds, spécialement les résultats de construction du cube OLAP. Ceux-là, nécessitent l'exécution du bloc de requêtes *HQL* et le traitement des résultats. La figure 5 compare les résultats des tests effectués sur les clusters à 4 puis à 6 nœuds :

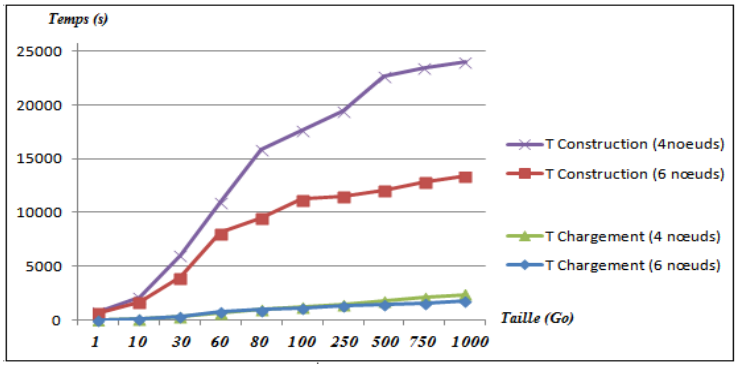


Fig. 5 – *Résultats de la montée en charge sur les clusters physiques à 4 et 6 nœuds.*

Les temps de chargement de *Hive* sont pratiquement les mêmes pour les deux clusters à 4 et 6 nœuds. Ces résultats suivent une évolution selon une droite avec une pente relativement faible. Il faut savoir que la phase de chargement ne nécessite ni la création ni l'exécution de *jobs MapReduce* par le compilateur de *Hive*. Les données sont découpées en petites unités et montées de façon identique sur les différents nœuds (*DataNodes)* du cluster via l'*HDFS*. Ainsi, plus la taille de l'entrepôt augmente plus il est nécessaire à *Hive* de consommer plus de temps. Par contre, l'évolution de ce temps reste relativement faible allant de 2 minutes pour un entrepôt de données de 10 Go à 20 minutes pour le chargement d'un entrepôt de données de 100 Go. Cependant, la variation du nombre de nœuds n'est probablement pas assez significative (de 4 à 6) pour mettre en évidence les différences entre les temps de chargement que des tests plus spécifiques, à l'échelle d'*Hadoop*, avec une dizaine voire une centaine de nœuds, pourront révéler.

Quant aux temps de construction du cube OLAP, les résultats des tests sont presque les mêmes pour les entrepôts de données qu'*Hadoop* et *Hive* considèrent comme relativement petits (inferieur à 10Go). Par exemple, le temps de construction du cube OLAP pour l'entrepôt de 1 Go de données est de 12 minutes 30 avec 4 nœuds et de 12 minutes avec 6 nœuds une différence donc de 30 secondes. L'écart entre les résultats des temps de construction du cube OLAP entre les deux clusters apparait bien à partir d'un entrepôt de 10 Go et les avantages d'ajouter deux nœuds au cluster sont appréciables au fur et à mesure que la taille

de l'entrepôt augmente. Par exemple, le temps de construction du cube OLAP, pour un entrepôt de 100Go, prend un peu plus de 3 heurs avec 6 nœuds et presque 5 heurs avec 4 nœuds une différence donc de 2 heures.

Les temps de construction du cube OLAP avec *Hive* sur le cluster à 6 nœuds sont nettement meilleurs que ceux obtenus avec celui à 4 nœuds. Dans ce cas, la construction du cube s'est faite en exécutant le groupe de requêtes *HQL*, définies précédemment. Des *jobs MapReduce* sont alors créés automatiquement par le compilateur de *Hive* et exécutés par les différents nœuds du cluster qui contiennent aussi les partitions de données. Ainsi, plus le nombre de nœuds augmente moins la taille des partitions de données est grande et plus les *Jobs MapReduce* retournent leurs résultats assez rapidement. Au final, le plan d'exécution reste le même entre les deux clusters mais le nombre de nœuds augmente ce qui minimise la taille des différentes partitions de données et accélère l'exécution des différents jobs accélérant ainsi la construction du cube OLAP.

6 Conclusion & perspectives

Beaucoup d'organisations ou d'institutions ont besoin de stocker de grandes quantités de données, mais la plupart des SGBD relationnels actuels ne permettent plus de répondre aux besoins de stockage et de traitement de ces énormes quantités de données accumulé jusqu'à aujourd'hui. En plus, et depuis l'apparition de *Facebook*, *Amazon* et *Google*, l'informatique dans les nuages, prend de plus en plus de place dans la société d'aujourd'hui. Cependant, peu de travaux mettent en évidences les technologies développées derrière ces noms. C'est dans ce contexte qu'apparaît le paradigme *MapReduce*. Notre motivation était de manipuler ces notions en créant un environnement avec un projet autour de ces technologies. Ce travail visait la création d'un environnement « type » sous *Hadoop*, en vue de se familiariser avec l'informatique dans les nuages et les grands entrepôts de données. Les expérimentations effectuées ont permis de manipuler et de monter différents entrepôts de données sous *Hive* et d'évaluer les performances de différentes variantes d'architectures de plateformes mises en place avec *Hadoop*. De plus, ces expérimentations nous ont permis de toucher le cœur d'*Hadoop* qui est son système de gestion de fichier *HDFS*, et surtout de comprendre la logique de son fonctionnement. De plus, l'environnement d'*Hadoop* sous *Ubuntu* (en plus de celui de la distribution *Cloudera* pour machine virtuelle), a permis de valider de nouvelles versions des projets de l'écosystème Hadoop tel que *Hive*. Celles-ci sont de plus en plus stables et relativement faciles à déployer pour pouvoir exploiter et profiter de la puissance de *MapReduce* dans la gestion des grands entrepôts de données. Ce travail nous a permis de découvrir le domaine de la gestion des grandes quantités de données avec *Hadoop* et *MapReduce* ainsi que l'introduction au concept de *Cloud Computing*. Les perspectives de ce projet sont, à court terme, de continuer ces expérimentations en augmentant le nombre de nœuds ainsi que la taille des données à l'échelle de plusieurs To voire Po pour mieux évaluer les performances de ces systèmes. A moyen terme, nous envisageons de développer et d'implémenter de nouveaux algorithmes avec les fonctions *Map* et *Reduce* de cet environnement et exploiter l'architecture du *Cloud* au niveau *PaaS*. Enfin, à long terme nous envisageons de développer des outils de solution BI (*Business Intelligence*) dans ce type d'environnement.

Références

A.Thusoo, et. Al. 2009 VLDB: Hive: A warehousing solution over a MapReduce framework.

E.Capriolo, D. Wampler, J. Rutherglen. O'Reilly 2012: Programming Hive.

J.Dean and S. Ghemawat. Communications of the ACM 2004: MapReduce: Simplified data processing on large clusters.

L.D'Orazio, S. Bimonte. EDA 2010: Intégration des Tableaux Multidimensionnels en Pig pour l'Entreposage de Données sur les Nuages.

M.Armbrust, A. Fox, R. Griffith. Technical Report UCB/EECS 2009: Above the Clouds: A berkeley view of Cloud Computing.

P.O'Neil, B.O'Neil, X.Chen: Star Schéma Benchmark. 2009, Web Page. http://www.cs.umb.edu/~poneil/StarSchemaB.PDF.

Q.Wang, et.al. 2007 VLDB: On The Correctness Criteria of Fine Grained Access Control in Relational Databases.

S.Chen, 2010 VLDB: Cheetah: A High Performance, Custom Data Warehouse on Top of MapReduce.

S.Genaud. Université de Strasbourg 2011: MapReduce: Un cadre de programmation parallèle pour l'analyse de grandes données.

T.White. O'Reilly 2009: Hadoop: The Definitive Guide.

V.Guana and J.Davidson. University of Alberta 2012: On Comparing Inverted Index Parallel Implementations Using Map/Reduce.

W.Inmon. Wiley, New York, USA, 1996: Building the Data Warehouse.

Summary

Large-scale data analysis has become increasingly important for many enterprises, and Cloud Computing, has recently endowed a special attention both in industry and academic researches. *Hadoop*, based on a new distributed computing paradigm, called *MapReduce*, has allowed facilitating the access to such environments, due to its impressive scalability and flexibility to handle structured as well as unstructured data. The goal of our work is to develop a Cloud Computing environment for exploiting data warehouses and perform online analysis. It consists of handling large non relational databases and supporting data warehouse with a new generation of database management systems (DBMS) such as *Hive*. Thus, to set up such an environment, we implemented a data warehouse under *Hadoop* and *Hive* and we used the *Map* and *Reduce* functions of this environment, then we compared the cost of loading the warehoused data and constructing OLAP cubes between a virtual and a physical cluster, as well as the rise in data loading on a physical cluster. Obtained results allows MapReduce developers to fully compare the performance, help in the choice of platform, in which a customer application can be developed to translate SQL requests to HQL (Hive-QL) requests, and check if a not-relational model is adequate or not.

Modélisation Multidimensionnelle des données textuelles: où en sommes-nous?

Sarah Attaf*, Nadjia Benblidia*

*Laboratoire LRDSI université Saad Dahlab Blida,
route de Soumaa BP 270 Blida(09000)
sarah.attaf@gmail.com
benblidia@yahoo.com

Résumé. La modélisation des données textuelles dans un but d'analyse consiste à donner à ce type de données non structurées, une représentation permettant de les décrire de façon suffisamment formelle pour pouvoir tirer profit des informations qu'elles contiennent. Notre étude des différents travaux portant sur la modélisation multidimensionnelle des données textuelles nous a permis de les classer, selon le type de concepts utilisés, en deux familles de modèles : (i)modèles extensifs basés sur les concepts de base des modèles d'entrepôts classiques et (ii) modèles à nouveaux concepts proposant une nouvelle manière de percevoir la complexité des données textuelles. Nous comparons l'ensemble de ces travaux par rapport à leur niveau de traitement des cinq aspects suivants: (1) la prise en compte de la structure des données textuelles, (2) la prise en compte de la sémantique véhiculée dans ces données,(3) la flexibilité d'analyse offerte par chaque modèle,(4)la prise en compte d'une mesure textuelle et (5) la définition de nouveaux opérateurs OLAP pour les données textuelles.

1 Introduction

Le document électronique représente aujourd'hui un vecteur et un support d'information que les organisations ne doivent pas négliger. En effet, il est entendu que plus de 80 % des données nécessaires au fonctionnement d'une organisation sont encapsulées dans des documents, et non uniquement dans les bases de données opérationnelles. Ces données textuelles restent hors de portée des systèmes décisionnels, ce qui induit qu'une grande partie de l'information demeure inaccessible. Pour répondre à cette problématique et afin de pouvoir prendre profit des informations contenues dans ces documents, il est devenu plus que nécessaire d'intégrer ces données textuelles dans des systèmes d'information décisionnels permettant leur analyse.

Les systèmes décisionnels classiques ont déjà fait leurs preuves dans le domaine de l'analyse des données simples. Or ces systèmes ne sont pas adaptés à l'analyse des documents textes, ce qui met en évidence la nécessité de créer de nouveaux modèles multidimensionnels pour les données textuelles. L'entreposage de ces dernières demeure encore aujourd'hui une des difficultés majeures, et implique de nombreux problèmes, notamment ceux de leur modé-

lisation et leur intégration d'une part et leur analyse d'autre part.

Les entrepôts de textes sont apparus comme une nouvelle solution, permettant une analyse multidimensionnelle des données textuelles. La nature complexe de ces données nécessite un traitement bien particulier, qui prend en compte leur sémantique. Dans la littérature, des méthodes de recherche d'information et de fouille de données ont donné de très bons résultats pour l'exploration des données textuelles. L'idée clef derrière ces entrepôts de textes est de faire un couplage entre les techniques de fouille de données et de recherche d'information d'un coté, et les techniques OLAP de l'autre coté.

Dans la littérature, La modélisation multidimensionnelle est définie comme une technique qui vise à organiser les données de telle sorte que les applications OLAP soient performantes et efficaces. Un modèle multidimensionnel permet d'observer des faits à travers des indicateurs (mesures) et des dimensions. Dans le domaine de l'analyse des données textuelles, de différents modèles multidimensionnels d'entrepôts de textes ont été élaborés. Nous distinguons deux familles de modèles : (i) modèles extensifs basés sur les concepts de base des modèles d'entrepôts classiques et (ii) modèles à nouveaux concepts proposant une nouvelle manière de percevoir la complexité des données textuelles. Nous présentons dans cet article un ensemble de modèles dédiés à l'analyse multidimensionnelle des données textuelles, ainsi qu'une étude comparative de ces modèles.

La suite de l'article est organisée comme suit. La section 2 présente un ensemble de modèles spécifiques à l'analyse multidimensionnelle des données textuelles, catégorisés en deux familles de modèles : (i) modèles extensifs et (ii) modèles à nouveaux concepts. La section 3 présente une étude comparative des modèles présentés selon cinq critères de comparaison : la prise en compte de la structure des données textuelles, celle de la sémantique, la flexibilité des modèles étudiés, la prise en compte d'une mesure textuelle et la définitions de nouveaux opérateurs OLAP pour les données textuelles. Finalement, la section 4 conclut l'article et évoque quelques perspectives.

2 Modèles Multidimensionnels d'analyse des données textuelles

La modélisation multidimensionnelle consiste à organiser les données de façon que les applications OLAP soient performantes et efficaces. Les modèles d'entrepôts existants offrent un cadre de travail pour effectuer une modélisation multidimensionnelle des données simples, mais ils ne sont pas adaptés aux données textuelles. Pour répondre à cette problématique, plusieurs travaux ont été élaborés. Certains ont proposé d'étendre les modèles d'entrepôts classiques, pour prendre en compte la complexité de l'analyse des données textuelles, ce qu'on désignera par **modèles extensifs**, tandis que d'autres ont proposé de nouveaux modèles basés sur de nouveaux concepts, ce qu'on désignera par **modèles à nouveaux concepts**.

2.1 Modèles extensifs

Les modèles d'entrepôts extensifs sont de nouveaux modèles multidimensionnels, dédiés à l'analyse des données textuelles. Ces modèles proposent des extensions du modèle d'entrepôt classique en se basant sur les deux concepts de base : fait et dimension. Parmi ces travaux nous citons le moteur multidimensionnel de recherche d'information (MIRE) Lee et al. (2002), qui est basé sur un modèle multidimensionnel de données textuelles. Leur approche consiste à construire un modèle multidimensionnel dédié aux données textuelles et permettre aux utilisateurs de faire des recherches à l'aide des techniques OLAP. Le système MIRE peut répondre à une requête multidimensionnelle tel que 'Trouver tout les documents contenant les termes modélisation multidimensionnelle, qui sont publiés en France pendant l'année 2009'. MIRE est une approche qui permet de construire un système de recherche d'information sur les bases des systèmes OLAP, où une table de faits contient la mesure (les mots apparaissant dans le document), et les tables de dimensions contiennent des données structurées d'une manière hiérarchique. MIRE intègre un index inversé pour les données textuelles et des méthodes d'accès multidimensionnels. Ces méthodes sont utilisées pour traiter les dimensions, et peuvent fournir des fonctionnalités d'OLAP tel que le *drill down* et le *roll up* sur les dimensions. Pour faire face à des problèmes d'évolutivité, MIRE construit un index inversé et utilise une structure multidimensionnelle d'accès unique *modified kdb tree* pour accéder aux données multidimensionnelles. La requête 'Trouver tout les documents contenant les termes modélisation multidimensionnelle qui sont publiés en France pendant l'année 2009 ' est traitée en deux phases. Les dimensions TEMPS et LOCALISATION sont retrouvées à partir de la structure d'accès multidimensionnelle. L'ensemble des documents retournés de cette structure seront enregistrés dans une collection de documents intermédiaire. A partir de l'indexe inversé, les documents qui contiennent les termes 'modélisation multidimensionnelle' seront enregistrés dans un autre ensemble intermédiaire. Un ensemble final de documents pondérés peut être obtenu par la fusion des deux ensembles intermédiaires obtenus précédemment.

Mothe et al. (2003) ont proposé un modèle basé sur un schéma en étoile nommé *Docube*, qui permet de produire des vues globales de grands corpus de documents, en utilisant la classification. Son élément de base est l'utilisation du concept hiérarchie afin de structurer les collections de documents, chaque hiérarchie correspond à une facette de documents 'dimension d'analyse' pour laquelle les utilisateurs peuvent être intéressés. Quelques exemples de dimensions sont : auteur, affiliation...*etc*. Ces dimensions ne sont pas différentes de celles utilisées dans les systèmes OLAP classiques. Tandis que le contenu de la table fait est différent. Celle-ci contient un lien qui associe une ligne de cette dernière à chaque document. Ce lien peut être pondéré par rapport au degré d'association du document avec les valeurs de la table de dimension. Ce poids est obtenu par l'application de la méthode de classification *vector voting*. Le lien est représenté par le fichier **Doc.Ref** qui correspond à l'identifiant du document ou à son URL. **DocCube** fournit deux ou trois dimensions de visualisation de sorte que l'utilisateur peut visuellement savoir combien de documents sont reliés les uns aux autres dans l'espace multidimensionnel et accéder directement à leurs contenus. Ils ont proposé aussi une fonction $score(Dd)$ qui retourne les tops documents, ces scores sont calculés par la moyenne des poids associés aux documents.

Tseng et al. (2006) ont proposé un entrepôt de documents pour l'analyse multidimensionnelle de documents textes. Dans leur modèle, une dimension est représentée par une structure d'arbre de m niveaux, utilisée pour représenter les relations hiérarchiques d'un ensemble de mots clefs issus des documents à analyser, des catégories de documents et des méta-données tels que *le titre, l'auteur, la date..etc*(chaque document est représenté par un ensemble de mots clefs). Toutefois, ils n'ont pas mentionné la façon dont les méta-données et les mots-clés pourraient être organisés sous forme hiérarchique. Ils utilisent comme mesure d'analyse une mesure numérique qui consiste à calculer le nombre de documents. Par exemple 'calculer le nombre de documents traitant la modélisation multidimensionnelle, entre l'année 2006 et 2013'. Nous pouvons à travers leur modèle, sélectionner un cube de documents à partir de l'entrepôt de documents pour permettre aux utilisateurs de naviguer dans les documents par un forage vers le haut *roll-up* et un forage vers le bas *drill-down* le long de certaines dimensions de différentes granularité.

Lin et al. (2008) ont proposé un cube de textes nommée *TextCube* dans lequel une dimension textuelle est représentée par une hiérarchie de termes. Cette hiérarchie spécifie les relations sémantiques entre les termes textuels extraits des documents, ce qui permet une navigation sémantique dans les données textuelles grâce aux deux opérateurs qui lui sont associés : *pull-up* and *push-down*. Ils définissent aussi dans leur cube, deux mesures d'agrégation adaptées aux données textuelles, fréquence des termes *term frequency TF* et l'index inversés *inverted index IV*.

Zhang et al. (2009) ont proposé un modèle nommé *Topic Cube* qui étend le cube de données traditionnel en intégrant une hiérarchie de thèmes '*Topics*' comme étant une dimension d'analyse, la figure 1 nous illustre un exemple d'une hiérarchie de *topics* (thèmes) extraits des rapports sur les anomalies enregistrées lors des vols. La racine représente l'agrégation de tous les thèmes (tous ce qui représente une anomalie), le niveau suivant comporte certaines anomalies générales, comme *Anomaly Altitude Diviation*. Un noeud enfant représente un événement spécialisé de l'événement représenté dans le noeud père, par exemple, *Undershoot* et *Overshoot* sont deux anomalies spécifiques à l'événement *Anomaly Altitude Deviaton*.

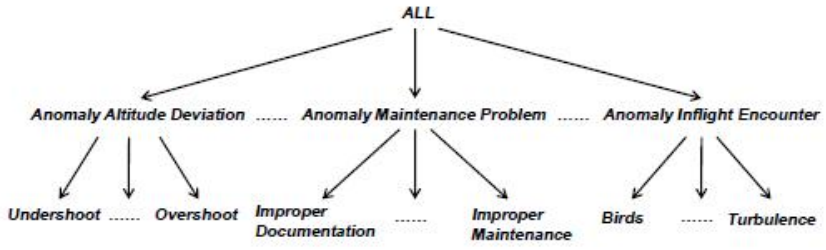


FIG. 1 – *Exemple d'un arbre d'hiérarchie de thèmes "cas d'anomalies" Zhang et al. (2009).*

Topic Cube propose deux mesures probabilistes : la distribution d'un mot dans un thème *word distribution of a topic* $p(wi)$ et la couverture d'un thème par les documents *topic coverage by documents* $p(topic.j)$. La couverture d'un *topic* est la probabilité qu'un document dj couvre le *topic*. Ainsi, nous pouvons facilement prédire quel est le sujet dominant dans l'ensemble des documents en agrégeant la couverture sur tous les documents dans l'ensemble. La figure 2 décrit le schéma en étoile d'un *Topic Cube*

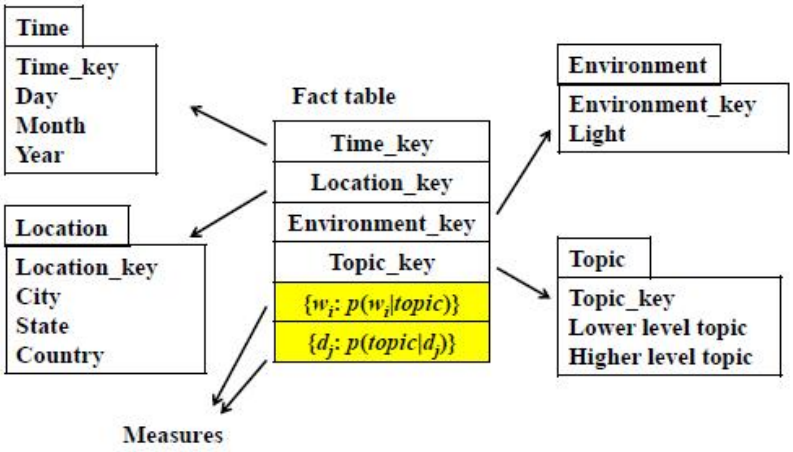


FIG. 2 – *schéma en étoile d'un Topic Cube Zhang et al. (2009).*

Bautista et al. (2010) ont proposé un modèle multidimensionnel qui prend en charge les informations textuelles dans un entrepôt de données, en introduisant une dimension textuelle *AP-Dimension* obtenue par la transformation des données textuelles en une structure sémantique *AP-Structure*. Cette structure est basée sur les items fréquents nommés *AP-sets(apriori sets)* obtenus par l'application de l'algorithme *apriori* sur les attributs textuels d'une base de données transactionnelle, Bautista et al. (2006). Le cube de données résultant est un cube de données classique qui intègre une dimension textuelle, tandis que la mesure définie pour ce cube reste une mesure numérique.

Zhang et al. (2011) ont proposé *MicroTextClusterCube*, un modèle basé sur un schéma en étoile. Ils ont proposé d'introduire une nouvelle mesure d'analyse *(meani, sizei)* qui représente respectivement le vecteur *mean* (un vecteur de termes pondérés) et la taille d'un *micro-cluster*, où un *micro-cluster* est une cellule texte qui permet de compresser les documents similaires (chaque cellule texte contient un certain nombre de documents). Cette compression (en *micro-cluster*) permet de retenir des informations sémantiques essentielles sur les cellules textuelles.

Les travaux dans cette catégorie ont permis d'étendre les modèles d'entrepôts classiques pour assurer l'analyse des données textuelles. Ils ont basé leurs modèles sur les deux concepts : fait et dimension. Ces modèles ont pu traiter la sémantique des données textuelles à travers

l'intégration d'une dimension sémantique. Tandis que l'aspect structurel n'a pas été pris en compte, ce qui ne permet pas de faire des analyses sur de différents niveaux structurels.

2.2 Modèles à nouveaux concepts

La complexité des données textuelles et la limite des modèles classiques ont poussés les chercheurs à proposer de nouveaux modèles basés sur de nouveaux concepts. Parmi ces travaux nous citons Khrouf et Dupuy (2001) qui ont proposé un modèle d'entrepôt de documents basé sur le paradigme objet. Leur modèle est basé sur deux concepts : la structure logique générique et la structure logique spécifique. La première décrit les structures logiques communes à un ensemble de documents. Elle regroupe ainsi toute une classe de documents ayant des structures logiques identiques ou similaires. La deuxième correspond à une spécialisation de la structure logique générique, elle est unique et correspond à un et un seul document. Leur processus d'analyse consiste à gérer à partir de l'entrepôt, le schéma du magasin de documents désiré. Ce processus se compose de quatre étapes : (1) le choix du type d'analyse qui permet à l'utilisateur de choisir un type d'analyse, il s'agit de décider de travailler sur des documents ayant des structures similaires ou différentes ou même sur un seul document, (2) la sélection des composants d'analyse, qui consiste à sélectionner les faits, les mesures ainsi que les dimensions d'analyse, (3) le filtrage qui permet à l'utilisateur d'affiner ses analyses en sélectionnant des valeurs précises, (4) la visualisation qui consiste à restituer le schéma du magasin des documents selon une représentation graphique facilitant les analyses multidimensionnelles, Khrouf et Dupuy (2005).

Tounier (2007), a proposé le modèle en Galaxie défini par un nouveau concept **galaxie**. Une galaxie est définie comme étant un regroupement de dimensions liées entre elles par un ou plusieurs noeuds centraux ; chaque noeud modélise les dimensions compatibles pour une même analyse. Son modèle est basé sur la généralisation du concept de constellation de Kimball (1996). Cette approche consiste à décrire un schéma multidimensionnel par l'unique concept de dimension où la notion de fait est supprimée. Afin de permettre l'analyse des documents textes, Tournier a introduit un nouveau concept hiérarchie structurelle de dimension documentaire qui permet de faire des analyses OLAP sur différents niveaux hiérarchiques des documents XML(section, paragraphe,...), il a aussi proposé deux fonctions d'agrégation pour les données textuelles : AVG-KW qui permet de regrouper des mots clefs en des mots clef plus généraux, à travers une ontologie de domaine et TOP-KWk, qui retourne une liste des termes les plus significatifs. Les termes sont pondérés par la méthode $Tf - Idf$, les K termes avec les plus grands poids sont retournés.

Dans leur modèle multidimensionnel d'objets complexes,Boukraa et al. (2011) se sont basés sur le paradigme objet grâce auquel il est possible de représenter les objets de l'univers et de capter la sémantique qu'ils véhiculent, notamment dans les liens avec les autres objets. Ainsi ils modélisent le monde réel par un ensemble d'objets complexes qui décrivent les entités de ce dernier. Le modèle d'objets complexes est un modèle à trois niveaux : le premier niveau est représenté par un diagramme de classe détaillé des faits candidats et des dimensions candidates. Dans le deuxième niveau, les classes décrivant le même objet complexe sont regroupées en un seul *package*, pour fournir à la fin un diagramme de *packages* décrivant des objets complexes. Le troisième niveau est représenté par un diagramme de *packages* qui ré-

sulte de la projection d'un *package* objet complexe du deuxième niveau comme étant un objet fait et de lui associer un ensemble d'objets dimensions décrites par des objets complexes liés à l'objet fait par des relations complexes. Chaque objet complexe dans leur modèle peut être défini grâce à leur opérateur de projection cubique comme étant un axe ou un sujet d'analyse. Donc l'objet fait n'est pas prédéfini au préalable ce qui offre une bonne flexibilité d'analyse. Leur modèle permet aussi une analyse sur de différents niveaux de granularité de chaque objet complexe. La figure suivante nous illustre les trois niveaux présents sur le modèle d'objet complexe, le premier niveau est représenté par la figure c, le deuxième par la figure b et le troisième par la figure a :

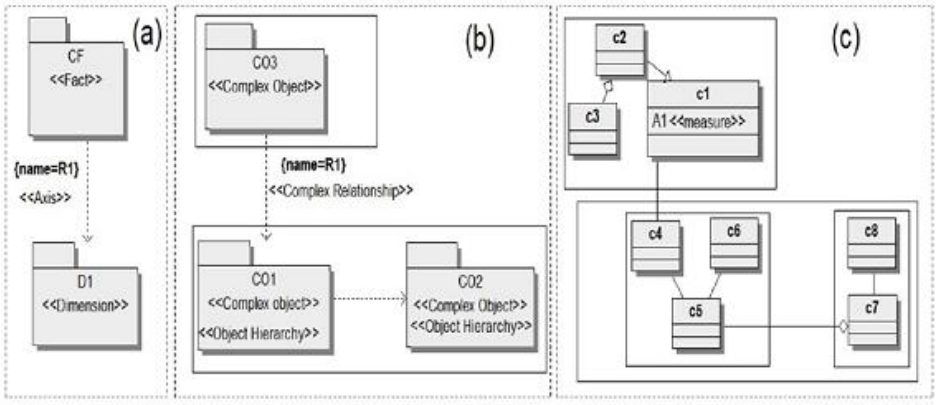


FIG. 3 – *Modèle multidimensionnel d'objets complexe à trois niveaux*
Boukraa (2013).

Les travaux dans cette catégorie ont proposé de nouveaux concepts pour répondre à la complexité des données textuelles. Leurs modèles prennent en compte l'aspect structurel des documents en considérant le document textuel comme étant un ensemble de mots ayant déjà une certaine structure, ce qui permet une analyse sur différents niveaux structurels. La sémantique des données textuelles est prise en compte grâce à l'utilisation d'une ontologie dans certains modèles, toutefois elle reste toujours peu exploitée. l'ensemble de ces modèles ne permettent pas une analyse sur de différents niveaux sémantiques.

3 Étude comparative

La modélisation des données textuelles dans un but d'analyse implique de nombreux problèmes notamment en ce qui concerne la prise en compte de leur structure et de leur sémantique d'une part et la flexibilité d'analyse d'autre part. Aussi les données textuelles comportent des mesures non numériques auxquelles il est nécessaire de définir de nouvelles fonctions d'agrégation. Nous présentons dans cette partie une étude comparative entre les différents modèles cités auparavant. Nous comparons l'ensemble de ces modèles par rapport à la prise en compte des cinq aspects suivants, qui nous ont permis de les étudier :

a. **L'aspect structurel :** la modélisation des données textuelles dans un but d'analyse peut considérer le document texte comme étant une donnée élémentaire. L'objectif consiste alors de structurer et de stocker les documents dans une base de documents textes et de les préparer à l'analyse, sans prendre en compte la structure interne des documents. Toutefois, cette approche de modélisation ne répond pas à toutes les exigences d'un décideur, tel que l'analyser des sections sportives d'un ensemble de journaux. Ce type d'analyse n'est pas supporté par cette approche car la structure interne des documents qui divise le document en plusieurs niveaux hiérarchiques, ce qui permet une analyse sur de différents niveaux de granularité, n'est pas prise en considération. Ainsi nous définissons un modèle qui prend en compte l'aspect structurel des documents, et permettant une analyse multidimensionnelle sur de différents niveaux structurels.

b. **L'aspect sémantique :** l'extraction et la représentation de la sémantique véhiculée dans les données textuelles présentent une problématique déjà traitée dans la littérature dans les domaines d'extraction de connaissances et de la recherche d'information. Tandis que dans les entrepôts de données, la prise en compte de cet aspect important dans la modélisation multidimensionnelle est une nouvelle problématique. Répondre à cette problématique reviens à trouver une manière d'incorporer la sémantique des données textuelles et de la modéliser au sein d'un cube de données.

c. **La flexibilité d'analyse** : dans les systèmes décisionnels classiques un fait représente un sujet d'analyse prédéfini. La définition d'un fait rend la spécification d'analyses peu flexible, car le décideur se voit contraint d'employer ces faits comme sujets, Tounier (2007). La flexibilité d'analyse est apparue comme un nouveau besoin exprimé par les décideurs. Elle réside dans le fait où le sujet d'analyse n'est pas prédéfini au préalable mais choisi au moment de l'analyse. Dans le domaine de l'analyse des données textuelles, nous percevons que le problème de flexibilité est assez complexe. Ainsi nous posons cette problématique autrement, lors d'une analyse textuelle, le contenu sémantique de ces données peut être vu comme étant une mesure d'analyse *(K-top keyword,Topic)*. Comme il peut être considéré comme étant un axe d'analyse. Donc assurer une bonne flexibilité revient à donner à ce contenu sémantique un double rôle.

d. **Mesure textuelle** : la modélisation reposant sur les concepts de fait et de dimension associés à des indicateurs numériques permet des analyses simples de documents textes. Ces analyses reposent principalement sur le comptage de documents. Une bonne analyse de contenu des données textuelles doit prendre en compte les mesures textuelles.

e. **Opérateur OLAP spécifiques au données textuelles** : les opérateurs OLAP appliqués aux données simples ne sont pas adaptés aux données textuelles. Les fonctions d'agrégation numériques telles que $somme$, $moyenne$ s'appliquent bien sur des données numériques, mais ne permettant pas d'agréger les données textuelles. Donc définir de nouveaux opérateurs OLAP s'appliquant sur les données textuelles s'avère nécessaire.

Nous présentons dans le tableau ci-dessous, une étude comparative des modèles présentés dans les sections précédentes.

Modèles d'entrepôts de textes	Familles de modèles		Mesure texte	Opérateurs OLAP		Aspect Sémantique	Aspect structurel	Flexibilité d'analyse
	Modèles extensifs	Modèles à nouveaux concepts		Fonctions d'agrégation	Opérateurs de navigation			
E.documents Khrouf et Dupuy (2001)		X	-	-	-	-	X	Bonne flexibilité
Mire Lee et al. (2002)	X		X	-	Drill down et Roll-up	-	-	Non flexible
DocCube Mothe et al. (2003)	X		-	Score(Dd)	Drill down et Roll-up	X	-	Non flexible
D.cube Tseng et al. (2006)	X		-	Count	Drill down et Roll-up	-	-	Non flexible
Galaxie Tounier (2007)		X	X	AVG-KW, Top-KW	Drill down et Roll-up	X	X	Bonne flexibilité
TextCube Lin et al. (2008)	X		X	-	pull-up et push-down	X	-	Non flexible
TopicCube Zhang et al. (2009)	X		X	-	Drill down et Roll-up	X	-	Non flexible
MMAP-structure Bautista et al. (2010)	X		-	-	-	X	-	Non flexible
M.Cube Zhang et al. (2011)	X		-	-	-	X	-	Non flexible
MMOC Boukraa et al. (2011)		X	X	utilisation du Top-KW	Drill down et Roll-up	-	X	Bonne flexibilité

Tableau comparatif

Malgré que les modèles extensifs ont permis d'effectuer des analyses multidimensionnelles sur les données textuelles, nous constatons qu'ils sont toujours limités et ne traitent que quelques aspects de complexité liés à l'analyse de ce type de données, tel que la sémantique qui a été représentée par une dimension. Les autres aspects comme la prise en compte de la structure des données textuelles ainsi que la flexibilité d'analyse sur ces derniers, restent toujours non traités. De plus, ces modèles ne sont pas génériques et ne permettant pas de représenter n'importe quelle données textuelles. Par contre, les modèles à nouveaux concepts ont permis de traiter d'autres problèmes d'analyse textuelle tel que la prise en compte de la structure ainsi que l'analyse du contenu des documents textes grâce à l'utilisation d'une mesure textuelle . Tandis que la flexibilité d'analyse restent toujours limitée, bien que le sujet d'analyse **Fait** n'est pas prédéfini au préalable dans les deux modèles (modèle en galaxie et à objets complexes). Nous constatons que le problème de flexibilité est assez complexe. Le contenu sémantique des données textuelles peut être vu comme étant une mesure d'analyse *K-top keyword,Topic*, comme il peut être considéré comme étant un axe d'analyse (hiérarchie de thèmes), les travaux actuels ne traitent pas ce double rôle.

4 Conclusion et perspectives

Analyser les données textuelles afin de pouvoir tirer profit des informations qu'elles contiennent est devenu essentiel, à cause de leurs volumes importants et de la quantité d'information qu'elles contiennent. Nous avons présenté dans cet article un survol de modèles dédiés à la représentation multidimensionnelle des données textuelles. Nous avons pu catégoriser l'ensemble de ces modèles selon le type de concept utilisé en deux familles : modèles extensifs basés sur les deux concepts de base des modèles classiques **fait** et **dimension**, et modèles à nouveaux concepts, qui ont proposé de nouveaux modèles basés sur de nouveaux concepts. Nous avons présenté aussi une étude comparative entre ces différents modèles, selon cinq critères de comparaison : la prise en compte de la structure des données textuelles, la prise en compte de leur sémantique, la flexibilité d'analyse, la prise en compte d'une mesure textuelle et la définitions de nouveaux opérateurs OLAP pour les données textuelles.

Ce survol de modèles multidimensionnels dédiés à l'analyse textuelle nous a permis de constater la diversité des modèles proposés et de décerner les problèmes majors liés à l'analyse des données textuelles.

Dans nos travaux, nous rejoignons l'ensemble des travaux proposant de nouveaux concepts. Ainsi nous proposons un nouveau modèle basé sur le paradigme objet qui intègre un nouveau concept **objet contenu sémantique** qui permet de représenter la sémantique des données textuelles et de l'organiser sous forme hiérarchique, pour assurer une analyse sémantique sur différents niveaux de granularité. Dans notre approche de modélisation nous nous intéressons en particulier aux trois aspects suivant : la prise en compte de la structure des données textuelles et de leur sémantique d'une part et la flexibilité d'analyse d'autre part. Dans ce contexte, nous proposons le modèle multidimensionnel sémantique d'objet texte. Celui-ci répond à des exigences de modélisation que les modèles existants ne satisfont pas ou alors partiellement, comme la prise en compte de la structure et de la sémantique des données texte à travers les

deux types de hiérarchies : hiérarchie structurelle et hiérarchie sémantique. Ainsi que la flexibilité d'analyse grâce aux deux relations (relation complexe et relation complexe étendue) sur lesquelles nous avons basé notre opérateur qui permet de projeter notre modèle orienté objet sous forme de modèle multidimensionnel.

Références

Bautista, M., C. Molina, E. Tejeda3, et A. Vila (2010). Using textual dimensions in data warehousing processes. *International Conference, IPMU, Dortmund, Germany*, 158–167.

Bautista, M., M. Prados, M. Vila, et S. Martinez-Folgoso (2006). A knowledge representation for short texts based on frequent itemsets. *Proceedings of IPMU, Paris, France*.

Boukraa, D. (2013). *Complex Object Data Warehouses: Multidimensional Modeling and Vertical Fragmentation*. Thèse de doctorat, Ecole Nationale Supérieure d?Informatique, Algerie.

Boukraa, D., O. Boussaid, F. Bentayeb, et D. Zegour (2011). Modèle multidimensionnel d'objets complexes. du modèle d'objets aux cubes d'objets complexes. *Ingénierie des Systèmes d'Information 16*.

Khrouf, K. et C. Dupuy (2001). Conception d'entrepôts de documents décisionnels. *INFORSID*, 387–401.

Khrouf, K. et C. Dupuy (2005). Docware : Vers l'entreposage et l'analyse multidimensionnelle de documents. *CORIA*, 405–420.

Kimball, R. (1996). *The data warehouse toolkit: Practical Techniques for Building Dimensional Data Warehouses,*. John Wiley and Sons.

Lee, J., O. Grossman, D .and Frieder, et McCabe (2002). Mire: A multidimensional information retrieval engine for structured data and text. *Proceedings of the International Conference on Information Technology: Coding and Computing, Las Vegas, NV, USA 14*, 224–229.

Lin, C. X., B. Ding, J. Han, F. Zhu, et B. Zhao (2008). Text cube: Computing ir measures for multidimensional text database analysis. *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining 54*, 905–910.

Mothe, J., B. Chrisment, C.and Dousset, et J. Alaux (2003). Doccube: Multi-dimensional visualisation and exploration of large document sets. *Journal of the American Society for Information Science and Technology, 54*, 650–659.

Tounier, R. (2007). *Analyse en ligne (OLAP) de documents*. Thèse de doctorat, Université Toulouse III . Paul Sabatier.

Tseng, A., S. Frank, Y. Annie, et Chou (2006). The concept of document warehousing for multi-dimensional modeling of textual-based business intelligence. *Decision Support 42*, 727–744.

Zhang, D., C. Zhai, et J. Han (2009). Topic cube: Topic modeling for olap on multidimensional text databases. *SDM '09: Proceedings of the 2009 SIAM International Conference on Data Mining, Sparks, NV, USA*, 1124–1135.

Zhang, D., C. Zhai, et J. Han (2011). Mitexcube:microtextcluster cube for online analysis of text cells. *The NASA Conference on Intelligent Data Understanding (CIDU)*, 204–218.

Summary

The modeling process for text data type is to give a special representation for this kind of non structural data. The given representation offers a formal description for text data so that the information that the text comprehends would be put to use. In our study of the different papers that dealt with multidimensional design for text data, we were able to identified two different classes for the models according to the concept used: (i)extensive models, based on the basic concepts of classical models of data warehouses, and (ii) models with new concepts offering a new way of perceiving the complexity of textual data. Our evaluation for the set of these works is based upon their treatment level for the following aspects: text data structure, semantic support, analytic flexibility, textual measure support and operator olap for textual data support.

Modèle multidimensionnel en diamant dédié à l'OLAP sémantique de documents

Maha Azabou*, Kaïs Khrouf*, Jamel Feki*, Chantal Soulé-Dupuy**, Nathalie Vallès**

* Laboratoire MIR@CL, Université de Sfax
Route de l'Aérodrome km 4, B.P. 1088, 3018 Sfax, Tunisie
Azabou.Maha@yahoo.fr, Khrouf.Kais@isecs.rnu.tn, Jamel.Feki@fsegs.rnu.tn

** IRIT, Université Toulouse 1 Capitole,
2 rue du doyen Gabriel Marty, 31042 Toulouse Cedex 9, France
{Chantal.Soule-Dupuy, Nathalie.Valles-Parlangeau}@ut-capitole.fr

Résumé.
Le document électronique représente aujourd'hui un support d'information que les entreprises ne peuvent plus négliger si elles veulent être certaines d'identifier et de gérer toutes les données qui leur sont utiles au quotidien. Plusieurs travaux ont proposé l'application des techniques OLAP (« On-line Analytical Processing ») aux informations documentaires. Dans cet article, nous présentons un nouveau modèle multidimensionnel dédié à l'OLAP de documents. Ce modèle, dit *en diamant*, est organisé autour d'une dimension centrale qui traduit la *sémantique* du contenu textuel du document.

1 Introduction

Les données issues des processus métiers d'une organisation constituent un capital précieux pour la prise de décisions. La technologie des entrepôts de données a largement contribué à tirer profit de ces données opérationnelles via une réorganisation multidimensionnelle adaptée aux décideurs. Cependant, outre les données factuelles dont la modélisation et l'exploitation est maîtrisée par les outils d'entreposage de données (« data warehousing »), de nombreux travaux de recherche considèrent que les documents constituent également une source très riche de données textuelles dont l'utilité n'est pas moins importante que celle des données factuelles. Ainsi l'architecture d'un entrepôt de données classique s'avère inadaptée pour le stockage et la manipulation des documents.

Depuis plus d'une décennie, certains chercheurs recommandent d'entreposer les documents (McCabe et al., 2000) (Sullivan, 2001). Un entrepôt de documents doit permettre le stockage de documents hétérogènes, sélectionnés et filtrés, ainsi que leur classification structurelle et sémantique à des fins d'analyses multidimensionnelles. La préparation des données textuelles pour ce type d'analyse passe par une phase obligée de modélisation

multidimensionnelle spécifique basée sur les concepts multidimensionnels (*fait, dimension et hiérarchie*). Bon nombre de chercheurs ont constaté que les modèles multidimensionnels classiques (tels que les modèles en étoile) exprimant une série d'observations par le biais d'indicateurs purement numériques et d'axes d'analyses ne sont pas adaptés pour les traitements analytiques en ligne (OLAP :« On-line Analytical Processing ») des données textuelles issues des documents. C'est pour apporter une contribution à cette problématique que nous proposons un modèle multidimensionnel spécifique aux documents ; il permet d'exprimer l'aspect sémantique du contenu textuel.

Cet article est structuré comme suit. La section 2 étudie les travaux abordant l'exploitation OLAP des documents. La section 3 présente le *modèle en diamant* proposé avec ses différents composants. Enfin, dans la section 4, nous décrivons une approche pour la génération semi-automatique de schémas multidimensionnels en diamant.

2 Etat de l'art

Plusieurs travaux se sont intéressés à l'exploitation OLAP des documents ; la majorité de ces travaux adoptent le schéma en étoile initialement proposé dans le contexte des entrepôts de données. Par exemple, les auteurs de (Mothe et al., 2003) ont proposé une analyse multidimensionnelle avec l'environnement DocCube qui représente un fond documentaire avec des nuages de sphères où chaque sphère correspond à un ensemble de documents. (Khrouf, 2004) a défini une approche d'analyse multidimensionnelle des données documentaires ; néanmoins; cette approche est basée sur la structure logique des documents. Les auteurs de (Tseng et Chou, 2006) proposent d'analyser des documents (emails, articles, pages Web…) selon des dimensions construites à partir des métadonnées définies par le Dublin Core (Dublin Core, 2007).

Dans (Boussaid et al., 2006), les auteurs ont proposé une modélisation en flocon de neige des données multidimensionnelles XML avec des méthodes de fouille de données. Plus précisément, ils ont défini une approche appelée X-Warehousing qui permet de concevoir un entrepôt, de représenter son schéma conceptuel à l'aide de schémas XML et enfin d'alimenter la structure multidimensionnelle à l'aide de données initialement stockées dans des documents XML. Ce modèle implique beaucoup de redondance dans les données des dimensions puisque pour chaque mesure du fait, il faut indiquer les valeurs des dimensions correspondantes. Une telle redondance peut impliquer des difficultés de maintenance.

D'autres travaux ont proposé et utilisé le schéma en galaxie (Ravat et al., 2008) (Tournier, 2007) qui repose sur un seul concept, celui de *dimension*. Une fonction permet d'agréger des données textuelles afin d'obtenir une vision synthétique des informations issues de documents. Les travaux de (Ben Messaoud et al., 2012) proposent une démarche pour l'unification des structures des documents à des fins d'analyses multidimensionnelles, sans proposer de fonctionnalités spécifiques pour une analyse OLAP intégrant des aspects sémantiques.

En résumé des travaux relatifs à la modélisation multidimensionnelle des documents, nous avons noté que trois types de modèles multidimensionnels ont été utilisés : le modèle en étoile, le modèle en flocon de neige et le modèle en galaxie. Cependant, tous ces modèles ne tiennent pas compte de l'aspect sémantique des données textuelles. Pour pallier cet

inconvénient, certains auteurs ont proposé des approches ou fonctions pour l'analyse du contenu textuel (Tseng et Chou, 2006) (Ravat et al., 2008). L'objet de ce papier est alors de proposer un modèle multidimensionnel dédié à l'OLAP de documents et qui englobe les dimensions de données factuelles (comme *date, auteur, éditeur*), ainsi que la sémantique des données textuelles (comme *résumé, contenu, paragraphe*).

3 Modèle en diamant

La modélisation multidimensionnelle classique vise à représenter les données en fonction de l'analyse prévue par les décideurs. Elle représente l'information à analyser, comme un point dans un espace à plusieurs dimensions appelées axes d'analyses.

Dans cet article, nous proposons un nouveau modèle multidimensionnel dédié à l'OLAP de documents, que nous appelons *Modèle en diamant*. Ce modèle peut être généré à partir des structures logiques décrivant une collection de documents, à travers :

- les DTD ou Xschema des documents.
- les structures génériques existantes dans un entrepôt de documents : une structure générique est une structure commune à un ensemble de documents ; elle est obtenue par classification structurelle d'un ensemble de structures spécifiques (Khrouf et al., 2012).

Ce nouveau modèle en diamant se compose de :

- un fait qui correspond à une observation non forcément numérique sur les documents (par exemple, liste de parties de documents ou de documents se rapportant à un thème d'analyse, nombre de documents, …) ;
- un ensemble de dimensions : Une dimension *Sémantique*, une dimension *Version* et des dimensions *Classiques* :
    - la dimension *Sémantique* occupe un emplacement central ; elle se compose de la hiérarchie suivante : Concept → Ontologie. Le paramètre *Concept* sera relié aux éléments textuels (comme *Section*, *Paragraphe*) des documents afin d'utiliser ces concepts lors de l'analyse multidimensionnelle de ces éléments textuels. L'affectation de concepts au contenu textuel des documents se base sur un calcul de degré de similarité entre contenus textuels et concepts d'une ontologie comme présenté dans (Ben Mefteh et al., 2012).
    - la dimension *Version* de document concerne les différentes versions des documents, ainsi que les métadonnées associées, comme par exemple: *Propriété*, *Date de création* et *Description*. Il est à noter que chaque version de document sera reliée à une ontologie de domaine (Ben Mefteh et al., 2012).
    - les dimensions *Classiques* sont les axes d'analyse constitués des éléments du premier niveau de la structure logique de documents. Pour chaque élément, ses descendants constituent les paramètres (organisés sous forme de hiérarchies) et les attributs faibles. La détermination des paramètres, hiérarchies et attributs faibles fera l'objet de la section suivante.

Modèle en diamant

Soit la structure logique *Article* de la Figure 1, le modèle en diamant correspondant est montré dans la Figure 2.

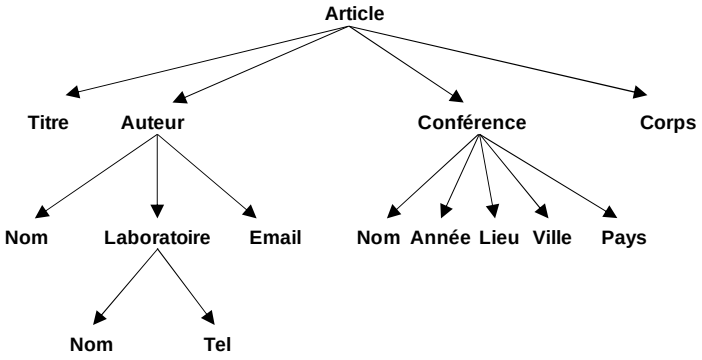


FIG. 1 – *Structure logique Article*

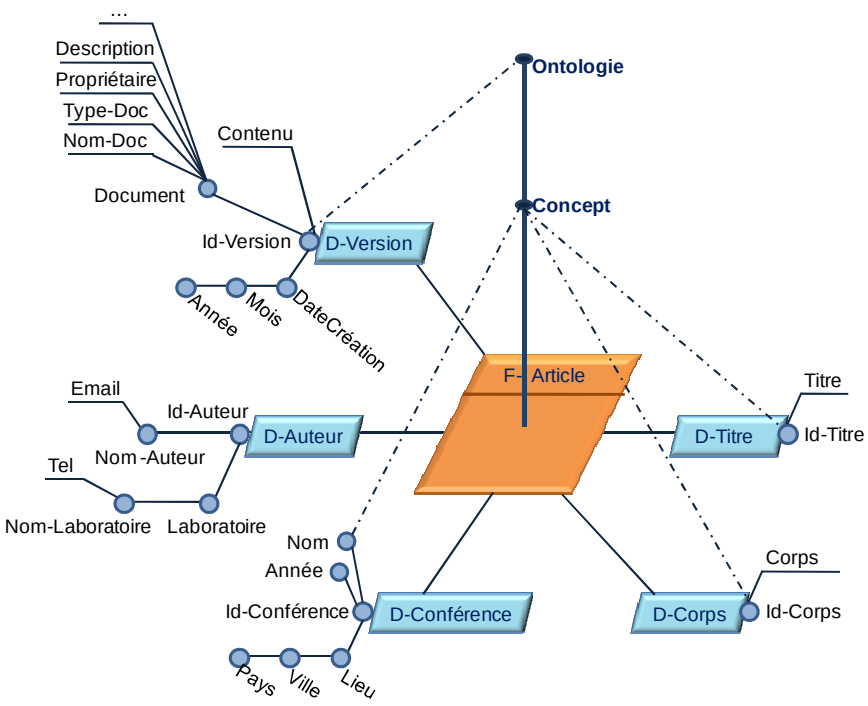


FIG. 2 – *Schéma en diamant obtenu pour la structure de la Figure 1*

4 Génération semi-automatique de modèles en diamant

Dans cette section, nous détaillons l'ensemble de règles que nous avons définies afin de générer un modèle en diamant à partir d'une structure logique décrivant une collection de documents. Ce modèle en diamant est composé d'un fait central, d'une dimension sémantique, d'une dimension version et de dimensions classiques.

Dans le reste de cet article, nous utilisons la structure logique *Article* de la Figure 1 pour illustrer les règles définies pour la génération d'un modèle en diamant.

- **Identification du fait**

Règle 1 : Le sommet *A* de toute structure logique contenant au moins une feuille constitue un fait *F-A.* (Hachaichi et al., 2010).

L'application de la *Règle 1* identifie le nœud *Article* comme un *fait* puisque la structure logique (cf. Figure 1), dont le sommet est *Article*, contient des descendants, ce fait est conventionnellement nommé *F-Article*.

- **Identification des dimensions classiques**

Règle 2 : Chaque nœud *d* descendant immédiat de la racine de la structure logique devient une dimension *D*.

Règle 3 : Chaque dimension *D* se verra affecter un identifiant *Id-D* (Ben Messaoud et al., 2011).

L'application de ces deux règles sur la structure logique *Article* identifie les quatre dimensions nommées *D-Titre*, *D-Auteur*, *D-Conférence* et *D-Corps avec les identifiants correspondants : Id-Titre*, *Id-Auteur*, *Id-Conférence* et *Id-Corps.*

Règle 4 : Chaque nœud, transformé en une dimension *D,* ne contenant pas de descendants aura un attribut faible ayant comme nom celui de la dimension *D* et directement relié à l'identifiant de *D* (*Id-D*).

Conformément à cette règle, la dimension *D-Titre* aura alors l'attribut faible *Titre* relié à l'identifiant *Id-Titre* (Idem pour *D-Corps*).

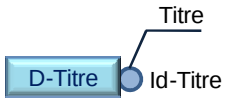


FIG. 3 – *Dimension D-Titre*

Pour les nœuds transformés en dimensions et possédant des descendants, nous définissons un ensemble de règles afin de dégager, à partir de ses descendants, les paramètres (organisés sous forme de hiérarchies) et leurs attributs faibles.

Modèle en diamant

- **Identification des hiérarchies**

Pour l'identification des hiérarchies, nous devons consulter le contenu des documents afin de déterminer les *Dépendances Fonctionnelles* (DF[1]) entre les éléments de la structure logique.

Soient *X* et *Y* deux descendants immédiats d'un nœud-dimension *D* (nous appelons nœud-dimension un nœud transformé en une dimension).

Règle 5 : S'il existe une DF de *X* vers *Y* (notée X → Y) non symétrique (*i.e.*, sans avoir *Y*→X) alors *X* et *Y* constituent deux paramètres consécutifs de *D, i.e.*, de rang *i* et *i+1* respectivement (en allant du plus fin vers le moins fin).

Règle 6 : Eliminer les DF transitives. Soient les DF suivantes : *X*→*Y*, *Y*→*Z* et *X*→*Z* ; alors la DF *X*→*Z* est transitive et doit être éliminée. Nous obtenons alors la hiérarchie suivante : *X*→*Y*→*Z.*

Règle 7 : S'il n'y a aucune DF entre *X* et les autres descendants du nœud-dimension *D*, alors *X* constitue un paramètre de la dimension *D* de rang 2 (*X* relié à l'identifiant *Id-D* de *D*).

Reprenons le sous-arbre *Conférence* de la structure *Article* (cf. Figure 1) avec l'échantillon de contenu indiqué dans la Figure 4.

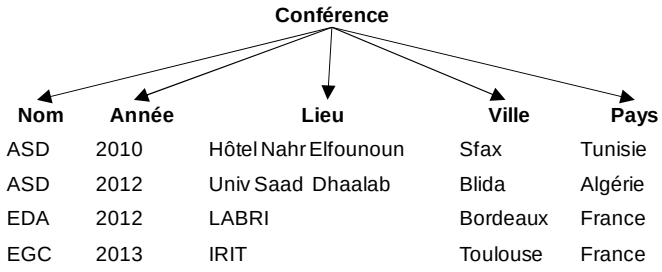


FIG. 4 – *Sous arbre Conférence*

Nous constatons à partir de ce contenu :

- L'existence des trois DF non symétriques (*Règle 5*) *Lieu*→*Ville*, *Lieu*→*Pays* et *Ville*→*Pays*. L'application de la *Règle 6* nous permet de déterminer la hiérarchie suivante : *Lieu*→*Ville*→*Pays.*

- Aucune DF entre *Nom* et les autres descendants de *Conférence* (*Année*, *Lieu*, *Ville* et *Pays*), *Nom* est alors un paramètre de rang 2 relié à *Id-Conférence* selon la *Règle 7*. (Idem pour *Année*).

[1] Une Dépendance Fonctionnelle (DF) de l'attribut *X* vers l'attribut *Y*, notée *X*→*Y*, exprime qu'à une valeur de *X* est associée une seule valeur de *Y* ; l'inverse d'une DF n'est pas forcément une DF.

La dimension *D-Conférence* sera définie de la manière suivante :

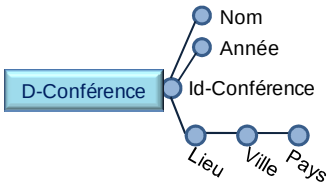


FIG. 5 – *Dimension D-Conférence*

- **Identification des attributs faibles**

Certains paramètres de dimension peuvent être accompagnés de descripteurs appelés attributs faibles.

> *Règle 8* : Etant donnés deux attributs *X* et *Y*, s'il existe deux dépendances fonctionnelles symétriques $X \rightarrow Y$ et $Y \rightarrow X$ et si *Y* n'a pas de descendants, alors *Y* constitue un attribut faible pour *X*.

Soit le sous arbre *Auteur* (cf. Figure 6) issu de la structure *Article* de la Figure 1.

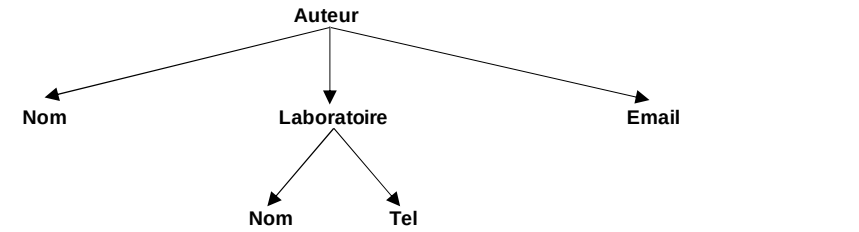


FIG. 6 – *Sous arbre Auteur*

Concernant la *Règle 8*, on remarque l'existence des deux dépendances fonctionnelles symétriques *Nom*$\rightarrow$*Email* et *Email*$\rightarrow$*Nom, alors Email* peut jouer le rôle d'un attribut faible pour le paramètre *Nom*.

Remarque : Soient *X* et *Y* deux descendants de la structure logique tels que $X \rightarrow Y$ et $Y \rightarrow X$, nous considérons arbitrairement que l'un d'entre eux est un attribut faible de l'autre attribut assimilé lui à un paramètre. C'est le concepteur qui vérifiera et validera ce choix par rapport à la sémantique des deux descendants.

Il convient d'itérer les règles 5, 6, 7 et 8 pour chaque élément ayant des descendants afin de déterminer tous ses paramètres à tous les niveaux et leur(s) attribut(s) faible(s).

Le paramètre *Laboratoire* possède des descendants (*Nom* et *Téléphone*) ; il s'agit d'un nouveau niveau. Selon la règle 8, il existe deux DF symétriques *Nom→Téléphone* et *Téléphone→Nom*, alors nous choisissons arbitrairement *Téléphone* comme attribut faible pour le paramètre *Nom*.

La dimension *D-Auteur* sera définie de la manière suivante :

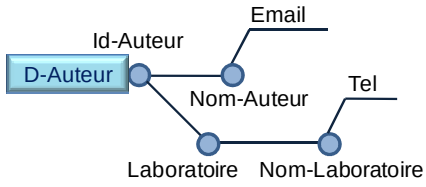


FIG. 7 – *Dimension D-Auteur*

Remarque : Si une dimension contient des paramètres ayant la même appellation (Exemple *Nom* dans *D-Auteur*), nous modifions le nom du paramètre par l'ajout de l'intitulé de son élément-père. Ainsi, dans notre exemple, nous aurons : Nom-Auteur et Nom-Laboratoire.

- **Ajout de la dimension *Version***

A ce niveau, nous ajoutons la dimension *D-Version* composée d'un identifiant *Id-Version*, de son *Contenu* et de deux hiérarchies : *i)* une hiérarchie temporelle se rapportant à la date de création du document organisée comme suit *DateCréation→Mois→Année* ; *ii)* une hiérarchie décrivant un ensemble de métadonnées (*Nom* physique du document, *Type* d'extension, *Propriétaire* créateur du document, *Description* sommaire du document, etc.).

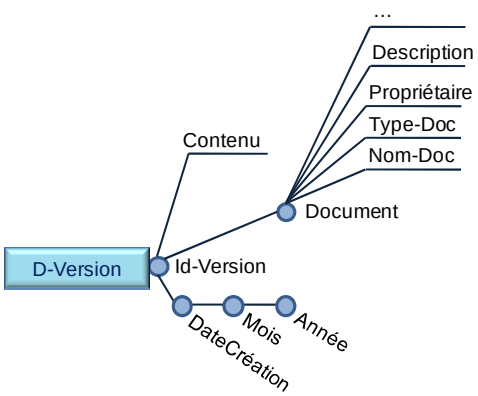


FIG. 8 – *Dimension D-Version*

- **Ajout de la dimension sémantique**

Les données textuelles véhiculent une sémantique. Elle est exprimée, dans le modèle en diamant, à travers la *dimension sémantique* faisant référence à un ensemble d'ontologies et de concepts décrivant les documents associés à la structure logique correspondante.

La détermination de la dimension sémantique se fait conformément à l'approche décrite dans (Ben Mefteh et al., 2012) et selon les étapes suivantes :

- Extraction des termes. Il s'agit d'extraire les mots-clés significatifs des éléments feuilles du document (fragments textuels associés aux éléments feuilles) selon un processus d'indexation classique tel que défini en recherche d'information (Baeza-Yates et al., 1999).
- Choix de l'ontologie. L'objectif est de déterminer, parmi un ensemble d'ontologies de domaines, celle qui convient le mieux pour décrire la sémantique du document, et ce à partir des mots-clés du langage d'indexation générés lors de l'étape précédente.
- Association de concepts aux éléments feuilles. Consiste à rechercher, dans l'ontologie de domaine retenue à la phase précédente, le concept le plus approprié à la description de la sémantique de l'élément feuille compte tenu des mots-clés qui le décrivent.
- Inférence de concepts aux éléments non-feuilles. Les concepts feuilles d'un nœud servent ensuite à inférer un concept à ce nœud à partir de l'ontologie sélectionnée.

La représentation de la dimension sémantique dans le modèle en diamant se fait de la manière suivante :

- relier le paramètre *Version* au paramètre *Ontologie* de la dimension sémantique puisque dans nos travaux, nous associons toute version d'un document de l'entrepôt à une ontologie de domaine.
- relier les paramètres textuels des dimensions classiques, « porteurs » d'un point de vue sémantique, au paramètre *Concept* de la dimension sémantique. Ces paramètres sont les éléments correspondants à la structure logique et dont les contenus sont reliés aux concepts des différentes ontologies (Ben Mefteh et al., 2012). A titre d'exemple, l'attribut *Nom* de la dimension *D-Conférence* sera relié au paramètre *Concept*, alors que l'attribut *Nom-Auteur* de la dimension *D-Auteur* ne sera pas relié. Ainsi, nous pouvons faire l'analyse soit par le nom de la conférence (ASD, par exemple), soit par le ou les concepts associés (Système décisionnel, par exemple) ; dans ce cas, les autres conférences appartenant aussi à ce concept (e.g., EDA) seront intégrées dans l'analyse.

En utilisant ces règles, le modèle en diamant obtenu à partir de la structure *Article* est celui présenté dans la Figure 2.

Notons que toutes les règles proposées ci-dessus sont automatisables. Une fois que la construction du modèle en diamant est terminée, c'est le concepteur (assisté du décideur) qui vérifie et valide le modèle en diamant obtenu. Il peut renommer, supprimer les éléments multidimensionnels et les liens entre les dimensions, réorganiser les paramètres d'une hiérarchie si nécessaire, …

Exemple : Supposons que le nom d'auteur *Olivier Dupond* est relié au concept *Oliviers* de l'ontologie *Agriculture*, le système relie alors le paramètre *Nom-Auteur* de la dimension *D-Auteur* au paramètre *Concept* de la dimension sémantique. Le concepteur doit intervenir pour supprimer ce lien.

5 Conclusion

Cet article présente un nouveau modèle multidimensionnel dédié à l'OLAP de documents, à savoir le *modèle en diamant*. Ce modèle se compose d'un fait, de dimensions *classiques* (construites à partir des structures génériques de l'entrepôt, DTD ou Xschema), la dimension version de document et la dimension sémantique. Cette dimension sémantique a pour rôle de passer du simple texte à un niveau plus sémantique, que sont les concepts associés. Nous avons défini aussi dans cet article un ensemble de règles en vue de la construction semi-automatique de modèles en diamant.

Plusieurs perspectives sont envisageables. Un prototype logiciel est en cours de développement supportant les différentes règles présentées dans cet article et la visualisation du modèle en diamant en 3D. En l'absence de Benchmarck de tests, nous comptons valider ces règles sur des exemples pris de la littérature, qui construisent des schémas multidimensionnels de documents. Il serait intéressant aussi de définir un processus d'instanciation de ces modèles en diamant à partir du contenu des documents et la visualisation des résultats sous forme de cubes ou tables multidimensionnelles.

Références

Baeza-Yates R. et Ribero-Neto B. (1999), *Modern Information Retrieval*, Addison Wesley, 1999.

Ben Mefteh S., Khrouf K., Feki J., Ben Kraiem M. et Soulé-Dupuy C. (2012). *Une approche d'extraction automatique de structures sémantiques de documents XML*, INFormatique des Organisations et Systèmes d'Information et de Décision (INFORSID 2012), p. 523-538, Montpellier, France.

Ben Messaoud I., Feki J. et Zurfluh G. (2011). *Modélisation multidimensionnelle des documents XML*. Journées francophones sur l'Entrepôts de Données et l'Analyse en ligne (EDA 2011), p. 55-70, Clermont-Ferrand, France.

Ben Messaoud I., Feki J. et Zurfluh G. (2012). *A First Step for Building a Document Warehouse: Unification of XML Documents* (S). International Conference on Research Challenges in Information Science (RCIS), Valencia, Spain, 2012.

Boussaid O., Ben Messaoud R., Choquet R. et Anthoard S. (2006). *Conception et construction d'entrepôts XML*. Journée francophone sur les Entrepôts de Données et l'Analyse en ligne (EDA'2006), p. 3–21, Versailles, France.

Dublin Core Metadata Initiative (DCMI) (2007): Dublin Core Metadata Element Set, Version 1.1, ISO Standard 15836, from http://dublincore.org/documents/dces/.

Hachaichi, Y., J. Feki, et H. Ben-Abdallah (2010). *Modélisation multidimensionnelle de documents XML centrés-données.* Journal of Decision Systems, vol 19/3, p. 313-345.

Khrouf K. et Soulé-Dupuy C. (2004). A *Textual Warehouse Approach: a Web Data Repository*. Chapter VII, Intelligent Agents for Data Mining and Information Retrieval, Idea Group Publishing, p. 101-124, 2004.

Khrouf K., Azabou M., Feki J. et Soulé-Dupuy C. (2012). *Towards a Multi-user Document Warehouse*. International Conference on Web Information Systems and Technologies, p. 149-154, Porto, Portugal.

McCabe, M. C., J. Lee, A. Chowdhury, D. Grossman et O. Frieder (2000). *On the design and evaluation of a multi-dimensional approach to information retrieval*. Annual International ACM SIGIR Conference, p. 363–365, Athens, Greece.

Mothe J., Chrisment C., Dousset B. et Alau J. (2003). *DocCube: Multidimensional visualization and exploration of large document sets*. Journal of the American Society for Information Science and Technology (JASIST), vol.54(7), Wiley Periodicals, p. 650–659.

Sullivan D. (2001). *Document Warehousing and Text Mining.* Wiley John & Sons, ISBN: 0471399590, 2001.

Ravat F., Teste O., Tournier R. et Zurfluh G. (2008). *Top_Keyword: an Aggregation Function for Textual Document OLAP*. International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2008), p. 55-64, Turin, Italy.

Tournier R. (2007). Analyse en ligne (OLAP) de documents. Thèse de doctorat, Université Paul Sabatier, Toulouse III, Décembre 2007.

Tseng F.S.C. et Chou A.Y.H. (2006). *The concept of document warehousing for multi dimensional modeling of textual-based business intelligence*. Decision Support Systems (DSS), Elsevier, p. 727-744.

Summary

Today, the electronic document represents an important support for information, therefore organizations cannot neglect these documents if they want to identify and manage all data useful for the decision support system. Several works have addressed how to apply On-line Analytical Processing (OLAP) techniques on documents. In this paper, we present a new multidimensional model called *diamond model;* it is for documents modeling and OLAP analyses. This model takes into account the semantics of the textual content of documents by means of a new central *semantic* dimension.

L'Analyse en Composantes Principales Normée : Une Nouvelle Approche pour la Fragmentation des Entrepôts de Données

Rachid Elmansouri*, Elhoussaine Ziyati**, Omar Elbeqqali*

*LIIAN / GRMS2I
Université Sidi Mohamed Ben Abdellah, FSDM
Fes, Maroc
elmansouri_rachid@yahoo.fr, omarelbeqqali@gmail.com
**GSCM_LRIT Université Mohammed V-Agdal
Rabat, Maroc
ziyatie@yahoo.fr

Résumé. Dans ce papier, nous présentons un état de l'art sur l'Analyse en Composantes Principales (ACP) et la possibilité de son utilisation pour la fragmentation horizontale et verticale des entrepôts de données (ED) dans le but de réduire le temps d'exécution des requêtes OLAP. Nous procédons à la projection des individus sur le premier plan principal et sur l'espace vectoriel à 3D engendré par les trois premières composantes principales et nous essayons de déterminer graphiquement des groupes homogènes d'individus et donc un schéma de fragmentation horizontale pour la table de données étudiée.
L'étude des corrélations entre les variables initiales et les composantes principales nous permettra de tracer le cercle des corrélations et déterminer graphiquement, sous certaines conditions, des variables candidates à être rassemblées dans des fragments verticaux.
Pour satisfaire un maximum de requêtes décisionnelles, notre approche est indépendante de tout ensemble de requêtes et cherche à exploiter les représentations graphiques fournies par l'ACP.
Nous terminons notre étude par une expérimentation sur un entrepôt de données qui montre l'intérêt et l'originalité de notre approche.

1 Introduction

Le concept d'entrepôt de données a été formalisé en 1990 par Bill Inmon comme « une base de données orientée sujet, intégrée et contenant des informations historisées, non volatiles destinées aux processus d'aide à la décision » Inmon (1996). Actuellement et grâce à l'évolution de l'informatique décisionnelle, les entreprises modernes exploitent de grands volumes de données issues de différentes sources hétérogènes et leurs volumes sont destinés à augmenter sans cesse. L'entreposage des données constitue un axe vital autour duquel tourne toute la stratégie commerciale et marketing. Les requêtes décisionnelles exécutées sur les ED sont très complexes. Elles contiennent des jointures, des sélections et des agrégations. Devant la complexité et un temps de réponse long (des heures sinon des jours) de ce type de requêtes, la tâche de l'administrateur des ED devient très difficile. En effet, l'administrateur requiert une bonne connaissance des structures d'optimisation (index, vues matérialisées,

fragmentation, etc.) et des méthodes de conception logique et physique afin de choisir une bonne politique de conception et d'optimisation Bouchakri et al. (2009).

Notre article s'articule de la manière suivante : La section 2 présente un aperçu sur la technique de fragmentation des entrepôts de données. La section 3 décrit brièvement l'ACP. La section 4 est consacrée à une étude expérimentale qui valide notre approche. Enfin la section 5 conclut le papier en récapitulant les résultats principaux et en suggérant des travaux futurs.

2 Aperçu sur la fragmentation des entrepôts de données

La fragmentation est une technique d'optimisation qui permet de décomposer les tables d'un entrepôt de données en plusieurs partitions Noaman (1999), Sanjay et al. (2004). Elle peut être combinée avec d'autres techniques d'optimisation comme les vues matérialisées Gupta (1999), les index Chaudhuri (2004) et le traitement parallèle Molina (1998).

Ces techniques peuvent être classées en deux catégories : des techniques redondantes comme les vues matérialisées et les index du fait qu'elles nécessitent un espace de stockage et un coût de rafraîchissement et des techniques non redondantes comme la fragmentation ne nécessitant pas de coût de stockage ni de rafraîchissement. Deux types de fragmentation peuvent être utilisés dans le contexte des entrepôts de données : la fragmentation horizontale et la fragmentation verticale.

La fragmentation horizontale se décline en deux versions Ozsu et Valduriez (1999): les fragmentations primaire et dérivée. La fragmentation primaire d'une relation R est effectuée grâce à des prédicats de sélection définis sur la relation. La fragmentation horizontale dérivée consiste à fragmenter une relation selon le schéma de fragmentation d'une autre relation. La fragmentation horizontale primaire favorise le traitement des requêtes de restriction portant sur les attributs utilisés dans le processus de la fragmentation. La fragmentation dérivée est utile pour le traitement des requêtes de jointure.

Dans ce papier, nous nous intéressons à la fragmentation mixte d'un entrepôt de données en se basant sur l'ACP : La fragmentation horizontale consiste à décomposer une relation R selon les lignes, la fragmentation verticale selon les colonnes. La fragmentation verticale favorise naturellement le traitement des requêtes de projection en limitant le nombre de fragments à accéder. La reconstruction de la relation R à partir de ces fragments verticaux est obtenue par l'opération de jointure de ces fragments. La fragmentation horizontale consiste à diviser une relation R en sous ensembles de n-uplets appelés fragments horizontaux, chacun étant défini par une opération de restriction appliquée à la relation. Les n-uplets de chaque fragment horizontal satisfont une clause de prédicats. La reconstruction de la relation R à partir de ces fragments horizontaux est obtenue par l'opération d'union de ces fragments.

3 Description sommaire de l'ACP

L'Analyse en Composantes Principales fait partie du groupe des méthodes descriptives multidimensionnelles appelées méthodes factorielles. Ces méthodes qui sont apparues au début des années 30 ont été surtout développées en France dans les années 60, en particulier par Jean-Paul Benzécri qui a beaucoup exploité les aspects géométriques et les représentations graphiques Duby et Robin (2006).

L'idée à la base de l'ACP est de pouvoir expliquer la variance observée dans la masse de données initiales en se limitant à un nombre réduit de composantes, définies comme étant des transformations mathématiques pures et simples des variables initiales. L'algorithme utilisé pour la détermination de ces composantes obéit à deux contraintes importantes : d'abord, les composantes extraites doivent capturer chacune une proportion de variance de moins en moins importante. Ensuite, les composantes doivent avoir des corrélations linéaires nulles (condition d'orthogonalité) Bouchier (2006).

On peut discerner différentes conséquences de cette analyse. Tout d'abord, la proportion de variance totale cumulée à travers les différentes composantes pourra éventuellement atteindre 100% si le nombre de composantes extraites équivaut au nombre de variables initiales. Mais sachant que l'objectif principal de l'ACP est de réduire la masse de données, Il s'avère donc inapproprié d'extraire le même nombre de composantes que de variables initiales. En d'autres termes, on devra prendre une décision judicieuse quant au nombre de composantes principales à extraire. Les critères d'extraction de ces composantes sont traités dans le paragraphe 3.2.

Dans des travaux précédents Elmansouri et al. (2012), nous avons étudié l'ACP simple, pour laquelle, tous les individus ont le même poids dans l'analyse et toutes les variables sont traitées de façon symétrique. Cela pose parfois des problèmes. Le premier reproche fait par des praticiens est que, si les anciennes variables sont hétérogènes, il est difficile de donner un sens aux composantes principales qui sont alors des combinaisons linéaires de variables hétérogènes. Le deuxième reproche est que, si on change d'unités sur ces variables, on peut changer complètement les résultats de l'ACP. Le dernier reproche vient du fait qu'une variable contribuera d'autant plus à la confection des premiers axes, que sa variance est forte Duby et Robin (2006).

Pour toutes ces raisons, nous avons décidé d'opter dans ce travail pour une ACP sur des variables centrées réduites, on parle d'ACP normée.

3.1 Etude des matrices de corrélation

Dans l'ACP on ne s'intéresse pas particulièrement aux variables individuelles, souvent très nombreuses, mais on s'intéresse à la présence d'intercorrélations entre ces variables pour essayer d'en extraire des dimensions plus globales.

Le nombre des coefficients de corrélation augmente rapidement et il est égal à $p(p-1)/2$. Si tous les coefficients de corrélation sont assez faibles, il n'y aurait absolument aucun intérêt de procéder à une ACP de ces données.

L'ACP s'accommode assez bien aux situations où un certain niveau de multi colinéarité existe entre les données. Cependant, il faut absolument se méfier de la condition dite de singularité où une variable serait parfaitement corrélée avec une autre variable ou avec une combinaison de plusieurs variables. Cette condition peut être détectée en calculant le déterminant de la matrice de corrélation. En effet, un déterminant égal à 0 indique que la matrice est singulière c'est-à-dire qu'il existe au moins un cas de dépendance linéaire dans la matrice ou, en d'autres termes, qu'une variable peut être entièrement expliquée ou prédite par une combinaison linéaire d'autres variables. À l'inverse, un déterminant égal à 1 correspond lui aussi à une condition impropre à l'ACP; il indique que la matrice de corrélation est une matrice d'identité.

Lorsque nous sommes en présence d'une variable qui n'est en corrélation avec aucune autre dans la matrice, il est recommandé de retrancher cette variable avant de procéder à une ACP.

3.2 Etude des critères d'extraction des composantes principales

Le nombre maximum de composantes principales qu'il est possible d'extraire est égal au nombre de variables initiales. Toutefois, le pourcentage de variance expliqué par chaque composante décroît systématiquement à mesure que l'on progresse dans le processus d'extraction et peut devenir tout à fait négligeable une fois que les composantes les plus importantes auront été extraites. Ceci nous amène à considérer différents critères qui nous aideront à déterminer combien de composantes il vaut la peine d'extraire.

3.2.1 Critère de Kaiser

Pour comprendre ce critère, il faut aborder brièvement la notion de variance présente dans les données : La variance totale dans un tableau de données correspond à la somme des variances de chaque variable.

Comment cette variance totale sera-t-elle répartie entre les différentes composantes que nous voulons extraire? La réponse s'obtient en calculant la valeur propre (eigenvalue) de chaque composante.

Le critère de Kaiser stipule qu'il ne vaut pas la peine de poursuivre l'extraction des composantes si les valeurs propres correspondantes sont inférieures à 1, ce qui correspond à moins de variance que celle associée à une variable initiale normalisée de la matrice de corrélation Kaiser (1960).

3.2.2 Critère de Cattell

En 1966, Cattell a proposé une méthode graphique pour décider du nombre de composantes à extraire. Le test d'accumulation de variance communément appelé « scree test » demande que l'on trace un graphique illustrant la taille des valeurs propres des différentes composantes en fonction de leur ordre d'extraction. ce critère nous amène à arrêter l'extraction des composantes à l'endroit où se manifeste le changement de pente dans le graphique Cattell (1966).

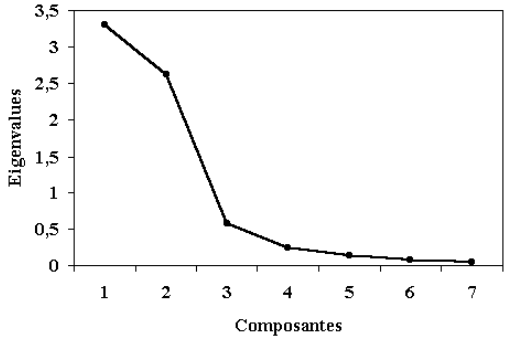


FIG. 1 – Illustration de l'accumulation de variance « scree test ».

La figue 1 correspond à des données fictives. On y constate que la pente change radicalement avec la composante C3, ce point appartient beaucoup plus au segment C3 à C7 qu'au segment C1 à C3. Selon ce critère on devrait donc se limiter à l'extraction des deux premières composantes.

3.2.3 Critère de Horn

L'approche suggérée par Horn s'appuie sur un raisonnement très différent des deux précédents. Horn indique qu'il est possible de découvrir par chance une composante pouvant expliquer une certaine proportion de variance, même en partant de données générées complètement au hasard et pour lesquelles aucune dimension réelle n'existe. Cette proportion de variance, expliquée par pure chance, pourrait donc servir comme point de comparaison afin de nous aider à décider si la variance que nous obtenons dans notre analyse est significativement plus importante que celle observable dans une matrice de données générées de façon aléatoire. L'analyse parallèle consiste donc à mener une ACP sur une matrice de corrélation générée au hasard mais comportant le même nombre de variables et d'individus que notre étude. La série décroissante des valeurs propres calculées sur ces données aléatoires sera alors comparée aux valeurs propres calculées sur les données réelles. Ainsi, Horn recommande de ne conserver pour extraction que les composantes dont les variances sont significativement supérieures à celles obtenues sur les données aléatoires Horn (1965).

La prise de décision est relativement facilitée si l'on trace un graphique représentant les deux séries de valeurs propres. L'inspection de la figure 2 permet de constater que cette méthode indiquerait deux composantes à extraire.

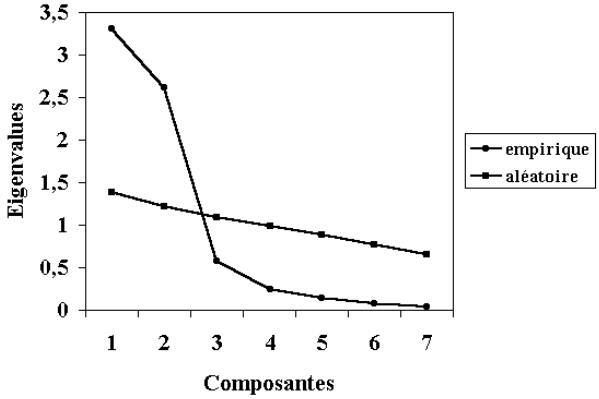


FIG. 2 – Illustration de l'analyse parallèle de Horn (1965).

3.2.4 Décision basée sur l'interprétation des composantes extraites

Généralement, la décision concernant le nombre de composantes à extraire doit aussi tenir compte de la capacité des chercheurs à interpréter les dimensions extraites. Il ne sert à rien d'extraire une composante en s'appuyant sur un critère aussi rigoureux soit-il, si par ailleurs cette composante défie toute compréhension. Par ailleurs, Wood et al. (1996) ont démontré qu'une surestimation du nombre de composantes était généralement moins dommageable qu'une sous-estimation. La décision quant au nombre de composantes à extraire

est difficile à prendre et comporte une part importante de subjectivité. Il est suggéré de confronter les différents critères plutôt que d'appliquer directement le critère de Kaiser.

4 Etude expérimentale

Notre expérimentation a été réalisée sur un entrepôt de données alimenté par deux bases de données sous Oracle 10g. Pour la réalisation de l'ACP normée et la construction des représentations graphiques associées nous avons utilisé le logiciel Revolution-R.

Le schéma en étoile de notre ED est constitué d'une table de faits et de trois tables de dimensions.

Table	Attribut	Commentaires
Journal (2 752 060 tuples)	Id_user (fk)	
	Id_nature_operation (fk)	
	Id_table (fk)	
	Num_operation	numéro séquentiel
	Code_agence	16 valeurs différentes
	Data_source	2 BD qui alimentent l'entrepôt (BD1 : 1 495 303 tuples et BD2 : 1 256 757 tuples)
Table (176 tuples)	Id_table (pk)	176 tables
	Table_name	
User (103 tuples)	Id_user (pk)	103 utilisateurs
	User_name	
	User_entity	
Nature_operation (3 tuples)	Id_nature_operation (pk)	3 valeurs possibles (Ajout, suppression et modification)
	Nature_operation	

TAB. 1 – *Schéma de l'entrepôt de données.*

Cette étude concerne uniquement la table de faits « Journal », la même démarche pourra être généralisée aux autres tables si nécessaire. Dans un premier temps, on va s'intéresser au centrage et à la réduction des variables pour réaliser une ACP normée et échapper à l'effet de masse que peuvent représenter certaines variables. Par la suite, nous allons effectuer la projection des individus en 2D et en 3D, L'étude des corrélations entre les variables initiales et les composantes principales nous permettra de tracer le cercle des corrélations et déterminer graphiquement des variables candidates à être rassemblées dans des fragments verticaux.

4.1 Calcul et interprétation des valeurs propres

En examinant la matrice des corrélations, on constate la présence d'une forte corrélation entre les trois variables (NUM_OPERATION, CODE_AGENCE, DATA_SOURCE).

	NUM_OP	AGENCE	USER	NAT_OP	TABLE	SOURCE
NUM_OP	1.000	-0.987	-0.012	0.030	0.038	0.993
AGENCE	-0.987	1.000	0.011	-0.023	-0.046	-0.993
USER	-0.012	0.011	1.000	0.026	-0.194	-0.013
NAT_OP	0.030	-0.023	0.026	1.000	0.089	0.033
TABLE	0.038	-0.046	-0.194	0.089	1.000	0.051
SOURCE	0.993	-0.993	-0.013	0.033	0.051	1.000

Le déterminant de cette matrice est égal à: 0.0001538589, sa valeur est supérieure à 0.00001, donc d'après Field(2000) il s'agit d'une situation propice pour l'utilisation de l'ACP.

Le tableau 2 présente les valeurs propres pour les données de la table de Fait « Journal ». On constate que la valeur propre associée à la première composante est de 2.99 ce qui correspond à 49.80 % de la variance totale qui est égale à 6. La composante C2 explique 1.20 unités de variance, ce qui correspond à 20.02 % de la variance totale. Nous pouvons donc dire qu'après avoir extrait deux composantes principales, on est en mesure de rendre compte de 70 % de la variance des données.

Donc nous avons pu réduire les données de 6 à 2 dimensions tout en réussissant à rendre compte de 70 % de la variance initiale.

La composante C3 explique 1.02 unités de variance, ce qui correspond à 16.98 % de la variance totale, et donc un total de 86,80 % de variance expliquée par les 3 premières composantes.

Le critère de Kaiser nous dit qu'il ne vaut pas la peine de poursuivre l'extraction puisque la composante C4 n'expliquerait que 0.77 unité de variance, ce qui correspond à moins de variance que celle associée à une variable initiale Kaiser(1960). Il est à rappeler que nous avons travaillé sur des variables centrées réduites, donc chaque variable possède 1.0 unité de variance.

Composante	Valeur propre	% de variance	% de variance cumulée
C_1	2.988024690	49.80041	49,80041
C_2	1.201345913	20.02243	69,82284
C_3	1.018818662	16.98031	86,80315
C_4	0.774942582	12.91571	99,71886
C_5	0.012538419	0.2089737	99,9278337
C_6	0.004329733	0.07216222	99,9999959
Total :	6.00000	100.00	

TAB. 2 – *Evolution des valeurs propres et pourcentages de variance*.

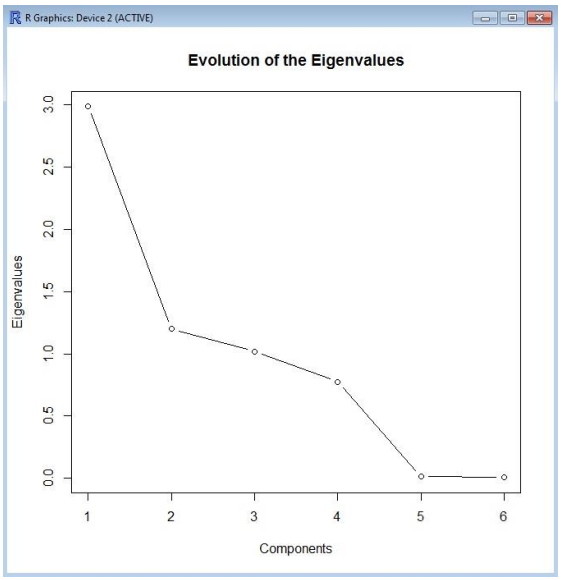


FIG. 3 – Graphe de l'évolution des valeurs propres.

4.2 Projection des individus sur le premier plan principal

Le premier plan principal est engendré par les deux premières composantes principales, il résume à lui seul 70 % de l'inertie totale contenue dans le tableau des données.

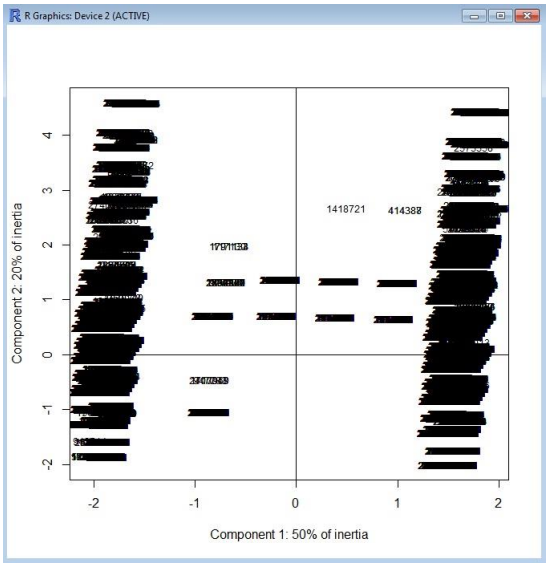


FIG. 4 – Projection des individus sur le premier plan principal.

On remarque que les individus appartiennent à deux groupes bien distingués : le premier groupe correspond aux individus dont c1>= 0 et le deuxième correspond à ceux dont c1<0.

On peut donc proposer un schéma de fragmentation horizontale composé de deux fragments.

4.3 Projection des individus sur l'espace vectoriel à 3D engendré par les trois premières composantes principales

La projection de 2 752 060 individus sur l'espace à 3D revient à calculer les nouvelles coordonnées de ces vecteurs/individus dans la nouvelle base des composantes principales, il s'agit d'un calcul lourd qui a été réalisé sur une machine puissante dotée de 32 Go de RAM.

L'espace vectoriel engendré par les trois premières composantes principales capture 86.80 % de l'inertie totale contenue dans le tableau des données. La visualisation des individus projetés sur cet espace nous apportera plus d'informations sur le comportement de ces derniers et donc une prise de décision meilleure par rapport à la projection en 2D.

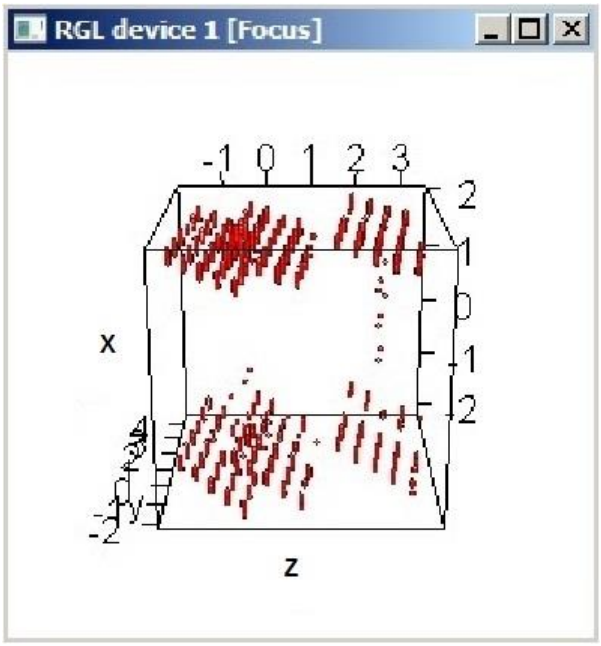


FIG. 5 – Projection des individus en 3 dimensions.

Dans cette projection, on s'intéresse particulièrement au positionnement des individus par rapport à la troisième composante principale (i.e: l'axe Z), d'après la Figure précédente, il est clair que cette projection fait apparaitre deux blocs d'individus bien séparés par rapport à l'axe Z (Bloc1 : individus pour lesquels z >=1.5 & Bloc2 : individus pour lesquels z <1.5), ce qui signifie des fragments horizontaux supplémentaires.

On peut donc recommander un schéma de fragmentation horizontale composé des quatre fragments suivants :

- Fragment horizontal 1 : l'ensemble des individus pour lesquels x>=0 & z >=1.5;
- Fragment horizontal 2 : l'ensemble des individus pour lesquels x>=0 & z < 1.5;
- Fragment horizontal 3 : l'ensemble des individus pour lesquels x< 0 & z >=1.5;
- Fragment horizontal 4 : l'ensemble des individus pour lesquels x< 0 & z < 1.5;

4.4 Cercle des corrélations : variables initiales – composantes C1 & C2

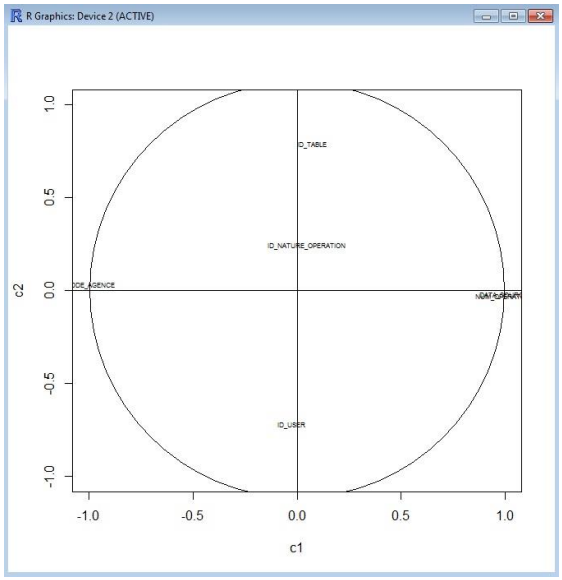


FIG. 6 – Cercle des corrélations.

En examinant le cercle des corrélation, on peut dire, d'une part, que les trois variables (DATA_SOURCE, NUM_OPERATION et CODE_AGENCE) sont bien représentées sur le plan (C1 ;C2) puisqu'elles sont proches du bord du cercle.

Les deux variables DATA_SOURCE et NUM_OPERATION sont très corrélées linéairement et positivement alors que la variable CODE_AGENCE et très corrélée linéairement mais négativement avec les deux premières variables.

On peut donc proposer un fragment vertical pour notre ED composé par les trois variables (DATA_SOURCE, NUM_OPERATION et CODE_AGENCE).

Les trois autres variables ne sont pas bien représentées (loin du bord du cercle) et donc on ne peut rien dire par rapport à ces variables. Cependant, il est clair que ces variables possèdent un coefficient de corrélation presque nul avec la composante C1.

On peut donc recommander un schéma de fragmentation verticale composé des deux fragments suivants :

- Fragment vertical1:(DATA_SOURCE, NUM_OPERATION et CODE_AGENCE);
- Fragment vertical2:(ID_TABLE, ID_USER et ID_NATURE_OPERATION);

5 Conclusion

Dans ce papier, nous avons utilisé l'ACP qui est une technique de description et de réduction des données. Elle comporte une série de décisions critiques portant sur les propriétés des variables soumises à l'analyse, les propriétés de la matrice d'intercorrélation et le nombre de composantes à extraire.

Nous avons mis en évidence la possibilité de l'utilisation de l'ACP pour la sélection de schémas de fragmentation horizontale et verticale des entrepôts de données.

La projection des individus en 3D nous a permis de mettre en évidence des fragments horizontaux supplémentaires, toutefois, une analyse approfondie au niveau de la qualité de représentation des individus et de l'interprétation des nouveaux axes semble nécessaire pour compléter l'étude.

L'étape suivante de notre travail consistera à réaliser la fragmentation sur l'ED sous ORACLE et à comparer le temps d'exécution d'un ensemble de requêtes OLAP sans et avec fragmentation basée sur l'ACP.

Références

Bellatreche. L, K. Boukhalfa, La fragmentation dans les entrepôts de données : une approche basée sur les algorithmes génétiques, Revue des nouvelles Technologies de l'information (EDA'2005), juin 2005, pages 141-160.

Bouchakri, R., L. Bellatreche, et K. Boukhalfa. Administration et Tuning des Entrepôts de Données : Optimisation par Index de Jointure Binaires et Fragmentation Horizontale. Doctoriales STIC'09. Msila. Décembre 2009.

Bouchier, A., Statistique et logiciel R : description théorique de l'ACP et mise en œuvre sous R avec le package ADE-4. Janvier 2006.

Boukhalfa, K., L. Bellatreche, and P. Richard, Fragmentation Primaire et Dérivée: Étude de Complexité, Algorithmes de Sélection et Validation sous ORACLE10g. Mars, 2008.

Cattell, R. B. (1966). The scree test for the number of factors. Multivariate Behavioral Research, 1, 245-276.

Chaudhuri, S., Index selection for databases : A hardness study and a principled heuristic solution. IEEE Transactions on Knowledge and Data Engineering, 16(11) :1313–1323, November 2004.

Duby, C., S. Robin. Analyse en Composantes Principales. Institut National Agronomique Paris-Grignon. Département O.M.I.P. 10 Juillet 2006.

Elmansouri, R., Ziyati, E., Aboutajdine, D., and Elbeqqali, O., The fragmentation of datawarehouses: an approach based on principal components analysis. ICMCS'2012.

Gupta, H., Selection and maintenance of views in a data warehouse. Ph.d. thesis, Stanford University, September 1999.

Horn, J. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30, 179-185.

Inmon,W. H. (1996). Building the Data Warehouse. John Wiley & Sons.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. Educational and Psychological Measurement, 20, 141-151.

Molina, H., W. J. Labio, J. L. Wiener, and Y. Zhuge. Distributed and parallel computing issues in data warehousing. Proceedings of the Seventeenth Annual ACM Symposium on Principles of Distributed Computing, page 7, June 1998.

Noaman, A. Y. and K. Barker. A horizontal fragmentation algorithm for the fact relation in a distributed data warehouse. CIKM'99, pages 154–161, November 1999.

Ozsu, M. T., and P. Valduriez. Principles of Distributed Database Systems : Second Edition. Prentice Hall, 1999.

Sanjay, A., V. R. Narasayya, and B. Yang. Integrating vertical and horizontal partitioning into automated physical database design. Proceedings of the ACM SIGMOD International Conference on Management of Data, pages 359–370, June 2004.

Wood, J. M., Tataryn, D. J., & Gorsuch, R. L. (1996). Effects of under- and overextraction on principal axis factor analysis with varimax rotation. Psychological Methods, 1, 354-365.

Ziyati, E., Aboutajdine D. and ElQadi A., Complete algorithm for fragmentation in data warehouse. Conf. AIKED'08. University of Cambridge. UK. Feb 20-22. 2008.

Sites web:

http://www.uqtr.uquebec.ca/cours/srp-6020/acp/acp.htm

http://eric.univ-lyon2.fr/~ricco/cours

Summary

In this paper, we present a state of the art on the principal components analysis (PCA) and the possibility of its use for horizontal and vertical fragmentation of data warehouses, in order to reduce the time of query execution. We focus on the study of correlation matrices, the impact of the eigenvalues evolution on the determination of suitable situations to achieve the PCA, and a study of criteria for extracting principal components. Then, we proceed to the projection of individuals on the first principal plane, and the 3D vector space generated by the first three principal components. We try to determine graphically homogeneous groups of individuals and therefore, a horizontal fragmentation schema for the studied data table.

The study of correlations between the original variables and the principal components allow us to draw the circle of correlations and define graphically, under some conditions, candidate variables to be collected in vertical fragments and thus make a decision on a vertical fragmentation schema.

To satisfy a maximum of decision queries OLAP, our approach is independent from any set of queries, and seeks to exploit the graphical representations provided by the PCA.

We conclude our study by an experiment on a data warehouse which shows the interest and the originality of our approach.

Nouvelle Approche Scalable Dédiée au Charges Volumineuses pour la Fragmentation des Entrepôts de Données

Amina Gacem*, Kamel Boukhalfa**

*Ecole Nationale Supérieure d'Informatique (ESI)
Oued Smar, BP 68, Alger, Algérie
a_gacem@esi.dz
**Université des Sciences et de la Technologie Houari Boumediene (USTHB)
BP 32, Bab Ezzouar, Alger, Algérie
kboukhalfa@usthb.dz

Résumé. La fragmentation horizontale est une technique largement utilisée pour améliorer la conception physique des entrepôts de données. Elle offre des avantages en termes de performances et de manageabilité sans consommation d'espace de stockage supplémentaire. Cependant, la sélection d'un schéma de fragmentation est un problème NP-complet. Ainsi, plusieurs algorithmes ont été proposés afin de gérer la complexité du problème et générer des schémas de fragmentation de bonne qualité. Ces travaux considèrent en entrée une charge de requêtes les plus fréquentes à optimiser. Néanmoins, la majorité des travaux ne prennent pas en considération la mise à l'échelle par rapport à la taille de la charge qui peut être très importante. En réalité l'entrepôt de données est soumis continuellement à des charges de requêtes volumineuses et cela nécessite la proposition d'une nouvelle approche pour gérer cette volumétrie sans détérioration de la qualité de la solution finale ni le temps d'exécution des algorithmes de sélection. Nous proposons dans ce papier, une approche scalable basée sur la classification et l'élection pour une fragmentation supportant des charges volumineuses. Nous avons mené une étude expérimentale sur le benchmark d'ABP-1 pour tester l'efficacité et le passage à l'échelle de notre approche. Les résultats obtenus sont encourageants

1 Introduction

Les entrepôts de données (ED) sont au cœur des systèmes décisionnels des entreprises. Ils sont accédés simultanément par plusieurs applications et sont interrogés par plusieurs transactions à des fins de prise de décision(Inmon, 1992). Pour les compagnies à grande activité stockant des téraoctets de données, il n'est pas rare qu'un entrepôt soit soumis à une charge de travail constituée de plusieurs jobs exécutés par plusieurs clients dépassant les *milliers de requêtes* (Feinberg, 2006). Ceci s'explique par l'objectif décisionnel lié à ce type de base de données. Les charges de travail volumineuses peuvent dégrader les performances du SGDB,

et ainsi, ralentir les applications et augmenter ainsi le temps de réponse au client, souvent exigeant dans les délais, en particulier lorsqu'il s'agit d'un décideur. Pour y remédier, diverses techniques d'optimisation ont été proposées, entre autres la fragmentation horizontale (FH) (Sanjay et al., 2004). La sélection d'un schéma de FH est un problème NP-complet (Boukhalfa, 2009). Des algorithmes et heuristiques ont été proposés afin de définir automatiquement le meilleur schéma de fragmentation. Ces algorithmes passent en général par 2 étapes : (1) Collecte et analyse des requêtes, métadonnées et statistiques et (2) Recherche d'un schéma de FH quasi-optimal.

Plusieurs travaux se sont intéressés au problème de sélection d'un schéma de FH (Ceri et al., 1982; Bellatreche, 2000; Rao et al., 2002; Mahboubi, 2008; Boukhalfa, 2009; Barr et Bellatreche, 2010; Karima et al., 2010; Rehme et Bruno, 2011). Ces algorithmes ont été testés sur des charges de requêtes de petite taille comportant quelques dizaines de requêtes, or, la fragmentation est appliquée afin d'optimiser les performances de l'entrepôt pour pouvoir exécuter un nombre important de requêtes. Par conséquent, pour pouvoir évaluer l'apport de ces travaux, il est impératif de les tester sur des charges volumineuses. Cependant, une charge comportant un nombre élevé de requêtes complique la tâche de conception physique car la volumétrie des requêtes allonge le temps d'accès à l'entrepôt de données et l'évaluation des solutions parcourues par les algorithmes de sélection. Une charge volumineuse comporte un nombre important de prédicats et conduit donc à une explosion de l'espace de recherche de l'entrepôt de données. Ces effets vont à leurs tour provoquer : (1) Une consommation d'un temps supplémentaire, d'abord pour analyser les requêtes, et ensuite pour rechercher un schéma de FH du fait de l'explosion du nombre de solutions possibles, (2) Une consommation d'un temps supplémentaire pour évaluer chaque schéma de fragmentation puisque cette évaluation est effectuée sur chaque requête de la charge et (3) Une dégradation de la qualité du schéma de FH obtenu à cause de l'exploration du nombre de solutions.

L'administrateur sera donc amené à faire un compromis entre la rapidité du temps de réponse de l'algorithme de sélection d'un schéma FH et la qualité du schéma de FH obtenu. Notre objectif est de concevoir et implémenter une nouvelle approche de sélection, capable de retourner un schéma de FH final de bonne qualité en un temps de recherche acceptable, même si la charge de requêtes est volumineuse. Pour y parvenir, nous proposons l'ajout de deux phases au processus classique de sélection d'un schéma de fragmentation horizontale : *la classification* et *l'élection*. La classification (MacQueen, 1967; Agrawal et al., 1998), très largement répandue dans plusieurs problématiques, pourrait être utile dans ce problème. Le regroupement des requêtes similaires en classes permet d'avoir un ensemble plus restreint de classes au lieu de milliers de requêtes. Une élection d'une requête par classe permet d'avoir une nouvelle charge de taille inférieure qui sera par la suite donnée comme entrée à un algorithme de sélection d'un schéma de FH. Les résultats d'une telle approche dépendent aussi du type de la charge de requête. En effet, la classification varie en fonction du degré d'homogénéité de la charge de requête.

Le présent papier est organisé comme suit : la section 2 aborde les différents travaux liés à la sélection d'un schéma de FH, la section 3 sera dédiée à la présentation d'une nouvelle approche de sélection basée sur la classification et l'élection. Nous détaillerons l'étude expérimentale dans la section 4. Enfin, nous conclurons le papier dans la section 5.

2 Travaux Connexes

La FH a été étudiée dans plusieurs environnements : *parallèle* (Valduriez, 1993; Rao et al., 2002), *distribuées* (Ceri et al., 1982; Valduriez et Özsu, 1999), *Grid* (Bellatreche, 2000; Fiolet et Toursel, 2005; Karima et al., 2010; Mahboubi, 2008) et *Cloud* (Bajda-Pawlikowski et al., 2011) et a été implémentée dans les principaux SGBD du marché : *Oracle* (Baer et al, 2011), *SQL Server* (Microsoft, 2012) et *IBM DB2* (Cain, 2006). Chaque SGBD dispose de plusieurs modes de fragmentation et fournit un ensemble de commandes pour créer et manipuler les partitions.
Le problème de sélection d'un schéma de FH a été démontré NP-Complet (Sacca et Wiederhold, 1983; Boukhalfa, 2009). Plusieurs travaux ont été proposés pour traiter ce problème. Ils peuvent être regroupés en quatre catégories : travaux guidés par *les prédicats*, travaux guidés par *l'affinité*, travaux guidés par *un modèle de coût* et travaux guidés par *la classification*.

Dans la première catégorie de travaux, la sélection est guidée par les prédicats de sélection extraits des requêtes (Ceri et al., 1982; Valduriez et Özsu, 1999). A partir de ces prédicats, l'ensemble des minterms est généré. Un minterm est une conjonction de plusieurs prédicats simples. Cette catégorie d'approches est simple à implémenter mais génère une complexité exponentielle (si n est le nombre de de prédicats, 2^n minterms sont générés).

Les travaux guidés par l'affinité associent un fragment à chaque groupe de prédicats possédants un certain degré d'affinité. L'affinité entre deux prédicats est exprimée par la somme des fréquences des requêtes référençant les deux prédicats en même temps. Deux algorithmes ont été proposés pour la génération des partitions. Le premier représente la matrice d'affinité par un graphe et génère des partitions par recherche de cycles dans le graphe grâce à l'algorithme de regroupement graphique (Navathe et al., 1984). Le deuxième applique d'une manière récursive l'algorithme BEA (McCormick et al., 1972) pour décomposer la matrice d'affinité en plusieurs sous-matrices représentant chacune une partition potentielle. L'approche par affinité a été proposée par (Navathe et al., 1984) pour la fragmentation verticale et a été adaptée dans le contexte de la FH par (Bellatreche et al., 1997).

Les approches guidées par un modèle de coût s'inspirent des modèles de coût utilisés par les optimiseurs des SGBDs afin de choisir le plan d'exécution le plus optimal. Dans ces approches, le problème de FH est modélisé par un problème d'optimisation avec une fonction objectif qui intègre le calcul du coût. L'exploration des solutions se fait en appliquant des méta-heuristiques. Certains travaux considèrent un modèle de coût mathématique (Bellatreche, 2000; Boukhalfa, 2009; Barr et Bellatreche, 2010), d'autres considèrent le coût de l'optimiseur du SGBD (Rehme et Bruno, 2011). Parmi les méta-heuristiques utilisées, nous pouvons citer : Hill Climbing (HC), Recuit Simulé (RS), Algorithme Génétique (AG)(Boukhalfa, 2009), Colonie de Fourmis (CF) (Barr et Bellatreche, 2010) et Branch & Bound (Rehme et Bruno, 2011). Les modèles de coût mathématiques évaluent en général le nombre d'accès E/S aux disques et utilisent des paramètres comme la taille de la page système ou du buffer. Notons que ces travaux considèrent que la FH consiste à découper le domaine de définition des attributs de fragmentation en plusieurs sous-domaines (Voir (Bellatreche et Boukhalfa, 2005) pour plus de détail).

Les travaux guidés par la classification s'inspirent essentiellement de l'algorithme K-means (MacQueen, 1967; Lloyd, 1982; Pham et al., 2004). Ces approches se basent sur la classification des prédicats extraits de la charge de requêtes à travers des algorithmes de classification puis associent à chaque cluster de prédicats, un fragment horizontal. L'algorithme K-means a

été utilisé dans les problèmes de classification dans le contexte des ED relationnels distribués (Karima et al., 2010) et les ED XML (Mahboubi, 2008). Dans (Mahboubi, 2008), les auteurs proposent une démarche de FH d'un ED XML basé sur K-means avant de répartir les différents fragments sur une grille. Dans (Karima et al., 2010), les auteurs utilisent K-means pour fragmenter l'ED relationnel avant d'allouer chaque fragment sur un site en fonction des fréquences d'accès de chaque nœud. Le Tableau 1 présente une étude comparative des travaux existants. Comme nous pouvons le constater dans ce comparatif, la majorité des travaux ont été testés sur des charges de requêtes comportant entre 20 et 60 requêtes. De plus, aucune approche citée plus haut ne définit son comportement lorsque la charge de requêtes est volumineuse. Sachant que le nombre et la complexité des requêtes décisionnelles croissent de jour en jour, il est impératif de développer une approche adaptée à ce type de charge qui puisse sélectionner un schéma de FH optimal efficacement. Nous présentons dans la section suivante notre approche de FH traitant cette problématique.

Travaux	Algorithmes de Sélection	Modèle de Coût	Type de BD	Taille de la Charge
Ceri et al. (1982)	Prédicats	-	BD Distribuées	-
Rao et al. (2002)	Affinités	-	BD Distribuées	-
Boukhalfa (2009)	HC, RS, AG	Mathématique	ED Centralisé	60
Barr et Bellatreche (2010)	CF	Mathématique	ED Centralisé	55
Rehme et Bruno (2011)	Branch&Bound	Optimiseur	ED Parallèles	50
Mahboubi (2008)	K-means	-	ED XML	20
Karima et al. (2010)	K-means	-	ED Distribuées	36

(-) : Non Défini

TAB. 1 – *Etude Comparative entre les approches de sélection d'un schéma de FH*

3 Notre approche de sélection d'un schéma de FH

En se penchant sur la littérature, nous pouvons constater deux insuffisances qui caractérisent les approches de FH proposées : (1) Les approches ignorent les charges de requêtes volumineuses et leurs incidences sur les performances de l'approche et (2) la réduction de la complexité du problème de sélection d'un schéma de FH repose généralement sur la réduction du nombre d'attributs de sélection et/ou des tables de dimension (Bellatreche et al., 2009; Bouchakri et al., 2010). A notre connaissance, aucun travail sur la FH ne propose une réduction de la charge de requêtes en entrée pour réduire la complexité du problème de sélection et le temps d'exécution des requêtes. En effet, une charge de requêtes volumineuse engendre deux problèmes : (1) le temps d'évaluation des solutions parcourues devient très important engendrant une explosion du temps d'exécution du processus de sélection et (2) le nombre de prédicats pris en compte par le processus de fragmentation devient important et augmente la taille de l'espace de recherche. Il est donc intéressant de proposer une approche de sélection d'un schéma de FH permettant de mieux gérer les charges volumineuses afin de réduire le temps d'exécution de la sélection tout en générant des solutions de bonne qualité.

L'approche que nous proposons dans le présent papier rentre dans cet objectif. Elle est basée sur deux principes : (1) *la classification* des requêtes pour réduire la taille de la charge et

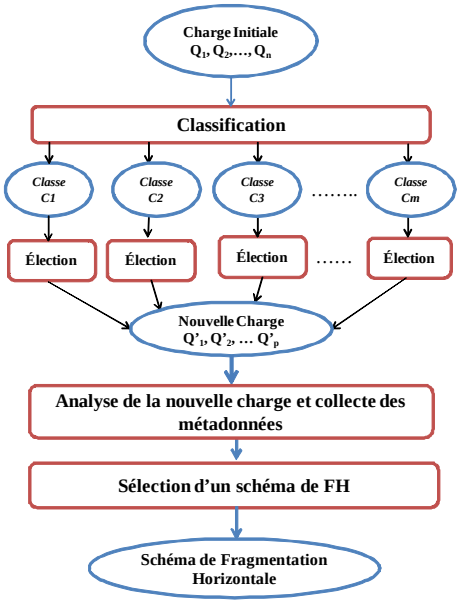


FIG. 1 – *Aperçu général de notre approche*

(2) *l'élection* pour produire une nouvelle charge représentative de la charge initiale. Évidemment, la réduction de la charge de requêtes doit se faire de manière judicieuse afin de ne pas perdre les caractéristiques de la charge d'origine. Nous avons été amenés à traiter deux problèmes : (1) sur quel critère le regroupement des requêtes dans une même classe doit se faire (critère de classification) et (2) sur quel critère une requête représentative de sa classe doit être sélectionnée (critère d'élection). La figure 1 illustre notre approche.

3.1 Phase de classification

L'objectif de la phase de classification est de regrouper les requêtes similaires dans une même classe. Elle prend en entrée une charge composée de N requêtes et génère en sortie R classes de requêtes tel que $R \leq N$. Dans le contexte de FH, nous avons considéré que deux requêtes sont similaires si elles référencent des données communes (tuples) dans la table de faits. Ce principe se base sur le fait que si deux requêtes partagent les mêmes données, alors un même schéma de fragmentation pourrait satisfaire les deux requêtes en même temps. Le cas idéal est constaté lorsque les données référencées par une requête sont incluses dans les données référencées par la deuxième requête.
Plusieurs algorithmes de classification ont été proposés dans la littérature, nous pouvons citer : *K-means* (MacQueen, 1967; Lloyd, 1982; Pham et al., 2004), *K-medoid* (Kaufman et Rousseeuw, 1990), *BIRCH* (Zhang et al., 1996), *DB-SCAN* (Ester et al., 1996) et *CLIQUE* (Agrawal et al., 1998). Nous avons choisi d'utiliser l'algorithme K-means dans le cadre de ce travail car il est simple à adapter pour supporter notre problématique. En effet, en utilisant cet

algorithme, nous pouvons créer les classes de requêtes à partir des clusters générés et effectuer l'élection par l'utilisation des centroïde générés pour chaque classe.

La classification des requêtes dans notre approches passe par quatre étapes : (1) extraction des prédicats de sélection, (2) construction de la matrice *requête-prédicat*, (3) calcul du nombre de classes et (4) génération des classes par K-means. L'extraction des prédicats de sélection se fait en examinant la clause *where* de chaque requête. Un prédicat de sélection a la forme suivante : $A\ \theta\ Valeur$ où A représente un attribut, $\theta \in \{<, >, \leq, \geq, =\}$ et $Valeur \in domaine(A)$. A partir de l'ensemble des prédicats obtenu dans l'étape précédente, nous construisons une matrice M appelée matrice *requête-prédicat*. Un élément m_{ij} de cette matrice est défini par : $m_{ij} = 1$ si la requête Q_i utilise le prédicat P_j, 0 sinon. Le nombre de classes à générer est calculé en utilisant le principe de similarité expliqué ci-dessus. Nous avons implémenté un module qui cherche les relations d'inclusion entre les prédicats de sélection utilisés par les requêtes de la charge. Le nombre de classes K correspond au nombre d'inclusions trouvées. L'algorithme K-means prend en entrée le nombre de classes K et la matrice requête-prédicats et génère en sortie K classes de requêtes $\{C_1, C_2, \cdots, C_K\}$.

3.2 Phase d'élection

L'algorithme K-means utilisé dans la phase de classification génère, en plus des classes, le *centre* de chaque cluster. Ce centre représente un ensemble de prédicats utilisés par la charge de requêtes. Plusieurs techniques peuvent être utilisées pour élire une requête représentative de sa classe comme la fréquence d'utilisation, le volume de données référencées, etc. Dans le cadre de ce travail, nous avons choisi d'exploiter les résultats du même algorithme utilisé pour la classification (K-means). L'approche consiste à prendre le centre de chaque cluster pour générer la requête élue. Notons que chaque centre est constitué d'un ensemble de prédicats où chacun est accompagné par un poids représentant son importance dans la construction de ce centre (figure 2). Le poids est d'un prédicat *j* d'une requête *i* est donnée par l'élément a_{ij} de la nouvelle matrice obtenue. Pour décider quels prédicats choisir pour générer la requête élue, nous avons défini un seuil α qui détermine à partir de quel poids le prédicat est choisi pour construire la requête élue. Ce seuil peut être paramétré par l'administrateur ou calculée d'une manière expérimentale. Un prédicat P_i est choisi pour la construction de la requête élue si et seulement si $Poids(Pi) \geq \alpha$. Par exemple, dans la figure 4, si le paramètre α a été fixé à 0,5 alors seuls les prédicats p_1, p_3, p_4, p_6 sont pris pour construire la requête élue correspondante au centre du cluster 2. Notons que la requête élue sera construite par conjonction des prédicats choisis.

4 Étude expérimentale

Nous présentons dans cette section, une évaluation de notre approche à travers une étude expérimentale qui se déroule sur 3 étapes : (1) tests guidés par la taille de la charge (moyenne, volumineuse), (2) tests guidés par le degré d'hétérogénéité de la charge (hétérogène, homogène) et (3) une validation sous Oracle. Nous avons utilisé l'algorithme génétique proposé dans Boukhalfa (2009) pour la sélection du schéma de FH puisque les auteurs ont montré qu'il est le meilleur parmi d'autres algorithmes. Les paramètres utilisés de l'AG sont : *population initiale 50, Taux de sélection 0.9, Taux de croisement 0.8, taux de mutation 0.1, nombre de*

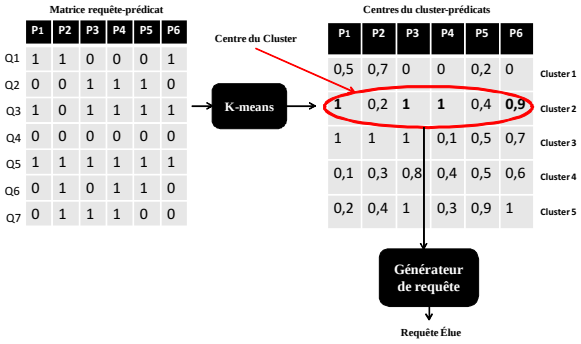


FIG. 2 – *Adaptation de K-means pour la classification et l'élection des requêtes*

générations 500 et nous avons fixé le nombre de fragments maximum à 1000. La qualité du schéma sélectionné est évalué par un modèle de coût proposé par les mêmes auteurs.
Nous avons utilisé l'ED fourni par le benchmark d'APB-1 (OLAP-Council, 1998) que nous avons généré et peuplé sous le SGBD Oracle 11g. Il est constitué d'une table de faits *Actvars* (24 786 000 tuples) et de quatre dimensions, *Prodlevel* (9 000 tuples), *Custlevel* (900 tuples), *Timelevel* (24 tuples) et *Chanlevel* (9 tuples). Nous nous sommes intéressés aux requêtes comportant des jointures et des agrégations. Nous avons développé notre solution en Java. Nous avons effectué nos tests sur une machine dotée d'un processeur *Intel* 2.10 GHz et d'une mémoire vive de 2 Go.

4.1 Test de l'approche

Dans la littérature consacrée aux approches de sélection d'un schéma de FH, les tests expérimentaux sont fait sur des charges d'au maximum 60 requêtes. Pour montrer que notre approche supporte les charges de requêtes plus volumineuses, nous avons testé notre approche sur une charge de taille moyenne constituée de 500 requêtes ensuite sur une charge comportant 1000 requêtes. Nous avons fixé les valeurs expérimentales de α à 1 et à 0,3.
Les résultats des expérimentations sur 500 requêtes sont montrés dans la figure 3. Notre approche réduit le coût d'exécution de la charge de 2% pour les deux valeurs de α par rapport à l'approche classique que nous appelons *approche sans classification* .

Pour montrer le comportement de notre approche dans le cas des charges volumineuses, nous avons généré une charge de 1000 requêtes et nous avons exécuté notre approche sur cette charge. Les résultats sont représentés dans la figure 4. Notre approche améliore de 20% la qualité du schéma de FH par rapport à l'approche classique pour $\alpha = 1$ et donne un schéma de qualité similaire à celui obtenu par l'approche classique pour $\alpha = 0,3$. Nous pouvons conclure de ces deux expériences que malgré l'utilisation d'une charge réduite, notre approche produit des schémas de fragmentation de qualité meilleure ou similaire par rapport à l'approche classique qui exploite une charge moins importante. Nous pouvons conclure aussi que la valeur du seuil α a une incidence sur la qualité du schéma de FH obtenu. Pour les expérimentations que nous avons effectué, la valeur $(\alpha = 1)$ a donné les meilleurs résultats.

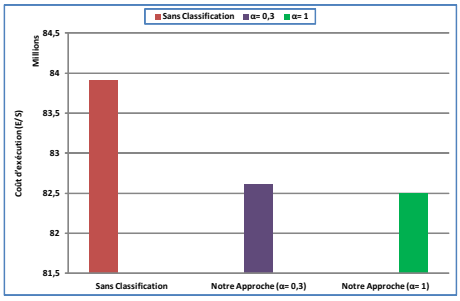


FIG. 3 – *Coût d'exécution des 500 requêtes*

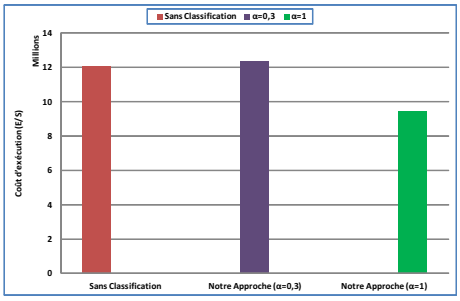


FIG. 4 – *Coût d'exécution des 1000 requêtes*

Montrer que l'approche donne de meilleurs résultats par rapport à la qualité de la solution obtenue n'est pas suffisant. Il est impératif de vérifier l'incidence de la nouvelle approche sur le temps d'exécution total du processus de FH. Ce temps exprime le temps d'exécution total depuis le chargement des requêtes de la charge jusqu'à l'obtention du schéma de FH. Ce temps est noté T_{total} et calculé comme suit :

$$T_{total} = T_{ch} + T_{an} + T_{cl} + T_{el} + T_{re}$$

où : *T_{ch} est le temps de chargement des requêtes à partir des fichiers sources, il ne dépend pas de l'approche appliquée, donc il n'a pas été comptabilisé, T_{an} est le temps de collecte des statistiques et des métadonnées (il dépend du volume de la charge), T_{cl} est le temps de classification des requêtes (il dépend de l'algorithme de classification), T_{el} est le temps d'élection (il dépend de l'algorithme d'élection) et T_{re} est le temps de recherche du schéma de FH (il dépend de l'algorithme de sélection, l'AG dans notre cas)*

Nous avons considéré une charge de 1000 requêtes et nous avons calculé les différents temps ci-dessus cités. Les résultats sont montrés dans la figure 5. Le temps d'exécution global de notre approche divise par 10 le temps d'exécution T_{total}. Nous expliquons ce gain de temps par la réduction de la charge de requêtes qui a diminué sensiblement le temps de recherche et d'évaluation des schémas de FH et le temps de collecte de métadonnées. Les taux de réduction du temps d'analyse de la charge de requête est de 70% lorsque α*=0,3* et de 87% lorsque α*=1*. Le temps d'exploration de l'espace des solutions a été amélioré de 95%. Les temps de classification et d'élection sont négligeables. De ce fait, une nette amélioration des performances est constatée. Nous pouvons conclure que notre approche réduit considérablement le temps d'exécution total de sélection d'un schéma de FH tout en générant un schéma meilleure ou de qualité similaire par rapport à l'approche classique.

4.2 Effet de l'homogénéité de la charge sur notre approche

La charge de requêtes utilisée dans les précédentes expérimentations a été générée d'une manière totalement aléatoire. Comme les processus de classification et d'élection reposent sur les similarités entre les requêtes, il est intéressant de montrer le comportement de notre approche sur des charges hétérogènes et sur des charges homogènes. Les charges hétérogènes contiennent des requêtes qui ne partagent que très peu de lignes de données. Les requêtes homogènes sont celles qui partagent un grand nombre de lignes de données. Pour faire des ex-

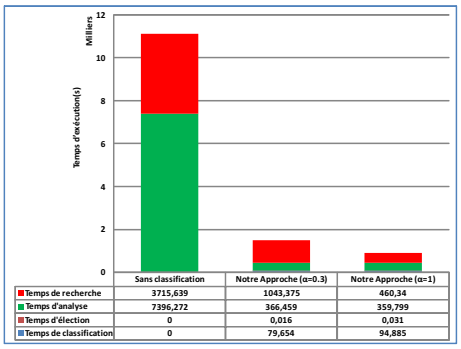


FIG. 5 – *Performances des algorithmes (1000 requêtes)*

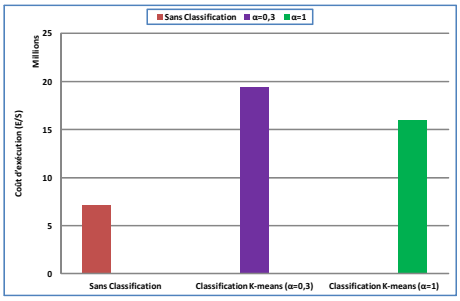


FIG. 6 – *Coût d'exécution des 730 requêtes hétérogènes*

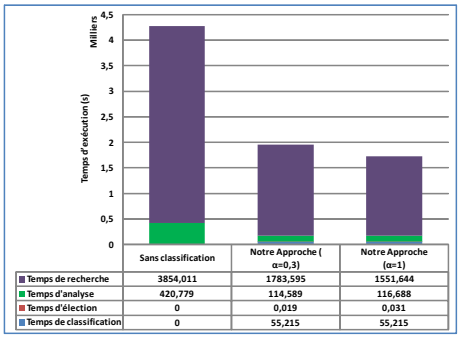


FIG. 7 – *Performances des algorithmes pour 730 requêtes hétérogènes*

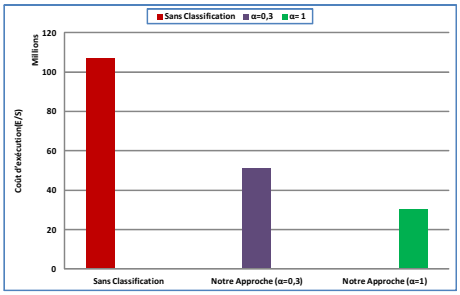


FIG. 8 – *Coût d'exécution des 1010 requêtes homogènes*

périmentations en fonction de l'hétérogénéité de la charge, nous avons implémenté un module qui génère des charges pouvant être, selon le choix de l'utilisateur, homogène ou hétérogène.

Charge Hétérogène : Nous avons généré une série de 730 requêtes hétérogènes, comportant des prédicats qui diffèrent d'une requête à une autre. Chaque prédicat apparaît dans 3 requêtes au plus. La figure 6 illustre les performances obtenues de notre approche. Nous ne constatons aucune amélioration particulière et une dégradation du coût d'exécution des requêtes hétérogènes. Nous avons observé le fait que l'algorithme de classification génère un nombre de classes très important proche du nombre de requêtes de la charge initiale. Par conséquent, la charge obtenue après classification est presque identique à la charge initiale. En effet, le nombre de classe est déterminé par la relation d'inclusion définie précédemment. Lorsque les requêtes sont hétérogènes entre elles, il existe peu de cas d'inclusion entre requêtes et donc, peu de requêtes font partie de la même classe, d'où le nombre important de classes. La figure 7 représente le temps d'exécution de notre approche comparé au temps de l'approche classique. Le temps d'exécution global est divisé par deux. Nous pouvons conclure que malgré la qualité non améliorée de la solution obtenue, le temps d'exécution nécessaire à son obtention est nettement amélioré.

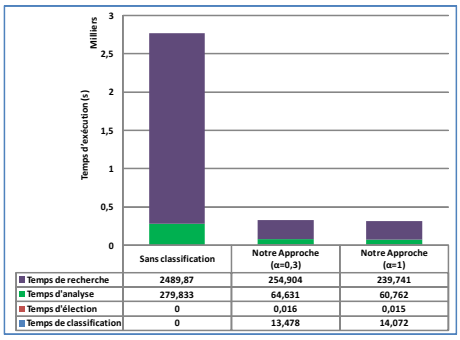


FIG. 9 – *Performances des algorithmes pour 1010 requêtes homogènes*

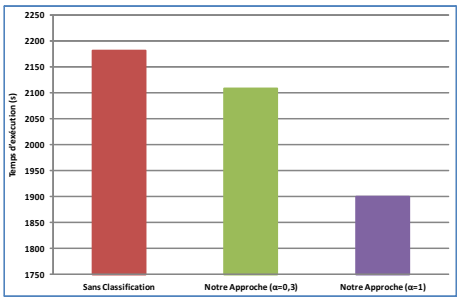


FIG. 10 – *Validation sous Oracle pour 1000 requêtes*

Charge homogène : Nous avons généré une charge de requêtes comportant 1010 requêtes. Pour que la charge soit homogène, nous avons veillé à ce que chaque prédicat soit présent dans au moins 200 requêtes. Les résultats de nos différentes approches sont représentés dans la figure 8 . Ces résultats montrent que notre approche améliore le coût de 52,33% lorsque α*=0.3* et de 70% lorsque α*=1*. De ce fait, la classification et l'élection des requêtes améliorent sensiblement le coût d'exécution lorsque les requêtes sont homogènes entre elles. Nous pouvons justifier ce comportement par le fait que l'algorithme de classification génère un nombre de classes réduit et l'élection génère une nouvelle charge qui représente efficacement la charge initiale grâce à son homogénéité. La figure 9 montre une baisse considérable du temps consommé à analyser les requêtes et à rechercher un schéma de FH, soit une baisse de 90%. Cela est dû à la taille très réduite de la charge obtenue après classification. Les meilleurs résultats de notre approche sont donc obtenus dans le ces des charges de requêtes homogènes.

Sur la base des résultats expérimentaux obtenus, nous pouvons formuler un ensemble de recommandations à l'administrateur de l'ED : (1) Pour les charges moyennes (moins de 500 requêtes), notre approche améliore nettement le coût d'exécution des requêtes (20%) et divise par 10 le temps d'exécution de l'approche. (2) Lorsque la charge de requêtes est hétérogène, notre approche produit des schémas de FH de même qualité que l'approche classique dans des délais plus raisonnables et divise le temps global de sélection par deux. Dans ce cas, si l'administrateur préfère la qualité de la solution, il pourra utiliser l'approche classique, par contre, s'il préfère avoir une solution de bonne qualité en un temps plus réduit, il pourra utiliser notre approche. (3) Pour les charges volumineuses homogènes, la réduction de la charge de requêtes apporte un gain considérable en termes de performances et réduit nettement le coût. Nous recommandons donc particulièrement notre approche en présence d'une charge de requête fortement homogène.

4.3 Validation sous Oracle 11g

Pour tester les performances de notre approche sur un SGBD commerçial, nous avons implémenté les schémas de fragmentation obtenus par l'approche classique et par notre approche sous Oracle 11g et avons exécuté la charge de requêtes sur le SGBD. Le temps d'exécution de la charge initiale est alors calculé sur le schéma généré par chaque approche. Nous avons

utilisé une charge de 1000 requêtes aléatoires pour effectuer ces tests. Les résultats montrent une baisse effective durant l'exécution des requêtes sur le schéma généré par notre approche de 3,4% pour $\alpha = 0,3$ et de 13% pour $\alpha = 1$ par rapport à l'approche classique (voir figure 10). Les résultats obtenus sous Oracle 11g confirment que notre approche apporte un gain considérable en temps d'exécution du processus de fragmentation d'un entrepôt de données tout en sélectionnant des schémas de fragmentation de meilleure qualité ou proches de ceux sélectionnés par l'approche classique.

5 Conclusion

Le présent travail traite le problème de conception physique d'un ED lorsqu'il est soumis à des charges de requêtes volumineuses. Plus précisément, nous nous sommes intéressés à une tâche importante de conception physique : la sélection d'un schéma de FH. Nous avons vu qu'il est difficile d'obtenir un schéma de FH de très bonne qualité en un temps raisonnable pour des charges volumineuses. Pour résoudre ce problème, nous avons proposé une approche basée sur la classification des requêtes suivie d'une élection d'une requête par classe. L'étude expérimentale et la validation sous Oracle ont montré que notre approche donne de meilleurs résultats par rapport à l'approche classique. En effet, elle réduit considérablement le temps de sélection d'un schéma de FH tout en générant des schéma de qualité meilleure ou similaires que ceux générés par l'approche classique notamment dans le cas de charges homogènes.

Nous pouvons améliorer ce travail par les points suivants : (1) l'amélioration du processus de classification et d'élection en intégrant d'autres critères comme la fréquence d'utilisation et le volume de données référencées par chaque requête, etc., (2) conduire des tests expérimentaux plus fins pour trouver les meilleures valeurs du paramètre α et (3) intégrer d'autres techniques d'optimisation dans l'approche comme les index et les vues matérialisées .

Références

Agrawal, R., J. Gehrke, D. Gunopulos, et P. Raghavan (1998). Automatic subspace clustering of high dimensional data for data mining applications. *SIGMOD 27*(2), 94–105.

Baer, H. et al (2011). Oracle database vldb and partitioning guide 11g release 2. Technical report, Oracle, Inc. Oracle White Paper.

Bajda-Pawlikowski, K., D. Abadi, A. Silberschatz, et E. Paulson (2011). Efficient processing of data warehousing queries in a split execution environment. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, New York, NY, USA, pp. 1165–1176. ACM.

Barr, M. et L. Bellatreche (2010). A new approach based on ants for solving the problem of horizontal fragmentation in relational data warehouses. In *Machine and Web Intelligence (ICMWI), 2010 International Conference on*, pp. 411–415.

Bellatreche, L. (2000). *Utilisation des vues matérialisées, des index et de la fragmentation dans la conception logique et physique d'un entrepôt de données*. Thèse de doctorat, Université de Clermont-Ferrand.

Bellatreche, L. et K. Boukhalfa (2005). An evolutionary approach to schema partitioning selection in a data warehouse. In *DaWaK*, pp. 115–125.

Bellatreche, L., K. Boukhalfa, P. Richard, et K. Y. Woameno (2009). Referential horizontal partitioning selection problem in data warehouses : Hardness study and selection algorithms. *IJDWM 5*(4), 1–23.

Bellatreche, L., K. Karlapalem, et A. Simonet (1997). Horizontal class partitioning in object-oriented databases. In *in Lecture Notes in Computer Science*, Berlin, Heidelberg, pp. 58–67. Lecture Notes in Computer Science.

Bouchakri, R., L. Bellatreche, et K. Boukhalfa (2010). Une sélection multiple des structures d'optimisation dirigée par la méthode de classification k-means. In *EDA*, pp. 207–222.

Boukhalfa, K. (2009). *De la Conception Physique aux Outils d'Administration et de Tuning des Entrepots de Donnees*. Thèse de doctorat, ENSMA.

Cain, M. (2006). Table partitioning strategies db2. Technical report, IBM.

Ceri, S., M. Negri, et G. Pelagatti (1982). Horizontal data partitioning in database design. In *Proceedings of the 1982 ACM SIGMOD international conference on Management of data*.

Ester, M., H. Kriegel, J. Sander, et X. Xu (1996). A density-based algorithm for discovering clusters in large spatial databases. *Data Mining Knowlege Discovery KDD 2*(2), 169–194.

Feinberg, D. (2006). Database management systems. Technology trends, Gartner.

Fiolet, V. et B. Toursel (2005). Intelligent database distribution on a grid using clustering. In *Proceedings of the Third international conference on Advances in Web Intelligence*, Berlin, Heidelberg, pp. 466–472. Springer-Verlag.

Inmon, W. (1992). *Building the Data Warehouse*. Hoboken : John Wiley.

Karima, T., A. Abdellatif, et H. Ounalli (2010). Data mining based fragmentation technique for distributed data warehouses environment using predicate construction technique. In *Networked Computing and Advanced Information Management (NCM), 2010 Sixth International Conference on*, pp. 63–68.

Kaufman, L. et P. Rousseeuw (1990). *Finding Groups in Data : an Introduction to Cluster Analysis*. John Wiley & Sons.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE Transactional Information Theory 28*(2), 129–137.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press*, pp. 281Ű297.

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

McCormick, W., P. Schweitzer, et T. White (1972). Problem decomposition and data reorganisation by a clustering technique. *Operation Research 20*(5), 993–1009.

Microsoft, C. (2012). Sql server 2012 performance white paper. Technical report, Microsoft Corporation.

Navathe, S., S. Ceri, G. Wiederhold, et J. Dou (1984). Vertical partitioning algorithms for database design. *ACM Trans. Database Syst. 9*(4), 680–710.

OLAP-Council (1998). Apb-1 benchmark. Technical report, OLAP Council. http ://www.olpacouncil.org/research/resrchly.htm.

Pham, D., S. Dimov, et C. Nguyen (2004). An incremental k-means algorithm. *Journal of Mechanical Engineering Science 7*(218), 783Ű795.

Rao, J., C. Zhang, N. Megiddo, et G. Lohman (2002). Automating physical database design in a parallel database. In *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, New York, NY, USA, pp. 558–569. ACM.

Rehme, R. et N. Bruno (2011). Automated partitioning design in parallel database systems. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, New York, NY, USA, pp. 1137–1148. ACM.

Sacca, D. et G. Wiederhold (1983). Database partitioning in a cluster of processors. In *Proceedings of the 9th International Conference on Very Large Data Bases*, San Francisco, CA, USA, pp. 242–247. Morgan Kaufmann Publishers Inc.

Sanjay, A., V. Narasayya, et B. Yang (2004). Integrating vertical and horizontal partitioning into automated physical database design. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 359–370.

Valduriez, P. (1993). *Parallel database systems : open problems and new issues*. Hingham, MA, USA : Kluwer Academic Publishers.

Valduriez, P. et M. Özsu (1999). *Principles of Distributed Database Systems : Second Edition*. New Jersey : Prentice Hall.

Zhang, T., R. Ramakrishnan, et M. Livny (1996). Birch : an efficient data clustering method for very large databases. *SIGMOD Rec. 25*(2), 103–114.

Summary

Horizontal Partitioning (HP) is a widely used technique for improving the physical design of data warehouses. It offers advantages in terms of performance and manageability without consumption of extra storage space. However, selecting a HP schema is NP-complete problem. Thus, several algorithms have been proposed to manage the complexity of the problem and generate fragmentation schema with good quality. These works consider as input a workload of most used queries. However, the majority of the work does not take into account the scaling concerning the size of the workload which can be very important. We propose in this paper a scalable approach based on classification and election for HP supporting large workloads. We conducted an experimental study on the ABP-1 benchmark to test the effectiveness and scalability of our approach. The results are encouraging.

Evaluation et comparaison de systèmes de recommandation

Latifa Baba-Hamed*, Lamia Ouardas*

*Laboratoire RIIR, Université d'Oran
lbabahamed@yahoo.fr

Résumé. Avec l'avènement d'internet, et l'avancement rapide des technologies, l'utilisateur se voit confronté à une surcharge d'informations, d'où la difficulté d'accès aux données pertinentes. Le filtrage d'informations est apparu afin de remédier à cette problématique en donnant naissance aux systèmes nommés *Systèmes de recommandation (SR).* Ce papier compare et évalue trois processus de recommandation en termes de précision. Le premier processus utilise le filtrage basé sur le contenu. Le second considère une hybridation du filtrage basé sur le contenu et du filtrage collaboratif. Quant au troisième, il tient compte des préférences négatives.

1 Introduction

De nos jours, nul ne doute de l'omniprésence de l'internet dans la vie quotidienne des gens. Son utilisation est devenue si grande que les utilisateurs se trouvent souvent perdus devant la masse de contenus qui se présente à eux. De ce fait, les systèmes de recommandation (SR) se sont imposés comme incontournables pour palier au problème de surcharge d'informations en fournissant aux utilisateurs les documents les plus pertinents. Ces systèmes se basent sur les préférences des utilisateurs pour filtrer les informations. L'ensemble de ces préférences constituent ce que l'on appelle le profil de l'utilisateur.

Les travaux dans le domaine des SR se divisent en quatre catégories : les SR classiques, les SR sensibles au contexte, les SR utilisant la diversité, et les SR bipolaires (utilisant les préférences positives/négatives). La première catégorie concerne les premières études sur les SR et englobe les systèmes FBC, les systèmes FC, et les systèmes hybrides. Dans les systèmes FBC Markov et Ivanova (1984), Shoval et al. (2007), les descripteurs de contenus sont directement appariés avec les profils utilisateurs pour estimer leurs utilités. Seuls les contenus ayant un score d'utilité important seront recommandés aux utilisateurs. L'idée principale des systèmes FC Breese et al. (1998), Herlocker (1999) est de recommander, à un utilisateur, les contenus consommés (appréciés) par les utilisateurs qui lui ressemblent. Ces systèmes n'ont besoin d'aucune description de contenu. Les systèmes hybrides Burke (2002), enfin, combinent les techniques du filtrage basé sur le contenu et celles du filtrage collaboratif pour améliorer la pertinence des recommandations.

La seconde catégorie introduit de nouveaux éléments dans le processus de recommandation tels que : l'heure à laquelle la recommandation est demandée, l'endroit où se trouvent les utilisateurs, le média à travers lequel la recommandation est consommée, etc. Cela mène à la définition des SR sensibles au contexte Abowd et al. (1999), De Carolis et al. (2009), Cantador et al. (2009), Domingues (2009). Une classification des principaux systèmes de cette catégorie est résumée dans Soltani et al. (2012). La troisième est consacrée à la diversi-

fication dans les recommandations Vee et al. (2009), Adomavicius et al. (2011), Angel et Koudas (2011), Yu et al. (2009). Cette catégorie de travaux a pour objectif de proposer à l'utilisateur non seulement l'ensemble des contenus les plus pertinents mais aussi les plus diversifiés possibles. Cela permet d'augmenter la satisfaction des utilisateurs et de diminuer la redondance des contenus. Enfin, la quatrième catégorie s'occupe de la prise en compte des préférences négatives en plus des préférences positives Koutrika (2005), Chao et al. (2005), Lee et Brusilovsky (2009), Bitarelli et al. (2007). Nous avons montré, dans l'un de nos travaux précédents, l'impact positif obtenu suite à l'utilisation de telles préférences Baba-Hamed et al. (2012).

Ce papier a une relation avec la première et la quatrième catégorie de travaux. Il s'agit de recommander des articles, à un utilisateur actif de différentes manières:

- par l'application du filtrage basé sur le contenu (FBC). Pour faire le matching, ce filtrage utilise dans un premier temps une similarité numérique, puis une similarité sémantique et enfin une combinaison linéaire de ces deux dernières.
- Recommander à l'utilisateur actif des produits pertinents obtenus par l'application du FBC en tenant compte des préférences positives et négatives. Les trois mesures de similarité citées ci-dessus sont également utilisées dans ce cas.
- Recommander à l'utilisateur actif des produits pertinents obtenus par l'application du filtrage hybride (combinaison du FBC et du FC).
- Puis d'évaluer et de comparer ces systèmes en termes de précision par la méthode des top-k (les k premiers produits de la liste recommandée).

Nous nous sommes intéressés, pour notre étude, au domaine cinématographique. L'évaluation a été réalisée en considérant un benchmark extrait des deux bases de données sur les films : IMDB qui fournit les caractéristiques des films (titre, genre, acteur, réalisateur,…etc.), et MOVIELENS qui fournit les notes attribuées par des utilisateurs à certains films (les notes sont comprises entre 1 et 5). Le reste de ce papier est organisé comme suit. La section 2 est consacrée à la définition de quelques concepts de base qui vont aider à comprendre la suite de l'article. La section 3 présente l'architecture englobant les différents systèmes considérés. Une évaluation en termes de précision et une comparaison de ces systèmes sont données dans la section 4. Enfin, la section 5 conclut cet article et présente quelques perspectives.

2 Concepts de base

Cette section inclut les concepts fondamentaux des SR. Elle inclut les profils de l'utilisateur, les préférences de l'utilisateur, et les opérateurs de matching.

Profile. Support de collecte et de sauvegarde de l'ensemble des caractéristiques et des goûts spécifiques à chaque utilisateur. Autrement dit c'est un modèle utilisateur et une source de connaissances qui contient des acquisitions sur tous les aspects liés à l'utilisateur. Une représentation multidimensionnelle d'un profil est donnée par Kostadinov (2007).

Préférence. Les préférences constituent une partie intégrale des profils utilisateurs. Une préférence est une expression qui permet de ranger/ordonner des items pour un utilisateur donné. Les préférences peuvent être classifiées selon plusieurs aspects, les plus connues sont les suivantes :

- positive : exprimant l'intérêt que porte l'utilisateur sur un item ou une caractéristique donnés (exemple : j'aime les plats du restaurant Calipso) Koutrika (2005);
- négative : exprimant le rejet d'un à l' égard d'une caractéristique ou d'un item donnés (exemple : je n'aime pas, ou une valeur négative) Bitarelli et al. (2007) ;
- quantitative (numérique) : exprime les préférences de façon numérique comme des scores ou des notes dans un intervalle donné Bitarelli et al. (2007) ;
- qualitative : préférence exprimée par des opérateurs binaires (exemple : fruits de mer >p viande) Bitarelli et al. (2007) ;
- contextuelle : la préférence est valide dans un contexte spécifique seulement. Par exemple, pour le contexte acteur dans le domaine des films : j'aime Julia Roberts dans les films du genre « Romance » Abbar et al. (2010), Soltani et al. (2012);
- non contextuelle : préférence ne dépendant d'aucun contexte ;
- simple : préférence pour un seul produit (exemple : j'ai une forte préférence pour les films du genre « Drame » Kießling (2002);
- complexe : une composition (conjonction/disjonction) de préférences simples ;
- conditionnelle : la préférence est valide uniquement si une ou plusieurs conditions sont satisfaites ;
- non conditionnelle préférence ne dépendant d'aucune condition ;
- implicite : préférence déduite par le système suivant les actions commises par l'utilisateur (comme le nombre de clic, le temps de lecture, fréquence, etc.) ;
- explicite : préférences collectées interactivement avec l'utilisateur via un formulaire (collecte directe des préférences).

Opérateurs de matching. Au cœur de la plupart des systèmes de recommandation, nous trouvons un opérateur de matching qui mesure la similarité de deux profils utilisateurs, de deux descripteurs de contenu ou bien la similarité entre un profil utilisateur et un descripteur de contenu. Comme les profils utilisateurs et les descripteurs de contenu sont souvent modélisés avec des vecteurs de mots clés pondérés, seules les mesures vectorielles comme Cosinus et corrélation de Pearson sont utilisées. Or, l'avènement du web sémantique et le développement des ontologies ont mis à notre disposition une panoplie de mesures de similarité sémantiques qui peuvent compléter les mesures vectorielles citées ci-dessus. Une classification de ces mesures est donnée dans Baba-Hamed et al. (2011).

3 Architecture générale du système

Comme nous l'avons déjà mentionné plus haut, nous nous proposons de recommander des items à un utilisateur donné par application de différentes techniques de filtrage (filtrage basé sur le contenu, filtrage basé sur le contenu en considérant les préférences négatives et filtrage hybride en combinant le FBC et le FC), et, de comparer les processus de recommandation correspondant à ces techniques entre eux en termes de précision (métrique très connue dans le domaine de la recherche de l'information). A cet effet, nous avons assemblé, dans un même système, ces différents sous-systèmes de filtrage. Nous consacrons cette section pour détailler les composants de ce système. A titre illustratif, nous avons choisi le domaine cinématographique (autrement dit, nous avons opté pour les films comme contenus), mais le même procédé reste valable pour un autre type de contenus comme : les livres, les articles de recherche, les restaurants, les chansons, etc.

L'architecture de notre système, est composée de sept modules principaux : le module conception de l'ontologie, le module qui réalise la recommandation basée sur le contenu (FBC classique ou pur), le module qui réalise le FBC tenant compte des préférences négatives (FBC Pos-Neg), le module qui réalise le filtrage hybride, le module de construction du descripteur de contenu, le module de construction du profil utilisateur et enfin le module qui fait l'affichage des recommandations et l'affichage des courbes d'évaluation (voir figure 1).

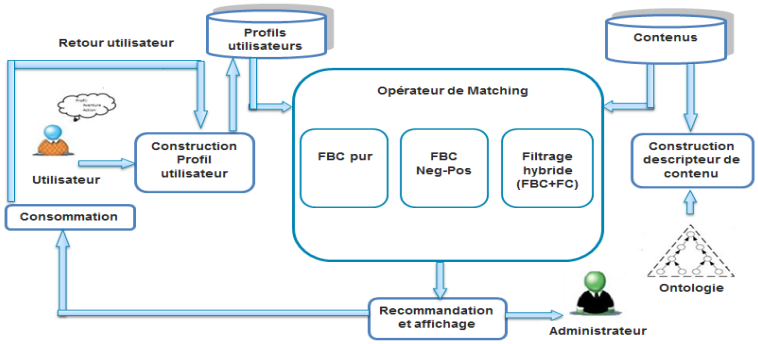


FIG. 1 – *Architecture générale du système.*

3.1 Module conception de l'ontologie

Notre système est un système de recommandation ontologique utilisant les trois types de filtrage. Afin d'augmenter le taux de pertinence des produits recommandés, plusieurs mesures de similarité (sémantique et numérique) ont été implémentées dans ce système. L'intérêt de la conception de l'ontologie dans le cas de notre système réside dans le calcul de la similarité sémantique et le calcul de la similarité sémantique positive et négative (que nous détaillerons plus loin). Ce module doit utiliser une ontologie de domaine. Dans notre cas d'étude, nous avons utilisé l'ontologie de films construite par Baba-Hamed et al. (2010).

3.2 Module Construction du descripteur de contenu

Un contenu est caractérisé par plusieurs propriétés. Certaines de ces propriétés constituent le descripteur du contenu. Dans notre cas, le contenu qui est un film est caractérisé par les propriétés (*titre, réalisateur, acteurs, genre, etc.).* Le descripteur que nous avons choisi se limite aux concepts *titre* et *genre*. Il est représenté par un vecteur dont les colonnes sont les genres des films et chaque cellule peut prendre la valeur 1 si le film a pour propriété le genre exprimé en colonne, sinon elle prend la valeur 0. Les genres que nous avons retenus sont : *Adventure, Action, Fantasy, Sci-Fi, War, Western, Biography, History, Drama, Comedy, Musical, Romance, Family, Animation, Sport, Thriller, Crime, Horror,* et *Mystery*.

3.3 Module Construction du profil de l'utilisateur

Sans perte de généralité, nous définissons le profil utilisateur comme étant un ensemble de préférences Pu = {p1, . . . , pn}. Chaque préférence *pi* est un couple de prédicat et de

poids (*pri, wi*) dans lequel le poids *wi* exprime le degré d'intérêt du prédicat *pri* pour l'utilisateur *u*.

Dans notre système, le profil utilisateur peut être explicite ou implicite. On parle de profil explicite lorsque c'est l'utilisateur qui renseigne ses préférences via notre plateforme d'acquisition. Ceci se fait en attribuant des poids d'importance aux différents genres de film reconnus par notre système. Lorsque le profil est directement calculé à partir des consommations (retour ou feedback) d'un utilisateur, alors on dit qu'il est implicite. Les préférences implicites sont alors calculées comme suit : si un utilisateur *u* consomme deux films d'*Action* avec les notes respectives de 5 et 4, alors la préférence de cet utilisateur pour le genre *Action* est égale à 4.5 (= (5+4)/2). Une normalisation des préférences de l'utilisateur s'avère nécessaire avant de les utiliser pour le calcul de la prédiction. Cette normalisation consiste à mettre des préférences comprise entre 1 et 5 dans un intervalle [0,1].

Afin de pouvoir appliquer une mesure de similarité vectorielle sur notre profil utilisateur, nous avons défini une représentation vectorielle des profils comme suit : soient *P* l'ensemble des préférences d'un profil et *Op* un ordre arbitraire mais fixé sur les préférences qui appartiennent à *P*. On se réfère au ième élément de *P* suivant l'ordre *Op* par *P[i]*. La représentation vectorielle du profil *P* est un vecteur numérique (réel) *V* de dimension *N* où l'ième élément de *V* représente le poids de la préférence *P[i]* Baba-Hamed et al. (2011).

Afin de prendre en compte les préférences négatives, nous pouvons scinder ce profil en deux sous-profils : le profil positif $Pu^+ = \{p1^+, \ldots, Pn^+\}$ qui contient toutes les préférences dont le poids est supérieur ou égal à 0,5 (il représente les préférences positives de l'utilisateur) et le profil négatif $Pu^- = \{p1^-, ..., Pn^-\}$ qui contient toutes les préférences dont le poids est inférieur à 0,5 (il représente les préférences négatives de l'utilisateur).

3.4 Modules de la recommandation par le FBC pur et le FBC Pos-Neg

Ces deux modules (FBC pur et FBC Pos-Neg) font l'appariement entre un descripteur de contenu (film dans notre cas) et un profil utilisateur.

Le matching se base sur un calcul de similarité. Nous avons utilisé la similarité globale présentée dans un travail antérieur Baba-Hamed (2011). Cette mesure combine linéairement la mesure de similarité sémantique de Jiang & Contrath (qui est une mesure sémantique hybride) avec la mesure de similarité vectorielle de Pearson, afin d'améliorer et d'augmenter la pertinence des réponses fournies à l'utilisateur. En effet, cette nouvelle mesure de similarité (nommée $Sim_{globale}$) se fait sur trois niveaux simultanément : la position des concepts dans la hiérarchie de l'ontologie (position des genres de film dans l'ontologie des films), le contenu informationnel des concepts (probabilité d'apparition d'un genre de film), et le poids attribué par l'utilisateur à ses préférences. Nous rappelons sa formule ci-dessous :

$$Sim_{globale}(Pu, C) = \alpha \times sim_{sém}(Pu, C) + \beta \times sim_{préf}(\overrightarrow{Pu}, \overrightarrow{C}) \quad (1)$$

Avec la somme des deux coefficients égale à 1 ($\alpha+\beta = 1$). $sim_{sém}$ représente la similarité sémantique entre le profil et le contenu en utilisant la mesure de Jiang & Contrath. $sim_{préf}$ représente la similarité obtenue en appliquant la corrélation de Pearson aux représentations vectorielles du profil et du contenu.

Etant donné que le profil utilisateur et le descripteur de contenu contiennent des ensembles de concepts ontologiques (ex. ProfilAli{(Action, 5), (Guerre, 4), (Thriller, 5)} et

Film1{Aventure, Policier}), et que la similarité sémantique de Jiang & Contrath calcule la similarité entre deux concepts seulement, nous avons proposé une mesure de similarité ensembliste basée sur celle de Jiang & Contrath qui calcule la similarité sémantique moyenne entre deux ensembles de concepts. Sa formule est la suivante :

$$sim_{sém}\left(P\vec{u},\vec{C}\right)=\frac{1}{M\times N}\sum_{j=1}^{m}\sum_{i=1}^{n}sim_{Jiang\&Contrath}(P\vec{u}[j],\vec{C}[i]) \qquad (2)$$

Où: $P\vec{u}$ et $\vec{C}$ sont les représentations vectorielles respectives du profil utilisateur et du descripteur de contenu, M et N représentent, respectivement, le nombre de concepts ontologiques dans le profil et dans le descripteur de contenu.

Pour décider si un contenu est à recommander ou pas à un utilisateur actif, il suffit de décider d'un score seuil de similarité globale (par exemple 0.5) en-dessous duquel un contenu n'est pas à recommander.

Concernant le calcul de la similarité globale pour la recommandation basée sur le FBC utilisant les préférences positives et négatives (FBC Pos-Neg), comme nous avons scindé le profil utilisateur en profil positif P^+ et profil négatif P^-, nous pouvons calculer donc :

- la similarité sémantique positive ($sim_{sém}^{+}$) : qui représente la similarité sémantique entre P^+ et le descripteur du contenu ;
- la similarité sémantique négative ($sim_{sém}^{-}$) : qui représente la similarité sémantique entre P^- et le descripteur du contenu ;
- la similarité numérique positive ($sim_{préf}^{+}$) : qui représente la corrélation entre P^+ et le descripteur du contenu en utilisant la mesure de Pearson ;
- la similarité numérique négative ($sim_{préf}^{-}$) : qui représente la corrélation entre P^- et le descripteur du contenu en utilisant la mesure de Pearson ;
- la similarité globale positive ($sim_{globale}^{+}$): qui combine les deux similarités positives selon la formule (1) en remplaçant Pu par Pu^+, $sim_{sém}$ par $sim_{sém}^{+}$ et $sim_{préf}$ par $sim_{préf}^{+}$;
- la similarité globale négative ($sim_{globale}^{-}$): qui combine les deux similarités négatives selon la formule (1) en remplaçant Pu par Pu^-, $sim_{sém}$ par $sim_{sém}^{-}$ et $sim_{préf}$ par $sim_{préf}^{-}$.

Pour déterminer les films pertinents, nous testons si la similarité globale positive est supérieure à la fois à un seuil (0.5 par exemple) et à la similarité globale négative. Dans le cas affirmatif, nous concluons que le film est pertinent. Il n'est pas à recommander sinon.

Concernant les coefficients α et β, nos expérimentations répétées sur plusieurs utilisateurs ont montré que α=0.75 et β=0.25 donnent de meilleurs résultats.

3.5 Module du filtrage hybride

Ce module réalise une hybridation entre les deux stratégies de filtrage FBC et FC. Il fait appel, dans un premier temps, au module FBC pur et reçoit une première liste (L1) des contenus à recommander. Ensuite, il utilise le filtrage collaboratif classique (FC pur) et reçoit une deuxième liste (L2). Enfin, il fait l'union des deux listes L1 et L2. La liste résultant de cette union constitue la liste des recommandations présentée à l'utilisateur actif.

Le filtrage collaboratif peut se faire en appliquant les méthodes basées sur la mémoire (comme la méthode des plus proches voisins KNN) ou les méthodes basées sur un modèle

(comme le clustering). Ces dernières sont beaucoup plus utilisées que les précédentes vu les avantages qu'elles offrent notamment en termes de temps de traitement.

Nous avons utilisé, dans notre système, le filtrage collaboratif basé sur un modèle qui est le clustering. Le traitement est effectué en deux étapes: la formation des communautés et la production des recommandations.

Formation de communautés. La formation des communautés est un processus qui consiste à créer des clusters d'utilisateurs positivement corrélés. Pour exécuter ce processus nous utilisons l'algorithme initial de k-means (k-moyennes) qui prend comme paramètres le nombre de clusters $K = 5$, le nombre d'itérations n et l'ensemble des utilisateurs.

Production de recommandations. Cette étape consiste à recommander à l'utilisateur actif, les contenus qui ont été appréciés par ses voisins dans la communauté.

3.6 Module Recommandations et Affichage

Ce module s'occupe du classement et de l'affichage de la liste des contenus envoyée par l'un des trois modules cités précédemment (FBC pur, FBC utilisant les préférences positives et négatives et le filtrage hybride). Il a, également, pour tâche d'évaluer la qualité des processus de recommandation correspondant à chacun des sous-systèmes en termes de précision, et, d'afficher les courbes. Nous détaillerons cette seconde tâche dans la section suivante.

4 Evaluation et comparaison

Comme nous l'avons souligné précédemment, l'évaluation du système s'est basée sur la plateforme de test réalisée au cours du projet APMD (Accès Personnalisé à des Masses de Données) à partir de laquelle un benchmark constitué de 21 utilisateurs qui ont regardés les mêmes 100 films a été extrait. De ce benchmark, nous avons extrait une partie (training set) pour l'apprentissage des profils des utilisateurs (ensemble de préférences). L'autre partie (result set) a servi à évaluer le processus de recommandation.

Cette évaluation a nécessité la création de tables dont les schémas sont les suivants :

T-usernote (userId, movieId, note)

Training-set (userId, movieId, genre, rating)

Profil-user (userId, genre, préférence)

Profil-user-positive (userId, genre, positive-preference)

Profil-user-negative (userId, genre, negative- preference)

Result-set (userId, movieId, note, predictionFBC, prédictionHyb, predictionFC, predictionglob-Pos-Neg)

Où *userId* représente l'identificateur de l'utilisateur, *movieId* représente l'identificateur du film, *note* représente la note qu'a attribuée l'utilisateur à un film qu'il a regardé, *genre* représente le genre d'un film, *préférence* représente la préférence (poids) que peut avoir un utilisateur pour un genre de film donné. Ce poids est obtenu implicitement par un calcul. *Positive-preference* représente la préférence positive, *négative-preference* représente la préférence négative, *predictionFBC, prédictionHyb, predictionFC,* et *prédictionglob-Pos-Neg* représentent, respectivement, la valeur de la similarité globale calculée par les sous-systèmes de recommandation FBC pur, hybride, FC et FBC utilisant les préférences négatives et positives.

La table T_usernote contient 21 utilisateurs ayant regardé les mêmes 100 films ce qui permet de créer des profils de bonne qualité. A partir de la table T_usernote, nous avons rempli les tables Training_set et Result_set de façon aléatoire en utilisant une procédure aléatoire que nous avons implémentée. Cette procédure permet de donner les répartitions suivantes : {(70,30) ;(60,40) ;(50,50)…etc.}. Par exemple, pour le premier couple (70,30), le premier composant constitue les 70 films qui servent à remplir la table Training_set (correspondant à la partie Training set) et le second, c'est-à-dire, les 30 restants parmi les 100 films initiaux, constitue l'ensemble des films qui servent à remplir la table Result_set (correspondant à la partie test).

La table Training_set, pour cet exemple contiendra 21 utilisateurs, 70 films pour chaque utilisateur, les genres associés à chacun des films et la note du film propagée sur les genres du film. La table Result-set, pour la répartition (70, 30) contiendra 21 utilisateurs, 30 films pour chaque utilisateur, les notes associées à chacun des films et les colonnes prédiction qui servent pour le calcul de l'utilité de chaque film pour chaque utilisateur selon la stratégie de filtrage adoptée. Cette utilité sera donnée par la similarité entre le profil utilisateur et le descripteur de chaque film dans le cas du FBC et FBC Pos-Neg, et par la prédiction centrée utilisateur (voir formule 3) dans le cas du filtrage FC. Ces colonnes sont initialement vides.

$$Pred\ (u_l, i_j) = \bar{X}_l + \frac{\sum_{v=1}^{k} r(u_l, u_v)(X_v^j - \bar{X}_v)}{\sum_{v=1}^{k} | r(u_l, u_v) |} \quad (3)$$

Avec :

Pred(u_l , i_j) : est la prédiction de l'item i_j pour l'utilisateur u_l.

$\overline{x_l}$ (respectivement $\overline{x_v}$) est la moyenne des notes attribuées par l'utilisateur u_l (respectivement u_v) à tous les items correspondants.

r(u_l, u_v) représente la corrélation de Pearson entre l'utilisateur actif u_l et son voisin u_v.

X^j_v : est la note de l'utilisateur *v* sur l'item *j*.

La table Profil-user permet de générer le profil d'un utilisateur à partir de la table Training-set. Elle contient les 21 utilisateurs, les genres des films tirés de la table Training-set correspondant à une répartition aléatoire donnée, et la valeur de la préférence d'un genre pour un utilisateur donné qui est obtenue par le calcul de la moyenne des notes de ce genre dans la table Training-set.

Table user-profile-positive (respectivement user-profile-negative) permet de générer le profil positif (respectivement le profil négatif) d'un utilisateur à partir de la table Training-set. Elle contient 21 utilisateurs, les genres de films extraits à partir de la table Training-set correspondant à une distribution aléatoire donnée, et la valeur de la préférence positive (respectivement préférence négative) d'un genre pour un utilisateur donné et qui est obtenue en calculant la moyenne des notes supérieures ou égales à 3 (respectivement inférieures strictement à 3) de ce genre dans la table Training-set.

Pour évaluer la précision de chacun des sous-systèmes de recommandation, nous avons mis en place le scénario d'extraction des TopK films les plus intéressants pour un utilisateur. Pour calculer la précision nous procédons comme suit :

Pour une partition (Training set/Result set) et un utilisateur donnés, on procède, dans une première étape, à la prédiction du score d'intérêt qu'aurait donné l'utilisateur à chaque film appartenant au Result set. Ces scores sont calculés par l'opérateur de matching correspondant à chacun des sous-systèmes de recommandation. A la fin de cette étape, les films appartenant

au Result set sont classés en fonction de leurs scores prédits. La seconde étape consiste à extraire les K premiers films (ceux ayant obtenus les meilleurs scores) et les mettre dans l'ensemble *TopKprédits*. Dans une troisième étape, les films de l'ensemble Result set sont classés en fonction des vraies notes données explicitement par l'utilisateur (données de MovieLens intégrées dans la plateforme d'APMD). Les K films, ayant reçus les meilleures notes de notre utilisateur, sont mis dans l'ensemble *TopKréels*.

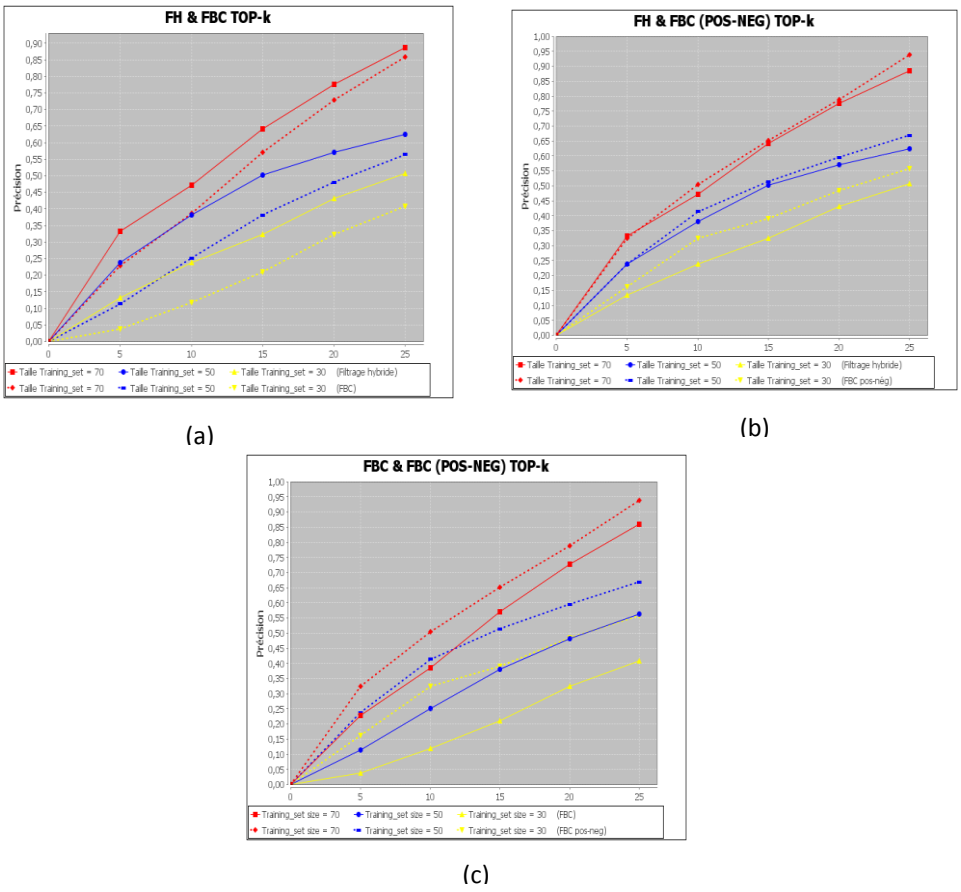


FIG. 2 – *Comparaison des sous-systèmes FBC, FH, et FBC Pos-Neg en termes de précision.*

La dernière étape consiste à calculer la précision qui est donnée par la proportion des films correctement prédits comme faisant partie des topK films pour l'utilisateur donné selon la formule 4.

$$Prévison = \frac{TopK\ Prédits\ \cap TopK\ réels}{K} \quad (4)$$

On répète ce procédé pour chacun des utilisateurs. La moyenne de ce calcul, pour tous les utilisateurs, nous délivre la valeur de la précision pour la valeur de K. La variation du nombre K nous permet d'obtenir d'autres valeurs de précision qui vont nous permettre de tracer la courbe de précision.

Les différentes valeurs de K considérées sont : 5, 10, 15, 20, 25 films. Nous avons représenté les résultats expérimentaux par des courbes. Chaque courbe donne la valeur de la précision en fonction de la variation de K pour une taille de Training-set précise. Les tailles de Training-set considérées sont : 30, 50, 70.

D'après les courbes de la figure 2, nous remarquons que le système obtient une meilleure précision lorsque la valeur de K est grande. Ceci s'explique par la difficulté qu'a le système à trouver le Top meilleur film.

Nous remarquons également, que la précision des prédictions (ou des recommandations) augmente en fonction de la taille de la table *Training-set*. Plus la taille *Training-set* augmente plus la précision est meilleure. Cette observation est importante, car elle nous permet de penser que notre approche de calcul de similarité implémentée dans l'opérateur de matching permet au système de recommandation de converger vers une meilleure précision dans le temps.

La figure 2 permet de comparer les sous-systèmes considérés entre eux et de faire les interprétations suivantes :

- les courbes de précision obtenues dans le filtrage hybride se trouvent au-dessus de celles obtenues par le filtrage basé sur le contenu, quelle que soit la taille de la table Traning_set et le nombre K. Nous en déduisons que la recommandation basée sur le filtrage hybride est meilleure que celle qui est obtenue par le filtrage basé sur le contenu (voir Fig. 2.a).
- La précision dans le FBC utilisant les préférences positives et négatives (FBC Pos-Neg) est supérieure à la précision dans le filtrage hybride, quelle que soit la taille de la table Traning_set et le nombre K. Nous en déduisons que la recommandation basée sur le FBC qui tient compte des préférences positives et négatives est meilleure que celle qui est obtenue par le filtrage hybride (voir fig. 3.b).
- La précision dans le FBC Pos-Neg est supérieure à la précision obtenue par le FBC pur, quelle que soit la taille de la table Traning_set et le nombre K. Par conséquent, nous pouvons conclure que les systèmes basés sur le FBC et utilisant les préférences négatives et positives fournissent une meilleure recommandation que celle fournie par les systèmes basés sur le FBC pur (voir fig. 3.c).

5. Conclusion

Dans ce projet, Nous avons réalisé un système qui rassemble les différents types de filtrage, à savoir le FBC pur, le FBC positif et négatif, le FC pur, et le filtrage hybride, dans le but de recommander des produits de différentes manières à l'utilisateur actif. Nous avons également évalué ces différents sous-systèmes, mesuré leur qualité en termes de précision par la méthode des TOP(K), et comparé ces évaluations pour déterminer le type de filtrage le plus efficace. La comparaison des résultats de l'évaluation des différents sous-systèmes en termes de précision a révélé que les recommandations obtenues par application du FBC Pos-Neg sont meilleures que celles obtenues par le filtrage hybride qui sont à leur tour meilleures que celles fournies par le FBC pur.

Nous pensons apporter des améliorations au système réalisé en envisageant quelques perspectives comme :

- l'évaluation du système avec d'autres métriques, comme la couverture ou bien la confiance, car il est clair que la précision à elle seule n'est pas suffisante pour mesurer la qualité d'un système de recommandation.
- L'intégration, dans l'architecture générale, des modules réalisant la contextualisation et la diversité respectivement.
- Apporter une amélioration à l'algorithme du filtrage collaboratif en adoptant une version plus performante de l'algorithme du k-means.

Références

Abbar, S., M. Bouzeghoub, and S. Lopes (2010). Service-Based Context-Aware Recommender System. *Actes des 26èmes Journées Bases de Données Avancées (BDA'10)*. Toulouse.

Abowd, G. D., A. K. Dey, P. J. Brown, N. Davies, M. Smith, and P. Steggles (1999).Towards a better understanding of context and context-awareness.*1st international symposium on Handheld and Ubiquitous Computing (HUC'99)*. London: Springer.

Adomavicius G., Y. OkKwon (2011).Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques. *IEEE Data Eng*. 1041–4347.

Angel, A., N. Koudas (2011).Efficient Diversity-Aware Search. *International conference on Management of data (SIGMOD '11*). New York, 781–792.

Baba-Hamed, L., R. Soltani, et K. Sabri (2010). Construction d'une ontologie pour la recommandation de films à un utilisateur. *Actes des Ateliers des 21es Journées Francophones d'Ingénierie des Connaissances (IC'2010)*. Nîmes.

Baba-Hamed, L., S. Abbar, R. Soltani, et M. Bouzeghoub (2011). Elaboration et Evaluation d'un Système de Recommandation Sémantique.*1st international Conference on Information Systems and Technologies (ICIST'11).* Tébessa, 515–523.

Baba-Hamed, L., S. Abbar, et A. Haouari (2012). The impact of negative preferences on a recommendation process. *3rd International IEEE Conference on Multimedia Computing and Systems (ICMCS'2012)*.Tangier, Morocco.

Bitarelli, S., M. S. Pini, F. Rossi, et K. Brent (2007).Venable Positive and negative preferences. Research report. Institute of computer science and telecommunication. CNR Pisa. Italy.

Breese, J. S., D. Heckerman, et C. Kadie (1998).Empirical analysis of predictive algorithms for collaborative filtering.Technical report, MSR-TR-98-12. Microsoft research, Redmond, WA 98052.

Burke R. (2002). Hybrid Recommender Systems: Survey and Experiments. *Journal of Personalization Research, User Modeling and User-Adapted Interaction*, 14: 331–370.

Cantador, I., and P. Castells (2009). Semantic Contextualization in a News Recommender System. *Cars*.

Chao, D. L., J. Balthrop, and S. Forrest (2005). Adaptive Radio: Achieving Consensus Using Negative Preferences. *GROUP'05*. Florida, USA.

De Carolis, B., I. Mazzotta, N. Novielli and V. Silvestri (2009). Using Common Sense in Providing Personalized Recommendations in the Tourism Domain. *Cars*.

Domingues, M., A. Mário Jorge, and C. Soares, Using Contextual Information as Virtual Items on Top-N Recommender Systems. *Cars*.

Herlocker, J. L., A. J. Konstan, A. Borchers, and J. Riedl (1999). An Algorithmic Framework for Performing Collaborative Filtering. *22th International ACM Conference on Research and Development in Information Retrieval (SIGIR'99)*, 230–237.

Kießling, W. (2002). Foundations of preferences in database systems. *28th Conference on Very Large Data Bases*, Hong Kong, China, 311–322. Morgan Kaufmann.

Kostadinov, D. (2007). *Personnalisation de l'information et gestion de profils utilisateur*. Thèse de doctorat, Université de Versailles Saint-Quentin-en-Yvelines.

Koutrika, G., and Y. E. Ioannidis (2005).Personalized Queries under a Generalized Preference Model.*21st International Conference on Data Engineering (ICDE 2005*). Tokyo, Japan, 841–852.

Danielle H. Lee, D. H., and P. Brusilovsky (2009). Reinforcing Recommendation Using Implicit Negative feedback. G.-J. Houben et al. (Eds.): UMAP 2009, LNCS 5535, 422–427. Berlin: Springer-Verlag.

Markov, K. and K. Ivanova (2007). An ontology-content-based filtering method. *Fifth International Conference "Information Research and Applications*, Varna, Bulgaria.

Soltani, R., L. Baba-Hamed, et S. Abbar (2012). Contextualisation des Préférences pour les Systèmes de Recommandation. *2nd International Conference on Information Systems and Technologies (ICIST'12),* Sousse, Tunisia.

Shoval, P., V. Maidel, and B. Shapira (2007). An Ontology-Content-Based Filtering Method. *I.Tech-2007, Information Research and Applications*.

Vee, E., J. Shanmugasundaram and S. Amer-Yahia (2009). Efficient Computation of Diverse Query Results. *IEEE Data Eng*., 32: 57–64.

Yu, C., and Laks V. S., Lakshmanan, S. Amer-Yahia (2009). It takes variety to make a world: diversification in recommender systems. *EDBT*. 368–378.

Summary

With the advent of Internet and the rapid development of web technologies, the user is faced with information overload problems, making it difficult to access relevant data. Filtering information appeared to overcome this problem by giving rise to systems called Recommendation Systems (RS). This paper compares and evaluates three recommendation processes in terms of precision. The first process uses the content-based filtering. The second considers a combination of the content-based filtering and the collaborative filtering. The third one takes into account the negative preferences.

Comparaison des approches de sécurité dans les WebHouses

Salma DAMMAK*, Faiza GHOZZI JEDIDI **, Faiez GARGOURI***
Laboratoire Mir@cl- ISIMS
Université de Sfax ISIMS, BP 242, 3021, Sfax, Tunisie

*damak.salma@gmail.com
**faiza_jedidi@yahoo.fr
***faiez.gargouri@isimsf.rnu.tn

Résumé. L'avènement du Web dans les systèmes décisionnels nécessite la gestion de très importants volumes de données. Dans ce contexte, deux points critiques sont à discuter : le besoin d'opter pour une nouvelle architecture pour WebHouse et la sécurisation des informations. Cet enjeu est accentué par l'apparition du Web dans les systèmes décisionnels avec la diversification et la multiplicité des utilisateurs et des communications qui augmentent les risques d'intrusion. Dans cet article, nous proposons un état de l'art et une étude comparative des différentes propositions dans le domaine de la conception et la sécurisation des entrepôts Web.

1 Introduction

Face à l'évolution des quantités d'informations et des sources Web qui ne cessent d'engendrer des flux de données de plus en plus complexes, de nouvelles technologies de stockage, d'accès et de traitement de l'information sont définies. Dans le domaine décisionnel, l'avènement du Web donne naissance à une nouvelle génération d'entrepôts de données : le WebHouse. L'architecture du WebHouse doit répondre aux nouvelles demandes des décideurs telles que la disponibilité de l'information à tout moment, l'analyse des besoins et la prise de décision rapide. Elle doit supporter des requêtes complexes et coûteuses en termes de temps de traitement et d'espace de stockage.

En outre, l'utilisation d'Internet lors d'un processus d'entreposage (acquisition, stockage et accès) augmente les risques d'attaques qui peuvent nuire à un système décisionnel et le rendent de plus en plus vulnérable. En effet, un WebHouse est caractérisé par la diversification et la multiplicité des utilisateurs et des communications, ce qui augmentent les risques d'intrusion et rend la sécurité un facteur déterminant pour un WebHouse. Cependant, les besoins de sécurité ne sont pas pris en considération lors de la conception des entrepôts de données et la plupart des méthodes de conception ne les intègrent pas. Ces aspects sont traités en aval de la création de l'entrepôt par un ingénieur qui n'a pas une description complète, ni des besoins, ni de l'importance des données et de la façon de les sécuriser.

Ainsi, la sécurité du WebHouse soulève deux problématiques. La première considère la définition d'une nouvelle architecture répartie en tenant compte du contexte Web et donnant accès à un grand nombre de ressources. La deuxième se focalise sur les traitements des besoins non fonctionnels de sécurité d'un WebHouse dès le niveau métier.

Dans cet article, nous dressons l'état de l'art des travaux portant sur ces deux problématiques et nous dégageons les différents défis posés pour le traitement de la sécurité dans les nouvelles architectures du WebHouse.

Cet article comporte, outre l'introduction, trois sections. La première introduit les nouvelles architectures et les technologies utilisées pour assurer la sécurité. Les deux dernières sections définissent les deux axes de notre approche : les nouvelles architectures et les solutions de sécurité pour les WebHouse. Et nous clôturons par une conclusion présentant nos futures recherches.

2 Architectures et standards de sécurité pour les WebHouse

2.1 Nouvelles architectures

L'entreposage Web offre la possibilité de stocker les données en provenance du Web et facilite leur exploitation en tant que sources décisionnelles. Les entrepôts Web utilisent, souvent, les services Web pour le stockage et la présentation des données. L'utilisation des services Web présente plusieurs avantages dont on peut citer l'interopérabilité, l'hétérogénéité et l'auto-descriptivité. Ces services Web, tels qu'ils sont définis par le World Wide Web Consortium (W3C), constituent une implémentation possible des architectures orientées services (SOA). SOA est une architecture d'intégration et de développement de systèmes dans laquelle les fonctionnalités sont regroupés autour de processus d'affaires et offerts sous la forme de services interopérables. Cette architecture permet la communication entre des systèmes qui n'ont pas été conçus dans cette optique, et leur participation conjointe dans des processus d'affaires. L'entreposage de données peut bénéficier de l'architecture SOA avec la possibilité de rejoindre les différentes actions (services) de différents domaines de DW pour créer des applications composites ou des services communs. Dans ce sens, plusieurs travaux ont adopté des architectures SOA : Marotta et al, (2012), Mehedintu et al. (2008), Hernández et al. (2010) et Abrahiem. (2007).
Le volume de données important et le besoin de réponse rapide aux requêtes mènent certains travaux à adopter des architectures réparties lors de l'entreposage de données : Nguyen et al. (2001), Kalnis et al. (2002), Abiteboul et al. (2008), Mehedintu et al. (2008) et Golfarelli et al. (2010).

2.2 Standards de sécurité :

La sécurité de l'information est définie, au sein de la norme ISO, comme la « préservation de la confidentialité, de l'intégrité et de la disponibilité de l'information ». L'ISO a défini des services de sécurité notamment l'authentification, le contrôle d'accès, la non-répudiation et la protection contre l'analyse du trafic. Différents types de mécanismes (chiffrement, signature numérique, …) servent pour assurer ces services.

En outre, des standards de sécurité sont définis pour les architectures Web tel que le langage XACML (eXtensible Access Control Markup Language) pour le contrôle d'accès, la circulation des règles et l'administration de la politique de sécurité des systèmes d'information. XACML est souvent utilisé pour assurer la fonction d'autorisation dans les architectures SOA. Ce langage est basé sur deux services principaux le PEP (Policy Enforcement Point) et le PDP (Policy Decision Point). Le PEP envoie une requête XACML demandant l'autorisation à un agent sur une ressource au PDP et le PDP formule la décision d'autorisation.

Comme politique de sécurité, les plus connues sont les politiques RBAC, MAC et DAC. Dans la sécurisation des architectures SOA de nouvelles politiques sont apparues telles que

Dynamic access control, Delegation of rights, Break Glass, Eyes Principle…. Le fonctionnement de ces politiques les rendent les adaptables aux critères spécifiques du Web tel que le dynamisme et le changement continus des données et des utilisateurs.

3 Nouvelles architectures des WebHouses

Dans cette section, nous présentons une synthèse des travaux proposant de nouvelles architectures pour les entrepôts de données dans le contexte du Web. Pour la modélisation des systèmes Web, plusieurs architectures mettent en œuvre un système d'entreposage ayant comme source des données en provenance du Web.

Dans ce contexte, Ravat et al. (2010) font l'étude de différents travaux adoptant XML comme langage d'entreposage de données. Selon les auteurs, XML est utilisé parfois comme source de données et parfois comme outil de stockage de donnée. Alors que dans d'autres travaux XML est utilisé pour l'interrogation OLAP. Ils classifient les entrepôts en deux types ceux orientés donnée et ceux orientés documents. Dans notre étude, nous sommes intéressés par le premier type d'entrepôt.

Abrahiem. (2007) suggère une nouvelle architecture pour les entrepôts de données. Il propose de combiner l'architecture SOA avec les entrepôts des données quasi-temps réel (near-real-time data warehouse) et donne naissance à une nouvelle architecture des systèmes décisionnels gérée par le Bus de Service de l'Entreprise (ESB : Enterprise Service Bus) qui assure la communication entre l'utilisateur et la couche de gestion des données dans les environnements distribués. Cette combinaison est proposée pour minimiser les coûts, prendre des décisions stratégiques et réutiliser des services.

Au-delà d'un système client/serveur, Hernández et al. (2010) proposent un méta modèle unifié modélisant les logs Web et un ensemble de règles de transformation formelles basées sur le standard QVT (Query/View/Transformation) afin de générer automatiquement le modèle conceptuel multidimensionnel à partir de ces logs. Le méta-modèle des logs Web est divisé en deux packages : un package des entrées et un package d'usage. Le package des entrées représente les entrées de n'importe quel format de logs Web, contenant les métaclasses *EntryField, AuthUser, WebObject, TimeTaken, Referer, Agent et Entry*. Alors que, le deuxième package sert à représenter les réactions des utilisateurs durant une session et produit des entrées dans les logs Web. L'indentification de l'utilisateur (personne ou programme qui utilise le site Web) est un défi dans l'analyse des Log Web, un utilisateur applique un ensemble de tâches dans une session et il peut avoir plusieurs sessions alors qu'une session est relative à un seul utilisateur.

Dans le même contexte, Liu et al. (2010) modélisent les données, analysent et conçoivent une architecture en temps réel des entrepôts de données en se basant sur l'architecture SOA. Ce travail réunit les services Web et les entrepôts de données et résout la problématique d'échange de données avec le serveur. L'entrepôt de données temps réel basé sur SOA, est principalement constitué de trois parties ; les données d'acquisition, les données de gestion et les données d'applications. La conception de l'entrepôt de données est composée de deux parties; les données de capture de changements (CDC : Change Data Capture) pour la capture en temps réel des données mises à jour et le processus ETL (Extract, transformation, chargement) pour le rafraîchissement des données mises à jour rarement. Dans cet article, les données temps réel sont capturées par des déclencheurs. Le processus ETL est basé sur l'architecture SOA, il est composé des Wrapper à l'extrémité de chaque source de données. Ces Wrapper contiennent le module d'extraction des méta-données et gèrent les données Meta.

Une plateforme est proposée par Marotta et al, (2012) offrant un support pour toutes les tâches nécessaires à la réalisation du système WW (Web Warehouse) notamment le processus ETL. Elle repose sur trois principaux processus : i) l'extraction des données, (ii) l'intégration des données et (iii) la transformation et le chargement des données dans le DW. Marotta et al, (2012) proposent le module DSI (Data Service Infrastructure) pour le premier processus ayant trois principales tâches : fournir les mécanismes pour extraire des données des différentes sources de données Web (WDS) en les exposant comme des Services de Données (DS) homogènes utilisant des technologies standards, surveiller les DSs à travaers des mesures de qualité périodiques et fournir des mécanismes d'adaptation et d'exécution. Le deuxième processus est supporté par le module intégrateur qui invoque la DS, réalise l'intégration de données et interagit avec les utilisateurs experts. Ces utilisateurs effectuent les principales décisions d'intégration de données, tels que l'identification d'objets et la résolution des conflits. Enfin, le module intégrateur applique des modules OLAP pour exercer les fonctions d'analyse classiques du contexte DW et effectue la propagation des métadonnées de qualité.

Vu la grande quantité d'informations provenant du Web, le coût élevé des requêtes et le besoin de résultat dynamique et rapide, certains travaux s'orientent vers les architectures distribuées.

Nguyen et al. (2001) proposent une approche de conception et de mise en œuvre d'un système d'entreposage Web basé sur XML. Les auteurs définissent un nouveau méta schéma XML appelé MetaCube-X pour supporter les entrepôts de données distribués. Deux types de modèle sont définies pour le stockage : le modèle MetaCube-X local qui est un méta-modèle pour décrire les données multidimensionnelles, de chaque magasin de données local et un modèle MetaCube- X global qui fournit des informations d'intégration des MetaCube-X locaux. Dans cette architecture, les auteurs définissent un médiateur pour chaque magasin de données local profitant de la médiation qui résout les problèmes d'interopérabilité sémantique (Perez et al. (2008)) et reconnaît l'autonomie et la diversité des entrepôts de données. Un médiateur est un module indépendant situé dans chaque magasin de données local et il supporte les interfaces applicatives flexibles, la réutilisation, la capacité de partager, et la maintenabilité.

Aussi, Mehedintu et al. (2008) définissent un WebHouse comme un système totalement distribué qui fournit les résultats des requêtes via les navigateurs distants et dans des délais raisonnables et qui nécessite une architecture bien distribuée comprenant les petits magasins de données. Selon Mehedintu et al. (2008), le navigateur Web est la clé de diffusion de l'information. Une architecture d'un entrepôt de données Web est présentée, elle consiste à capturer les clickstreams de tous les visiteurs du site Web et à assurer toute les fonctions du DW traditionnels. Avant l'extraction des données, l'approche impose la vérification de l'exactitude des données sources provenant du Web. L'architecture proposée est composée de deux parties : un référentiel des caractéristiques fonctionnelles d'un entrepôt de données et un référentiel des clickstreams décrivant les données Web. Les auteurs proposent d'accéder aux données soit via l'intranet, pour limiter l'accès aux données pour les utilisateurs internes, soit via l'extranet, pour ouvrir les données à des tiers ayant l'autorisation appropriée.

Kalnis et al. (2002) proposent une architecture PeerOLAP pour traiter les requêtes OLAP. Le système est complètement distribué et peut se reconfigurer dans le but de diminuer le coût de la requête et la charge de travail observée. Le réseau PeerOLAP est complémentaire pour les entrepôts de données distribués. Il est composé d'un ensemble de pairs qui accèdent aux entrepôts de données et posent des requêtes OLAP. Chaque pair Pi possède une cache locale et met en œuvre un mécanisme pour la publication de son contenu cache. D'autres pairs peuvent se connecter à Pi et demander des résultats. Pi peut soit répondre à la requête au

niveau local, s'il possède les données nécessaires, soit propager la requête à ses voisins. Dans les deux cas, tous les résultats reviennent directement à l'hôte qui a initié la requête. Le but de PeerOLAP est d'agir comme une cache combinée virtuelle, où tous les composants offrent des ressources visant à atteindre le faible coût de la requête.

D'autre part, Abiteboul et al. (2008) proposent une plate-forme WebContent pour la gestion des référentiels distribués de données Web XML et leur stockage dans des entrepôts de données en se basant sur une architecture pair à pair. La fonctionnalité de stockage est mise en œuvre par les pairs dans le réseau, dont chacun peut devenir responsable de stocker une partie des ressources. Par exemple, dans les ressources de WebContent, les documents XML sont stockés dans un répertoire XML local pour chaque pair, par contre les autres types de fichiers sont stockés directement dans un fichier système local. Pour développer cette plateforme, Abiteboul et al. (2008) se basent sur un langage de composition ActiveXML. Un document ActiveXML est un document XML spécifiant quels services à appeler, comment construire les messages d'entrée, et la façon dont les appels devraient être ordonnés. Pour stocker des ressources XML sur un pair, toute base de données XML peut être utilisée. Ce travail intègre la base eXist et MonetDB / XQuery.

Golfarelli et al. (2010) présentent une architecture pair à pair pour l'entreposage de données dont chaque pair est équipé d'un système d'entrepôt de données indépendant s'appuyant sur un schéma multidimensionnel. Pour améliorer le processus de décision, les utilisateurs accèdent aux données décisionnelles réparties via le réseau en formulant une requête OLAP. Cette requête sera transmise au réseau et reformulée sur les autres pairs en fonction de leurs propres schémas multidimensionnels. Chaque pair traite la requête reformulée et renvoie ses résultats. Enfin, les résultats sont intégrés et renvoyés à l'utilisateur. Golfarelli et al. (2010) décrivent un langage pour la définition des correspondances sémantiques entre les schémas multidimensionnels et les pairs. Le langage proposé exprime la façon dont le schéma local multidimensionnel des pairs cibles est relié au schéma locale multidimensionnel des pairs sources.

Toutes les propositions vues ont comme source de donnée le Web, certaines présentent une architecture dès le niveau conceptuel et certaines présentent juste le niveau physique. Mais Aucun de ces travaux n'abordent le sujet de sécurité lors de l'entreposage Web.

4 Sécurité et Web

Les sources de données hétérogènes, l'utilisation des nouvelles technologies de stockage et de communication via le Web rendent la sécurité des informations du processus décisionnel un défi majeur qui doit être levé afin de protéger ces informations sensibles des utilisateurs non autorisés. La sécurité de l'information sensible est une condition primordiale pour la survie du système nécessitant une réflexion approfondie. Par conséquence, de nombreux travaux ont été consacrés à la recherche de solutions pour remédier à ce problème. Nous allons, dans cette partie, présenter les principaux travaux de recherche qui ont proposé des solutions pour assurer la sécurité, en premier lieu, dans les entrepôts de données et puis, dans les WebHouses.

4.1 Sécurité dans les entrepôts de données

Plusieurs travaux proposent des solutions de sécurisation d'entrepôts de données qui diffèrent en fonction des niveaux d'abstraction traités.

Katic et al. (1998) décrivent un modèle d'entrepôts sécurisés basé sur des métadonnées. Ce modèle assigne une vue des entrepôts de données réduite à chaque groupe d'utilisateurs et limite la portée de leurs requêtes aux données. Dans ce modèle, le chef de la sécurité définit pour chaque utilisateur (groupe d'utilisateurs) des secteurs de données auxquelles il peut accéder. Ce modèle donne l'impression à l'utilisateur qu'il accède à toutes les données de l'entrepôt afin d'éviter toute tentative d'inférence : C'est une mesure de sécurité.

Triki et al. (2011) propose une approche pour la sécurisation des entrepôts de données par la prévention des inférences. A partir du modèle conceptuel des sources de données représenté par un diagramme de classes, l'approche proposée assiste le concepteur dans l'identification des données sensibles et celles qui peuvent être soumises à des inférences. Cette approche se compose de trois phases dont la première, réalisée par le concepteur de la sécurité, identifie les éléments qui doivent être protégés dans la conception de l'entrepôt. A la deuxième phase, le graphe d'inférences est construit automatiquement, permettant la détection des éléments qui peuvent conduire à des inférences. Le concepteur distingue alors les éléments qui conduisent à des inférences précises engendrant la détection de la valeur exacte des données et celles qui conduisent à des inférences partielles où l'utilisateur retrouve une vision partielle des données. Dans la troisième phase, le modèle de schéma en étoile de l'entrepôt est enrichi automatiquement par des annotations UML mettant en évidence les éléments soumis à ces deux types d'inférences.

Rosenthol et al. (2000) étendent la politique d'accès de SQL. Ils définissent la permission d'accès d'un utilisateur comme un quadruplet (sujet, opération, objet, mode) où le sujet est l'utilisateur, l'opération est le droit d'accès SQL, l'objet est la table et le mode est le type de permission. Ils ont spécifié dans leur travail deux types de permission : *permission d'information* indiquant qui a le droit d'accéder une information et *permission physique* indiquant quel utilisateur à le droit d'accéder quelle table. Le premier type exprime une permission au niveau métier avec la possibilité de présenter des conflits entre permissions. Par contre, le deuxième type est précis et ne pose pas de difficultés d'implémentation.

De leur côté, Villoarell et al. (2006), Soler et al. (2008) et Medina et al. (2008) ont proposé une méthode de conception d'un entrepôt de données sécurisées couvrant différent niveaux en se basant sur l'architecture MDA. Soler et al. (2008) proposent le modèle SMD CIM (*Secure Multidimensional Computation Independent Model*) représentant les deux exigences d'un entrepôt de données : les exigences d'information qui permettent de définir le sujet~~s~~ et les axes d'analyse et les exigences de QoS qui analysent les problèmes de sécurité via le modèle SR de la stratégie i*. Villorroel et al. (2006) définissent le modèle sécurisé de l'entrepôt de données SECDW du niveau PIM (*Platform Independent Model*) qui modélisent les exigences définies précédemment en se basant sur une extension du profil UML appelée *SECure Data Warehouse* (SECDW) pour résoudre les problèmes de confidentialité de modélisation conceptuelle des DW et une extension OCL (*Object Constraint Language*) qui permet de spécifier les contraintes de sécurité des éléments de l'entrepôt de données. Medina et al (2008) utilisent, au niveau PSM (*Platform Specific Model*), des mécanismes d'extension propres fournis par le CWM (*Common Warehouse Metamodel*) et étendent le paquet relationnel pour obtenir le modèle SECRDW qui est défini par un schéma ROLAP en étoile représentant toutes les mesures de sécurité et d'audit capturées pendant la phase de modélisation conceptuelle de l'entrepôt de données.

4.2 Sécurité dans le Web

Seul le travail de Kimball et al. (2000) discutent des aspects de sécurité dans les WebHouses. Les auteurs présentent une solution technique à intégrer après la conception de l'entrepôt. La prise en comptes des aspects de sécurité est réalisée par l'administrateur en amant de la construction du WebHouse sans considérer les besoins de sécurité métiers des décideurs. La solution proposée assure la sécurité via quatre éléments qui sont l'authentification à deux facteurs, la sécurisation des connexions à travers soit des VPN (*Virtuel Private Network*) soit des communications cryptées, la définition des rôles utilisateurs et le contrôle d'accès aux objets de data WebHouse.

La mise en œuvre d'un entrepôt de données dans le Cloud représente une avancée technologique mais apporte aussi des problèmes de sécurité qu'il faut prendre en compte. Karkouda et al. (2012) proposent de partager chaque donnée stockée dans l'entrepôt sur plusieurs fournisseurs de Cloud. Leur solution consiste à sécuriser le stockage et l'exploitation d'un entrepôt de données dans le Cloud par le stockage d'un n-uplet chez plusieurs fournisseurs. Cette façon de répartir les données permet; d'une part; de stocker au niveau de chaque fournisseur une partie de l'information, celles-ci sont alors non compréhensibles et non exploitables par un utilisateur malveillant en cas d'intrusion et; d'autre part; de ne pas dépendre d'un seul fournisseur, ce qui minimise le risque de non disponibilité des données.

Le travail de Vela et al. (2013) redéfinit la couche PSM, présentée dans les travaux de Villoarell et al. (2006), Soler et al. (2008) et Medina et al. (2008), en adoptant le langage XML pour le développement d'un entrepôt de données sécurisé. Un ensemble de règle de transformation QVT est développé pour la dérivation semi automatique de l'entrepôt XML à partir du modèle PIM.

Loin du contexte d'entreposage, plusieurs travaux traitent la problématique de sécurité dans le Web. Gallino et al. (2010) présentent une approche basée sur des modèles MDA pour développer une architecture orientée Web services avec l'intégration de la sécurité et du contrôle d'accès. Cette approche est composée d'un ensemble de modèles, qui sont le modèle de système fonctionnel, le modèle des ressources de contrôles d'accès, le modèle des aspects non-fonctionnels (modèle politique de sécurité) et le méta-modèle intermédiaire (IMM) et propose de développer chaque modèle comme un modèle indépendant. Après le développement, ces modèles constituent un modèle complet assurant la sécurité du système. Le méta-modèle IMM comporte trois packages : le premier s'occupe de la conception du système, il décrit les composants de service Web, le deuxième modélise le contrôle d'accès en se basant sur différentes techniques (DAC, MAC, …) et le troisième comporte les métas données relatives à la politique.

Dans un contexte industriel, Best et al. (2007) présentent une analyse du moteur de recherche basée sur le MBSE (*Model-Based Security Engineering*) dans l'intranet d'une entreprise de construction d'automobiles allemandes. Le MBSE fournit une approche de développement des logiciels de sécurité critique où les exigences de sécurité (tels que le secret, l'intégrité, l'authenticité …) peuvent être spécifiées dans une spécification UML, ou dans le code source (Java ou C) sous forme d'annotations. Le MBSE se base sur l'extension UmlSec qui définit les outils d'extension UML tel que les stéréotypes, les tags pour formuler les exigences de sécurité et les Contraintes en utilisant une sémantique formelle. L'extension UMLsec est donnée sous la forme d'un profil UML en utilisant les mécanismes standards d'extension d'UML. Un protocole d'authentification pour le processus de connexion est modélisé comme un diagramme de séquence UML comportant neuf étapes à établir entre le client (client PC) et le serveur (authServer).

Aussi, Hafner et al. (2009) présentent les concepts de base de la sécurité dans les architectures SOA. Ils identifient les politiques de sécurité de base : la disponibilité, la confidentialité et l'intégrité, les modèles de politique : MAC, DAC et RBAC et les politiques de sécurité avancées : contrôle d'accès dynamique, la délégation des droits...
En outre, ce travail définit les attaques, les menaces et les vulnérabilités et les contrôles de sécurité en particulier dans le contexte de la SOA. Comme standards de sécurité spécifiques aux services Web, Hafner et al. (2009) mentionnent l'*eXtensible Access Control Markup Language* (XACML), la spécification *XML Key Management* (XKMS), le WS-Trust, *WS-Secure Conversation* et *WS-Federation*. Ces travaux touchent uniquement la sécurité au niveau des services Web et n'intègrent pas cet aspect dans les systèmes décisionnels Web.

Le problème de sécurité dans les architectures SOA Web n'est pas considéré par les chercheurs seulement, Oracle (2008) traite aussi ce sujet. Ce rapport décrit les normes essentielles pour la sécurité des environnements SOA tels que le protocole *SSL Transport-Level Security*, le framework XML : *Application-Level Security*… Le framework XML assure la confidentialité (*XML Encryption*) et l'intégrité et l'authenticité (*XML Signature*). WS-Security est la spécification la plus importante pour assurer la SOA et comprend plusieurs outils tels que le langage SAML (*Security Assertion Markup Language*), Web *Services Addressing* (WS-Addressing), *Web Services Policy* (WS-Policy) …
Dans le contexte de cryptage, Oracle. (2012) propose un nouveau standard TDE (*Transparent Data Encryption*) qui offre la possibilité de chiffrer les données d'application sensibles sur des supports de stockage. Un ensemble de nouveaux concepts est défini avec ce nouveau standard tel que le *Master encryption key*, *Unified encryption key*, *Table key*, *TableSpacekey*….

5 Comparaison

Dans cette section, nous proposons de comparer les travaux de recherche liés à la modélisation des systèmes décisionnels Web et à la sécurité. Au niveau de la modélisation des systèmes décisionnels Web, notre étude comparative est basée sur les critères suivants :

- *Les entrées :* Ce critère est relatif au type des données sources considérées par l'approche. Nous distinguons les données opérationnelles (*Op.*), les *Cleackstream* et les données *Web* ;
- *Les sorties :* Ce critère décrit le(s) modèle(s) résultat(s); nous distinguons les travaux qui génèrent un entrepôt Web (ED Web) et ceux un entrepôt temps réel (ED TR).
- *Le(s) niveau(x) de modélisation :* Trois niveaux d'abstraction sont considérés : conceptuel (C), logique (L) ou physique (P), un travail peut couvrir un niveau ou plus.
- *Le langage :* définit le langage de modélisation,
- *L'architecture (Arch.) :* indique si l'architecture de l'entrepôt est distribuée (Dis) ou bien centralisée (Cen).
- *La sémantique (Sém.) :* indique si la sémantique des données Web est prise en compte.
- *La sécurité*: indique si le travail propose une solution pour sécuriser l'ED.

	Entrée	Sortie	Niv_mod	Langage	Arch.	Sém.	sécurité
Nguyen (2001)	Web	ED Web	C	XML	F	oui	non
Abrahiem. (2007)	Web &Op.	ED TR	C	XML	Cen	Non	non
Mehedintu (2008)	Clickstream	ED Web	L		Dis	Oui	donnée
Liu et al. (2010)	Web	ED TR	L	XML	Cen	Non	non
Hernández (2010)	Log Web	ED Web	C	UML/ QVT	Cen	Oui	non
Marotta (2012)	Web	ED Web	L		Cen	oui	non

TAB. 1 – *Tableau comparatif des travaux traitant la conception des entrepôts Web.*

Nous constatons que tous les travaux traitent des entrées Web mais certains proposent de les enrichir par les informations extraites des fichiers log.. Nous mettons l'accent sur l'absence d'une démarche de conception couvrant tous les niveaux d'abstraction. Pour le choix de l'architecture, la plupart des travaux étudiés se basent sur une architecture centralisée alors que Nguyen et al. (2001) et Mehedintu et al. (2008) font recourt à une architecture distribuée. La sémantique des données, qui posent un conflit dans l'entreposage des données Web Perez et al. (2008), est traitée dans la plupart des travaux.

En outre, les besoins non fonctionnels de sécurité ne sont pas traités dans aucune des démarches de conception à l'exception du travail de Mehedintu et al. (2008) qui a mis l'accent sur le besoin de vérifier l'exactitude des sources de données.

Dans le contexte de sécurité, notre étude nous a amené à distinguer les critères de comparaison suivants :

- *Niveau de modélisation* : nous constatons que les besoins de sécurité sont modélisés à différent niveaux d'abstraction; métier (M), conceptuel (C) et physique (P);
- *Critère de sécurité* : chaque travail supporte certains critères de sécurité tel que la confidentialité (C), l'intégrité (I), la disponibilité (D), l'authentification (Au) et l'audit(A);
- *Aspect de sécurité* : tel que la définition des rôles (R), des contraintes de sécurité (Cte), des permissions (P);
- *La couverture* : L'expression des exigences de sécurité concerne trois axes: les profils Utilisateur (U), les données (D) et les communications (C).

	Niv_mod	Crit. Sécurité	Aspect de sécurité	Couverture
Katic et al. (1998)	P	D	R; P, Séc. réseaux	D, U, C
Triki et al. (2011)	C	C, I, D	Contrôle d'inférence	D
Rosenthol et al. (2000)	P	C, D	P, Contrôle d'inférence	D
Trujillo et al	M, C, P	C, I	R (hiérarchies), Cte	D, U
Kimball et al. (2000)	P	I, D, Au	R (hiérarchies), P Séc. Réseau	D, U, C
Karkouda et al (2012)	C, P	C, D	Contrôle d'accès	D
Gallino et al. (2010)	C	C, A	Contrôle d'accès	D
Best et al. (2007)	C, P	Au, I	Séc. transactions client / serveur	D, C
Hafner et al. (2009)	P	C, D, Au	Politiques et Standards de sécurité	D, C, U
Vela et al. (2013)	C, P	C, I	R (hiérarchies), Cte	D, U

TAB. 2 – Tableau comparatif des travaux traitant le sujet de sécurité

D'après ce tableau comparatif, quelques approches semblent être pertinentes et assurent un niveau de sécurité acceptable mais qui reste insuffisant. La sécurité est exprimée selon

différents aspect. La plupart des solutions existantes pour sécuriser le transfert et le stockage des données sont basées sur la cryptographie des données qui n'est pas toujours une solution complète pour protéger les données surtout en présence du Web. La sécurité n'est pas traitée dès le niveau métier. Seul le travail de Trujillo et al a traité la sécurité dès le niveau métier en prenant compte des exigences exprimées par les utilisateurs. La définition des besoins de sécurité doit être faite dès le niveau métier pour pouvoir concevoir un système sein et protéger de tout risque d'intrusion. Tous les travaux couvrent la sécurité des données stockés mais la plupart ignorent la sécurité des profils utilisateurs et surtout des communications. La problématique de sécurité dans les WebHouses est traitée par un seul travail Kimball et al. (2000) qui présente une solution tardive à intégrer après la conception du système sans prise en compte des besoins métiers. Les autres travaux étudiés présentent de diverses solutions dans le contexte de sécurité Web mais ils ne touchent pas la sécurisation du WebHouse.

6 Futurs travaux

L'avènement du Web comme source de données pour les WebHouse, nécessite d'être capable de gérer de très importants volumes de données. En outre, l'architecture du WebHouse doit supporter des requêtes d'analyse décisionnelles complexes et coûteuses. Une architecture répartie apporte une solution efficace à ces problématiques.

Ainsi, nous proposons une architecture pair à pair pour un WebHouse réparti qui donne accès à un grand nombre de ressources et dispose d'une administration transparente. Les qualités de cette architecture (robustesse, disponibilité, performances...) augmentent avec le nombre d'utilisateurs, et présentent de nombreux avantages (décentralisation, pas de coûts d'infrastructure...).

Néanmoins, l'émergence du Web entraine le traitement d'une grande quantité d'information provenant de différentes sources qui ne sont pas toujours seine et entraine aussi la diversification et la multiplicité des utilisateurs et des communications qui augmentent les risques d'intrusion. En plus, l'architecture répartie pose des problèmes de sécurité à cause de certaines parties malveillantes non écartées du réseau. Par conséquence, la sécurité des informations dans le processus de prise de décision, accentuée par l'apparition du WebHouse, représente une question cruciale qui doit être traitée.

Divers travaux ont discuté le sujet de sécurité en proposant des solutions de sécurisation des systèmes Web et généralement des Web services. Le travail de Kimball et al (2000) est une exception, qui propose une solution tardive de sécurité relative au WebHouse à intégrer après la conception du système sans prise en compte des besoins métiers. L'analyse de la sécurité doit être intégrée dans tous les niveaux de développement du système décisionnel; lors de la phase d'analyse des besoins métiers en incluant les besoins non fonctionnels de sécurité jusqu'au la phase d'implémentation en spécifiant les solutions et les techniques de sécurité. Notre objectif consiste à assurer la sécurité en partant des sources de données jusqu'à arriver au processus OLAP. Ainsi, la sécurité doit couvrir le niveau d'extraction vu que les données extraites doivent êtres seines. Aussi, nous devons assurer le stockage des données autrement assurer les pairs du WebHouse reliés entre eux et présentant chacun un sous système de stockage. Le choix d'une architecture pair à pair pour le WebHouse offre un avantage à la sécurité du WebHouse vu que la répartition des données rend inutile d'accéder à tout le système pour répondre à une requête. Il suffit, donc, d'interroger les pairs concernés. En outre, en cas d'accès malveillant, seul le pair contenant les données est attaqué et non pas tout le système. Et finalement, le processus OLAP doit être contrôlé et couvert par les

aspects et mécanisme de sécurité pour interdire tout accès malveillant surtout dans un contexte Web.

7 Conclusion

Dans cet article, nous avons dressé l'état de l'art des travaux se rapportant à deux axes : la conception des systèmes décisionnels Web et la modélisation de la sécurité dans ces systèmes. Nous avons également réalisé une comparaison des différents travaux selon plusieurs dimensions. La sécurité des informations dans le processus de prise de décision, accentuée par l'apparition du WebHouse, représente un enjeu crucial qui doit être traité. En effet, la principale question est : Comment peut-on choisir les solutions (outils, mécanismes, protocoles…) de sécurité convenables à un système décisionnel Web (WebHouse) réparti, à chaque niveau de modélisation et pour chaque partie du système (donnée du WebHouse, client Web et réseau) tout en tenant compte des spécificités du Web tel que le dynamisme, le besoin de réponse rapide ?

Références

Abrahiem. R, (2007), *A New Generation of Middleware Solutions for a Near-Real-Time Data Warehousing Architecture*. Chicago: Electro/Information Technology, 192-197

Best. B., J. Jan et N. Bashar, (2007), *Model-based Security Engineering of Distributed Information Systems using UMLsec*. Minneapolis: ICSE , 581-590

Gallino, S.J.P., Miguel, M.A., Briones, J.F. et Alonso, A. (2010), *Model-Driven Development of a Web Service-Oriented Architecture and Security Policies*. Madrid, Spain: ISORC '10, 92-96.

Golfarelli M, F. Mandreol, W. Penzo, S. Rizzi et E. Turricchia (2010), *Towards OLAP Query Reformulation in Peer-to-Peer Data Warehousing*. Canada: DOLAP, 37-44.

Hernández P, I. Garrigós et J-N. Mazón (2010), *Model-Driven Development of Multidimensional Models from Web Log Files*. Canada: E*R Workshops, 170-179.*

Kalnis P, W.S. Ng, B.Ch. Ooi, D. Papadias et Kian-lee Tan (2002), *An Adaptive Peer-to-Peer Network for Distributed Caching of OLAP Results*. Madison, Wisconsin: ACM SIGMOD international conference Management of data, 25-26

Karkouda K., N. Harbi, J. Darmont et G. Gavin (2012), *Confidentialité et disponibilité des données entreposées dans les nuages*. Bordeaux, EGC-FDC 2012.

Katic, N., G. Quirchmayr, J. Schiefer, M. Stolba et A.M. Tjoa (1998), *A prototype model for data warehouse security based on metadata.* Vienna, Austria: DEXA, 300-308.

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

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

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

Medina. E.F., J. Trujillo, E.M. Soler (2008), *Building a secure star schema in data warehouse by an extension of the relational package from CWM*. Amsterdam: Computer Standards and Interfaces, 30, 341-350.

Mehedintu A., I.s Buligiu et C. Pirvu (2008), *Web-enabled Data Warehouse and Data Webhouse*. Bucharest: Revista Informatica Economică nr.1(45), 96-103.

Nguyen Th. B., A Min Tjoa (2001), *An XML Metadata Foundation for Interoperability Search among Web Warehouses*. Switzerland: 3rd International Workshop on Design and Management of Data Warehouses.

Oracle (2008), *Web Services Security: What's Required: To Secure A Service-Oriented Architecture*, In Oracle White Paper.

Oracle (2012), *Oracle Advanced Security Transparent Data Encryption Best Practice.*

Pérez J M., R.B. Llavori, M.J Aramburu et T.B Pedersen (2008), *Integrating Data Warehouses with Web Data: A Survey*. IEEE Trans. Knowl. Data Eng. 20(7): 940-955

Ravat F., O. Teste, R. Tournier, G. Zurfluh (2010), *Finding an application-appropriate model for XML data warehouses*, Inf. Syst. 35(6): 662-687

Rosenthal A. et E.Sciore (2008), *View security as the basic for data warehouse security*, Sweden: DMDW, Workshop on Design and Management of Data Warehouse.

Abiteboul S., T. Allard, P. Chatalic, G. Gardarin, A. Ghitescu, F. Goasdou, I. Manolescu , B. Nguyen, M. Ouazara, A. Somani, N. Travers et G. Vasile (2008), *WebContent: Efficient P2P Warehousing of Web Data*. Auckland, New Zealand: VLDB Endowment, 1, 1428-1431.

Soler. E., V. Stefanov, J-N. Mazon, J. Trujillo, F-M. Eduardo et M. Piattini (2008), *Towards Comprehensive Requirement Analysis for Data Warehouses*. Barcelone, Espagne: 3rd International Conference on Availability, Reliability and Security, 104-111.

Triki, S., H. B, Abdallah., N, Harbi. et O, Boussaid (2011), *Securing Data Warehouses: A Semi-automatic Approach for Inference Prevention at the Design Level.* Portugal*:* First International Conference MEDI, 71-84.

Vela, B., J.N. Mazón, C. Blanco, E. Fernández-Medina, J. Trujillo et E. Marcos (2013), *Development of Secure XML Data Warehouses with QVT*, Information and Software Technology doi: http://dx.doi.org/

Villorroel, R., E. Medina, J. Trujillo, M. Pattini (2006), *A UML 2.0/OCL Extension for designing Secure DataWarehouse.* Journal of researchs and Practice in information Technology, 38.

Summary

The advent of the Web in decisional systems requires the management of very large volumes of data. In this context, two critical points are discussed: the need to choose a new architecture for WebHouse and to secure information. This issue is compounded by the advent of the Web in decisional systems with the diversification and the multiplicity of users and communications that increase the risk of intrusion. In this paper, we propose a survey on design and secure WebHouse.

Qualitative data warehouse modeling – urban sites annoyance analysis use case

Fatiha Amanzougarene*, **, Karine Zeitouni*
Mohamed Chachoua**

* PRISM, Université de Versailles- SQ 45 Avenue des Etats-Unis 78035 Versailles, France
(fatiha.amanzougarene, karine.zeitouni)@prism.uvsq.fr
**EIVP, École des ingénieurs de la ville de Paris, 15 Rue Fénelon, 75010 Paris, France
(fatiha.amanzougarene, chachoua)@eivp-paris.fr

Abstract. Decision support systems based on data warehouses constitute a powerful framework for multidimensional data analysis. In these systems, facts are analyzed through indicators (numerical measures) according to various analysis perspectives (dimensions). In real world applications, facts are not always numerical, and can be of qualitative nature. In addition, sometimes a human expert or prediction model provides a qualitative evaluation of phenomenon based on its different parameters. Conventional data warehouses are thus not adapted to qualitative reasoning and have not the ability to deal with qualitative data. In this paper, we propose an original approach of qualitative data warehouse modeling, which permits integrating qualitative measures. Based on computing with words methodology, classical multidimensional data model will be extended to allow aggregation and analysis of qualitative data in OLAP environment. To illustrate the problematic and our proposal, we will consider throughout this paper the case of urban building sites annoyance.

1 Introduction

Data warehouses and on-line analytical processing (OLAP) constitute the main elements of decision support systems. A data warehouse means a decision support database allowing integration, organization, historization, and management of data from heterogeneous sources, with the aim of exploiting them for decision-making Inmon (2005), Kimball (2002). OLAP refers to the technology that allows users to efficiently retrieve the information stored in a data warehouse. To conceptualize data in a data warehouse, the multidimensional model is used. This model organizes data into facts (subjects of analysis) and dimensions (perspectives of analysis). A fact is composed of numerical measures and dimensions which characterize it. A dimension is organized into hierarchical levels of detail. Based on the navigation and aggregation mechanisms offered by OLAP tools, facts can be analyzed according to the desired level of detail. In some real world applications, the subject of analysis may be subjective and consequently its measures are provided in qualitative fashion. In addition, sometimes a human expert or a prediction model such as a decision tree can be used to provide a qualitative evaluation of some phenomenon based on its different parameters. This arises in many applications such as customer satisfaction, process control, consumer products, and annoyance evaluation.

Conventional data warehouses are thus not adapted to human reasoning and have not the ability to deal with qualitative data. To meet this difficulty, in this paper we will present an

original work that aims at making it possible to handle raw qualitative data and providing a more flexible method for the multidimensional analysis over that type of data. Based on computing with words methodology, we will introduce qualitative measures and aggregates as an extension of multidimensional data model of a data warehouse. Using these measures and aggregates, OLAP queries allow the decision maker to manipulate data in a qualitative fashion using linguistic values. Although there is no precision about the type of measures in the original model Inmon (2005), to the best of our knowledge, there exists no multidimensional data warehouse offering ordinal qualitative measures. To illustrate the problematic and our proposal, we will consider throughout this paper the case of urban sites annoyance.

This paper is structured as follows. In the second section we present our work motivation and the use case related to urban building sites annoyance evaluation and analysis. In the section 3, we explore the used approach for qualitative data aggregation. In the section 4, we describe our proposition extending the classical multidimensional model to allow qualitative data analysis. In the section 5 we present the experimentation framework that consists of the Spatial Decision Support System (SDSS) designed to the annoyance analysis. Finally, in the last section we conclude and we present our future work.

2 Motivation and use case : urban building sites annoyance

Although indispensable for the development and renovation of cities, urban building sites are often a source of various kinds of nuisance[1]. These nuisances have not negligible impacts on quality of life of urban citizens. This issue is crucial and becomes more complex in cities with high population density. The main objective of our work is to develop a spatial decision support system (SDSS) dedicated to reducing the annoyance generated by urban building sites. We make the observation that, in human reasoning, the annoyance is evaluated subjectively and qualitatively by using an ordinal scale of linguistic degrees. Therefore, for a perfect match with the human expert reasoning, we propose in this paper a qualitative model of annoyance evaluation. In our previous studies Amanzougarene et al. (2011), we have presented a quantitative model that allows evaluating urban people annoyance due to the noise. By comparison, in the present work, we generalize our previous model by privileging a qualitative data handling of annoyance. We also extend our previous model of annoyance evaluation to other types of nuisance than noise, which strengthens the interest of multidimensional analysis. Indeed, an urban building site is generally likely to cause many nuisances.

2.1 Qualitative representation of annoyance

Notion of annoyance. In Amanzougarene et al (2011), we have defined the notion of annoyance as follow.

Definition1. In a spatiotemporal environment, annoyance is subjective relationship between an individual and a harmful phenomenon.

In other words, an individual can be only annoyed, in the presence of one or more harmful phenomena for this individual. Thus, a human expert can evaluate subjectively the degree of annoyance, according the various factors presented in Amanzougarene et al. (2011). The most relevant factors can be classified in three categories: (1) factors related to individual,

[1] We denote by nuisance, a disturbance of environment having a negative effect.

(2) factors related to nuisance, (3) factors related to environment. Table 1 below shows these three categories, with the main factors.

Factors related to individual	Factors related to nuisance	Factors related to environment
Age	Nuisance type	Space
Health condition	Nuisance intensity	Time
Gender	Exposure duration	
Socio-professional category	Frequency	
Acceptability		

TAB. 1 – *Main factors of annoyance.*

Dimensions of annoyance. In practice, the choice of factors to be considered for the annoyance evaluation depends on the human experts' appreciation. In our case study, these experts have retained some factors related to individual, nuisance and environment. The latter is actually a combination of space and time dimensions. This leads to a multidimensional representation described by Figure 1 below and including the dimensions: (1) category of population grouping the factors related to the individual, (2) nuisance grouping the factors related to the nuisance, (3) space, and (4) time. Notice that the choice of the dimensions is application-dependant, and could add or ignore some factors such as the building type or gender. Our model adapts to other schemas as well.

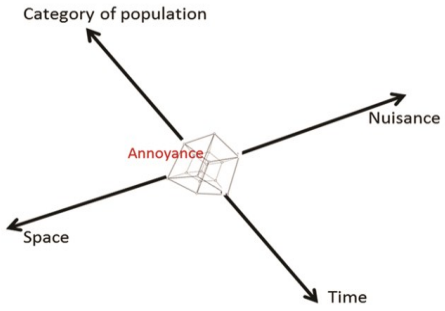


FIG. 1 – *Multidimensional representation of annoyance.*

2.2 Annoyance evaluation

In human reasoning, the subjective evaluation of annoyance is done qualitatively by using a finite scale of linguistic degrees, such as, "*little annoyed*", "*very annoyed*"…. Generally, the human subject uses ordered scales with 5 or 7 linguistic degrees. To reproduce this same mechanism and to represent qualitatively these linguistic degrees, we will use the following algebraic structure:

Let $M \geq 2$ be an integer ($M \in \mathbb{N}$). Let us consider a set $\mathcal{L}_M = \{\tau_1, \tau_2, \dots, \tau_M\}$ of elements totally ordered by the relation " $\leq$ " such as: $\tau_\alpha \leq \tau_\beta \Leftrightarrow \alpha \leq \beta$. Thus, $\mathcal{L}_M$ is an ordinal scale in which the smallest element is $\tau_1$and the largest element is τ_M. Basic operations on $\mathcal{L}_M$ are:

$Max(\tau_\alpha, \tau_\beta) = \tau_{Max(\alpha,\beta)}$; $Min(\tau_\alpha, \tau_\beta) = \tau_{Min(\alpha,\beta)}$.

Example1*:* For a scale with 5 degree ($M = 5$), we can introduce the following ordered set of linguistic degrees $\mathcal{L}_5 = \{$*Not at all*, *Slightly, Moderately, Very, Extremely}*

These linguistic degrees correspond respectively to the degrees $\tau_\alpha - Annoyed\ with\ \alpha \in [1,5]$. Thus, we can represent the linguistic scale of the qualitative evaluation of annoyance, as shown by the Figure 2 below.

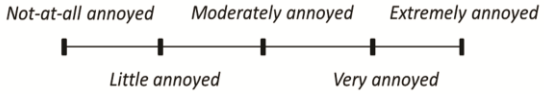


FIG. 2 – *Linguistic scale of annoyance evaluation.*

Note: thereafter, in the interests of simplifying notations, we will represent these linguistic degrees respectively by τ_α instead of $\tau_\alpha - Annoyed$.

Application: example of annoyance evaluation. Let us consider a given location L_1 where one has three nuisances, noise, odor, and dust. An extract of the annoyance evaluation carried out by the human experts is shown in Table 2 below. We note that, this evaluation takes into account only the following factors: (1) socio-professional category (SPC), (2) age, (3) health condition, (4) type, (5) intensity, (6) frequency of nuisance, (7) time of day, and (8) period of year. In this evaluation, 5 levels of nuisance intensity are considered. Level 1 corresponds to the absence of nuisance, which means that the degree of annoyance is $\boldsymbol{\tau_1}$ i.e. not at all annoyed. This table is an extract of the decision matrix carried out by the experts based on different dimensions of annoyance. This matrix will serve as knowledge base to populate the data warehouse designed to contain data related to annoyances. This warehouse constitutes the core of our SDSS.

Space = L_1				Category of population							
				SPC	Inactive resident						
				Age	Young		Adult		Old people		
				Health condition	N	Y	N	Y	N	Y	
Nuisance	Type	Inten-	Frequen-cy								
	Noise	1	Con.		τ_1	τ_1	τ_1	τ_1	τ_1	τ_1	
			Dis.		τ_1	τ_1	τ_1	τ_1	τ_1	τ_1	
		⋮	⋮		⋮						
		5	Con		τ_5	τ_5	τ_5	τ_5	τ_5	τ_5	
			Dis.		τ_5	τ_5	τ_5	τ_5	τ_5	τ_5	
				Time	Morning		Evening		Night		
				Period	Non-rainy period						
				Time							

Legend associated with Table 2

Con.: Continuous
Dis.: Discontinuous
Y: Yes i.e. in good health condition
N: No i.e. not in good health condition
SPC: Socio-professional category

TAB. 2 – *Qualitative evaluation of annoyance.*

3 Qualitative data aggregation

Annoyance aggregation consists in defining indicators to facilitate analysis and obtain an overview of different annoyances generated by urban building sites. Example of these indicators can be "*annoyance aggregation over time*" and "*multi-sources annoyance aggregation"*. Given the qualitative nature of annoyance values, we propose to use the "computing with words" methodology.

3.1 Used Approach: Linguistic Approach "Computing with words"

Linguistic approach is a technique appropriate to dealing with qualitative aspect of problems. In this approach, variables which participate in the problem are assessed by means of linguistic terms instead of numerical values Zadeh (1975), Herrera et al. (2009). The ordinal linguistic approach is a particular case of linguistic approach where values of linguistic variable are totally ordered, more specifically they are said to constitute an ordinal scale. In any linguistic approach it is necessary to define some operators that can be applied for the data processing. In our case, we are interested in the class of ordinal aggregation mean operators Yager (2007). In which follows, we will present the definitions of this class of operators. In all these definitions we will use the ordinal scale introduced in section 2.2.

3.2 Mean Operators

These operators are characterized by being bounded and generally have self-identities.
Definition 2. An aggregation operator $F: \mathcal{L}_M{}^n \rightarrow \mathcal{L}_M$ is called a mean operator if it has the following proprieties:

- Symmetry: means the order of the elements does not matter i.e.
  $F(\tau_1, \tau_2, \dots, \tau_n) = F\left(\tau_{\sigma(1)}, \tau_{\sigma(2)}, \dots, \tau_{\sigma(n)}\right)$ where $\sigma(i)$ is any permutation of $\tau_{(i)}$.
- Monotonicity: $F(\tau_1, \dots, \tau_n) \geq F(\varphi_1, \dots, \varphi_n)$ if $\left(\tau_i \geq \varphi_j\right)$
- Boundedness: which is the key propriety of these operators, means:
  $Min(\tau_i) \leq F(\tau_1, \tau_2, \dots, \tau_n) \leq Max(\tau_i)$
- Idempotency: $F(\tau_i, \tau_i, \dots, \tau_i) = \tau_i$

This category of operators includes mainly the three following operators: (1) ordinal median, (2) ordinal OWA (Ordered Weighted Averaging) without degree of importance, and (3) ordinal OWA (Ordered Weighted Averaging) with degree of importance. Thereafter, we note these operators respectively by $\mathrm{OMed}, \mathrm{OOWA}, \text{and } \mathrm{OOWA_d}$

Ordinal median. Generally, median is described as the numerical value separating the higher half of a sample, or a probability distribution from the lower half. According to Yager (1998), ordinal median can be defined as follows:
Definition 3. Let $\mathcal{L}_M = \{\tau_1, \tau_2, \dots, \tau_M\}$ be an ordinal scale. Let us consider $\mathcal{C} = \{\tau_1, \dots, \tau_n\}$ a collection of elements of $\mathcal{L}_M$. $OMed(\tau_1, \dots, \tau_n) = OMed(\varphi_1, \varphi_2, \dots, \varphi_n)$. Where φ_j is the *j*th largest element of $\mathcal{C}$.

Thus, $OMed(\mathcal{C}) = \varphi_k$ that such as $k = (n+1)/2$ if n is odd, and $k = n/2$ if n is even.

Ordinal OWA (Ordered Weighted Averaging). Ordered weighted averaging (OWA) operator was first introduced by Yager (1988) to provide a method for aggregating multiple criteria that lie between the min and the max operators. The particularity of this operator is the weighting vector associated to it. For more details on semantic and calculation of this vector, see Ahn (2006), Yager (1988), Yan et al. (2011). Note that the concept of weighting vector associated with OWA operator is completely independent from the concept of degree of importance vector that may be associated with the criteria to aggregate if those are not of equal importance.

Definition 4. Let $\mathcal{L}_M = \{\tau_1, \tau_2, \dots, \tau_M\}$ be an ordinal scale. Let us consider $\mathcal{C} = \{\tau_1, \dots, \tau_n\}$ a collection of elements of $\mathcal{L}_M$ and $\mathcal{W} = \{w_1, \dots, w_n\}$ a vector of associated weights.

$$\text{OOWA}(\mathcal{C}) = Weighted\ Median\left((w_1, \tau_1), \dots, (w_n, \tau_n)\right)$$

To calculate the weighted median, proceed as follows:

- Reorder the elements of $\mathcal{C}$ to obtain a new collection $\mathcal{C}' = \left((u_1, \varphi_1), (u_2, \varphi_2), \dots, (u_n, \varphi_n)\right)$ such as φ_j is the *j*th largest element of $\mathcal{C}$ and u_j is the weight that is associated with τ_i becomes φ_j.
- Calculate the sum of the first *i* weights:
  $T_i = \sum_{j=1}^{i} u_j$
- $Weighted\ Median(\mathcal{C}) = \varphi_k$ where k is such that:
  $T_{k-1} < 0{,}5\ \ and\ T_k \geq 0{,}5$.

Ordinal OWA with degree of importance. We have seen the definition of Ordinal OWA operator that can be used to aggregate criteria which are of important equal. In the following we will give a definition for including the ability to handle different importance in ordinal OWA operator.

Definition 5. Let $\mathcal{L}_M = \{\tau_1, \tau_2, \dots, \tau_M\}$ be an ordinal scale. Let us consider $\mathcal{C} = \{\tau_1, \dots, \tau_n\}$ a collection of elements of $\mathcal{L}_M$, $\mathcal{W} = \{w_1, \dots, w_n\}$ a vector of associated weights, and $\mathcal{D} = (d_1, d_2, \dots, d_n)$ a vector of degrees of importance. Such as $d_i \in [0,1]$, $\sum_1^n d_i = 1$.

To calculate the $OOWA_d(\mathcal{C})$, proceed as follows:

- Reorder the elements of $\mathcal{C}$ to obtain a new collection $\mathcal{C}' = \left((d'_1, u_1, \varphi_1), (d'_2, u_2, \varphi_2), \dots, (d'_n, u_n, \varphi_n)\right)$. Such as φ_j is the *j*th largest element of $\mathcal{C}$, u_j is the weight that is associated with τ_i becomes φ_j, and d'_i is the degree of importance that is associated with τ_i becomes φ_j
- Calculate: $T_i = \sum_{j=1}^{i} u_j, S_i = \sum_{j=1}^{i} d'_j$
- Search the index i for which: $T_{i-1} < 0{,}5\ and\ T_i \geq 0{,}5$
- Denote $\alpha = \frac{i}{n}$
- Search the index k for which: $S_{k-1} < \alpha\ and\ S_k \geq \alpha$.
- $OOWA_d(\mathcal{C}) = \varphi_k$

3.3 Application: annoyance aggregation

The choice of the aggregation operator to be used to aggregate annoyances depends on the user aggregation strategies. In this section we will present some examples.

Annoyance aggregation over time. Annoyance aggregation over time expresses the global or average annoyance of a given category of population over a given period of time. As

degrees of annoyance are qualitative, one cannot use a classical average operator to aggregate these degrees. In this case, one can use ordinal median presented in section 3.2.

Example 2. Let $\mathcal{L}_5 = \{\tau_1, \tau_2, \tau_3, \tau_4, \tau_5\}$ be an ordinal scale of 5 linguistic degrees. Let us consider $\mathcal{C} = \{\tau_2,\ \tau_2, \tau_3\}$ the annoyance degrees corresponding to the three moments of the day 29/11/2012. To calculate the average annoyance of this day one proceeds as follows:

$\varphi_1 = \tau_3, \varphi_2 = \tau_2, \varphi_3 = \tau_2$, since $n = 3$ then $OMed(\mathcal{C}) = \varphi_2 = \tau_2$

Multiple-sources annoyance aggregation. In real world, categories of population are often subjected at the same time to several sources of different type of nuisances. For example noise, odors, air pollution, and traffic congestion. The problem of calculating the global annoyance generated by several sources of different type of nuisances can be considered as a multi-criteria aggregation problem. Where the sources of nuisances constitute predictive criteria. In this case, one can use $OOWA$ to calculate global annoyance.

Example 3. Let $\mathcal{L}_5 = \{\tau_1, \tau_2, \tau_3, \tau_4, \tau_5\}$ be an ordinal of scale with 5 linguistic degrees. Let us consider $\mathcal{C} = \{\tau_3, \tau_2,\ \tau_4,\ \tau_5,\ \tau_4, \tau_2, \tau_5\}$ the annoyance degrees corresponding to seven sources of nuisance (noise, odor, air pollution …).

Let $\mathcal{W} = \{0.04,\ 0.07, 0.11, 0.14, 0.18, 0.21, 0.25\}$ be a vector of weighting associated. Global annoyance corresponding to these sources is $OOWA(\mathcal{C}) = Weighted\ Median(\mathcal{C})$

$OOWA\ (\mathcal{C}) = \varphi_3 = \tau_4$

φ_i	u_i	T_i
τ_5	0.14	0.14
τ_5	0.25	0.39
τ_4	0.11	0.50
τ_4	0.18	0.68
τ_3	0.04	0.72
τ_2	0.07	0.79
τ_2	0.21	1.00

TAB. 3 – *Example of multiple-source annoyance aggregation using $OOWA$ operator.*

4 Towards a qualitative multidimensional data model

Now that the qualitative ordinal data model has been introduced, the next step is to define a qualitative model of multidimensional of data warehouse and operations for OLAP based of this model. In this section, first, we will describe the multidimensional model of annoyance that will be used as a running example for the rest of the paper. Then, we will present the multidimensional structure of our proposed model, followed by the operations of qualitative data manipulation.

4.1 Multidimensional data model of annoyance

In our case study concerning urban building sites, the subject of analysis corresponds to the annoyance. This subject is analyzed according to dimensions which we have presented in

section 2.1. Namely: *Nuisances, Categories of population*, *Time*, and *Space*. To model data of urban building sites, we have used a star schema represented by figure 3. It concerns a schema of spatiotemporal data warehouse. To represent it, we use the graphical formalisms proposed by Malinovski and Zimanyi (2008). We have defined a fact table Annoyance. Thus, data from this table are analyzed according to the dimensions: *Nuisances*, *Categories of population, Time,* and *Space.* Measures associated with the Annoyance fact table are: *egree of annoyance, and population size.*

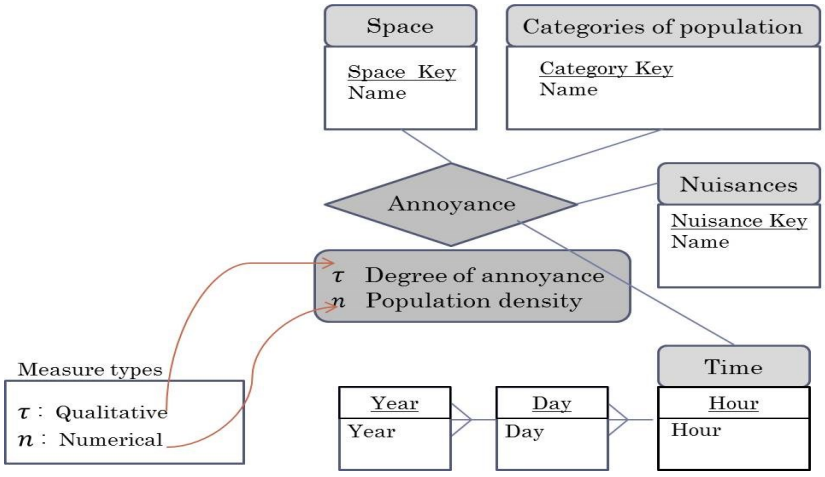


FIG. 3 – *Multidimensional data model of annoyance*

The qualitative nature of the measure *degree of annoyance*, poses a problem for the process of the analysis of data, which consists essentially in carrying out the operations of generalization and aggregation. Indeed, the current multidimensional data models do not allow the integration and the management of this kind of data.

To deal with this problem, we propose to extend the classical multidimensional model.

In the following sections we will present our proposition to extend the conventional multidimensional model to allow integrating and processing qualitative measures.

4.2 Multidimensional structure of qualitative data

The multidimensional structure is characterized by two different types of data dimension (including hierarchies) and facts.

– *Dimension:* a dimension is a set of attributes that can be structured using one or more hierarchies.

Definition 6. An attribute is a set of values $Att_i = \{a_{i1} \ldots, a_{ik}\}$.

Example4.In the figure 3above, the attribute Name of the *Nuisance* dimension can be represented as: $Att_{\text{Name}} = \{\text{noise, odor, dust, pollution, traffic congestion}\}$.

Definition 7. A hierarchy is a tuple $H = (l, \leq_d, l_0, l_{All})$, where $l = l_i, i = 1, \ldots n$ set of attribute levels, l_0, l_{All} are special levels, called base level and top level respectively. The partial order relation $\leq_d$ gives the hierarchical relation between levels.

Example 5. In the figure 3 above, the hierarchy defined on *Time* dimension can be represented as: $H\_Time = (Hour, Day, Year, All, \leq_{Time}, Hour, All)$ where:

$Hour \leq_{Time} Day \leq_{Time} Year \leq_{Time} All.$

Definition 8. A Dimension is a tuple $d = (N, Att_i, H_k)$

- N The name of the dimensions.
- Att_i Set of attributes.
- H_k Set of hierarchies defined on the dimension d.

Definition 9. For each dimension d the domain is $dom(d) = \bigcup_i Att_i$.

– *Fact:* A fact is characterized by set of measures.

Definition 10. A Fact is a set of measures $F = \{m_{i1} \dots, m_{ik}\}$.

Note: we recall that in classical multidimensional data model these measures are numerical. To extend the classical multidimensional data model to deal with qualitative measures, we introduce the concepts of Element, Qualitative Cube, and Collection of Elements.

Definition 11. An Element is a 2-tuple value $e = (v, s)$, where:

- v is a numerical value.
- s is a positive integer such as

$$\begin{cases} if\ s = 0, v\ is\ numerical\ measure \\ else\ \ v\ is\ \ qualitative\ measure, s\ is\ the\ ordinal\ scale\ size \end{cases}$$

Definition 12. A Collection of Elements is a set of elements $\mathcal{C} = \{e_1 \dots, e_k\}$. This structure of data will be used to contain the elements of data cube that we want to aggregate.

Aggregation Operator: When we apply an aggregation operator, we summarize the information about a collection of elements into a single element.

– *Data Cube*: A data cube can be considered as a space representation composed by a set of cells. A cell is associated with a measure and identified by coordinates represented by one member from each dimension level. Each cell in a cube represents a precise fact.

Definition 13. A qualitative data cube is a 4-tuple $C = (\mathcal{D}, \mathcal{L}_b, E, A)$, such that $\mathcal{D}$ is a set of dimensions d_i, $\mathcal{L}_b = (l_{1b}, \dots, l_{nb})$ is a set of levels such that l_{ib} belongs to d_i, E is an element measure, A is an application defined as $A: l_{1b} \times \dots \times l_{nb} \rightarrow E$. A cell of C is noted as $\vec{a} = (a_1, \dots, a_n)$. In order to access the content of cell we shall use $\vec{a}.v$ for the measure value, and $\vec{a}.s$ for the ordinal scale size in the case of qualitative measure. Figure 4 shows an example of qualitative data cube.

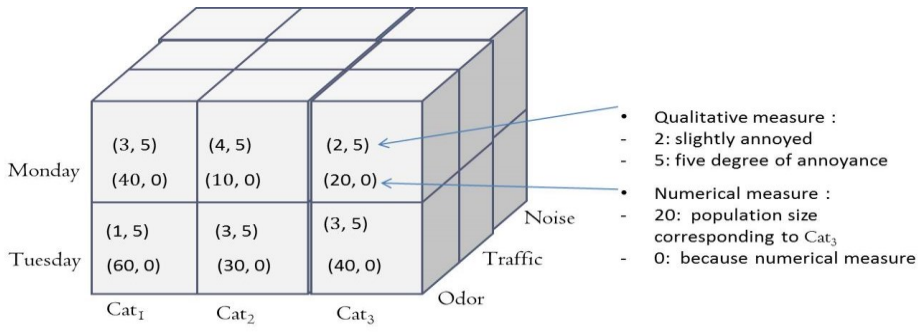


FIG. 4 – *An example of qualitative data cube.*

4.3 Manipulation of qualitative multidimensional data

The operators of multidimensional data manipulation are of two types. Aggregation and navigation operators:

- Operators of navigation: The navigation in the data cube allows the user to move from detailed data to summarized ones e.g. moving up in a hierarchy. This kind of processing requires to aggregate data.
- Aggregation operators: these operators are used to summarize data cube to a single value. The most commonly used are Sum(), Min(), Max(), and Avg(). However, these functions cannot be applied to aggregate degrees of annoyance which are qualitative data. We shall integrate the aggregation operator of qualitative data (presented in section 3.2) as user defined aggregates forming part of advanced functionalities of DBMS.

User defined aggregation functions (UDA). To create a UDA, one must implement four functions. (The exact names of the functions, and the class to which they must belong, differ depending on the particular database system) Cohen (2006).

1. *Init ()*: This function is used to initialize any variables needed for the computation later on.
2. *Terminate ()*: This function is used to end the calculation and return the final value of the aggregate function. It may involve some calculations on variables which were defined for use with the aggregate function.
3. *Accumulate ()*: This function is called once for each value aggregated. Generally, this function will "add" the value to the running total computed so far.
4. *Merge ()*: This function is called to merge intermediate results from two different computations of the aggregate function. This can be used, for example, to enable parallel computation of the aggregate function.

5 Spatial decision support system of annoyance analysis

Our experimentation framework consists on developing a decision support system for analyzing the annoyances of urban building sites. Figure 5 below shows its global architecture.

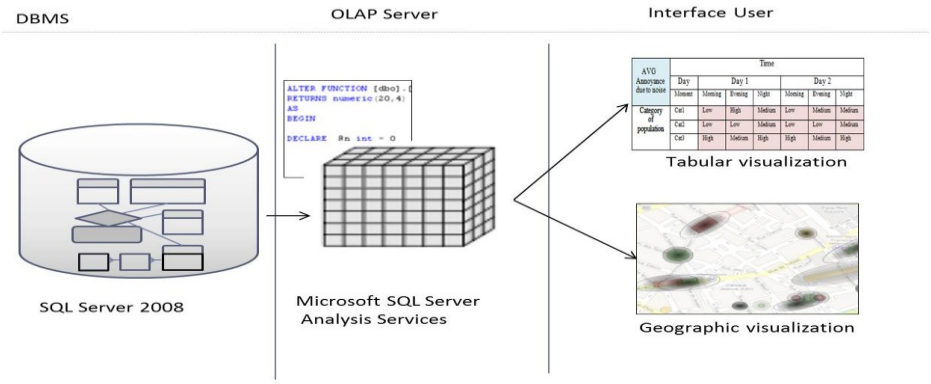


FIG. 5 – *System architecture.*

This architecture consists of three main components:

- The database management system (DBMS) for managing the fact and dimensions tables of annoyance ;
- OLAP server for the multidimensional analysis of annoyance, that consists of multidimensional data cubes constructing and exploitation;
- User interface allows visualizing the annoyance aggregation data using tabular or cartographic representation.

Implementation of annoyance aggregation operators. To ease navigation through the various dimensions of annoyance, and to analysis data aggregation, we have implemented the UDA corresponding to annoyance aggregation operators using C# functions. These functions have been integrated within the data server (see figure 5 above). We have also extended OLAP server to support these functions during the multidimensional data analysis.

6 Conclusion and future work

The main objective of work that we have presented in this paper is the extension of conventional data warehouses to allow integration and processing of qualitative measures. Based on computing with words methodology, we have introduced qualitative measures and aggregates as an extension of multidimensional data model of a data warehouse. Using these measures and aggregates, OLAP queries allow the decision maker to manipulate data in a qualitative fashion using linguistic terms. To illustrate the problematic and our proposal, we have considered the case of urban building sites annoyance. We have proposed an original approach which allows managing the annoyance evaluated qualitatively, as in commons sense reasoning, by using linguistic expressions.

In our future work, we will extend our approach of evaluation in order to include spatial and temporal extent of annoyance as measures in the multidimensional data model of data warehouse. We will also, define aggregation operations allowing data processing of these extents (for example: fusion of annoyance influence areas, and concatenation of exposure time interval). Indeed, that will improve the decisions of managers of public spaces concerning urban building sites planning. So far, we have focused on the measurement and aggregation in a qualitative fashion for the annoyance from previous inputs. We also intend to extend our approach to the prediction of annoyance, so that it helps predict the best place and time for new building site.

References

Ahn, B. S. (2006). On the properties of OWA operator weights functions with constant level of orness. *IEEE Transactions on Fuzzy Systems*, *14*(4), 511-515.

Amanzougarene, F., Sadoun, I., Hankach, P., Chachoua, M., and Zeitouni, K. (2011). A new approach for annoyance assessment of the urban building sites. *International Review on Modelling and Simulations I.RE.MO.S.* Italy.

Cohen, S. (2006). User-defined aggregate functions: bridging theory and practice. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data* (SIGMOD '06). ACM, New York, NY, USA, 49-60.

Herrera, F., Alonso, S., Chiclana, F., and Herrera-Viedma, E. (2009). Computing with words in decision making: foundations, trends and prospects. *Fuzzy Optimization and Decision Making.* Springer 11, 44, 337–364.

Inmon, W. H. (2005). *Building the Data Warehouse, 4th Edition*. Wiley Publiching Inc.

Kimball, R. and Ross, M. (2002). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, 2nd*. John Wiley & Sons.

Malinowski, E. and Zimányi, E. (2008). *Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications*. Springer, Data-Centric Systems and Applications.

Yager, R. R. (1988). On ordered weighted averaging operators in multi-criteria decision making. *IEEE Transactions on Systems, Man and Cybernetics.18*(1), 183–190.

Yager, R. R. (1998). Fusion of ordinal information using weighted median aggregation. *International Journal of Approximate Reasoning*, *18*(1-2), 35-52.

Yager, R. R. (2007). Aggregation of ordinal information. *Fuzzy Optimization and Decision Making* 6, 3), 199-219

Yan, H. B., Huynh, V. N., Nakamori, Y. and Murai, T. (2011). On prioritized weighted aggregation in multi-criteria decision making. *Expert Systems with Applications*, *38*(1), 812-823. Elsevier Ltd. doi:10.1016/j.eswa.2010.07.039.

Zadeh, L. A. (1975). The concept of a linguistic variable and its application to approximate reasoning. *Information Sciences*, 8(3), 199-249.

Résumé

Les systèmes d'aide à la décision basés sur les entrepôts de données constituent un outil puissant pour l'analyse de données multidimensionnelles. Dans ces systèmes, les faits sont analysés à travers des indicateurs (des mesures numériques) selon diverses perspectives d'analyse (dimensions). Cependant, dans les domaines d'applications émanant du monde réel, les faits ne sont pas toujours numériques, et peuvent être de nature qualitative. Par ailleurs, il arrive parfois qu'un modèle expert ou appris comme un arbre de décision fournisse une estimation qualitative d'un phénomène en fonction de différents paramètres. Par conséquent, les entrepôts de données conventionnels ne sont pas adaptés au raisonnement humain et n'ont pas la capacité de traiter des données qualitatives. Dans ce papier, nous proposons une approche originale de modélisation d'entrepôt qualitatif de données. En se basant sur l'approche linguistique "Computing with words", nous proposons une extension du modèle multidimensionnel classique pour permettre l'agrégation et l'analyse des données qualitatives dans un environnement OLAP. Pour illustrer la problématique et notre solution, nous considérons dans ce papier le cas des gênes des chantiers urbains.

A framework for scalable NoSQL storing moving objects' trajectories

Azedine Boulmakoul *, Lamia Karim **

FST Mohammedia, Département informatique
BP 146 Mohammedia 20650, Mohammedia, Morocco
*azedine.boulmakoul@gmail.com
** lkarim.lkarim@gmail.com

Abstract. Size of spatio-temporal data generated by positioning systems and researches dealing with trajectories data grows so large that it is difficult to store, manage, analyze and visualize using heterogeneous models and relational data bases. This work provides a performance and scalable framework to gather, store and visualize different kinds of geographical data based on the unified moving object trajectories' Meta-model. The proposed framework offers components to collect spatio-temporal data from GPS enabling devices using .Net sockets, store and process trajectories "big data" in a NoSQL database "MongoDB". Furthermore, it displays complex spatio-temporal trajectories in a scalable manner that allows users to easily interpret and identify moving objects trajectories' features for further analysis. In order to test the scalability of the proposed collector system, a vehicle tracking simulator is developed to produce spatio-temporal data outputs of different moving objects.

1. Introduction

Nowadays, it becomes more and more important to combine different applications fields with spatial and time related information. "80% of All Information is Geospatially Referenced" (Fitzke et al, 2010) and there is a large diffusion of mobile devices, mobile services, and location based services. Providing location based services (LBS) have multiple challenge as scalability, performance, query processing, high-precision positioning, and privacy preservation. Therefore, LBS growth and need unified model to deal and explore captured data, e.g. a system for destination and future route prediction based on trajectory mining (Chen et al.2010), real-time monitoring of water quality using temporal trajectory of live fish (Heng et al. 2010), analyzing bird migrations trajectory (Spaccapietra et al. 2008), automatic monitoring of vehicles (Kaplan et al., 1996), etc. Spatio-temporal datasets grow so large that is difficult to capture, store, manage, share, analyze and visualize using classical models and databases. Also, visualizing the implicit information of moving objects trajectories data is very important for analyzing human activities and is of great value in the decision making process.

The main objective of this paper is to present a performance and scalable geographical collecting, storing and visualizing framework for the unified moving object trajectories meta-model that benefits from a general conceptual model. Although the storage and processing difficulties associated with tremendous trajectories raw data are overcome using alternative database management model called as NoSQL.

The remainder of the paper is organized as follows: section 2 will provide an overview of the Unified Moving Object Trajectories Meta-Model. Section 3 will present performance database benchmarking results for storing trajectory's data at scale and section 4 propose the best suitable file format for visual exploration of trajectory data. Whereas section 5 highlighting the scalable data collector conception and the adopted technology. Section 6 will provide an evaluation of the proposed framework when dealing with a very large amount of spatio-temporal data in real time. Finally, section 7 will provide conclusions of our proposed framework.

2. Unified Moving Object Trajectories Meta-Model Overview

The Unified Moving Object Trajectories' Meta-model (Boulmakoul et al., 2012) describes a general meta-model that could be used by different application domains; it can also use an object approach and integrates previous trajectories models described in literature (Yan et al., 2010), (Giannotti et al., 2007), (Meng et al., 2003) and (Wolfson et al., 1998). Using the space-time event ontology, the meta-model models Space according to OGC Spatial Data Model (OGC, 2008), Observation domain of trajectory, according to OGC Sensor Meta Model and OGC Feature Type, Physical and virtual activities between the beginning and the end of Space Time Path (Shaw, 2011) and (Faisal et al., 2007), sensors used for collecting moving object's traces, and Movement patterns using composite Region of Interest.

3. Performance database benchmarking for storing trajectory's data at scale

The rapid increase in the use of inexpensive indoor and outdoor location sensors has led to unprecedented excitement about location based services. To build these services, it is necessary to collect tremendous amounts of data about users' trajectories and activities at each space time event. However generated data from indoor and outdoor location sensors (*GPS*, *RFID* and *WIFI*) are not natively in the same structured format but used for the same target applications.

Current relational database storage systems are designed for static application data model in which data volumes were small and the database lived on a single server in one data center. Hence, traditional database is inadequate for manipulating the volume, velocity and variety of all dynamic spatio-temporal datasets, required to support such services, when we favor performances rather than guarantee of writing data. In order to decide whether relational or *NoSQL* database will be used to provide a powerful and scalable framework for collecting and visualizing moving object's trajectory's data. We analyse performance and do benchmark testing the geospatial performance of the most popular open sources in *NoSQL* Databases and relational systems.

3.1. NoSQL databases

The acronym *NoSQL* signifies "*Not Only SQL*" (NoSQL Wiki, 2013) and (Mike, 2012). It is designed for storing data in a much simpler, flatter, and non-relational manner that allows data repositories to be scaled up. In a *NoSQL* database, there is no fixed schema so we can store, in the same entity, heterogeneous spatio-temporal data and activities generated by

different kinds of locations sensors. Also, they are often open source, non relational, distributed and often don't guarantee *ACID* of relational database (Atomicity, Consistency, Isolation and Durability). Relational database scales up by getting faster hardware and adding memories whereas *NoSQL*, on the other hand, can take advantage of scaling out by spreading the load over many commodity systems. Consequently, *NoSQL* is an inexpensive database for scaling trajectories datasets. Companies like (*Google*, *Facebook*, *Twitter*, *Amazon*, *Twitter*, *Adobe*, *Viadeo*) have left the relational world and all use *NoSQL* in one way or another because they have seen their needs in terms of load and data volume grow exponentially. Existing *NoSQL* solutions can be grouped into 4 main families: *Key-values Stores*, *Column Family Stores, Document Databases*, and *Graph Databases*.

3.2. Databases benchmarking

Performance of storing process can really only be understood once the databases are running at scale but that is too late to be discovering problems. Our main interest is how to best represent and store spatio-temporal data for efficient processing, analysis, and visualization. Generated data by location sensors are usually in the form of *Extensible Markup Language XML* (W3C, 2008), *GPS Exchange Format GPX* (GPX, 2007), *Comma-Separated Values* CSV (Shafranovich, 2005) or *JavaScript Object Notation JSON* (ECMA-262, 2011) documents. Furthermore, we need to generate documents such as *XML*, *Keyhole Markup Language KML* (OGC KML, 2012) or *JSON* for moving objects trajectories visualisation. Thus, we choose in following to analyse the most popular *NoSQL* database document "*MongoDB*" with the object-relational database management system *PostgreSQL* with *PostGIS* extension.

3.3. What is MongoDB?

MongoDB (from "*humongous*") is a scalable, high-performance, open source *NoSQL* database developed by 10gen in 2009 (10g10, 2013). It is written in C++, document-oriented storage, full Index, rich document-based queries, and flexible aggregation and data processing. MongoDB may contain several databases. Using *JavaScript* for its query language, *MongoDB* supports both single and complex queries. Storing *JSON* documents, the basis documents format of many modern geospatial applications, makes it easy to build on top of *MongoDB*. MongoDB database benefits from ascending, descending, unique and geospatial indexes. To makes performance better, *JSON* is stored by *MongoDB* in an efficient binary format called *BSON* (BSON, version 1.0). *BSON* is a binary serialization of *JSON* documents and stands for Binary *JSON*.

3.3.1. PostgreSQL with the PostGIS extension

PostgreSQL (PostgreSQL, 2012) was selected as the object-relational database management system for of exciting features. Indeed, it is highly regarded reputation as an enterprise-class open source relational database management system. Also, *PostGIS* extension to the *PostgreSQL* server, allowing it to be used as a backend spatial database for geographic information systems (*GIS*) that conform to the *OpenGIS* Simple Features Specification for *SQL*. Furthermore, it supports for unlimited rows and indexes (3D and 4D indexing) per table, as well as a very large (32 TB) maximum table size (PostgreSQL, 2012).

3.3.2. Geospatial databases benchmarking

Managing *GIS* data with *NoSQL* or Relational systems in circumstances where performances and scalability are a major issue could be the way for the win in Locations Based Services. We started our benchmark by developing a testing application in .NET environment using C#. Both databases are evaluated in the same 64-bits machine, using latest versions of *PostgreSQL* with *PostGIS* extensions (version 9.2.2) and *MongoDB* databases (version 2.2.3). The intention of this benchmark is to test databases performance for their geospatial data store while saving raw data generated by our tracking simulator. Each message sent from *GPS* enabled devices is an object with a location and multiple tags (key-value-pairs). In our test, we used files generated by our tracking simulator; the structure of a collected tracking message is as follows: *moving_object_id; latitude; longitude; date; time; observation*.

From geospatial point of view, *MongoDB* allows geospatial indexing points natively (without extension); contrary to *POSTGIS* allowing uses of not only geometric points but also other more advanced geographical objects (e.g. line string, polygons).The testing scenario is for each test we created a new table for both databases in order to keep the same environment conditions. Then, we run program by using indexes and geometric objects points. In fact, we are interested in recording raw data trajectories which is evidently geometric points that are supported by both *PostgreSQL* with the *PostGIS* extension and *MongoDB* databases. A similar benchmark between *MongoDB* and *PostGIS* was performed in a 32-bit environment (Suter, 2012) proves that *MongoDB* is more scalable than *PostgreSQL* with *PostGIS* extension for inserting data. This benchmark, shown in Fig 1, proves that MongoDB is much faster for geographical data storage. We explain this performance by the fact that *MongoDB* supports geospatial indexes with no dedicated geospatial data type. Whereas *PostGIS* uses two tables in the background, *spatial_ref_sys* and *geometry_columns* (Ramsey,2012) to retrieve and learn about the types of geometry available in a specific database. Moreover, MongoDB's architecture supports horizontal scalability, and high availability through replica sets. Auto-sharding allows distributing load across multiple servers and keeping data balanced across those servers.

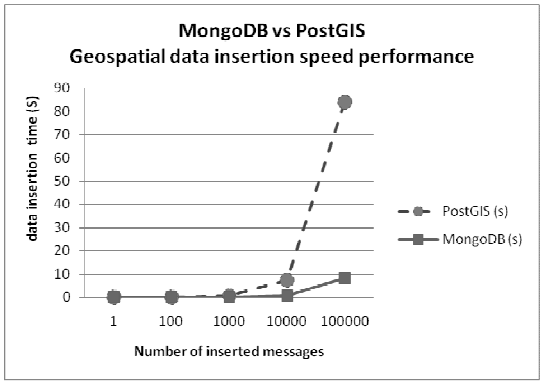


Fig. 1. *Comparison of geospatial data insertion performance of MongoDB and PostgreSQL with PostGIS extension.*

4. Exporting trajectory's data from MongoDB data to geographic visualization tools

The second objective of this paper is to propose the best suitable file format for visual exploration of trajectory data on the web using existing visualization libraries. As we have mentioned in previous paragraph, *MongoDB* is a *NoSQL* oriented document database for saving spatio-temporal points in a scalable manner. In fact, another reason whereby we have chosen this type of database "oriented document" is the requests output format that is as document type, this *JSON* document could be directly used by visualization libraries. Before deciding to use this file format, as an intermediary between database and visualization tools, we seek if it is supported by majority of visualization tools and then evaluate it using following general criteria: speed performance, availability of documentation, size variability, and the speed of debugging program.

Several new versions of web interactive visualization libraries are available to put a dynamic map and display trajectories in any web page. The most useful and popular libraries are *Openlayers* (OPENLAYERS, 2012), *SIMILE* (SIMILE, 2009), *Timemap* (URL 4), *Google MAP API* (Google MAP API, 2011), *Processing* (Processing, 2012), *ESRI JavaScript API* (ESRI, 2009) and they run on most of the recent browsers (e.g. *FireFox*, *Opera*, and *Internet Explorer*). These libraries include *JavaScript* libraries and allow user to load multiple datasets in *GML/KML*, *GeoRSS* format (which can only deal with points) and generic vector formats as *JSON* and *GPX*. Each of these libraries has its own advantages and disadvantages for a case of study (Hoang, 2010). In (Li et al., 2012), they proposed to convert raw data to *GML* format for storage and transmission, and then using *XSLT* and *XPATH* technologies, *GML* is converted to *KML* to achieve the rapid visualization of trajectory data in browser. (Krizek, 2012) analyzed and compared data formats *Extensible Markup Language* (*XML*) and *JavaScript Object Notation JSON* for exchanging data. They found that speed performance is the best for *JSON* documents because *JSON* size overhead is much smaller and can be in the most cases rewritten to *XML* and vice versa. In fact *JSON* requires less tag than *KML – KML* items must be wrapped in open and close tags whereas *JSON* just name the tag once. Also, we prefer *JSON* as it encapsulates any *JavaScript data*. Contrary to *KML* and *GeoRSS*, data must fit in predefined *XML* elements (Udell, 2009) and if we want to present other information like activities of a moving object we blocked.

Since results format returned from *MongoDB* are in *JSON* (*JavaScript Object Notation*) with no needed conversion, and also *JSON* is much faster than other *XML* based technologies. We use *JSON* format, in the proposed framework, to display the trajectories in browsers.

5. Scalable data collector system for the Unified Moving Object Trajectories Meta Model

Using Sockets interfaces aim to make real-time data collector applications possible between data sources (*GPS* equipped mobile phones, device camera, vehicles with navigational equipment or location based services, etc.) and data collector server.

5.1. Data collector server socket

Sockets are a low-level network programming features used in a standard protocol running on the same network. In data collector server, .Net sockets API are used to create an endpoint bidirectional communication between the collector server and client data sources. *Sockets* are used as a transport mechanism for creating high-performance communication link between two endpoints (client & server). *Sockets* can be classified into synchronous and asynchronous mode depending on the type of operation performed on that socket. On one hand, when blocking (synchronous) sockets, programs are "blocked" and waiting to be dealt with until the operation on the socket is fully completed. On the other hand, for asynchronous sockets, the server application is permitted to respond events (non-blocking) upon the completion and execution of the process.

In our collector system, we used asynchronous mode with a pool of sockets objects and reusing them to collect geographical data and get notification to recognize error locations and successful operations. In reality, a socket pool increases performance on a server in a situation where there are many clients connecting and disconnecting quickly. In order to develop the scalable system collector for the unified trajectories meta-model, .Net sockets classes are likely used. Both classes, *System.Net.Sockets.Socket* and *System.Net.Sockets.SocketAsyncEventArgs* are used by specialized high-performance socket applications and specifically designed for network server applications that require high performance (MSDN Library). In addition, communication between data sources simulators and server data collector is managed by event programming to collect in a non-blocking mode on different thread. As a result, none of the clients' sockets is suspended while waiting for the network operation to complete. The following diagram shows classes diagram that make up our data collector system.

5.2. Data collector class diagram

Class diagram, shown in Fig 2, describes potential utility classes of the Socket data collector server. The *SocketAsyncEventArgs* class is a context object for the *System.Net.Sockets.Socket* class, which supports non-blocking network I/O in Net framework.

Using this advanced class, a high performance data collector system is provided. *SocketListener* class manages connection and messaging with multiple mobile clients who use the *acceptAsync* method of the *SocketAsyncEventArgs* class and may have a list of connections objects to store the accepted sessions. The *SocketClient* class uses the *connectAsync* method of the *SocketAsyncEventArgs* for the sake of connection. *SocketListener* class's main role is to listen and response to the messaging transactions of each socket session. Each connection has a *SocketAsyncEventArgs* instance for the messaging operations (receive and send). So, the *SocketAsyncEventArgs* pool manager would provide the instance library of many *SocketAsyncEventArgs* instances. There are two pools in context of *SocketAsyncEventArgsPool* to deal with accepting connection and receiving/sending socket operations. Once the accept operation completes, a reference to the socket transferred to another *SocketAsyncEventArgs* object as fast as required. *SocketListenerSettings* class holds a lot of settings that are passed to the *SocketListener*, Prefix Handler manages how to get data in the *TCP* buffer, whereas *MessageHandler* class creates the array where the complete message will be stored and handled with collected messages data whether it requires one or many receives operations.

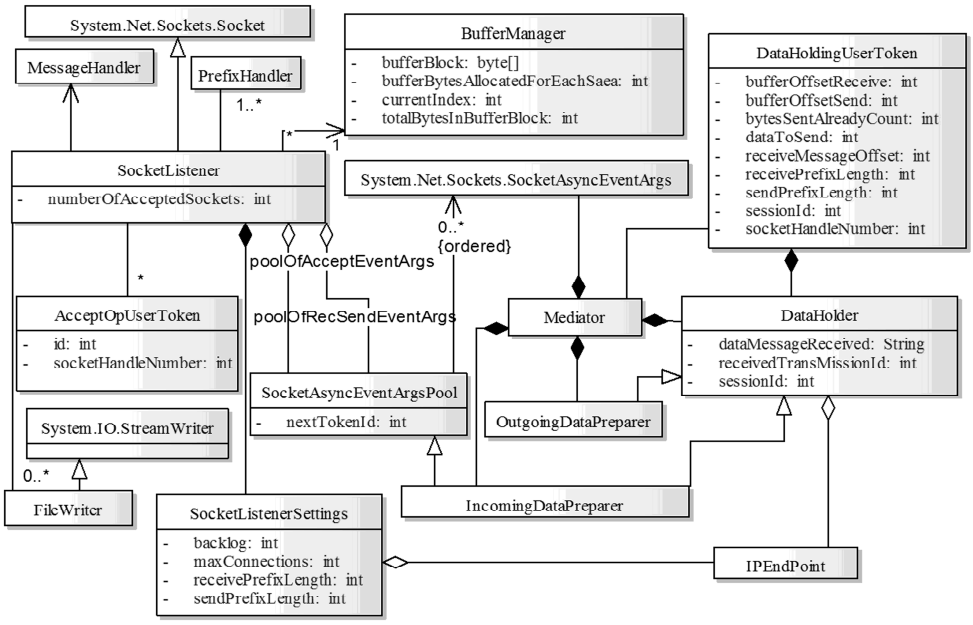


Fig. 2. *Data collector class diagram.*

6. Experimentation

In this section, there will be a brief analysis of the framework intended as a scalable moving object's data collector system for the unified moving object meta-model.

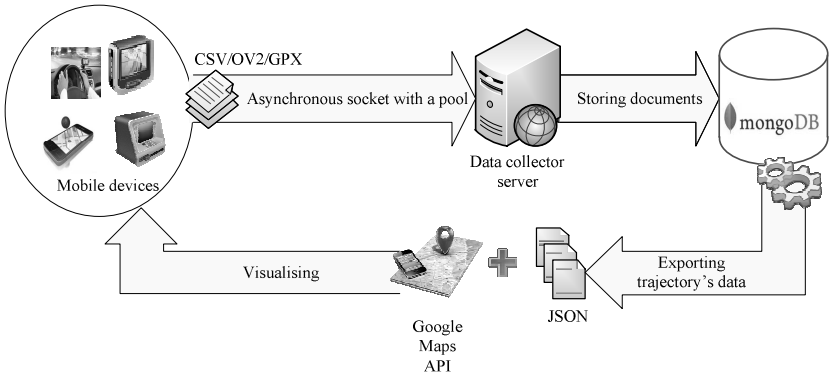


Fig. 3. *Trajectory's data collector system architecture.*

Using *.Net* sockets, the collector framework, shown in Fig 3, offers components to collect trajectories data sets as *GPX*, *OV2*, or *CSV* files from different *GPS* enabling devices. Then, using sockets, data collector system gathers different form of geographical data and stores theme in *NoSQL* document database *MongoDB*, to benefit from the database scalability, and

speed performance. Finally, querying results of stored trajectory's data are exported as JSON documents in order to be visualized using *Google Maps API* in mobiles devices.

In order to test the scalability of the proposed collector system, we developed vehicle tracking simulator to generate tremendous spatio-temporal data that *GPS* devices generate. In the following, these components are described and the worst-case testing is analyzed.

6.1. Vehicle tracking simulator

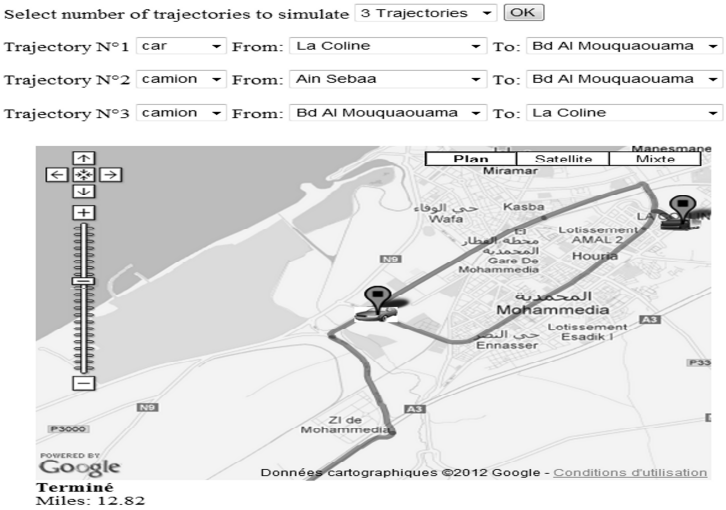


Fig. 4. *Vehicle tracking simulato*r.

Simulator, given in Fig 4, is developed to provide a case study for evaluating and testing the collector framework. It allows generating real vehicles' spatio-temporal data. Users launch the simulator by choosing number of vehicles to track, time delay, and also define geographic points where the trajectory of each vehicle (moving object) starts and ends. This simulator developed using *GoogleMaps Javascript API* and *C#* as programming language with *ASP.NET* Web Pages to generate driving directions and display them in the map.

6.2. Worst-case testing

The goal is to test the performance and scalability of our proposed spatio-temporal data collector framework under the worst-case conditions. The testing strategy is as follows: First, the collector server socket is listening to connection and the performance monitor is launched to capture the system performance. Then, from the client's side, moving objects to track is chosen. After that, the number of moving object's trajectories is grown to simulate and analyze the performance of the collector server. Fig 5 describes the time consumed to collect and save collected data, in *NoSQL* documents database *MongoDB*, like when collecting one message from 1 till 10 000 moving objects, and when taking 0.5s to 1s as time delay in a *Windows* 7 *Ultimate computer*, 64 bit Operating System, *Processor core i5* CPU, and 4 *GB* installed Memory (*RAM*).

The total time consumed for collecting and saving messages in *MongoDB* varies between 9 ms when collecting 1 message from 1 moving object, 67s when collecting and saving messages from 10000 moving objects tracked, and about 4s to collect and save 50 000 messages from one moving object.

The structure of each message sent from *GPS* enabled devices is as follows: *moving_object_id; latitude; longitude; date; time; observation.* Whereas structure of messages collected using banking transactions or Electronic payment terminals are as follows: *moving_object_id; electronic_terminal_id; date; time; amount; credit_card_kind; credit_card_crypted_number; debit_or_credit. (From electronic_terminal_id we recognize the geographic position of moving object).*

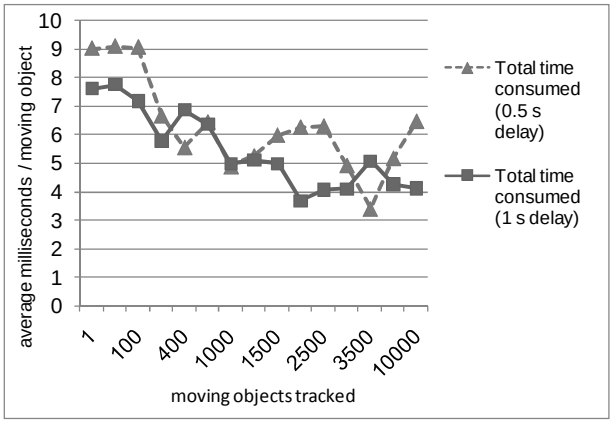


Fig. 5. *Data collector system total time consumed (collecting and saving data).*

As regards our proposed framework scalability, tests demonstrate that variation of system resources like *Memory*, *Physical Disk*, and *Processor* is lightly and controlled. Therefore, our collector framework is able to handle a growing amount of work in a capable manner.

As a result, our collector system is scalable and efficient for different types of applications such as: Tourist Guides Community; fleet management, tracking of goods, trucks, and taxis; advertising; urban cleanliness and sanitation; protection of goods, vehicles and antitheft; Behavioral studies of human beings.

6.3. Trajectories visualization

We visualize trajectory's data stored in *MongoDB* using the newest version of *Google Maps Javascript API* and *Jquery*, to replay dynamically trajectories on maps. As discussed previously, we use *JSON* files as *MongoDB* request output and also as *Google Maps Javascript API* data source (Fig 6).

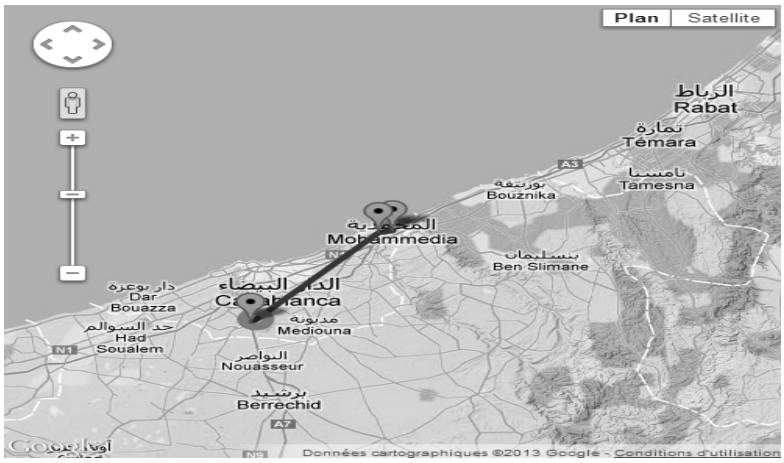


Fig. 6. *Trajectories visualization using Google Maps Javascript API and Jquery.*

7. Conclusion

The intent of this study was to develop a scalable prototype framework to deal trajectories big data, with five primary goals. First, it is based on unified conceptual Meta model for trajectories representation. Furthermore, it simulate the production of *GPS* tracking outputs to test the performance and scalability of the proposed data collector framework and also to see how the relational/*NoSQL* databases scale when receiving a growing amount of spatio-temporal data. Using Sockets interfaces make real-time data collection possible between data sources (*GPS* enabled devices, Database transactions, IP Cameras) and data collector server.

Despite the fact that *NoSQL* databases have a number of significant advantages, they also have a number of setbacks. Critics point to *NoSQL's* lack of maturity, standards, business intelligence limitations, and possible instability issues as they are in pre-production versions with many key features yet to be implemented. Indeed, for performance reasons, we used *MongoDB* as a *NoSQL* document database rather than *POSTGRES* with *POSTGIS* extension for its geospatial data store as well as *JSON* files rather than *GML*/*KML* files for transferring data to visualization libraries.

References

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

Binary JSON, version 1.0. Available from: http://bsonspec.org/#/specification.

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

Chen, L., L. Mingqi, G. Chen (2010). *A system for destination and future route prediction based on trajectory mining*. In:Pervasive and Mobile Computing, Elsevier Science Publishers; p. 657-676.

ECMA Standard-262 version 5.1(2011).

ESRI JavaScript-API VERSION-1.6. http://resources.esri.com/help/9.3/arcgisserver/apis /javascript/arcgis/help/jshelp_start.htm#jshelp/overview_api.htm

Faisal, I., A. Khokhar,and D. Schonfeld (2007). *Object Trajectory-Based Activity Classification and Recognition Using Hidden Markov Models*. Image Processing, IEEE Transactions; p. 1912 – 1919.

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

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

Google MAP API v3 (2011). Available from: http://code.google.com/apis/maps/.

GPS eXchange Format (2007). Available from: http://www.topografix.com/gpx.asp.

Heng, M., TF. Tsai , and C. Cheng (2010). *Real-time monitoring of water quality using temporal trajectory of live fish*. Expert Systems with Applications; p. 5158–5171.

Hoang Long, N. (2010). *Web Visualization of Trajectory Data using Web Open Source Visualization Libraries.* Thesis submitted to the International Institute for Geo-information Science and Earth Observation.

Kaplan, E.D. (1996) . *Understanding GPS Principles and Applications*. Artech House Publishers.

Krizek, V. (2012). *Intelligent Visualization of data from Multi-agent Simulations. Thesis. Czech Technical University in Prague.* faculty of Electrical Engineering

Li. J., J. Wang, L. Yu, R. Qi, and J. Zhang (2012). *Visualization Method of Trajectory Data Based on GML, KML*. Advances in CSIE, Springer-Verlag Berlin Heidelberg; Vol. 2, AISC 169, p. 479–484.

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

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

MongoDB C# Driver (2012). Available from: http://docs.mongodb.org/ecosystem /drivers /csharp

MSDN Library; URL: http://msdn.microsoft.com/en-US/en-us/library/system.net.sockets-.socketasynceventargs.aspx

NoSQL Wiki (2013). http://fr.wikipedia.org/wiki/NoSQL

OGC 07-022r1 Version: 1.0(2008). Available from: http://www.opengeospatial.org/legal/.

OGC KML (2012). Available from: http://www.opengeospatial.org/standards/kml/.

Openlayers, version 2.12. (2012). Available from: http://www.openlayers.org/.

PostgreSQL, versions 9.2.2 (2012). Available from:http://www.postgresql.org/"

Processing (2012). Available from: http://processingjs.org/.

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

Shafranovich, Y., (2005). Common Format and MIME Type for Comma-Separated Values (CSV) Files. RFC 4180. Available from: http://www.ietf.org/rfc/rfc4180.txt

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

SIMILE (2009). Available from: http://code.google.com/p/simile-widgets/.

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

Suter, R., (2012). *MongoDB An introduction and performance analysis*. Seminar Thesis. University of Applied Sciences Rapperswil.

Udell, S. (2009). Beginning Google Maps mashups with mapplets, KML, and GeoRSS : from novice to professional. Berkeley; chapter 10 p. 255–264.

W3C (2008). Extensible Markup Language (XML) 1.0 (Fifth Edition). Available from: http://www.w3.org/TR/REC-xml/.

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

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

Résumé

La taille des données, spatio-temporelles générées par les systèmes de positionnement et les recherches traitant les données spatio-temporelles des trajectoires, augmente au point qu'il est difficile de les stocker, les gérer, les analyser et les visualiser. En se basant sur le méta model unifié des trajectoires, nous fournissons dans ce travail un système scalable et performant pour collecter, stocker et visualiser les différents types de trajectoires. Le Framework proposé offre des composants pour collecter les données spatio-temporelles générés par les systèmes de localisation. Il utilise les sockets asynchrones avec pool de threads pour la programmation de bas niveau en environnement *.Net*, stocke et traite les données dans une base de données *NoSQL* de catégorie documents "*MongoDB*". En outre, il traite les données spatio-temporelles des trajectoires d'une manière évolutive permettant aux utilisateurs d'interpréter et d'analyser les trajectoires des objets mobiles. Afin de tester la scalabilité du system proposé, nous avons développé un simulateur produisant les données spatio-temporelles générées par différents objets mobiles. Les résultats obtenus de cette évaluation montrent que le système est scalable et performant pour différents types d'application.

Un système d'aide à la décision spatiale de groupe : Couplage analyse multicritère et théorie des jeux satisfaisants

Djamila Hamdadou*,
Sarah Oufella**, Karim Bouamrane ***Fouzia Amrani****
Laboratoire d'Informatique d'Oran (LIO). Département d'Informatique,
Faculté des Sciences ,Université d'Oran, BP 1524. El Mnaouer, Algérie
ENPO (Ecole Nationale Polytechnique d'Oran) Laboratoire LIO.S
dzhamdadoud@yahoo.fr
osarah84@yahoo.fr
kbouamranedz@yahoo.fr
amranibf@yahoo.fr

Résumé. Notre contribution, par la présente étude, porte sur la conception et l'élaboration d'un système interactif d'aide multicritère à la décision collective dans l'objectif de soutenir efficacement la problématique de localisation spatiale en Aménagement du Territoire (AT).La démarche décisionnelle multicritère de groupe proposée, dans cet article, permet de prendre en compte à la fois la dimension spatiale ainsi que les intérêts spécifiques et divergents des différents décideurs afin de parvenir à un accord acceptable par les différents acteurs du territoire. Dans cette optique, nous concevons un système d'aide à la décision spatiale basé sur un couplage de deux représentations de la réalité : les Systèmes Multi Agents (SMA) et les Systèmes d'Information Géographique (SIG). Nous dotons le module SMA d'un protocole de négociation basé sur la médiation et l'Analyse Multicritère d'Aide à la Décision (AMCD) et exploitant les avantages offerts par la théorie des jeux.

1 Introduction

Dans la présente étude, nous contribuons à la résolution des problèmes de décision relatifs à la gestion du territoire en proposant une approche méthodologique pour un outil interactif d'aide à la décision multi décideurs basée sur la négociation multilatérale, l'analyse multicritère et la théorie des jeux en combinant

- **Les systèmes d'information géographiques (SIG)** : permettant de représenter le territoire. Grâce à leur capacité de gestion, d'analyse, de modélisation et d'affichage de données à référence spatiale, se présentent comme l'outil le plus adéquat pour appréhender les problèmes de décision spatiaux.
- **Les Systèmes Multi Agents (SMA):** sont très adaptés pour modéliser des entités complexes pouvant coopérer, collaborer ou négocier afin de parvenir à un accord. Dans le contexte de notre étude, les SMA permettent de représenter la diversité des décideurs concernés par la décision en Aménagement du Territoire (AT), leurs comportements ainsi que leurs interactions.

A cette fin, nous nous intéressons à formuler une démarche générique et interactive d'aide multicritère à la décision de groupe , pour faire face a une problématique qui consiste en la recherche d'une surface satisfaisant au mieux certains critères pour une construction donnée (tels que celui de la localisation d'une infrastructure : centre commercial, centre hospitalier, etc.).L'article décrit en section 2 notre contribution dans sa globalité. En section 3, les principaux travaux en aide à la décision spatiale sont présentés. La section 4 décrit, quant à elle, le système d'aide à la décision de groupe SIGMAS proposé et le protocole de négociation proposé est décrit, en détail, en section 5. Les caractéristiques de ce dernier sont énumérées en section 6 et en section 7, nous justifions l'intérêt de l'utilisation de la théorie des jeux dans le processus de négociation proposé. La section 8 est consacrée à une étude de cas réel qui constitue en soi une première étape de validation. Enfin, nous concluons notre propos, en section 9, en donnant quelques perspectives.

2 Contribution

Le système d'aide à la décision de groupe SIGMAS proposé dans cet article s'appuie sur la représentation du territoire grâce aux fonctionnalités du SIG et sur la représentation de la multiplicité et de la diversité des acteurs grâce aux fonctionnalités du SMA. Ces derniers dotés de protocoles de négociation constituent une approche pour l'aide à la décision de groupe (la négociation et participation de plusieurs acteurs). Dans la présente étude, nous visons à satisfaire au mieux les points suivants :

- Permettre une analyse détaillée des problématiques d'AT en exploitant les SIG;
- S'appuyer sur une stratégie d'intégration SMA-AMCD-Théorie des jeux contribuant à optimiser les résultats de l'aide à la décision de groupe ;
- Permettre l'intégration des points de vue divergents de plusieurs acteurs et faciliter ainsi les stratégies de négociations en adoptant une stratégie d'intégration SIG-SMA.

Plus précisément, nous visons à mettre en place une démarche décisionnelle dont l'objectif principal est de représenter la multiplicité des acteurs, leur diversité, leur comportement ainsi que leur interaction. Dans cette perspective, l'approche proposée a l'avantage de simuler les conséquences de l'agrégation multicritère, en utilisant la méthode ELECTRE III Roy et Bouyssou (1993), selon différents critères quantitatifs et (ou) qualitatifs dans un contexte spatial multi- acteurs et multi- échelles et permet ainsi de concevoir un instrument interactif d'aide à la négociation en AT mettant en scène un ensemble d'agents participants (des agents négociateurs ainsi qu'un agent initiateur (coordinateur)) qui tente de trouver un compromis satisfaisant au mieux les participants.

3 Travaux Connexes

Plusieurs systèmes d'aide à la décision en AT ont retenu notre attention. Dans Joerin (1997), MEDUSAT est proposé pour la localisation de l'emplacement d'une usine de traitement des déchets en Tunisie. Le système MEDUSAT allie un outil SIG à l'analyse multicritère. Dans Boulemia et al. (2000), les auteurs présentent quelques outils d'aide à la décision dans des collectivités locales à partir de système d'information géographique pour répondre aux problèmes de la gestion des eaux et dans Laouar (2005), l'objectif principal est d'apporter aux décideurs du territoire une aide adaptée à la prise en compte des nouveaux enjeux en

matière de déplacement. Tous ces systèmes intègrent à divers niveaux les outils d'analyse multicritère couplés aux SIG, toutefois ils considèrent les critères comme indépendants et sont incapables de modéliser une quelconque interaction entre eux (interchangeabilité, corrélation, dépendance préférentielle, etc.). Dans Hamdadou (2007), nous avons déjà abordé de manière significative la prise en compte de corrélation entre critères en introduisant l'intégrale de Choquet (au lieu de la somme arithmétique) comme opérateur d'agrégation dans les méthodes d'AMCD particulièrement Electre Tri. La discipline de l'AT est amenée, aujourd'hui à concilier un nombre croissant d'objectifs souvent divergents, défendus par une grande diversité d'acteurs (individus ou organisations). La littérature associée à l'aide à la décision spatiale et territoriale, offre peu de systèmes qui considèrent cet aspect. Dans cette optique deux principales études sont à l'origine du présent article à savoir Hamdadou et Libourel (2009) et Hamdadou et Bouamrane (2012).

4 Le système SIGMAS proposé

Par soucis de simplicité, nous choisissons un **"couplage lâche"** entre les modules SMA et SIG qui restent indépendants et communiquent uniquement en s'échangeant des données. L'approche préconisée est générique, la vision spatiale du territoire est injectée dans le SIG, et le SMA assure la négociation multi-acteurs. Nous détaillons, dans ce qui suit le système SIGMAS:

4.1Le module SIG (modèle du territoire)

Etant intrinsèquement un outil de gestion des informations géographiques, le SIG aura pour fonction essentielle de permettre la gestion des connaissances du domaine. Grâce à ses fonctionnalités, il est possible de :Manipuler et interroger les bases des données géographiques; fournir une représentation spatiale des systèmes étudiés; Mettre en forme et visualiser les données. Dans le contexte de note étude, nous exploitons aussi les fonctionnalités du SIG pour préparer les entrées (inputs) nécessaires pour l'analyse multicritère. Lorsque les décideurs parviennent à identifier les actions[1] et les critères[2], grâce aux capacités analytiques du SIG, une valeur (note) est affectée à chaque critère. L'ensemble des jugements des actions relativement aux différents critères constitue la matrice d`évaluation (Tableau de performances[3]).

4.2 Le module SMA (modèle de la négociation)

Nous déléguons au SMA le choix final de la ressource élue après négociation, la ressource choisie sera visualisée grâce aux fonctionnalités du SIG. Le module SMA implique deux types d'agents :

1. **L'agent initiateur (coordinateur):** c'est l'agent responsable de la gestion de la négociation, de la modification de contrat et du choix final concernant la ressource élue.

[1] Une action est une solution possible au problème décisionnel considéré, une alternative ou encore une ressource.

[2] Une expression quantitative ou qualitative permettant d'examiner les actions.

2. **Les agents participants (contractants):** ce sont les agents concernés par la décision en AT, l'objectif de chacun de ces agents est que sa ressource préférée soit choisie.

Il est indispensable que les agents participants passent par une phase de négociation, selon un protocole bien structuré que nous proposons et détaillons dans les sections suivantes. La figure 1 illustre une vue d'ensemble de SIGMAS.

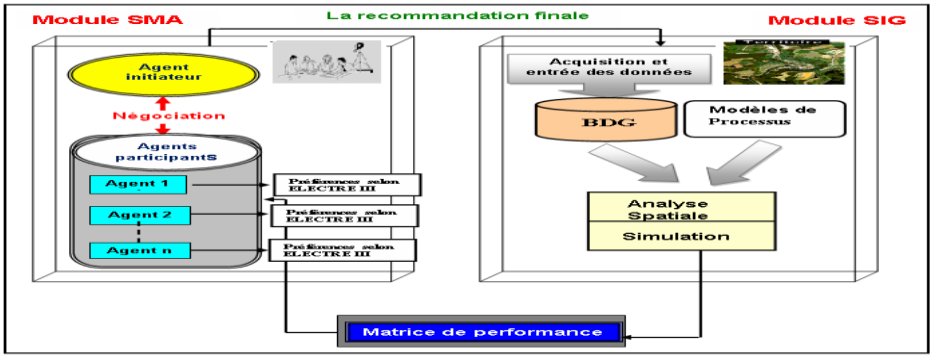


FIG. 1 – Le système d'aide à la décision de groupe SIGMAS : une vue d'ensemble

5 Le protocole de négociation proposé

Le protocole de négociation que nous proposons s'inspire du Contrat Net Protocol et constitue une optimisation du protocole proposé dans Hamdadou et Bouamrane (2012)**.** Il se caractérise par une suite de messages échangés entre un agent initiateur et des agents participants. Il procède en cinq phases:

1. **La phase d'initialisation :** les participants sont appelés à exprimer leurs préférences concernant les ressources, chaque participant effectue un classement des ressources de la meilleure (qui lui ait la plus bénéfique) à la moins bonne.
2. **La phase de proposition** : l'initiateur propose un contrat à tous les participants concernant une ressource donnée, les participants vont soit accepter ce contrat soit le refuser en se référant à leur vecteur de préférence[4] construit précédemment.
3. **La phase d'évaluation :** lorsque l'initiateur reçoit toutes les réponses des participants concernant la proposition du contrat, il comptabilise le nombre d'agents ayant accepté sa proposition. Si ce nombre est supérieur ou égal à un certain seuil[5] alors la négociation est un succès sinon il doit procéder à une modification du contrat.
4. **La phase de modification :** durant cette phase, l'initiateur est amené à faire une modification du contrat en s'inspirant des propositions des agents. Il doit établir une-synthèse à partir des réponses reçues lors de la phase d'évaluation puis revient à la phase de proposition.

[4] Classement des ressources de la meilleure à la moins bonne relativement à un certain nombre de critères en utilisant une approche d'agrégation multicritère.

[5] Seuil indispensable pour que la négociation soit un succès, c'est le nombre d'accords.

5. **La phase de décision :** une décision est prise par le coordinateur selon les réponses des participants aux propositions qu'il leur a fait. Elle peut s'achever soit par un succès ou par un échec où une décision finale est prise par l'initiateur.

6 Les caractéristiques du protocole proposé

Lors du déroulement de la négociation, un certain nombre d'aspects doit être pris en compte tels que : les ressources et la cardinalité de la négociation, le langage de la négociation [6] utilisé par les agents pour échanger des informations pendant la négociation ainsi que les stratégies adoptées par les agents au cours du processus de négociation. Dans ce qui suit, nous décrivons en détail les différentes caractéristiques du protocole de négociation que nous proposons dans cette étude**.**

Les ressources de la négociation : Les ressources sont les objets de la négociation. Elles peuvent être soit personnelles, soit communes. Dans notre cas, ce sont des ressources communes (les îlots vierges destinés pour une construction donnée).

Le nombre d'accords : Pour que la négociation soit un succès, l'initiateur doit se référer à un paramètre fixé au départ représentant le nombre minimal d'accords nécessaires pour arriver à un consensus.

Le nombre de tours de parole des participants : Afin d'assurer la convergence du processus de négociation, nous devons définir un nombre maximal de tours de négociation, lorsque ce nombre de tours est dépassé, la négociation est un échec. Dans ce cas, l'agent coordinateur doit prendre une décision finale concernant la ressource choisie en exploitant les avantages de la théorie des jeux.

La cardinalité de la négociation :C'est une notion importante pour les SMA. Il s'agit de savoir combien d'agents négocient entre eux .

6.1 Le langage de la négociation : les primitives

Pour mener à terme un processus de négociation entre agents, il est nécessaire de définir des primitives spécifiques à l'agent coordinateur et d'autres primitives spécifiques aux agents participants.

6.1.1. Les primitives spécifiques à l'agent coordinateur

Les messages envoyés par le coordinateur sont destinés à tous les agents participants (broadcast), nous lui associons, par conséquent, trois primitives de négociation :

- **Request () :** l'agent coordinateur envoie un message aux participants pour les inciter à établir leur vecteur de préférences ;
- **Propose () :** l'agent coordinateur propose un contrat aux agents participants concernant une ressource donnée ;
- **Confirm () :** l'agent coordinateur envoie un message à tous les agents pour les informer que la négociation a été un succès et que la ressource a été trouvé ;

[6] Défini par un ensemble de primitives spécifiques.

- **Failure ():** l'agent coordinateur envoie un message à tous les agents participants pour leur indiquer que la négociation a été un échec et que le nombre de tours de paroles est dépassé.

6.1.2. Les primitives spécifiques aux agents participants

Les messages envoyés par les participants sont uniquement destinés à l'initiateur. Les autres participants n'ont pas connaissance de ces messages. Le participant dispose de trois primitives de négociation :

- **Inform () :** après avoir établit un classement des ressources de la meilleure à la moins bonne, chaque agent participant informe l'initiateur qu'il peut leur faire une première proposition;
- **Accept ()** : ce message répond à la proposition du contrat faite par l'initiateur. Le participant indique par ce message à l'initiateur qu'il accepte le contrat ;
- **Refuse () :** Le participant indique à l'initiateur qu'il refuse sa proposition.

6.2 Les stratégies des agents lors de la négociation

Le protocole proposé distingue deux rôles : initiateur et participant. La stratégie de négociation n'est pas la même, elle diffère selon le rôle de l'agent.

6.2.1. Les stratégies du participant

Nous associons à chaque agent participant trois stratégies :
Stratégie d'établissement de préférences : Chaque agent participant doit établir un classement des ressources depuis la meilleure (celle qui lui ait la plus bénéfique) jusqu'à la à la moins bonne en se référant à un certain nombre de critères. Pour atteindre cet objectif, il exploite les avantages qu'offre la méthode d'aide multicritère à la décision ELECTRE III. Après que chaque participant ait établit son vecteur de préférence, il associe à chaque ressource un rang, la ressource classée première aura le rang le plus grand (elle représentera la préférence du participant lors du premier tour). Ce rang est, à chaque fois, décrémenté de 1 pour les ressources suivantes. Pour pouvoir être conduite, la méthode ELECTRE III nécessite en entrée un ensemble de paramètres intra critères **(seuil de préférence, seuil d'indifférence et seuil de veto)** et de paramètres inter critères (la pondération de chacun des critères) qui représente l'importance de chaque critère selon le participant considéré. Après avoir accéder à la matrice de performance normalisée préalablement, chacun des agents participants exprime ses préférences en terme de critère. La pondération des critères par les agents participants s'avère une tâche délicate, pour cela, nous proposons d'exploiter la méthode multicritères AHP (Analytic Hierarchy Process) Hamdadou, D. Bouamrane, K. (2012).

Stratégie d'acceptation : A chaque nouveau tour, le participant reçoit une nouvelle proposition. Si cette dernière correspond à sa préférence au tour t, il l'accepte. Sinon, il vérifie si la proposition correspond à l'une de ses préférences antérieures, si c'est le cas, il accepte le contrat tout en indiquant sa préférence actuelle.

Stratégie de refus : Lorsque le participant reçoit une proposition et que celle-ci ne correspond ni à sa préférence au tour t, ni à aucune de ses préférences antérieures, il la refuse.

6.2.2. Les stratégies du coordinateur

Nous associons au coordinateur deux stratégies :
Stratégie de modification : Lorsque les participants n'ont pas été assez nombreux à accepter la proposition du coordinateur, ce dernier doit modifier son contrat pour le prochain tour en s'inspirant de toutes les modifications envoyées par les participants au tour t.Afin de trouver une nouvelle possibilité pour le contrat, le coordinateur associe un score à chaque ressource en prenant en considération le poids de l'agent participant ainsi que le rang de la ressource. Pour calculer le score de chaque ressource R_i (i= 1,.. n) lors d'un certain tour t, nous avons utilisé la formule suivante $_m$
$SCORE(R_i)=\sum_{j=1}$ POID(participant[j])*RANG(Ri, participant[j]) (n, m : nombre de ressources et participants respectivement)
Avec :
POID (participant[j]) : compte tenu que dans la réalité, les représentants politiques, par exemple, n'ont pas le même poids que les associations de protection de l'environnement lors d'une décision en AT, nous associons à chaque participant j un poids différent.
RANG (R_i, participant[j]) : le rang associé à la ressource (action) i par le participant j dans son vecteur de préférence (rangement fourni par ELECTRE III) ;
Comme dans les méthodes de scorages, la ressource qui a obtenu le score le plus élevé lors du tour t, sera la ressource gagnante et le coordinateur la proposera dans le nouveau contrat. Ce score est remis à jour à chaque fois que les participants n'ont pas été assez nombreux à accepter le contrat et la ressource gagnante sera visualisée grâce aux fonctionnalités des SIG.

Stratégie de choix : Lorsque le nombre de tours défini initialement a été atteint sans que les agents participants n'arrivent à un consensus, l'agent coordinateur doit prendre une décision qui reflète les préférences de chacun. Pour cela, nous proposons d'exploiter les avantages qu'offre la théorie des jeux et plus précisément la théorie des jeux satisfaisant qui permet l'élaboration d'une préférence collective à partir des préférences individuelles (celles établit précédemment par des agents participants).

La figure 2 représente les différents échanges entre l'agent coordinateur et les agents participants à travers les primitives proposées sous la forme d'un diagramme de séquences UML.

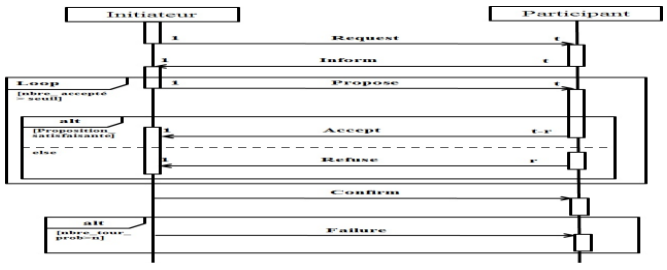


FIG. 2 – Diagramme de séquences UML

7 Pourquoi la théorie des jeux

La plupart des méthodes de surclassement de synthèse ont été développées pour résoudre des problèmes de décision impliquant plusieurs critères mais un seul décideur. L'application de ces méthodes dans les travaux relatives à la décision de groupe n'est pas très recommandée d'où la nécessité de se tourner vers des outils permettant la prise en compte d'éventuels avis contradictoire par rapport a l'importance à accorder aux différents critères. Dans le contexte de la présente étude, ils nous a semblé convoité d'utiliser la théorie des jeux satisfaisants où l'aide à la décision est multicritères et multidécideurs .
Un jeu satisfaisant consiste en la donnée du triplet <U, Ps, Pr> où :

- U est l'ensemble (discret) des options ou alternatives possibles ;
- Ps et Pr représentent des fonctions masses sur U mesurant pour chaque option $u \in U$, respectivement le degré de sélectabilité et le degré de rejetabilité de cette option.

La théorie des jeux satisfaisant permet à l'agent coordinateur de se prononcer et de prendre une décision concernant la ressource choisie et ceci par rapport aux critères plutôt que par rapport aux alternatives Stirling,V.C, **(**2003). A cet effet, le coordinateur doit définir les mesures de sélectabilité et de rejetabilité à partir des mesures des différents critères et l' intégration des jugements des différents agents participants qui n'ont pas la même sensibilité par rapport à l'importance à accorder à chaque critère lors de la décision finale. Dans l'objectif d'aboutir à une décision finale en se référant aux préférences individuelles des agents participant, l'agent initiateur procède comme suit :

a. Le calcul des degrés de sélectabilité $P_S(U)$ et de rejetabilité $P_R(U)$: L'agent coordinateur récupère les pondérations $(\propto_j^k, \beta_j^k)$ qui reflète , respectivement, l'importance de chacun des critères positif et négatif relativement à chaque participant. Avant de calculer le degré de selectabilité et de rejtabilité de chaque ressource, il calcule les pondérations agrégées pour chaque critère j ainsi que le degré de selectabilité et rejetabilité exprimés par les formules suivantes :

Le Degré de séléctabilité **Le Degré de rejetabilité**

$$Ps(u) = \frac{\sum_j \omega_j^s c_j^n(u)}{\sum_{x \in U} \sum_j \omega_j^s c_j^n(x)} \qquad P_R(u) = \frac{\sum_j \omega_j^R c_j^n(u)}{\sum_{x \in U} \sum_j \omega_j^R c_j^n(x)}$$

Les pondérations agrégées

$\omega_j^s = \sum_k \propto_j^k$ et $\omega_j^R = \sum_k \beta_j^k$

Avec : $C_j^n(u)$ désigne la performance de l'alternative u du critère j.

$C_j^n(x)$ désigne la performance de l'alternative x du critère j

b. La construction de l'ensemble des alternatives (ressources) satisfaisantes : Les alternatives intéressantes selon l'agent coordinateur sont celles qui peuvent être qualifiées de satisfaisantes c'est-à-dire pour qui la sélectabilité dépasse la rejetabilité.

c. La construction de l'ensemble des alternatives (ressources) qui domine chaque alternative u : D(u)

Une ressource satisfaisante peut être dominée en ce sens qu'il peut exister une autre alternative dont la sélectabilité est plus grande et la rejetabilité plus faible que la précédente. A cette fin, les choix ou alternatives qualifiables d'assez bonnes seront celles qui sont à la fois satisfaisantes et non dominées. Pour les qualifier, appelons D(u) l'ensemble des options qui dominent l'option u ; D(u) est alors donné par :

$D(u) = D_s(u) \cup D_R(u)$

Où :

$D_s(u) = \{\gamma \in U : P_R(\gamma) < p_R(u) \text{ et } P_s(\gamma) \geq P_s(u)\}$

$D_R(u) = \{\gamma \in U : P_R(\gamma) \leq p_R(u) \text{ et } P_s(\gamma) > P_s(u)\}$

d. La construction de l'ensemble des ressources (alternatives) équilibres non dominées *E* : Nous appelons alternatives ou actions équilibre E, les alternatives non dominées ie : E = $\{u \in U : D(u) = \emptyset\}$.

L'ensemble AB des alternatives assez bonnes avec indice de confiance q est alors donné par $AB_q = \sum_q \cap E$

e. Le choix de la ressource finale : L'alternative ou ressource finale u* choisie par le coordinateur parmi les alternatives de l'ensemble AB_q est celle telle que :

- La sélectabilité maximale $u^* = \max\{P_s(u)\}$
- La rejetabilité minimale $u^* = \min\{p_R(u)\}$

La discrimination maximale $u^* = \max\{P_s(u) \geq qp_R(u)\}$.

8 Etudes de Cas : Résultats Expérimentaux

Le développement d'un module multi agents est un problème complexe. Par conséquent, il est préférable d'utiliser une plateforme multi agents existante que nous adaptons à nos besoins. Notre choix a porté sur la plate forme **JADE** pour servir de base au module multi agents. Cette plateforme de développement est gratuite et implémentée en JAVA. Pour le développement du module SIG, nous avons opté pour l'utilisation du logiciel **MAPINFO :** C'est un outil de type SIG, il a permis, dans notre étude, de visualiser et de modifier les différentes bases de données géographiques utilisées.

8.1 Délimitation de la région d'étude : identification des ressources

Dans le cadre du projet d'extension de la ville d'Oran (Algérie) vers l'est, une nouvelle gare routière est prévue comme l'un des équipements projetés.Nous disposons de sept (07) îlots vierges pouvant convenir pour cette construction. Afin de pouvoir visualiser les îlots vides, nous exploitons les diverses avantages qu'offrent les systèmes d'information géographiques en terme d'affichage.

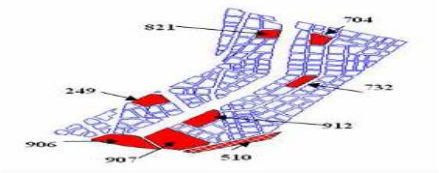


FIG. 3 – Affichage des îlots vides

8.2 Identification des critères et Identification des décideurs

Selon la disponibilité des données ainsi que les caractéristiques particulières de la zone d'étude, nous avons identifié les critères suivants :

- **C1 :** Superficie (à maximiser) ;
- **C2 :** Distance autoroute (à minimiser) ;
- **C3 :** Accessibilité (à maximiser) ;
- **C4 :** Nuisance sonore (à minimiser)
- **C5 :** Extrémité ville (à maximiser) ;

Dans cet exemple, les critères **C1, C3,C5** sont positifs et les autres critères sont négatifs. La matrice de performance générée en utilisant les fonctions d'analyse spatiale du SIG est illustrée par la figure 4.Pour cette étude de cas, les décideurs concernés par ce projet d'AT sont **les associations d'environnement, les politiciens, les économistes et le public.** Chacun de ces acteurs est représenté par un agent. Nous associons à chacun des agents participants, un poids pour exprimer son importance lors du déroulement de la négociation. La création des agents est illustrée par la figure 5.

Actions équilibre & assez bonnes		Action Finale	Matrice Normalisée & Pondérée		
Selectabilité & Rejetabilité			Actions Sasisfaisantes		
Matrice de performance		Normalisation	Répartition des critères		
	C1	C2	C3	C4	C5
A1	2670	820	80	700	54
A2	1145	710	50	820	460
A3	3510	530	90	120	400
A4	2180	700	70	300	880
A5	1450	1040	54	800	120
A6	1145	240	40	420	200

FIG. 4 – Matrice de performance

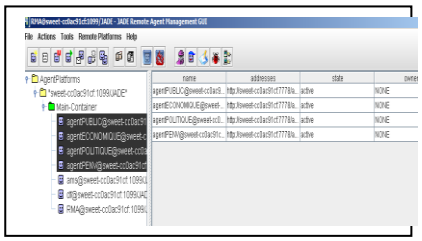


FIG. 5 – Création des agents sous JADE

8.3 Simulation de la négociation

Chacun des participants est amené à établir son vecteur de préférence où il classe les ressources de la meilleure à la moins bonne .Avant de lancer le processus de négociation, nous fixons :le seuil d'acceptation (le nombre d'accords nécessaires pour l'acceptation d'un contrat) à 60% et le nombre de tours de parole à deux tours.

Il est possible de visualiser les différents messages échangés entre l'agent initiateur et les agents participants en utilisant l'outil SNIFFER de JADE. Après deux tours de négociation, les agents participants reçoivent le message **Failure** de la part du coordinateur. Par conséquent, la négociation est un échec, la ressource finale doit être choisie par l'agent initiateur en exploitant les avantages qu'offre la théorie des jeux. Les différents messages échangés lors de la phase de négociation sont présentés par la figure 6.

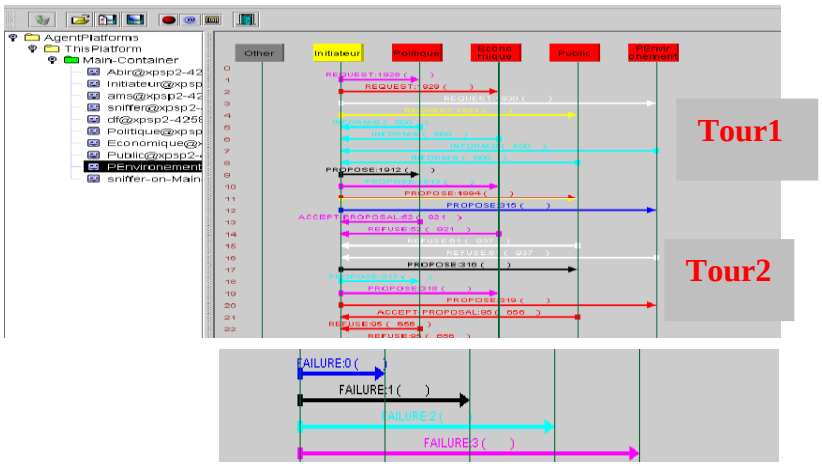


FIG. 6– Visualisation des messages échangés ; échec de la négociation

En utilisant les pondérations (calculées précédemment via la méthode AHP) propre à chaque agent participant ($\propto_j^k$, β_j^k) reflétant l'importance de chacun des critères positif et négatif par rapport à chaque participant, l'agent initiateur calcule les pondérations agrégéepour chaque critère j puis évalue le degré de selectabilité(Ps) et le degréde rejtabilité(Pr) de chaque action(ilot) comme le montre les figures 7 et 8 respectivement.
En se basant sur le degré de selectabilité et de rejetabilité précédemment calculés, l'agent initiateur construit l'ensemble des alternatives (ressources) équilibres non dominées **E** et l'ensemble des alternatives assez bonnes $\mathbf{AB_q}$. Parmi ces alternatives, la décision finale prise par l'agent initiateur (la ressource choisie) reposera soit sur la séléctabilité maximale, soit sur la rejetabilité minimale, ou encore sur la discrimination maximale comme l'indique la figure 9. La décision finale reflétant la préférence collective des participants sera visualisée grâce aux SIG comme illustré dans la figure 10 (celle dont la rejetabilité est minimal îlot 510)

Critères	Pondération
C 1	0.0593
C 2	0.5486
C 3	0.1402
C 4	0.1779
C 5	0.0742

FIG. 7– Les pondérations agrégées pour chaque critère

Action	Ps	Pr
249	0.06	0.2751
510	0.1686	0.0
704	0.0775	0.3093
912	0.1888	0.2193
821	0.1236	0.1913
906	0.1782	0.0026
907	0.2038	0.0026

FIG. 8– Le calcul des degrés de selectabilité et de rejetabilité

Stratégies	Emplacement sélectionné
Sélectabilité maximale	907
Rejetabilité minimale	510
Discrimination maximale	907

FIG. 9– Résultat final (alternatives proposées)

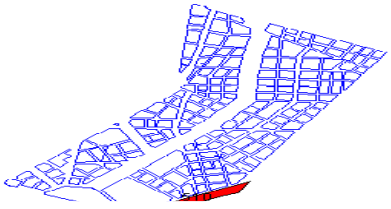


FIG. 10– Visualisation de l'ilot choisi

9 Conclusion

A travers cet article, notre effort a porté sur la proposition de solutions méthodologiques mise en œuvre par intégration de SIG, SMA et AMC afin de fournir de véritables outils d'aide à la décision à référence spatiale. En effet, le thème de cette recherche se situe au point de rencontre des domaines suivants: l'AT, les SIG, l'Aide Multicritère à la Décision et les SMA. Ainsi, nous avons initié dans le cadre de l'aide à la décision spatiale une nouvelle approche alliant:

D'une part, l'analyse multicritère aux modèles décisionnels spatiaux classiques dans le but d'optimiser la qualité de décision prise dans un tel contexte;
D'autre part, l'analyse multicritère et la théorie des jeux aux modèles à base d'agents afin de traiter la multiplicité et la diversité des acteurs dans un projet d'AT. La stratégie proposée :

- Permet une intégration complète dans le sens où les fonctions d'évaluation multicritère sont définies individuellement de manière générique, et peuvent facilement être incorporées dans un SIG;
- Assure la représentation de la multiplicité des acteurs, leur diversité, leur comportement ainsi que leur interaction.

Nous terminons cette conclusion en évoquant les différentes perspectives de recherche que nous envisageons aborder dans le futur. le développement d'une version Web du concept de la carte décisionnelle dans le but de "démocratiser" et généraliser l'utilisation de la technologie SIG; ainsi que la p rise en compte des dimensions spatiale et temporelle dans la modélisation multicritère;

Références

Boulemia,C., Henry, G., Pecqueur, O (2000), *Eléments de proposition à la mise en place d'ne base de données urbaines dans une collectivité locale, Congrès de l'AUGC*, Lyon.

Hamdadou D., Bouamrane, K (2007), *A Multicriterion SDSS for the Space Process Control: Towards a Hybrid Approach, MICAI 2007: Advances in Artificial Intelligence, LNCS, Springer*, ISSN 0302-9743 (Print) 1611-3349 , 2007, p. 139 -150, Mexico.

Hamdadou, D., Libourel, T (2009), *Couplage approche multicritère et négociation pour l'aide à la décision en aménagement du territoire, SAGEO, Paris.*

Hamdadou D., Bouamrane, K (2012), *Towards a Multicriteria Spatial Group Decision Support System, Application: Territory Planning*, ICEMCS 2012, IEEE, Tanger, Maroc.

Laouar,R. (2005), *Contribution pour l'aide à l'évaluation des projets de déplacements Urbains* . Thèse de doctorat, Laboratoire d'Automatique, de Mécanique, et d'Informatique industrielles et Humaines (LAMIH), Université de Valenciennes et du Hainaut Cambrésis, France.

Roy, B., Bouyssou, D (1993), *Aide multicritère à la décision: méthodes et cas*, Economica, Paris.

Stirling, W.C. (2003). *Satisficing Games and Decision Making: With Applications to Engi neering and Computer Science*, Cambridge University Press.

Summary

The work presented in this article, is a part of the vision that seeks to eliminate or lessen the impact of unsuccessful attempts to establish diagnostic tools on a functioning business. The Developing Multicriterion System to Support Diagnosis is an answer to the problem. It is decisional integrated tool for choosing the most relevant diagnosis system. This tool was developed using the multicriteria approach PAMSSEM I while conducting a sensitivity analysis and robustness of results. Based on a set of criteria and a set of diagnostic tools (alternative) carefully selected and implemented for the occasion. The developed tool allows us in first part to guide the maintenance expert to choose the diagnostic system to adopt, and secondly, to make a quickly and efficiently diagnosis with developed tools.

Infrastructure logicielle intégrant un système spatial décisionnel pour la géo-gouvernance des réseaux urbains

Aziz Mabrouk*, Azedine Boulmakoul**

*Faculté Polydisciplinaire de Larache
Département de Mathématiques et Informatiques - B. P. 745 Larache Maroc
aziz.mabrouk@gmail.com
http://www.fpl.ma
**Faculté des Sciences et Techniques de Mohammedia
FSTM – Département informatique - B.P. 146 Mohammedia Maroc
azedine.boulmakoul@yahoo.fr
http://www.fstm.ac.ma/

Résumé. Suite à la conception et l'implémentation des algorithmes existants et des processus de calcul que nous avons proposé dans des travaux antérieurs, nous avons développé un système logiciel spatial à base Diagramme de Voronoï de type réseau crisp et/ou flous constitué d'un ensemble de composants logiciels appropriés. Cette infrastructure logicielles offre aux experts de l'espace urbain des interfaces interactives leurs permettant d'exploiter des données spatiales réelles et de disposer des supports spatiaux fiables pour l'aide à la décision et par conséquence la géo-gouvernance des réseaux urbains.

1 Introduction

L'équité territoriale est considérée comme un concept et un principe d'aménagement permettant d'interpréter les situations réelles marquées par l'injustice spatiale. Elle désigne une configuration géographique qui garantirait à tous les mêmes conditions d'accès aux équipements et services publics. Ceci ne se réalise qu'à travers une géo-gouvernance des réseaux urbains, fondée sur une évaluation objective de l'accessibilité spatiale à des lieux d'intérêt, de tout point appartenant à l'aire urbaine déterminant la zone d'étude; et inversement, sur une identification réaliste de l'aire de desserte de ces lieux d'intérêt. Ainsi, l'‘intérêt croissant envers les systèmes d'aide à la décision, tant en planification urbaine que dans d'autres domaines provient d'un besoin stratégique d'outils d'aide à la décision basés sur une infrastructure d'informations spatiales. En effet, les experts de l'espace urbain requièrent des méthodes et outils de l'analyse spatiale ainsi que des supports spatiaux mettant à leur portée une information territoriale pertinente pour décrypter la complexité des systèmes territoriaux, à mettre en lumière leurs enjeux Mabrouk, A.et Boulmakoul A., (2012).

En effet, dans ce papier, nous proposons une infrastructure logicielle intégrant un système spatial décisionnel pour la géo-gouvernance des réseaux urbains. En fait, nous avons développé cette solution spatiale suite à des travaux antérieurs dont nous avons conçu et implémenté des algorithmes existants et proposer des processus de calcul des diagrammes de Voronoï basés sur les concepts de la théorie des ensembles flous et de la théorie des graphes.

Les fonctions de base de ce système spatial calculent les diagrammes de Voronoï de type planaire et de type réseau crisp et/ou flous d'une part, et d'autre part, elles estiment l'accessibilité spatiale de Voronoï. Les autres fonctions consistent des outils de prétraitement des données spatiales. Toutes ces fonctions sont exécutées par des composants logiciels constituant le système.

2 Problématique

À l'instar de la rationalité, pour Bosc et Prade (1997), il paraît légitime d'avancer que nulle information n'est parfaite. Concrètement, cette imperfection peut provenir d'une incertitude et/ou d'une imprécision, voire d'une inconsistance ou incohérence lorsque les données disponibles sont en conflits les unes par rapport aux autres Bosc et Prade (1997). Or, si dans la pratique ces différentes formes d'imperfection ne sont pas exclusives les unes des autres, il est pourtant indispensable de bien les discerner, au moins sur un plan formel, et notamment dans le but de sélectionner les traitements adéquats.

Les diagrammes de Voronoï de type réseau classique (DVR) sont discutés par Okabe et al. (2002a) et également dans une revue concernant les problèmes d'optimisation de localisation Okabe et Suzuki (1997). Ainsi, plusieurs algorithmes ont été proposés dans la littérature Erwig (2000), Takehiro et al. (2005), Margot Graf et Stephan Winter (2003) et qui permettent de calculer le diagramme de Voronoï de type réseau dont le poids des chemins parcourus est considéré une valeur réel. Or, cette estimation ne reflète pas parfaitement la réalité.

Par ailleurs, SANET est un outil développé par Okabe et al.(2002a) pour construire les DVRs crisp. Cet outil ainsi que le Système Logiciel Spatial à base Diagramme de Voronoï (SLSDV) que nous avons développé et présenté dans des travaux antérieurs et dans ce papier, ne permettent pas de traiter ni analyser spatialement des données floues. En effet, les supports spatiaux générés par ces outils ne fournissent pas des données fiables et réelles décrivant et analysant d'une manière complètes les réseaux urbains et par conséquences ils ne permettent pas de prendre des bonnes décisions.

3 Systèmes existants calculant les Diagramme de Voronoï crisp

3.1 L'outil SANET d'Okabe

SANET est un outil développé par Okabe et al.(2002a). Il comporte un ensemble de fonctions capables d'éditer et d'analyser les réseaux spatiaux y compris le calcul des Diagrammes de Voronoï. En effet, il reçoit en entrée et génère en sortie des fichiers de formes comportant des données crisp décrivant les réseaux spatiaux.

FIG. 1 – *L'outils SANET développé par Okabe et al.(2002a)*

4 Système Logiciel Spatial à base Diagramme de Voronoï (SLSDV)

4.1 Description du SLSDV

La programmation par composants représente une évolution technologique appuyée par de nombreuses plateformes (composants EJB, CORBA, .Net, WSDL, . . .). Ce type de programmation se base sur la réutilisation du composant et l'indépendance de son évolution vis-à-vis des applications qui l'utilisent. En outre, l'utilisation de composants est assimilable à une approche objet, non pas au niveau du code, mais au niveau de l'architecture générale du logiciel.

En effet, le système spatial proposé, qui est une application MFC que nous avons développé sous Visuel C++, intègre un ensemble de composants logiciels capables de calculer les diagrammes de Voronoï crisp et d'évaluer l'accessibilité spatiale à des lieux d'intérêt.

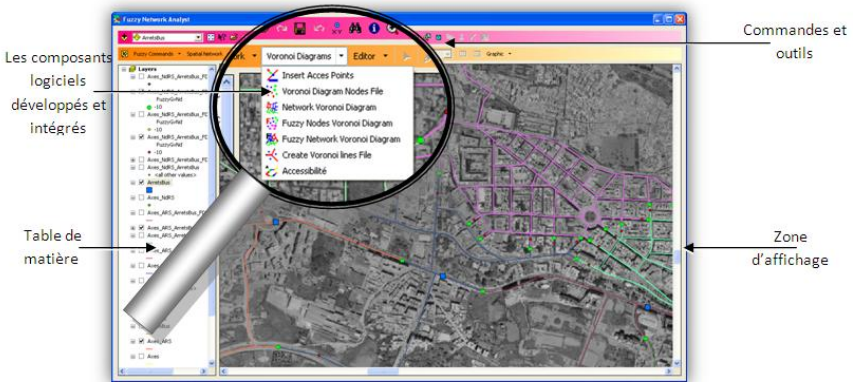


FIG. 2 – *Notre système spatial intégrant les composants logiciels développés*

4.2 Les composants logiciels intégrés dans le SLSDV et leurs fonctions

Les composants logiciels intégrés dans le SLSDV, sont des objets COM (Component Object Model) indépendants (sous forme de dynamic link library DLL) que nous avons développés avec Visual C++ en utilisant la bibliothèque de modèles actifs ATL (Active Template Library) vu que la technologie COM a été développée pour permettre la communication interprocessus.

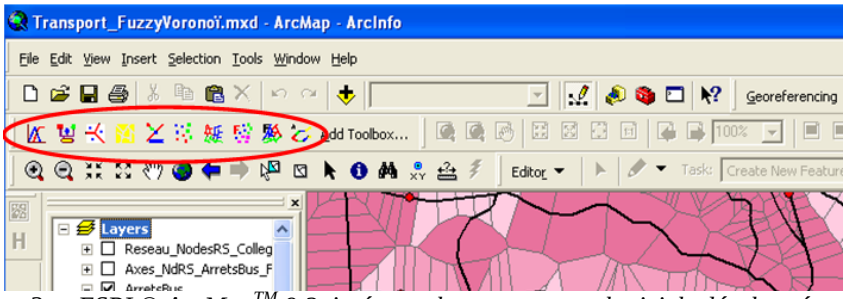


FIG. 3 – *ESRI ® ArcMap™ 9.2 intégrant les composants logiciels développés*

Ils reposent sur un système de communication via interface composé en fait d'une série de fonctions accessible aux autres programmes. En effet, ces composants sont également réutilisables par des applications SIG à savoir le logiciel SIG ESRI ® ArcMap 9.2 (0 4).

Après la modélisation de l'espace urbain qui consiste à préparer les données spatiales d'entrées, chaque composant assure une fonction.

En effet, les composants logiciels sont développés pour assurer les fonctions suivantes :

- La génération des digrammes de Voronoï planaires
- La création de la topologie réseau
- L'insertion des points d'accès dans les arcs du réseau spatial
- La génération des nœuds du DVR
- La génération des arcs du DVR
- L'évaluation de l'accessibilité spatiale à des lieux d'intérêt

Afin de réaliser une de ces fonctions, Chaque composant logiciel exécute trois processus principaux :

- Réception des données spatiales
- Traitement des données spatiales
- Génération des données spatiales

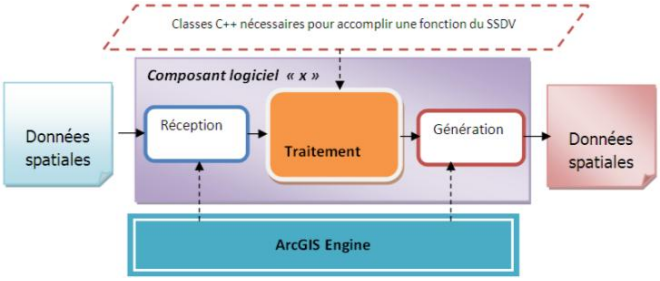


FIG. 4 – *L'architecture d'un composant logiciel réalisant une fonction du SSDV*

4.2.1 Processus de réception et génération des données spatiales

Ces deux processus sont assurés par ArcGIS Engine Runtime. Ceci constitue une bibliothèque permettant d'accéder et gérer des données spatiales. En utilisant les classes et les interfaces de cette bibliothèque, le composant logiciel accède aux propriétés géométriques et alphanumériques des données spatiales stockées dans des fichiers de formes ou des classes d'entités d'une base de données géographique. En se basant sur ces données reçues et selon le traitement nécessaire à la réalisation de la fonction demandée, le composant logiciel construit des nouvelles données spatiales et les stocke ensuite dans des nouveaux fichiers de formes ou des classes d'entités d'une base de données géographique.

4.2.2 Processus de traitement des données spatiales

Ce processus reçoit en entrée les propriétés géométriques (forme, distance, surface, nombre d'éléments,…) et alphanumériques (attributs) des données spatiales reçues par le processus « réception ». Et selon la fonction du SSDV demandée et en se basant sur des méthodes que nous avons proposées et des classes C++ que nous avons implémentées et qui

sont nécessaires pour accomplir cette fonction, le processus met en sortie les éléments nécessaires pour construire ou mettre à jour des données spatiales par le processus « génération ».

5 Diagrammes de Voronoï basé sur la modélisation Floue du Réseau spatial (DVR-Flou)

5.1 Définition

Nous définissons le Diagramme de Voronoï d'un Réseau Spatial Flou (DVR-Flou) par la division du réseau spatial en sous réseau de Voronoï dont chacune contient les points les plus proches à chaque générateur de Voronoï en parcourant les plus courts chemins flous pour mettre en relation ces composantes spatiales , A., Boulmakoul, A. et Bielli, M. (2009).

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

Chaque poids valuant un arc, est représenté par un nombre flou. Considérons v et w deux sommets appartiennent à S. Nous allons utiliser $P_{flou}(v,w)$ pour représenter le poids flou du plus court chemin de v à w dans le réseau N. Le DVR-Flou pour G divise le réseau N en n sous réseaux de Voronoï flou $Vor_{flou}(1), \dots, Vor_{flou}(n)$ avec :

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

5.2 Notre processus de calcul du DVR-Flou

Dans des travaux antérieurs Mabrouk, A., Boulmakoul, A. et Bielli, M. (2009) et Mabrouk, A. et Boulmakoul A., (2008c), nous avons prolongé les algorithmes proposés dans la littérature et qui permettent de calculer le diagramme de Voronoï de type réseau. Cette extension est basée sur une modélisation floue d'un réseau réel (réseau routier, réseau de transport, etc.…). Ceci implique la recherche des plus courts chemins flous pour mettre en relation les nœuds de ce réseau spatial à leurs plus proches générateurs de Voronoï (Ecoles, hôpitaux, etc. …). En effet, nous avons proposé un pseudo code de cette méthode dans Mabrouk, A., Boulmakoul A., (2008b) dans lequel nous avons utilisé $R()$ une fonction de tri des nombres dont :

$Ppcc_{flou}(w) \geq \Delta_{flou}$ *si et seulement si* $R\,(Ppcc_{flou}(w)) \geq R\,(\Delta_{flou})$.

Avec $Ppcc_{flou}(v)$ dénote le poids flou du plus court chemin, du plus proche générateur de Voronoï temporaire $Gt(v)$au nœud v.

6 Système logiciel spatial à base DVR-Flou (SLSDV-Flou)

6.1 Description du SLSDV-Flou

Ce système Constitue une extension du Système logiciel spatial à DV que nous avons développé et présenté dans la partie 4. C'est un système qui se compose des composants logiciels permettant d'une part de préparer des données spatiales floues modélisant des réseaux spatiaux réels ainsi que les poids flous des déplacements (la fuzzification des données spatiales). D'autre part, et en s'alimentant de ces données spatiales floues, de construire les

Diagrammes spatiaux flous de Voronoï. En effet, ces composants logiciels que nous avons développés avec les mêmes technologies (C++, ATL, ArcGIS Engine) et qui sont utilisées pour développer les composants logiciels du SLSDV, sont les suivants:

- Composant logiciel pour le calcul du temps flou du déplacement sur un réseau spatial flou
- Composant logiciel pour l'affectation des nœuds d'un réseau spatial flou à leurs plus proches générateurs de Voronoï (les nœuds du DVR-Flou) ;
- Composant logiciel pour l'affectation des arcs d'un réseau spatial flou à leurs plus proches générateurs de Voronoï (les arcs du DVR-Flou).

6.2 Composant logiciel pour le calcul du temps flou du déplacement sur un réseau spatial flou

6.2.1 Les nombres flous et les opérations arithmétiques floues

La conception orientée objet de l'arithmétique floue devrait couvrir les opérations arithmétiques, qui pourraient être mises en application d'une manière différente selon une classe concrète de nombre flou. Le modèle créé couvre les nombres flous trapézoïdaux et les nombres flous triangulaires.

La classe « TrapeziodalFuzzyNumber » représente la description orientée objet du nombre flou trapézoïdal. Dans la classe, quatre attributs $a, b, c\ et\ d$ représentant un nombre flou trapézoïdal $A\ (a1, a2, a3, a4)$ sont déclarés.

La classe « TriangularFuzzyNumber » représente la classe dérivée de la classe des nombres flous trapézoïdaux « TrapeziodalFuzzyNumber ». Dans cette classe, nous déclarons seulement trois attributs a, b, c en posant $c = d$. Cette classe implémente égalementles méthodes arithmétiques floues.

Les méthodes arithmétiques floues (addition, soustraction, multiplication et la division) déclarées dans cette classe représentent des opérations arithmétiques surchargées. En utilisant C++, nous avons redéfinit les opérateurs arithmétiques standards (+, -, * et /) pour les nombres flous. La redéfinition des ces opérateurs est un moyen extrêmement efficace pour écrire des programmes compréhensibles.

Dans la littérature, plusieurs fonctions permettant le tri des nombres flous. Rank (), RV () et la formule correcte du point centroïde Wang et al. (2006) et Shieh (2007) sont les fonctions que nous avons implémenté en déclarant les méthodes de leurs calculs dans la classe « TriangularFuzzyNumber ». Nous avons choisi ces fonctions vu la facilité de leurs implémentations.

L'objectif de cette implémentation c'est l'utilisation de ces fonctions de classement pour la redéfinition des opérateurs relationnels ($<, \leq, >, \geq et\ ==$) qui permettent à leurs tours de comparer des nombres flous.

6.2.2 Conception du composant logiciel

Le composant logiciel reçoit en entrée des fichiers de forme comportant les données spatiales et descriptives du réseau spatial comportant trois champs $V_{réelle}, \alpha\ et\beta$ décrivant les valeurs floues des vitesses de déplacement $Vitesse_{floue}\ (\alpha, V_{réelle}, \beta)$.

En utilisant la valeur de la distance de chaque arc et la valeur floue de la vitesse associée à cet arc, le composant logiciel calcule le temps de déplacement flou $T_{flou}\,(LTime, Time, RTime)$.

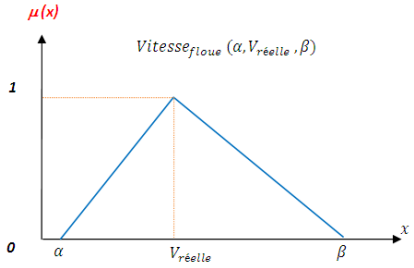


FIG. 5 – *Vitesse de déplacement exprimée par une valeur floue* $Vitesse_{floue}\ (\alpha, V_{réelle}, \beta)$

Il génère ensuite un fichier de formes semblable au fichier de formes des arcs reçu en entrée, mais avec trois champs supplémentaires pour stocker les bornes des fonctions d'appartenance relatives au temps flou de déplacement sur chaque tronçon du réseau spatial (0). Ce fichier constituera la base de calcul du DVRS-Flou.

	FID	Shape	Id_Arc	Nd1	Nd2	DIRECTION	Poid_Arc	Vitesse	alpha	beta	LTime	Time	RTime
	0	Polyline	0	0	1	0	403,960867	40	5	10	0,484753	0,605941	0,692504
	1	Polyline	1	2	3	0	416,412272	40	5	10	0,499695	0,624618	0,71385
	2	Polyline	2	4	5	0	813,953495	40	5	10	0,976744	1,22093	1,395349

TAB. 1 – *La table associée au fichier de forme des arcs du réseau spatial après l'ajout et le calcul des trois champs en fonction des distances et des vitesses floues.*

7 Composant logiciel pour l'affectation des nœuds d'un réseau flou à leurs plus proches générateurs de Voronoï

La construction du DVR-Flou des nœuds par ce composant logiciel est similaire à la construction du DVR des nœuds en prenant en considération la modélisation floue du réseau spatial.

Ce composant logiciel commence alors par la construction du graphe flou modélisant le réseau spatial. Après la réception en entrée des données relatives aux générateurs de Voronoï, il procède à la recherche parallèle des plus courts chemins flous en utilisant un tas flou pour mettre en relation ces générateurs avec le reste des nœuds Mabrouk, A., Boulmakoul A., (2008b). Ceci implique que ce composant logiciel est réalisé à travers :

- Conception et l'implémentation des classes modélisant un réseau spatial flou
- Conception & Implémentation du processus de calcul du DVR-Flou

7.1 Conception & Implémentation des classes modélisant un réseau spatial flou

Dans Mabrouk, A., Boulmakoul, A. et Bielli, M. (2009), nous avons démontré qu'un réseau spatial peut être modélisé par un graphe flou valué ou pesé $G(\Omega, \mu, \phi)$dont les sommets $\Omega: \{1, \ldots, n\}$ sont associés aux infrastructures nodales et les arêtes $(i, j) \in \Omega^2$ sont associées aux infrastructures linéaires dont leurs poids ϕ sont des nombres flous. D'autre part, un

graphe flou peut être représenté comme une liste d'étoiles floues, chaque étoile étant composée d'un sommet et d'une liste des arêtes floues qui ont ce sommet comme origine.

7.2 Conception & Implémentation du processus de calcul du DVR-Flou

Le processus de calcul du DVR-Flou reçoit en entrée d'une part, le graphe flou construit dans la section précédente et d'autre part, les générateurs de Voronoï. Considérant ces générateurs comme des sources multiples, il cherche en parallèle les plus courts chemins flous entre ces générateurs et le reste des nœuds du réseau spatial. Cette opération est concrétisée par la définition du constructeur $FuzzyNetworkVoronoi(FuzzyGraph*, int*, int)$ de la classe « $FuzzyNetworkVoronoi$ » permettant le calcul du DVR-Flou.

7.3 Composant logiciel construisant les nœuds du DVR-Flou

Ce composant logiciel reçoit en entrée le fichier de formes comportant les nœuds du réseau y compris les générateurs de Voronoï, le fichier de formes des points d'origine (Ecoles, hôpitaux, etc. …) et le fichier de formes des arcs comportant les trois champs stockant les bornes des fonctions d'appartenance qui sont relatives au temps flou de déplacement sur chaque tronçon du réseau spatial réel Mabrouk, A. et Boulmakoul A. (2010).

En utilisant l'algorithme que nous proposons en haut , ainsi publié dans Mabrouk, A., Boulmakoul, A. et Bielli, M. (2009), le composant logiciel construit le graphe flou et puis procède parallèlement à la recherche des nœuds qui sont plus proche à chaque générateur de Voronoï en parcourant les plus courts chemins.

FID	Shape *	Id_Nd	Nd_X	Nd_Y	Nd_Xt	Nd_Yt	Type	FuzzyGvNd	Precd_Nd	Dist_GvNd	Dist_GvA	Dist_GvB	Dist_GvD
0	Point	0	501679,669132	551913,597673	50167967	55191360		310	310	26,46071	0,317529	0,529214	0,793821
1	Point	1	501530,588968	551538,63755	50153059	55153864		310	0	430,421577	5,165059	8,608432	12,912647
2	Point	2	501088,457173	551676,091965	50108846	55167609		311	3	476,240365	5,714884	9,524807	14,287211
3	Point	3	501104,861354	552077,26629	50110486	55207727		311	162	59,828092	0,717937	1,196562	1,794843

TAB. 2 – *4 champs supplémentaires sont ajoutés au fichier des nœuds pour stocker les générateurs de Voronoï associés et les poids flous des plus courts chemins flous trouvés.*

Il génère alors en sortie un fichier de formes semblable au fichier de formes des nœuds mais avec quatre champs supplémentaires pour stocker pour chaque nœud le générateur associé et le total du temps flou de déplacement (trois valeurs réelles) pour atteindre ce générateur (0).

8 Composant logiciel pour l'affectation des arcs d'un réseau flou à leurs plus proches générateurs de Voronoï

D'une manière similaire au composant logiciel affectant les arcs d'un réseau crisp à leurs plus proches générateurs de Voronoï, Ce composant logiciel marque chaque arc du réseau spatial selon l'affectation de son nœud de début et celle de fin aux générateurs de Voronoï. Il reçoit en entrée le fichier de forme des arcs et le fichier de forme des nœuds généré par le composant logiciel présenté dans la section précédente (les nœuds du DVR-Flou). Ce fichier des nœuds comporte des champs contenant des données sur les générateurs de Voronoï associées aux nœuds et les poids flous des plus courts chemins parcourus pour atteindre ces générateurs (0).

Si un nœud de début s et celui de la fin s' sont affectés au même générateur «g», ou s'ils appartiennent aux différents générateurs mais l'arc déterminé par ces deux nœuds est unidirectionnel, ce composant logiciel affecte cet arc à un générateur de Voronoï.

Si le nœud de début et celui de la fin appartiennent aux différents générateurs, et l'arc déterminé par ces deux nœuds est symétrique et ne peut être accédé qu'à partir de ses extrémités (s ou s'), dans ce cas, ce composant logiciel appelle à la fonction de classement déclarée dans la classe « $TriangualerFuzzyNumber$ » pour comparer les poids flous des plus courts chemins associés à ces nœuds. Il assigne alors l'arc à l'arbre des plus courts chemins où appartient le nœud (s ou s') qui a un poids flou de déplacement minimal pour atteindre les générateurs les plus proches.

Si le nœud de début et celui de la fin appartiennent aux différents générateurs, et l'arc déterminé par ces deux nœuds n'est pas nécessairement symétrique et il est accessible par n'importe quel point qui lui appartient. Dans ce cas l'arc doit être divisé. Le point de division est le point $C(X_c, Y_c)$ auquel le poids flou de déplacement $Poids_{flou}\big(C, g(s)\big)$ au générateur $g(s)$ associé au nœud de début "s", est égal au poids flou de déplacement $Poids_{flou}\big(C, g(s')\big)$ au générateur $g(s')$ associé au nœud de fin "s". Alors:

$$Poids_{flou}\big(C, g(s)\big) = Poids_{flou}\big(C, g(s')\big) \quad (2)$$

La première part $[sC]$ (du nœud de début « s » au point de division « C ») sera assigné à l'arbre des plus courts chemins auquel appartient le nœud de début « s », et la deuxième part, $[Cs']$ (du point de division « C » au nœud de fin « s' ») sera assignée à l'arbre des plus courts chemins auquel appartient le nœud de fin « s' ».

8.1 Calcul des références spatiales floues d'un point de division

Soit A l'ensemble des arcs d'un réseau spatial flou

Soit $Q \subset A$ l'ensemble des arcs dont les générateurs de leurs extrémités sont différents.

Soit $[NM]$ ϵ Q l'un de ces arcs et qui se compose d'un ensemble de segments.

Soit $[AB]$ ϵ $[NM]$, le segment où le point flou de division $C(X_C, Y_C)$ coupe $[NM]$ en deux parts : $[NC]$et $[CM]$ Avec :

$$NM = NC + CM \qquad et \quad P_{flou}\big(N, g(N)\big) + NC = P_{flou}\big(M, g(M)\big) + CM \quad (3)$$

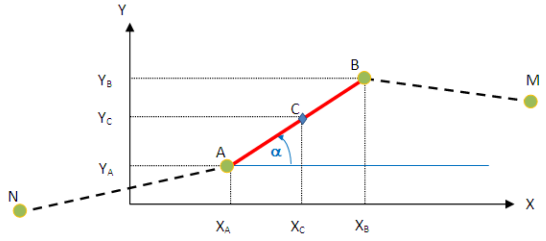


FIG. 6 – *Le point de division C coupe le segment [AB] en 2 parts [AC] et [CB] et coupe l'arc [NM] en deux arcs [NC] et [CM]*

Alors : $$NC = \frac{Poids_{flou}\,(M,g(M)) - Poids_{flou}\,(N,g(N)) + NM}{2} \quad (4)$$

et $$C\,M = \frac{P_{flou}\,(N,g(N)) - P_{flou}\,(M,g(M)) + NM}{2} \quad (5)$$

et $$AC = NC - NA \; et \; \text{CB} = \text{CM} - \text{BM} \quad (6)$$

Par ailleurs,

$$Cos \propto = \frac{X_C - X_A}{AC} = \frac{X_B - X_A}{AB} \quad \Rightarrow \quad X_C = \frac{AC}{AB}(X_B - X_A) + X_A \qquad (7)$$

$$Sin \propto = \frac{Y_C - Y_A}{AC} = \frac{Y_B - Y_A}{AB} \quad \Rightarrow \quad Y_C = \frac{AC}{AB}(Y_B - Y_A) + Y_A \qquad (8)$$

$$C \in [AB] \quad \Leftrightarrow \quad C \in [NM] et NA \leq NC \leq NB \qquad (9)$$

D'après (4), (5), (6), (7) et (8) :

$$X_C = \left(\frac{Poids_{flou}\left(M, g(M)\right) - Poids_{flou}\left(N, g(N)\right) + NM}{2} - NA\right)\left(\frac{X_B - X_A}{AB}\right) + X_A$$

$$Y_C = \left(\frac{Poids_{flou}\left(M, g(M)\right) - Poids_{flou}\left(N, g(N)\right) + NM}{2} - NA\right)\left(\frac{Y_B - Y_A}{AB}\right) + Y_A$$

X_C et Y_C sont alors des sous-ensembles flous déterminés respectivement par les bornes (x_{CL}, x_C, x_{CR}) et (y_{CL}, y_C, y_{CR}) avec $x_{CL}, x_C, x_{CR}, y_{CL}, y_C \; et \; y_{CR}$ sont des nombres réels.

Le point flou $C(X_C, Y_C)$ est alors un ensemble de points $c(x, y)$ avec $x \in IR \; et \; y \in IR$.

$Soit \; \mu X_C(x) \; et \mu Y_C(y)$ respectivement les fonctions d'appartenance de X_C et Y_C

$\forall \propto \in [0,1]$:

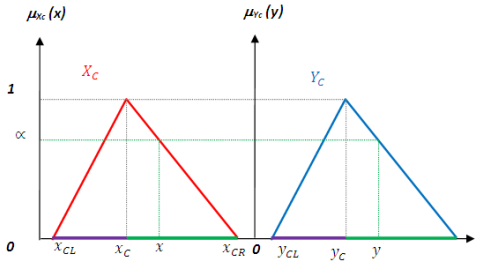


FIG. 7 – *Les références spatiales floues d'un point de division*

$$Si \; x_{CL} < x < x_C \; alors \propto = \mu X_C(x) = (x - x_{CL}) / (x_C - x_{CL})$$
$$\Rightarrow x = \propto . (x_C - x_{CL}) + x_{CL}$$
$$Si \quad x_C < x < x_{CR} \; alors \propto = \mu X_C(x) = (x_{CR} - x) / (x_{CR} - x_C)$$
$$\Rightarrow x = x_{CR} - \propto . (x_{CR} - x_C)$$

D'une manière similaire,

$$Si y_{CL} < y < y_C alors \propto = \mu Y_C(y) = (y - Y_{CL}) / (Y_C - Y_{CL})$$
$$\Rightarrow y = \propto . (y_C - y_{CL}) + y_{CL}$$
$$Si \quad y_C < y < y_{CR} \; alors \propto = \mu X_C(y) = (y_{CR} - y) / (y_{CR} - y_C)$$
$$\Rightarrow y = y_{CR} - \propto . (y_{CR} - y_C)$$

Pour chaque $0 \leq \propto < 1$, on distingue deux type de DVR-Flou pour le même réseau spatial et les mêmes générateurs de Voronoï (figure 9) :

- DVR-Flou Gauche calculé en fonction des valeurs appartenant au support de la partie gauche des poids flous des plus courts chemins et qui sont représentés par des NFTs ;
- DVR-Flou Droit calculé en fonction des valeurs appartenant au support de la partie droite des poids flous des plus courts chemins et qui sont représentés par des NFTs.

Pour $\propto = 1$, le Diagramme de Voronoï résultant est un diagramme de Voronoï de type réseau classique (figure 9).

- La (figure 8) montre une interface du composant logiciel permettant à l'utilisateur de déterminer les champs relatifs aux poids flous de déplacement, le type de DVR-Flou et la valeur de $\propto$.

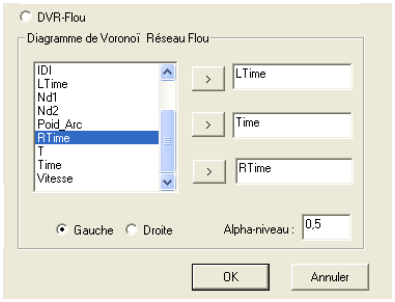


FIG. 8 – *Interface utilisateur pour déterminer les champs des poids flous de déplacement et les paramètres du DVR-Flou*

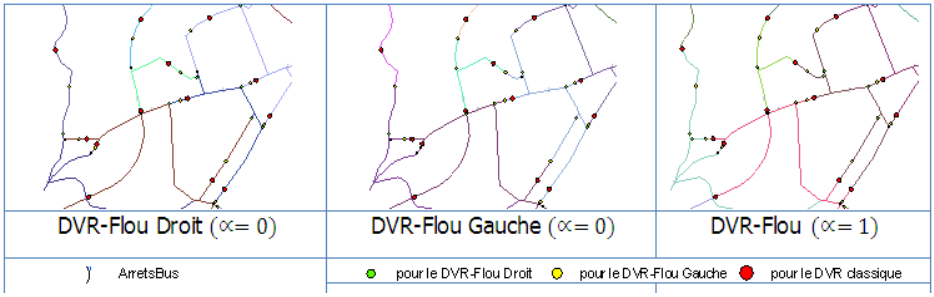


FIG. 9 – *Extraits des DVR-Flou selon le type et la valeur de $\propto$, pour les arrêts de bus de la ville de Tétouan*

9 Conclusion

Les outils et systèmes existants calculant les Diagrammes de Voronoï de type réseau classique ne fournissent pas des supports spatiaux fiables capables de participer dans la géo-gouvernance des espaces urbains, vu que la modélisation de la réalité d'un réseau spatial exploite des quantités précises et certaines. Dans ce papier, nous avons présenté une nouvelle approche basée sur les concepts de la théorie des ensembles flous et de la théorie des graphes, et ce pour modéliser et analyser les réseaux spatiaux réels. Cette approche est concrétisée par le développement d'une infrastructure logicielle d'aide à la décision interactive basée sur les diagrammes de Voronoï des réseaux spatiaux flous. Les fonctions de trie des nombre flous utilisées par notre système, malgré la facilité de leur implémentation, restent encore limitées aux nombres flous triangulaires et trapézoïdaux.

Références

Bosc, P. et H. Prade, (1997), An Introduction to the Fuzzy Set and Possibility Theory-based Treatment of soft Queries and uncertain or imprecise Database, in Uncertainty Man-

agement in Information Systems: From Needs to Solutions, Smets, P. et Motro, A (Ed.), Kluwer, Dordrecht, pp. 285–324.

Okabe, A.; Okunuki, K.; Funamoto, S., 2002a: SANET: A Toolbox for Spatial Analysis on a Network, Center for Spatial Information Science, University of Tokyo.

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

Takehiro Furuta, Atsuo Suzuki and Keisuke Inakawa (2005) The Kth Nearest Network Voronoi Diagram And Its Application To Districting Problem Of Ambulance Systems, Nanzan University, No.0501, 2005

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

Mabrouk, A., Boulmakoul A., (2012), Modèle Spatial basé sur les Diagrammes Spatiaux de Voronoi pour la géo-gouvernance des espaces urbains. INTIS'12, FST –Mohammedia

Mabrouk, A., Boulmakoul A., (2010), Système spatial interactif d'aide à la décision basé sur les diagrammes de Voronoï flous. la sixième conférence internationale SITA'10, Systèmes intelligents théories et applications, ENSIAS –Rabat, Mai.

Mabrouk, A., Boulmakoul, A. and Bielli, M. (2009) Fuzzy spatial network Voronoï diagram: a spatial decision support for transportation planning, Int. J. Services Sciences, Vol. 2, Nos. 3/4, pp.265–280.

Mabrouk, A., Boulmakoul A., (2008b) Composant logiciel pour l'intégration des réseaux spatiaux de Voronoï et application à l'évaluation de l'accessibilité, in Systèmes intelligents théories et applications, SITA- INPT, pp.118-126 (2008). ISBN : 978-2-909285-55-3, Europe IA, France.

Mabrouk, A., Boulmakoul A., (2008c), Processus de calcul du diagramme de Voronoï de type réseau basé sur la modélisation floue des réseaux spatiaux. journées JOSTIC'08, Association ACTIF & le laboratoire LIMIARF, Faculté des sciences de Rabat, Novembre

Wang, Y.H., Yang, J.B., Xu, D.L., & Chin, K.S. (2006). On the centroids of fuzzy numbers. Fuzzy Sets and System, 157, 919-926.

Shieh, B.S. (2007). An approach to centroids of fuzzy numbers. International Journal of Fuzzy Systems, 9 (1), 51-54.

Summary

Following the design and implementation of existing algorithms and calculation process we proposed in earlier work, we have developed a software system based crisp and/or fuzzy Voronoï spatial network consists of a set of components appropriate software. This software infrastructure offers experts of urban interactive interfaces allowing them to exploit spatial data and dispose of real media reliable space for decision support and consequence geo-governance of urban networks.

Intelligence Distribuée et Modèle de Décision des déplacements des Piétons

Meriem MANDAR*, Azedine Boulmakoul**

*École Nationale des Sciences Appliquées, Bd Béni Amir, BP 77, Khouribga - Maroc
meriem.mandar@gmail.com
**Faculté des Sciences et Techniques de Mohammedia
FSTM – Département informatique - B.P. 146 Mohammedia Maroc
azedine.boulmakoul@yahoo.fr

Résumé. La recherche dans le domaine de la dynamique collective des piétons croit de plus en plus avec la croissance de son applicabilité dans plusieurs systèmes civils. Dans la première partie de ce travail, nous avons présenté les tendances de comportements individuels et collectifs des piétons proposées dans la littérature. Nous présentons également un modèle microscopique de navigation des piétons virtuels dans des environnements contraints. Ce modèle s'est établit en deux phases. La première couple la métaheuristique ACO et les automates cellulaires, tout en adoptant une représentation floue de certains paramètres imprécis Boulmakoul and Mandar (2011). La deuxième phase étend le modèle en utilisant la méthode de champs de potentiel artificiels. La superposition des forces appliquées par les composants statiques de l'environnement, permet de guider les piétons vers leurs destinations tout en étant repoussés par les obstacles. Nous présentons également une formulation d'un indicateur d'exposition au risque d'accidents mutuels entre piétons et véhicules. Le système logiciel développé permet la simulation des trafics des piétons virtuels et des véhicules dans diverses configurations spatiales. Les indicateurs fournis par le modèle proposé sont en conformité avec les lois de transport. Dans de futurs travaux, la solution logicielle dans sa phase de déploiement sera intégrée à l'analyse des accidents des piétons dans les réseaux de transport urbain.

1 Introduction

La recherche dans le domaine de la dynamique collective des piétons vise à offrir aux décideurs urbains des moyens de prédiction des flux de piétons dans diverses situations. Par ailleurs les déplacements collectifs de piétons présentent des traits de similarité avec ceux des essaims naturels et notamment celui des fourmis. L'optimisation par colonies de fourmis dénote un paradigme de conception des systèmes multi-agents, fondé sur la coopération et la communication distribuées de nombreuses entités élémentaires. Ces mécanismes d'interaction contribuent à l'émergence d'une intelligence collective distribuée. Dans ce travail, nous proposons les résultats de nos recherches menées au laboratoire de Mohammedia. Ces travaux concernent le projet de thèse de Meriem Mandar, sous la supervision du professeur Boulmakoul, Mandar(2012).

La structure du papier est organisée comme suit. Après une introduction, la section 2 aborde un état de l'art non exhaustif en matière de modélisation et de simulation des piétons virtuels. La section 3 présente le modèle microscopique que nous avons élaboré en deux phases. La section 4 présente une formulation d'un indicateur d'exposition aux risques d'accidents mutuels entre piétons et véhicules. La section 5 présente un scénario de simulation du trafic de piétons, ainsi que les résultats obtenus. Enfin la section 6 propose les conclusions et les perspectives de ce travail.

2 Piétons virtuels : Etat de l'Art

Face à un problème de navigation suivant une motivation propre, les piétons explorent leur environnement et procèdent à une planification du chemin à entreprendre. Cette décomposition en étapes nécessite trois niveaux d'analyses : (i) stratégique, (ii), tactique et (iii) opérationnel. La figure 1 illustre ces derniers en prenant comme exemple (schéma à gauche) le cas d'un responsable qui veut rejoindre son bureau après une réunion : au niveau stratégique, il a deux chemins possibles dont il choisit le plus proche : niveau tactique, et puis il s'y déplace. Et ce bien sur en interagissant avec d'autres piétons sur son chemin et en évitant les obstacles : niveau opérationnel. Ce dernier s'avère suffisant pour l'étude du comportement collectif des piétons. Il regroupe trois échelles de modélisation selon le degré de détails étudié : microscopique, mesoscopique et macroscopique.

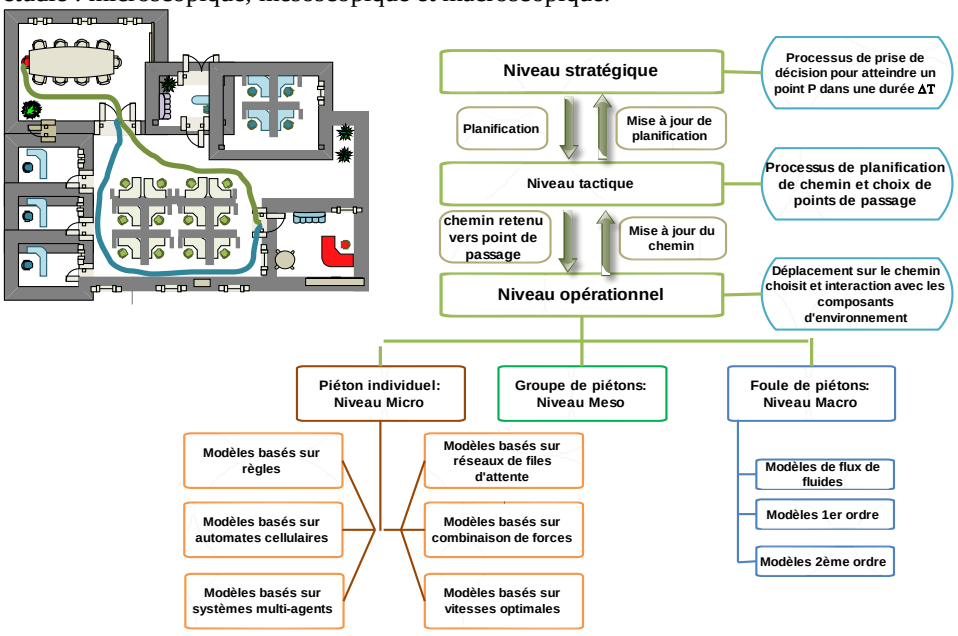


FIG. 1 – Décomposition du déplacement des piétons virtuels.

2.1 Approche macroscopique

Les modèles de la dynamique des foules de piétons appartenant à cette échelle représentent une généralisation de ceux du trafic routier, en considérant la nature multidimensionnelle de la dynamique et les motivations des piétons à se déplacer vers des objectifs spécifiques Bellomo et Dogbè (2011). Ces modèles s'appuient sur une analogie entre le déplacement collectif des piétons et celui des fluides et des écoulements granulaires Cusack (2002), Helbing et al. (2000). L'approche macroscopique utilise l'équation de la conservation de la masse et l'équilibre de la quantité de mouvement Bellomo et Dogbè (2011). Cependant Il faut rappeler que les piétons possèdent une souplesse de déplacement dans deux dimensions, avec une grande flexibilité de s'arrêter et d'avancer selon l'intervalle des vitesses admissibles. Or que l'approche hydrodynamique se réfère à des quantités moyennes au niveau local, et par conséquent les fluctuations locales de la vitesse ne sont pas modélisées explicitement. Par ailleurs les piétons ne se comportent pas uniquement selon les lois de la physique. De plus, les particularités et hétérogénéités des déplacements des piétons ainsi que leurs caractéristiques individuelles, ne sont pas prises en considération Still (2000), Lerner et al. (2007).

2.2 Approche mesoscopique

Cette approche adopte le principe de peloton du trafic routier en s'intéressant aux groupes de piétons ayant des caractéristiques comportementales communes Hanisch et al. (2003). Elle est utilisé lorsque l'état du système peut être identifié par les positions et les vitesses des entités microscopiques, tandis que leurs représentation est donnée par une distribution de probabilité convenable sur l'état microscopique.

Les modèles servant cette échelle se différencient de la manière dont ils modélisent les interactions entre les particules. Ces interactions peuvent être localisées, comme dans le cas de l'équation de Boltzmann, ou à portée moyenne, comme pour l'équation de Vlasov. La différence par rapport à la théorie cinétique classique réside dans le fait que les interactions ne suivent pas les règles de la mécanique classique, mais plutôt la stratégie de conduite est exprimée par les règles comportementales des groupes de piétons formés. Cependant il faut noter que les variables d'espace et de vitesse sont définis dans plus d'une dimension spatiale.

2.3 Approche microscopique

Cette échelle décrit la dynamique collective des piétons et ses formes d'auto-organisation, à partir d'une analyse détaillée de leur déplacement individuel. A l'inverse des modèles macroscopique et mesoscopiques, ces modèles tiennent en compte des motivations propres des piétons et leurs interactions Helbing et al. (2000), Teknomo et Millonig (2007). Cependant ils s'affrontent aux problèmes d'analyse, de calcul et de coût. Ils peuvent être classés en plusieurs approches correspondant aux diverses façons de décrire le terme d'accélération sur la base d'une interprétation détaillée des comportements individuels. Des approches qui se présentent comme suit :

- **Modèles basés sur des règles.** Ils ont été largement utilisés pour simuler les troupeaux d'animaux et les foules de piétons Pelechano et Malkawi (2008). Les deux exemples clés sont le modèle de Boids Reynolds (1999) et la métaheuristique PSO

d'intelligence en essaims. Ces modèles présentent des règles simples permettant de simuler un groupe d'entités virtuelles se déplaçant collectivement en évitant des obstacles et toute collision entre elles Izquierdo et al. (2009).

- **Modèles basés sur les automates cellulaires (AC).** Ils adoptent une approche d'intelligence artificielle pour la modélisation de simulation de piétons, qui se base sur de simples formulations des systèmes physiques, dans des conditions discrétisées en termes d'espaces, de temps et de valeurs des quantités physiques. La discrétisation spatiale représente l'espace de déplacement des piétons sous forme de treillis uniforme de cellules discrètes. Tandis que la discrétisation temporelle représente la fréquence de changement de positions des piétons, sous forme de transitions entre cellules. Ces transitions sont régit par une fonction des états des cellules du voisinage des piétons ainsi que celles qu'ils occupent.
- **Modèles basés sur des forces physiques.** Ils sont motivés par l'observation que le mouvement des piétons dévie d'une trajectoire rectiligne en présence des autres piétons ou d'obstacles. Les tendances de comportement social des piétons sont formalisées à l'aide d'une combinaison de forces socio-psychologiques et physiques couvrant la motivation individuelle des piétons et leur évitement des obstacles. D'où provient le terme "forces sociales" (accélération, répulsion, attraction) aux quelles les piétons sont soumis à longue portée Helbing et al. (2000), Moussaid et al. (2009).
- **Modèles basés sur des réseaux à files d'attente.** Ces modèles représentent l'environnement des piétons sous forme de réseau, et décrivent comment les piétons se déplacent d'un nœud à un autre Lovas (1994). Ils s'appuient fortement sur les principes de bases de la théorie d'attente. Or que les hypothèses de cette dernière conditionnent et restreignent son applicabilité dans la modélisation des situations du monde réel.
- **Modèles basés sur des systèmes Multi-agents.** Ils sont particulièrement bénéfiques lorsqu'il s'agit de modéliser une population d'agents hétérogène ayant des comportements complexes. Des exemples importants de modèles basés sur les systèmes multi-agents peuvent être repérés dans les travaux suivants Teknomo et Millonig (2007), Pelechano et Malkawi (2008). Toutefois malgré leurs avantages évidents, ces modèles sont critiqués pour leurs manque (ou rareté) d'intégration d'éléments psychologiques et physiologiques pour les rendre plus réalistes et permettre des prises de décision semblables à celles des êtres humains.

3 Modèle proposé

3.1 Modèle de piétons virtuels flous à base des automates cellulaires

Nous nous sommes basés sur le paradigme des automates cellulaires à deux dimensions et la métaheuristique ACO dans sa version simple, sans se focaliser sur un aspect d'optimisation qui diverge de notre objectif. Ce dernier consiste à élaborer un modèle de déplacement des piétons virtuels, sans leur doter d'une quelconque formule d'intelligence personnelle. Le caractère bio-inspiré de la technique d'intelligence en essaims, traitera les piétons virtuels comme un essaim, pouvant agir et interagir avec les composants de leur environnement, tout en ayant une intelligence collective artificielle distribuée, leur permet-

tant de produire des structures auto-organisées globales qui ne sont même pas envisagées au niveau local.

Les piétons se déplacent avec une préférence de marche personnelle. Une ”matrice de préférence" contient les possibilités de déplacement aux 8 cellules voisines. L'élément central décrit la position actuelle du piéton (voir figure 2). Nous supposons qu'un piéton se déplace à une cellule par pas de temps, et les éléments de la matrice de préférences sont choisit par une superposition de la destination des piétons et de leurs champs de vision. Nous avons modifié le terme de visibilité des fourmis dans la métaheuristique ACO pour l'adapter à la visibilité ou la désirabilité des piétons virtuels. Et ce de manière à inclure la matrice de préférence, les champs de sol dynamique et statique, ainsi que l'occupation de la cellule cible. Les trois premiers paramètres sont considérés comme des nombres flous puisque leur connaissance n'est pas déterministe et reste imprécise pour les piétons. Tandis que nous avons choisit de ne pas fuzzyfier le paramètre d'occupation des cellules, car il ne contient pas un degré d'incertitude. Chaque cellule ne peut être occupée que par un seul piéton à la fois.

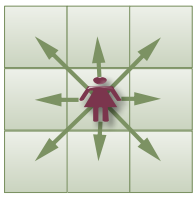


FIG. 2 – *Voisinage de Moore pour un piéton avec les transitions possibles*

Pour chaque piéton se trouvant à une cellule (i) et désirant se déplacer à une cellule (j):

1. Une matrice de préférence floue $\tilde{P}_{ij}$ est assignée, et reflète la possibilité de mouvement aux huit cellules voisines;
2. Une influence statique floue $\tilde{S}_j$ *des objets fixes de l'environnement*;
3. La possibilité de déplacement à une cellule dépend de son état d'occupation ($O_j = 1$ ou $O_j = 0$) ;
4. L'utilité générale floue de déplacement d'une cellule (i) à une cellule (j) est donnée par :

$$\tilde{U}_{ij}(t) = \frac{\left[\tilde{\tau}_{ij}(t)\right]^{\alpha}\left[\tilde{\eta}_{ij}(t)\right]^{\beta}}{\sum_{l \in J_i^T}\left[\tilde{\tau}_{il}(t)\right]^{\alpha}\left[\tilde{\eta}_{il}(t)\right]^{\beta}} \times \mathrm{I}_{V_B^i}(j) \qquad \textbf{(1)}$$

Où $\tilde{\tau}_j$: quantité de phéromone; et α : paramètre de contrôle de l'influence de $\tilde{\tau}_j$; β : paramètre de contrôle de l'influence de $\tilde{\eta}_j$.

La désirabilité ou la visibilité est donnée par: $\tilde{\eta}_j = \tilde{P}_j \times \tilde{S}_j \times (1 - O_j)$ **(2)**

La règle de mise à jour de la phéromone est donnée par: $\tilde{\tau}_j = \rho\tilde{\tau}_j + \Delta\tilde{\tau}_j$ .Où ρ est le taux d'évaporation de la phéromone. $\Delta\tilde{\tau}_{ij} = \sum_{k=1}^{m} \Delta\tilde{\tau}_{ij}^{k}(t)$ est la somme des phéromones dépo-

sées par tous les piétons dans une itération. Or vu que pendant une itération t, un seul piéton occupe une cellule, d'où $\Delta\tau_j^k(t) = O_j(t)$. Nous définissons les termes ρ et $\Delta\tilde{\tau}_{ij}$ comme des nombres flous déterminés par inspection. Les piétons se déplacent à la cellule une utilisé générale floue maximale $\tilde{U}_{ij}(t) = \max_k\left(\tilde{U}_{ik}(t)\right)$

Gestion des conflits. Une situation de conflit se produit si deux piétons ou plus désirent se déplacer à la même cellule au prochain pas de temps. La solution proposée est basée sur l'utilité générale floue de déplacement vers la cellule partagée. Le piéton ayant la plus grande utilité générale est en mesure d'exécuter son pas. Soit $\tilde{R}_1$ (respectivement $\tilde{R}_2$) la valeur de l'utilité générale floue de déplacement du piéton 1 (respectivement piéton 2) pour la cellule partagée. Si $\tilde{\tilde{R}}_1 > \tilde{\tilde{R}}_2$ alors c'est le piéton 1 qui peut se déplacer, sinon c'est le piéton 2.

3.2 Extension du modèle avec les champs de potentiel artificiels

La méthode de champ de potentiels artificiels considère chaque piéton comme une particule se déplaçant dans un champ de potentiel artificiel pour atteindre un objectif donné. Ce dernier agit comme une force d'attraction sur le piéton et les obstacles connus agissent comme des forces répulsives. La superposition de toutes les forces influe la trajectoire du piéton, en le guidant vers son objectif tout en évitant les obstacles.

En général, le champ potentiel scalaire est défini comme la somme du champ d'attraction potentiel de l'objectif, et le potentiel de champ répulsif sur les obstacles Khatib (1986):

$$U = U_{rep} + U_{att} \quad \textbf{(3)}$$

Où U_{att} et U_{rep} sont les champs de potentiel attractif et répulsif respectivement. Similairement le vecteur des forces artificielles $F(p)$ agissant à la position $p = (x, y)$ est donné par :

$$F(p) = F_{att}(p) + F_{rep}(p) \quad \textbf{(4)}$$

Où $F_{att}(p) = -\nabla U_{att}$ et $F_{rep}(p) = -\nabla U_{rep}$. Le terme ∇U désigne le gradient du vecteur potentiel U à la position du piéton $p = (x, y)$ dans un espace à deux dimensions.

3.2.1 Champ de potentiel attractif

La forme la plus utilisée des champs de potentiel attractif a été introduite par Khatib (1986). Elle est définit par l'expression suivante :

$$U_{att} = \frac{1}{2}\xi \ d_g{}^2 \quad \textbf{(5)}$$

Où $d_g = \|p - p_g\|$ désigne une distance euclidienne; p est la position actuelle du piéton et and p_g est la position du point d'attraction. Le terme ξ est une constante positive ajustable. La force attractive $F_{att}(p)$ peut être calculée puisque le potentiel correspondant est dérivable. Nous avons donc :

$$F_{att}(p) = -\nabla U_{att} = \xi \ \|p - p_g\| \quad \textbf{(6)}$$

Par conséquent, plus le piéton s'approche de sa destination, plus la force de champs de potentiels artificiels qui lui est appliquée tends vers zéro, et inversement.

3.2.2 champ de potentiel répulsif

Ge and Cui ont proposé une formulation du champ de potentiel répulsif, afin de résoudre problème des objectifs non atteints dans une certaine configuration de l'espace comme suit Ge et Cui (2000) :

$$U_{rep} = \begin{cases} \frac{1}{2}\eta\left(\frac{1}{d_o} - \frac{1}{\rho_0}\right)^2 \left\|p - p_g\right\|^n & if \quad d_o \le \rho_0 \\ 0 & if \quad d_o > \rho_0 \end{cases} \tag{7}$$

Le terme $\left\|p - p_g\right\|^n$ assure que le potentiel total atteint son minimum global 0, si et seulement si $p = p_g$. Ainsi la force répulsive correspondante est donné par:

$$F_{rep} = -\nabla U_{rep} = \begin{cases} F_{rep1} n_{OR} + F_{rep2} n_{OG} & if \quad d_o \le \rho_0 \\ 0 & if \quad d_o > \rho_0 \end{cases} \tag{8}$$

Où

$$F_{rep1} = \begin{cases} \eta\left(\frac{1}{d_o} - \frac{1}{\rho_0}\right)\frac{d_g^n}{d_o^2} & d_o \le \rho_0 \\ 0 & d_o > \rho_0 \end{cases} \tag{9}$$

Et

$$F_{rep2} = \begin{cases} \frac{n}{2}\eta\left(\frac{1}{d_o} - \frac{1}{\rho_0}\right)^2 {d_g}^{n-1} & if \quad d_o \le \rho_0 \\ 0 & if \quad d_o > \rho_0 \end{cases} \tag{10}$$

Avec $n_{OR} = \nabla d_o$ et $n_{OG} = -\nabla d_g$ sont deux vecteurs unitaires pointant de l'obstacle vers le piéton, et de celui ci vers l'objectif respectivement. La force F_{rep} repousse le piéton de l'obstacle avec sa composante F_{rep1} et l'attire vers l'objectif avec sa composante F_{rep2}. Par conséquent.

Nous avons modifié l'expression de désirabilité (eq. 3) des piétons pour inclure les champs de potentiels artificiels comme suit :

$$\tilde{\eta}_{ij}(t) = \exp\left[-\left(U_{att}(c_j) + U_{rep}(c_j)\right)\right] \times (1 - O_{ij}(t)) \tag{11}$$

Où U_{att} désigne le champ de potentiel attractif, s'exprimant par : $U_{att}(c_j) = \frac{1}{2} * K_{att} * {d_g}^2$

Et U_{rep} désigne le champ de potentiel répulsif, s'exprimant par :

$$U_{rep}(c_j) = \begin{cases} \frac{1}{2} * K_{rep} * \left[\frac{1}{d_o} - \frac{1}{\rho_0}\right]^2 * {d_g}^2 & if \quad d_o \le \rho_0 \\ 0 & if \quad d_o > \rho_0 \end{cases}$$

Alors que ρ_0 est un paramètre spécifique pour chaque obstacle dans la grille de simulation. Les termes k_{att} et k_{rep} sont des constants strictement positives. d_o et d_g sont les distances de *Manhattan* entre la position du piéton et celle du plus proche obstacle et objectif respectivement. Or, vu que dans le monde réel, les piétons ne peuvent pas toujours se déplacer en ligne droite, nous choisissons alors de remplacer la distance de Manhattan. Nous

avons choisi d'appliquer l'algorithme de Brush Fire dans le calcul de la distance entre la cellule cible suivante et un obstacle, et l'algorithme Front Wave pour celle entre la cellule prochaine cible et l'objectif. Ces algorithmes sont appliqués à l'aide d'un voisinage Von Neumann quartier. Ces algorithmes utilisent un gradient de distance dans une grille pour déterminer le potentiel des cellules McGough et al. (2010). Le champ de potentiel total résultant s'obtient par superposition des deux champs attractif et répulsif (voir figure 3).

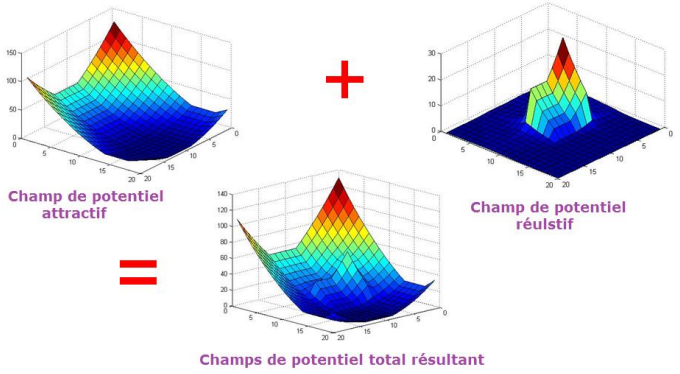


FIG. 3 – *Champ de potentiel total résultant*

4 Exposition des piétons au risque d'accidents

Nous supposons que les piétons sont exposés à des risques quand ils traversent la route. Cette hypothèse est presque réaliste compte tenu du faible taux d'accidents ailleurs que sur les routes. En outre, les trajectoires de traversée de piétons peuvent avoir des formes différentes. Le choix de trajectoire représente généralement un compromis entre la perception de risque par le piéton et son habilité à traverser avec le plus de confort possible. Sur le passage piéton, la trajectoire prends habituellement la forme de d'une ligne plutôt perpendiculaire à la route, alors qu'hors passage piétons, ces derniers ont tendance à arrondir les angles et choisir des lignes obliques Wakim (2005). Quelque soit la trajectoire et la section de route choisit, les piétons tentent d'adapter leur vitesse en fonction de la situation à laquelle ils sont exposés. .

L'exposition au risque pour des entités données, est généralement définit comme étant le produit de leur débit et la durée de leur exposition. Dans notre contexte, La mesure d'exposition des piétons aux risques d'accidents est définit comme suit :

$$Exp_{p/V} = q_V \cdot t_c \quad \textbf{(12)}$$

Où q_V et t_c sont le débit des véhicules et le temps de traversée des piétons respectivement. Nous supposons que les piétons traversent en ligne rectiligne.une route d'une largeur donnée avec une vitesse v_p . Par ailleurs le débit des véhicules peut être exprimé en fonction de leur densité et de leurs vitesses. Ces dernières atteignent leur valeur maximale si la densité est nulle, et s'annule dans le cas d'une densité maximale (voir figure 4), selon l'équation :

$$\upsilon = \upsilon_{\max}\left(1 - \frac{\rho}{\rho_{\max}}\right) \quad \textbf{(13)}$$

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

$$Exp_{p/V} = \rho_V \cdot \upsilon_{\max} \cdot t_p - \frac{\rho_V^2 \cdot \upsilon_{\max} \cdot t_p}{\rho_{\max}} \quad \textbf{(14)}$$

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

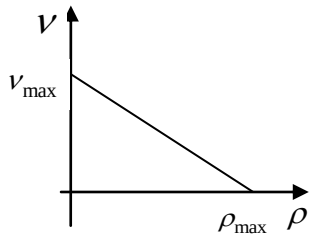


FIG. 4 – Variation de la vitesse en fonction de la densité

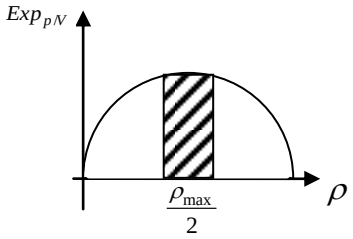


FIG. 5 – variation de l'indicateur de risque d'accidents des piétons en fonction de la densité des véhicules.

Par ailleurs, les routes et intersections se voient équipées par des feux rouges pour les piétons, dont le changement d'état dépend de l'état de ceux des véhicules. Ainsi selon les stéréotypes de sexe, les piétons transgressent ou pas leurs règles de passage sur la route. Le passage au feu rouge est populaire. De plus, quand quelqu'un réussi à traverser dans l'écart entre les deux véhicules, d'autres vont accélérer et le suivre, imposant aux véhicules de leurs céder le passage. Ce qui nous pousse à penser qu'il existe un risque d'accident pour les véhicules, imposé cette fois par les piétons. Cette exposition prend la même forme que dans le cas des piétons

$$Exp_{V/p} = q_p \cdot t_V \quad \textbf{(15)}$$

Où cette fois q_p et t_V sont le débit des piétons et le temps de passage des véhicules respectivement. Nous avons utilisé d'une part le modèle du conducteur intelligent IDM Treiber et al. (2000) pour simuler les déplacements longitudinaux des véhicules. Et d'autre part le modèle MOBIL Kesting et al. (2007) pour gérer leurs changements de voies.

5 Résultats

Le scenario de simulation est représenté sous forme de centre commercial avec deux entrées sur lesquelles sont placés les générateurs de piétons. Des centres d'intérêts tels que les magasins, les cafés, lieux de repos… ces derniers sont représentés par différents objectifs placés

sur la carte de simulation. Les piétons entament leurs visites avec une motivation personnelle de visiter un objectif définis et bien sur tout en gardant la possibilité d'être attirés par d'autres coin du centre commercial. Lors de la simulation (Fig. 6), La densité de piétons qui augmente jusqu'à atteindre une valeur maximale lorsque le trafic devient congestionné (Fig. 7), tandis que le flux piétons diminue dans la même phase (Fig. 8)

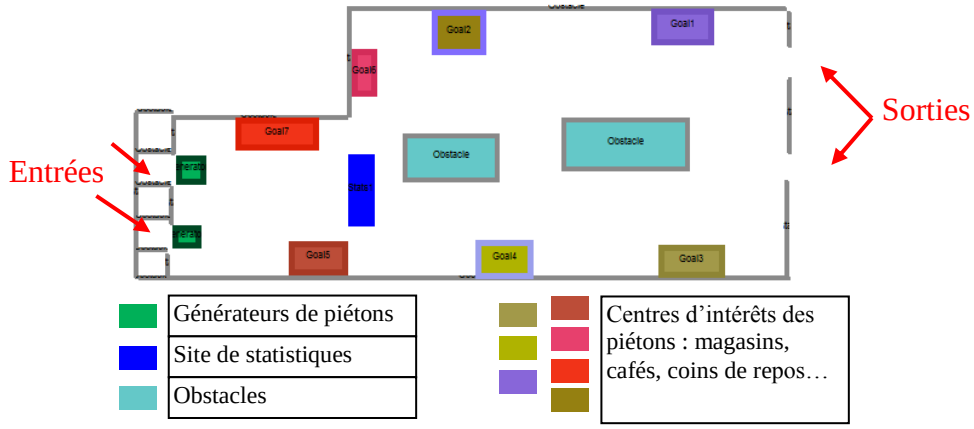


FIG. 6 – Scénario de simulation sous forme de centre commercial

Paramètres d'influence des champs de sols $\alpha = 1 \quad \beta = 1$

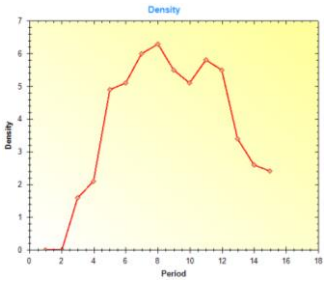


FIG. 7– Diagramme de densité des piétons

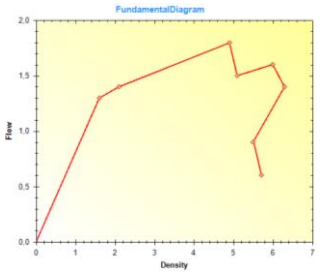


FIG. 8 – Diagramme fondamental du trafic des piétons.

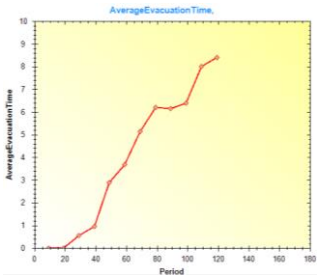


FIG. 9 – diagramme du temps moyen d'évacuation des piétons durant la période de simulation.

Le diagramme fondamental (Fig. 8) illustre les phases de trafic piéton libre et congestionné. Par ailleurs, le temps moyen d'évacuation pour les piétons augmente tandis que la densité de piétons augmente (Fig. 9). En effet, pour de très faible densité, les piétons peuvent s'évacuer facilement et rapidement que dans le cas d'une densité maximale où la circulation des piétons devient pratiquement impossible.

6 Conclusion

Nous avons présenté dans ce papier, un modèle microscopique de simulation floue des déplacements des piétons virtuels. Et ce en s'inspirant de la métaheuristique ACO de l'intelligence en essaim. Nous avons également tiré profit des avantages du paradigme des automates cellulaires, et de la méthode des champs de potentiels artificiels pour guider les piétons dans leurs navigations. La fuzzyfication de l'utilité de déplacement des piétons ne concerne que la perception spatiale (obstacles, la quantité de phéromone, etc.) Notre objectif dans cette approche est d'avoir un modèle simple intégrant la modélisation floue et le paradigme colonie de fourmis artificielle et le concept des champs de potentiel. Ce modèle assure une navigation des piétons en les attirant automatiquement à leurs objectifs tout en les repoussant d'obstacles. Certes, d'autres facteurs cognitifs et comportementaux seront pris en compte dans nos travaux futurs. En effet nous nous pencherons sur la dangerosité de la traversée des intersections par les piétons. La perception des vitesses du véhicule par les piétons et d'autres facteurs psychologiques peuvent être intégrés. L'architecture logicielle du simulateur permet cette extension. L'utilité floue générale proposée ici, peut être interprété comme une probabilité floue, prolongeant ainsi la probabilité de transition nette donnée par le paradigme colonies de fourmis. Les résultats des simulations confirment les prédictions données par la théorie du premier ordre des flux de trafic. La validation du modèle de simulation vers les données du monde réel est recommandée pour une étude plus approfondie. Dans nos travaux futurs, nous prévoyons de mettre à jour la solution logicielle développée pour compléter l'étude d'exposition de risque d'accident pour les piétons et les véhicules, afin d'estimer le risque de traversée des intersections du réseau urbain.

Références

Bellomo, N., and Dogbè C. (2011). SIAM Rev 53 :409–463.

Boulmakoul, A., and M. Mandar. (2011). The Open Operational Research Journal, [DOI: 10.2174/1874243201105010019]. ISSN : 18742432 :19-29.

Cusack, M. A. (2002). Mathematical and Computer Modelling of Dynamical Systems 8:33–48.

Ge, S., and Y.J Cui (2000). IEEE Transactions on Robotics and Automation 16 :615- 620.

Hanisch, A., J. Tolujew, K. Richter, and T. Schulze (2003). Winter Simulation Conference, USA .

Helbing, D., I.J. Farkas, and T. Vicsek (2000). Nature, 407 :487-490.

Henderson, L (1974). Transportation Research 8 :509-515.

Hughes, R (2003). Annual Review of Fluid Mechanics 35 :169–182.

Izquierdo, J., I. Montalvo, R. Pérez, and V.S. Fuertes (2009). Physica A :Statistical Mechanics and its Applications 388 :1213–1220.

Kesting, A., Treiber M., and Helbing D. (2007). Transportation Research Record: Journal of the Transportation Research Board 1999:86-94.

Khatib, O (1986). *Real-Time Obstacle Avoidance for Manipulators and Mobile Robots*. The International Journal of Robotics Research. Vol. 5, No. 1, pp. 90-98, 1986

Lerner, A., Y. Chrysanthou, and D. Lischinski (2007). Eurographics 26:655- 664.

Lovas, G (1994). Transportation Research B 28:429–443.

McGough, J., and R.C. Hoover (2010). Proceedings of the IASTED International Conference Cambridge, Massachusetts, USA.

Mandar M. 2012. Intelligence artificielle distribuée pour la micro-simulation floue des piétons virtuels. PhD thesis Université Hassan II Mohammedia.

Moussaid, M., D. Helbing, S. Garnier, A. Johanson, M. Combe, and G. Theraulaz (2009). Proc. Roy. Soc. BBiol. Sci. 276 :2755-2762.

Pelechano, N., and A. Malkawi (2008). Automation in Construction 17:377–385.

Reynolds, C (1999). Game Developers Conference:763-782.

Still, K (2000). *Crowd Dynamics*. PhD thesis University of Warwick.

Teknomo, K., and A. Millonig (2007). *A navigation algorithm for pedestrian simulation in dynamic environments*. Proceeding of the 11th World Conference on Transport Research (WCTR) University of California, Berkeley.

Treiber, M., Hennecke A., and Helbing D. (2000). Physical Review E 62 :1805-1824.

Wakim, C (2005). Etude de la prédiction de chocs véhiculepiéton. PhD thesis Université Paris Sud - Paris XI.

Summary

Research in the field of collective dynamics of pedestrians grows with the growth of its applicability in many civilian systems. In the first part of this work, we presented the trends of individual and collective behavior of pedestrians in the literature. We also present a microscopic model of pedestrian navigation in virtual environments constrained. This model was established in two phases. The first couple ACO metaheuristic and cellular automata, while adopting a fuzzy representation of some imprecise parameters Boulmakoul and Mandar (2011). The second phase extends the model using the method of artificial potential field. The superposition of the forces applied by the static components of the environment, guides pedestrians to their destinations while being repelled by obstacles. We also present a formulation of a mutual exposure indicator to the accidents risk between pedestrians and vehicles. The software system developed allows the simulation of traffic of pedestrians and vehicles in Virtual various spatial configurations. The indicators provided by the proposed model are in accordance with the laws of transport. In future work, the software solution in the deployment phase will be integrated into the analysis of pedestrian accidents in urban transport networks.

Vidéo surveillance : Analyse des déplacements de personnes dans un environnement clos

Boutaina Hdioud*
Rachid Oulad Haj Thami*, Mohammed El Haj Tirari[+]*

*Equipe RIITM, ENSIAS
Université Mohamed V Souissi
hdioud.boutaina@hotmail.fr, oulad@ensias.ma

[+]Institut National de Statistique et d'Economie Apliquée
Rabat, Maroc
tirari@insea.ac.ma

Résumé. Dans cet article, nous nous intéressons à l'analyse de la trajectoire des personnes dans un environnement clos à partir du flux vidéo d'une caméra de surveillance. L'intérêt de ce travail est de développer des outils d'analyse du comportement des personnes. Ce travail trouve ses applications dans divers domaines comme : la surveillance de personnes à mobilité réduite dans un environnement clos, les clients dans une galerie marchande, le comportement des voyageurs dans une gare ou les piétons dans un carrefour.

1 Introduction

L'acquisition des données par différents outils informatiques permet de mettre à la disposition des analystes un nombre important de données. Sans l'intelligence, l'archivage de ces données n'aura aucun intérêt. En effet, l'intelligence permet d'établir des règles, de déduire des résultats, de faire des prédictions, etc...

La fouille de données (Data Mining) (Jambu, 1999 ; Tufféry, 2005), ou encore analyse intelligente des données, désigne l'ensemble de méthodes destinées à l'exploration et l'analyse des données informatiques. Elle permet d'ajouter de l'intelligence aux archives de données afin d'être capable de prendre des décisions.

Dans le cas de la vidéo, l'acquisition des séquences d'images par des caméras permet de les analyser afin de contrôler et de suivre le déplacement des personnes dans une scène. Elle permet également d'obtenir plusieurs caractéristiques de ces séquences d'images comme par exemple, les trajectoires des personnes en mouvement. Ainsi, ces trajectoires peuvent être utilisées pour analyser les déplacements des personnes dans une scène à travers le regroupe-

ment de celles-ci dans des classes homogènes construites à l'aide des méthodes de classification.

Le suivi des trajectoires est un problème basique mais essentiel dans de nombreux domaines. Il consiste à déterminer le chemin parcouru par chaque personne afin d'avoir une idée sur les trajectoires empruntées par les personnes. Par exemple, le suivi des trajectoires peut être utilisé pour contrôler les patients dans les hôpitaux ou les personnes dans les centres commerciaux.

Les techniques de suivi de trajectoires sont issues du domaine de radar dans lequel on établit des pistes à partir des mesures obtenues en les associant aux mesures précédentes afin de les mettre à jour ou de terminer certaines qui seraient sorties de la zone de surveillance ou d'initialiser d'autres.

Plusieurs approches ont été développées dans le domaine de suivi des trajectoires des personnes. Par exemple, Berclaz et al. (2006), Rota (1998), Han et al. (2007). En effet, l'objectif de Rota (1998) était de détecter, de reconnaître et de suivre plusieurs personnes qui parcourent une scène au cours du temps. Le système de suivi doit suivre leurs pistes, c'est-à-dire l'ensemble de points qui correspondent aux positions des objets au cours du temps.

Quant à Berclaz et al. (2006), ils ont proposé un modèle simple et robuste pour suivre les personnes avec la trajectoire globale. Ils ont traité les trajectoires individuellement et séparément sur le long de la séquence vidéo dans le but d'éviter de confondre les individus.

Han et al. (2007) proposent un suivi multi hypothèses qui intègre le processus de détection dans le processus de suivi, la trajectoire globale est recherchée dans les multiples hypothèses. Ces hypothèses sont détectées et générées. Ainsi, un modèle d'observation autorise le suivi de multiples trajectoires.

Une fois les trajectoires des personnes sont détectées et enregistrées, la classification vise regrouper ces trajectoires en plusieurs groupes (clusters) composés d'individus à comportement similaire, c'est-à-dire des groupes de trajectoires similaires. Pour cela, les travaux réalisés peuvent être classés en deux grandes catégories : l'étude de similarité entre trajectoires, et la conception d'algorithmes de classification adaptées aux trajectoires.

La mesure de similarité est un outil indispensable pour les documents vidéo numériques. Dans le cas de la mesure de similarité entre trajectoires, plusieurs distances ont été proposées. Par exemple, Berndt et Clifford (1996) ont présentés la distance DTW (Dynamic Time Warping) qui permet de comparer des trajectoires de longueurs différentes. Lin et Su (2005) proposent la distance OWD (One-Way Distance) pour comparer des trajectoires en se basant seulement sur leur forme spatiale. Vlachos et al. (2002) exploitent le principe de LCSS (Longuest Common Subsequence) pour proposer un ensemble de distances et de mesures de similarité qui sont robustes face à la présence de données aberrantes dans les trajectoires analysées.

Pour détecter les trajectoires similaires, on fait recours à des méthodes de classification dont l'objectif est de regrouper un ensemble de trajectoires en des classes homogènes. Plusieurs approches ont été proposées dans la littérature, que ce soit des méthodes de classification par partitionnement (ex. K-Means), des méthodes de classification hiérarchiques (ex. BIRCH) ou encore des méthodes de classification basées sur la densité (ex. DBSCAN et OPTICS).

Dans cet article, nous développons une méthode permettant de suivre plusieurs personnes en même temps dans un environnement fermé à partir de leurs trajectoires et de les regrouper dans des classes homogènes selon la similitude de leurs déplacements dans cet environnement. Le suivi de l'évolution de la position des personnes en temps réel nous a permis

d'avoir des informations sur les endroits occupés par chaque personne dans une scène. Afin de pouvoir regrouper les trajectoires obtenues dans des classes homogènes, nous proposons un nouvel algorithme qui permet d'adapter la technique de classification hiérarchique ascendante (CHA) au cas de la mesure de similarité entre trajectoires.

La figure1 suivante schématise le processus que nous allons détailler dans ce papier :

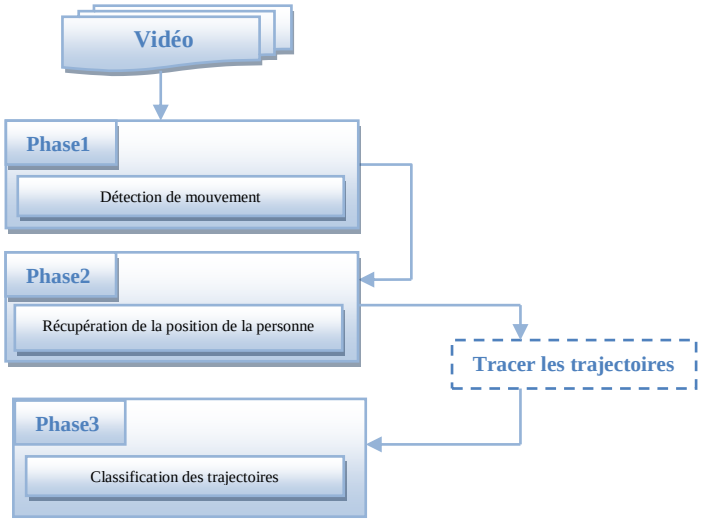


FIG. 1 – *Processus mis en place*

Ce papier est organisé de la manière suivante : la section 2 est consacrée à la présentation des techniques utilisées dans la méthode que nous proposons pour détecter une trajectoire tout en les regroupant dans des classes homogènes. La méthode de classification utilisée pour regrouper les personnes dans des classes homogènes selon la similarité de leurs trajectoires, est présentée à la section 3. Une illustration des résultats obtenus avec les techniques proposées est donnée à la section 4. Enfin, la section 5 est consacrée à une conclusion générale montrant l'intérêt des travaux réalisés ainsi que les perspectives en termes de pistes de recherches relatives aux méthodes de détection et de classification proposées dans ce papier.

2 Détection et extraction de la trajectoire d'une personne

Pour pouvoir suivre des personnes en mouvements dans une vidéo filmée dans un environnement clos, la première étape consiste à les détecter. Cette étape joue un rôle très important dans le système de vidéo surveillance car le résultat de celle-ci va influencer toutes les étapes suivantes.

2.1 Détection des objets en mouvement

L'identification des objets en mouvement est une tâche importante pour de nombreuses applications. Elle fournie une classification de pixels, soit en avant-plan (mobile) ou arrière-plan (statique). En raison des changements dynamiques dans des scènes naturelles, plusieurs

techniques ont été développées dans ce domaine, les plus utilisées sont la soustraction du fond, le flux optique, etc. Dans notre cas nous avons choisi l'une des techniques statistiques de soustraction du fond qui utilisent les caractéristiques statistiques de chacun des pixels. Ces techniques font aussi la mise à jour dynamiquement des statistiques de pixels qui appartiennent à l'image de fond. Les pixels d'avant-plan sont identifiés en comparant chaque pixel avec les statistiques du modèle de fond.

2.1.1 Modèle de mélange gaussienne (GMM)

Pour détecter les objets en mouvement dans une vidéo, nous avons choisi d'appliquer le modèle de mélange gaussienne (GMM) qui est une technique très répondu dans le cas des arrières plan dynamique comme par exemple les vagues sur la mer, les branches d'un arbre agitées par le vent, les écrans de télévision ou d'ordinateur etc. Cette méthode consiste à modéliser les valeurs de chaque pixel à l'aide de densités de probabilités Gaussienne et ceci afin de classifier les pixels, c'est-à-dire les mesures dont la probabilité d'être observées est élevée correspondent à des pixels qui seront étiquetées comme arrière-plan, tandis que celles dont la probabilité d'être observées est faible correspondent à des pixels qui seront étiquetés comme avant-plan.

Dans Stauffer et Grimson (1999) chaque pixel de l'arrière-plan est modélisé par un mélange de *K* gaussiennes. La probabilité d'observer la valeur *x* à l'instant *t* est représentée par :

$$P(x_t) = \sum_{k=1}^{K} \omega_{k,t} \eta\left(x_t, \mu_{k,t}, \Sigma_{k,t}\right)$$

où $\{x_1, \ldots, x_t\}$ représente l'historique des valeurs des pixels sur les *t* dernières images de la vidéo, $\omega_{k,t}$ est le poids de la $k^{\text{ième}}$ gaussienne et $\eta\left(x_t, \mu_{k,t}, \Sigma_{k,t}\right)$ est le $k^{\text{ième}}$ modèle gaussien défini par :

$$\eta\left(x_t, \mu_{k,t}, \Sigma_{k,t}\right) = \frac{1}{\sqrt{(2\pi)^n \left|\Sigma_{k,t}\right|}} e^{-\frac{1}{2}\left(x_t - \mu_{k,t}\right)\Sigma_{k,t}^{-1}\left(x_t - \mu_{k,t}\right)^T}$$

La matrice de covariance $\Sigma_{k,t}$ est estimée par une matrice diagonale $\hat{\Sigma}_{k,t} = \sigma_k^2 \mathbf{I}$

FIG. 2 – *Exemple de personne détectée par GMM*

Les régions de l'avant-plan détectées par cette technique correspondent à une région compacte avec des pixels isolés et ceci est dû aux fausses détections. Pour supprimer les pixels isolés et combler les trous, nous avons choisi d'utiliser un ensemble d'opérations morphologiques. Ensuite, les pixels de l'avant-plan sont regroupés sous forme de composantes connectées.

2.1.2 Les composantes connectées

L'étiquetage par composantes connectées consiste à regrouper les pixels d'une image au sein de classes basées sur la connexité des pixels et ceci afin de déterminer les différentes régions qui la compose. Les régions constituées de pixels adjacents possèdent la même valeur.

Dans notre cas, nous avons filtré l'image détectée par le modèle GMM et ensuite, nous avons regroupé les pixels sous forme de composantes connectées.

Les étapes à suivre pour trouver les composantes connectées dans l'image binaire sont :

1. La recherche du pixel *p* non étiqueté,
2. L'étiquetage de tous les pixels dans la composante connectée contenant le pixel *p* en utilisant l'algorithme flood fill.
3. La répétition de l'étape 1 et 2 jusqu'à ce que tous les pixels soient étiquetés.

2.2 La position de la personne dans la scène

Pour extraire la position d'une personne dans une scène, nous avons suivi la démarche suivante : dès que la personne en mouvement est détectée, on procède par l'étiquetage des composantes connectées relatives à cette personne, c'est-à-dire les pixels sont regroupés sous forme de composantes connectées, et ceci afin d'extraire le centre de gravité de chacune de ces composantes.

Le déplacement de la personne dans la scène est repéré par le centre de gravité de ses composantes connectées. Ainsi, si $X_1, X_2, \ldots, X_n$ sont les n points de la composante connectée détectés dans chaque séquence de vidéo, le centre de gravité de ces points dont les coordonnées sont (x_i, y_i) est le point G dont les coordonnées $(\bar{x}, \bar{y})$ sont données par :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{et} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Où n est le nombre de points de chaque composante connectée.

Vidéo surveillance : Analyse des déplacements de personnes dans un environnement clos.

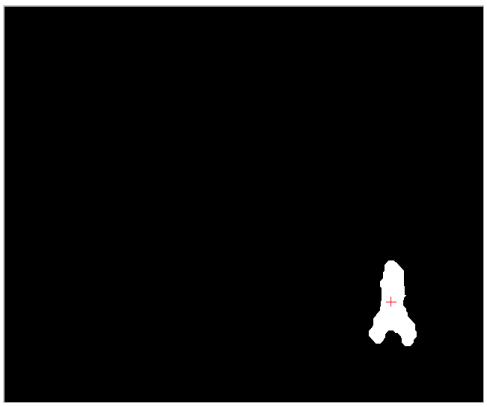


FIG. 3 – Le *centre de gravité d'une personne détectée*

2.3 Extraction des trajectoires des personnes

Après avoir détecté les positions de chaque personne aux différents instants, ces positions ont été utilisées pour tracer le chemin emprunté par les personnes en mouvement. De plus, pour conserver l'historique de déplacement de chaque personne, ses positions aux différents instants sont enregistrées dans un vecteur.

Ainsi, les étapes à suivre pour détecter les trajectoires des personnes en mouvement dans une séquence de vidéo peuvent être résumées dans le diagramme suivant :

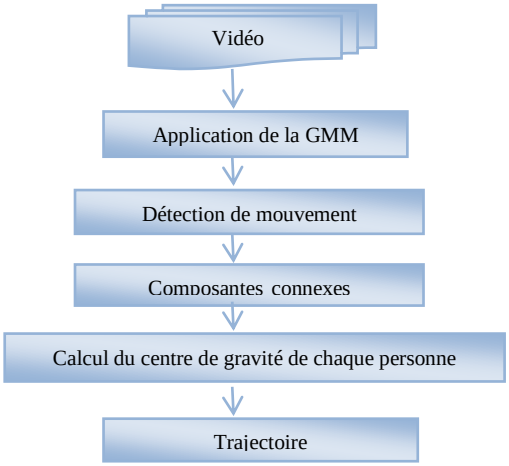


FIG. 4 – *Processus du calcul des trajectoires des personnes en mouvement*

Un exemple de détection en temps réel de la trajectoire des personnes en mouvement dans une séquence de vidéo est donné à la figure 5.

FIG. 5 – *Les trajectoires des personnes en mouvement tracées en temps réel*

3 Classification des trajectoires

L'objectif des méthodes de classification est de partitionner un ensemble d'objets en *k* sous-ensembles homogènes selon les caractéristiques utilisées dans la classification avec *k* est le nombre de regroupements attendus par l'utilisateur.

Pour regrouper un ensemble d'objets dans des classes homogènes, plusieurs algorithmes de classification ont été proposés. Ces algorithmes sont utilisés pour classer des objets en se basant sur des caractéristiques unidimensionnelles. Dans le cas des trajectoires, les caractéristiques de classification sont bidimensionnelles car il s'agit des coordonnées des positions occupées à différents instants. C'est pour cela que les algorithmes de classification existant doivent être adaptés pour tenir compte de l'aspect bidimensionnel des caractéristiques de classification des trajectoires. Dans ce travail, nous nous sommes intéressés à l'adaptation l'algorithme correspondant à la technique de classification hiérarchique ascendante (CHA) au cas où les objets à classer sont des trajectoires.

3.1 Similarité entre trajectoires

Pour comparer la similitude des trajectoires des personnes en mouvement, plusieurs fonctions de distance ont été proposées. On se place dans le cas où les trajectoires pour lesquels on mesure la similitude sont obtenues à partir de la détection des personnes en mouvement dans un environnement fermé contenant une caméra fixe. Pour chaque personne, sa trajectoire est déterminée à travers l'enregistrement de ses positions occupées à des intervalles de temps qui sont les mêmes pour toutes les personnes détectées.

Comme les caractéristiques de classification des trajectoires (les coordonnées des positions occupées à différents instants) sont quantitatives, nous avons choisi d'utiliser la distance Euclidienne pour comparer deux trajectoires ou deux groupes de trajectoires.

On note que pour deux trajectoires quelconque T_i et T_j, La distance euclidienne est définie par :

$$\text{distance}(T_i, T_j)_{Eucli} = \sqrt{\sum_{k=1}^{n}\left[\left(x_{ik} - y_{ik}\right)^2 + \left(x_{il} - y_{il}\right)^2\right]}$$

où (x_{ik} , x_{il}) et (y_{ik} , y_{il}) sont respectivement les coordonnées des positions occupées (au même intervalle de temps) par les centres de gravité des deux personnes correspondant aux trajectoires T_i et T_j.

3.2 Algorithme de classification hiérarchique ascendante

La technique de classification hiérarchique ascendante (CHA) est parmi les techniques de classification les plus utilisées pour regrouper des objets en classes homogènes. Il s'agit d'une démarche algorithmique itérative dont l'objectif est de chercher à effectuer des regroupements d'objets statistiques les plus proches selon plusieurs caractéristiques. Dans notre cas, les objets statistiques sont les trajectoires des personnes en mouvement. Les étapes à suivre pour adapter l'algorithme de CHA au cas de la classification des trajectoires sont les suivantes :

1. Lors de la première étape, chaque trajectoire est considérée comme une classe à part entière. Nous avons donc, à ce niveau du processus, autant de classes que de trajectoires (*n* classes pour *n* trajectoires).
2. L'algorithme de CHA commence par calculer une distance entre toutes les classes en utilisant la distance euclidienne : plus cette distance sera petite, plus les classes seront proches (similaires).
3. Une fois l'ensemble des distances entre les trajectoires sont calculées, l'algorithme va fusionner les deux trajectoires (ou les deux classes de trajectoires) ayant la distance la plus petite (donc les plus semblables) pour ne constituer qu'une seule classe.
4. L'algorithme repart à zéro puisqu'il recalcule, à nouveau, toutes les distances entre les classes de trajectoires, pour fusionner deux nouvelles classes, selon le même principe que précédemment.
5. Ce processus est répété jusqu'à ce qu'il ne reste plus qu'une seule classe de trajectoires. En d'autres termes, toutes les classes finissent, en fin d'algorithme par ne constituer qu'une seule classe (1 classe pour *n* trajectoires).

4 Résultats expérimentaux

Afin d'illustrer les étapes à suivre pour réaliser la méthode proposée dans ce travail, permettant de détecter et de classer ensuite les trajectoires des personnes en mouvement dans une vidéo, nous avons utilisé une séquence de vidéo surveillance issue de la base de données de l'Université d'EdinBurgh. Il s'agit d'une séquence de vidéo filmée par une caméra fixe où la scène comporte plusieurs scénarios de personnes qui se déplacent devant la caméra.

Les images contenues dans les figures 6 à 8 présentent les résultats obtenus pour la détection des personnes en mouvement. En effet, dans la figure 6, on dispose de l'image réelle ne

contenant aucune personne en mouvement. La première image de la figure 7 montre que la méthode GMM permet bien de détecter deux personnes en mouvement. Tandis que, dans la deuxième image de la figure 7, les personnes en mouvement sont détectées automatiquement par un MBR (Minimum Boundary Rectangle).

FIG. 6 – *L'image réelle ne contenant aucune personne en mouvement*

FIG. 7 – *La première image est celle détectée par la méthode GMM. La seconde est celle qui contient les personnes détectées automatiquement par un MBR*

Pour la détermination des trajectoires des personnes en mouvement, l'image de la figure 8 contient les résultats obtenus suite à l'utilisation de l'algorithme permettant de suivre les déplacements des personnes en mouvement en se basant sur le suivi de la position du centre de gravité des composantes connectées de ces personnes.

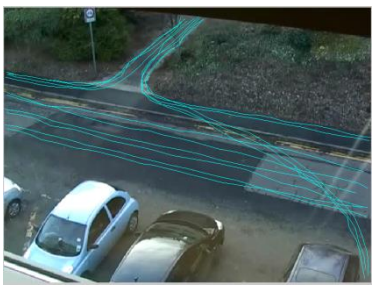


FIG. 8 – *L'ensemble des trajectoires des personnes ayant traversées la zone filmée*

Pour pouvoir appliquer notre algorithme de classification de trajectoires, nous disposons d'un échantillon de 21 trajectoires qui correspondent aux personnes ayant traversées la zone filmée. Ainsi, la figure 9 contient l'arbre de la CHA (Dendogramme) obtenue pour les 21 trajectoires. Il s'agit d'un graphique qui retrace toutes les étapes de regroupement des trajectoires dans des classes.

On note qu'à chaque étape, on agrège les deux classes le plus proches, ce qui augmente l'hétérogénéité à l'intérieur des classes. La longueur des axes horizontaux de l'arbre de classification représentent les pertes en termes d'homogénéité à l'intérieur des classes enregistrées à chaque étape de l'agrégation des classes. Ainsi, on arrête de fusionner les classes lorsqu'on enregistre une perte importante d'homogénéité à l'intérieur des classes.

Par conséquent, l'examen de l'arbre de la CHA obtenue pour les 21 trajectoires montre qu'on peut construire trois classes de trajectoires.

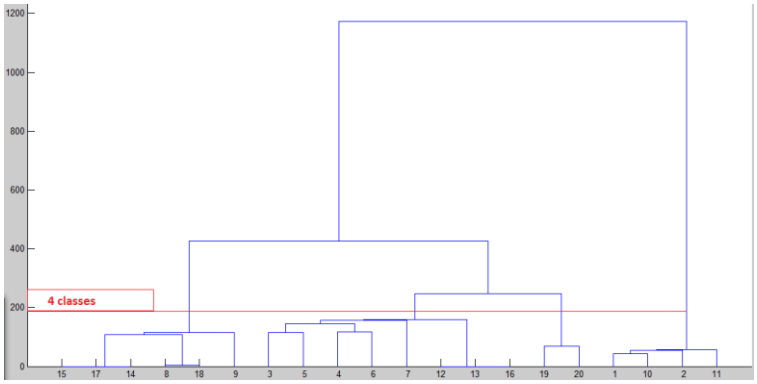


FIG. 9 – L'arbre de la CHA (Dendrogramme)

La figure 10 contient les résultats de la CHA appliquée sur les 21 trajectoires regroupées en quatre classes où les trajectoires appartenant à une même classe sont représentées par la même couleur.

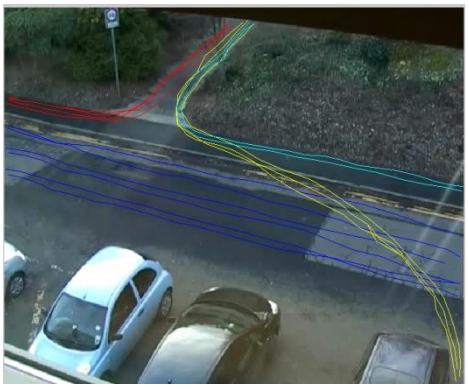


FIG. 10 – *Résultat de la classification des trajectoires des personnes en mouvement*

5 Conclusion

Pour détecter et classer en temps réel les trajectoires des personnes en mouvement dans une séquence de vidéo, la méthode que nous avons proposée dans ce travail est basée sur trois phases. La première phase a pour objectif de détecter les personnes en mouvement. Pour cela, nous avons utilisé la méthode GMM pour extraire les objets mobiles. La deuxième phase est celle de la détermination des trajectoires des personnes en mouvement détectées à la première phase. Ces trajectoires sont déterminées en se basant sur la position du centre de gravité des composantes connectées de chaque personne en mouvement.

Enfin, la troisième phase a pour but de regrouper les trajectoires des personnes en mouvement dans des classes homogènes selon la similitude des trajectoires. Pour cela, nous avons adapté l'algorithme de classification hiérarchique ascendante (CHA) au cas de la classification des trajectoires où les caractéristiques de classification sont les positions des personnes à différents instants (bidimensionnelle).

La méthode proposée dans ce travail a été testée sur une séquence vidéo et les résultats obtenus montrent que celle-ci permet bien de détecter les trajectoires des personnes en mouvement tout en les classant dans des groupes homogènes (trajectoires similaires).

6 Références

Ankerst M., Breunig M., Kriegel H.-P., and Sander J. (1999) *Optics : ordering points to identify the clustering structure. SIGMOD Rec*., 28 (2) : 49–60.

Berclaz J., Fleuret F.and . Fua P. (2006). *Robust people tracking with global trajectory optimization EPFL* - CVLAB, CH - 1015 Lausanne, Switzerland.

Berndt D. J. and Clifford J. (1996), *Finding patterns in time series : a dynamic programming approach*, pages 229–248.

Chen L. and Ng R. (2004), *On the marriage of Lp-norms and edit distance*. In VLDB '04 : Proceedings of the Thirtieth international conference on Very large data bases, pages 792–803. VLDB Endowment.

Chen L., Özsu M. T. and Oria V. (2005), *Robust and fast similarity search for moving object trajectories*. In SIGMOD '05 : Proceedings of the 2005 ACM SIGMOD international conference on Management of data, pages 491–502, New York, NY, USA.

Ester M., Kriegel H.-P. and Xu X.. (1996) *A density-based algorithm for discovering clusters in large spatial databases with noise*, pages 226–231,

Han M., Xu W. and Gong Y. (2007). *Multi-object trajectory tracking*, Machine Vision and Applications, Special Issue Paper, Vol. 18, pages 221-232. Springler-Verlag

Jambu M., (1999) *Introducion au Data Mining*, Editions Eyrolles, Paris.

Lin B. and Su J. (2005), *Shapes based trajectory queries for moving objects*. In GIS '05 : Proceedings ofthe 13th annual ACM international workshop on Geographic information systems, pages 21–30,New York, NY, USA.

Roh G.-P. and Hwang S.-W. (2010). *Nncluster : An efficient clustering algorithm for road network trajectories*. In Kitagawa, H., Ishikawa, Y., Li, Q., and Watanabe, C., editors, Database Systems for Advanced Applications, volume 5982 of Lecture Notes in Computer Science, pages 47–61. Springer Berlin Heidelberg.

Rota N., (Septembre 1998) Rapport de DEA : *Système adaptatif pour le traitement de séquences d'images pour le suivi de personnes*. Sous la direction de Monique Thonnat et Nicolas Chleq, Projet ORION, INRIA, Sophia-Antipolis.

Stauffer C. and Grimson W.E.L. (1999) Adaptive background mixture models for real-time tracking. international conference on Computer Vision and Pattern Recognition, 2.

Vlachos M., Gunopoulos D. and Kollios G. (2002). *Discovering similar multidimensional trajectories*. In ICDE '02 : Proceedings of the 18th International Conference on Data Engineering, page 673, Washington, DC, USA. IEEE Computer Society.

Tufféry S., (2005) *Data Mining et statistique décisionnelle L'intelligence dans les bases de données*, Editions TECHNIP, Paris.

Zhang T., Ramakrishnan R. and Livny M. (1996) *Birch : an efficient data clustering method for very large databases*. SIGMOD Rec., 25 (2) : 103–114.

Summary

Abstract. In this paper, we developed a method to track multiple people even in a closed environment from their paths. Monitoring the evolution of the position of people in real time has allowed us to get information on areas occupied by each person in a scene. In addition, in order to consolidate the trajectories obtained in homogeneous classes, we proposed a new algorithm which allows to adapt the technique of hierarchical bottom-up (CHA) to the case of similarity measure between trajectories.

Keywords: Trajectory tracking, GMM, CHA.

Un modèle de fouille de données Cloud basé sur le principe Map/Reduce de Google

Abdelfettah Idri*, Azedine Boulmakoul

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

Résumé. Les données à traiter pour extraire la connaissance depuis les entrepôts de données deviennent de plus en plus volumineuses et complexes vu l'évolution rapide des technologies de stockage et de traitement dans les différents domaines. La nécessité de stockage et de calcul intensif a incité plusieurs chercheurs de la communauté scientifique à développer plusieurs algorithmes et techniques pour répondre aux besoins de ce contexte. Ce processus de fouille de données s'accompagne souvent d'une complexité exponentielle aussi bien d'espace que du temps. En plus d'algorithmes performants et robustes, la fouille de données nécessite des architectures dynamiques et scalables facilement adaptables à son contexte. Du moment que notre démarche repose sur le treillis de Gallois pour la prospection des données, on propose dans ce papier une projection de notre approche parallèle distribuée de la construction du treillis de Galois basée sur CORBA vers le modèle Cloud de Google *Map/Reduce* qui s'apprête à implémenter l'architecture déjà proposée et implémentée dans nos travaux antérieurs.

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. Dans ce papier, on s'intéresse au 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 de Treillis de Galois en préconisant une approche parallèle distribuée Idri 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 de treillis de Galois a fait l'objet de plusieurs recherches, spécialement dans les domaines d'analyse de 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 la distribution intégrale du processus de la fouille de données et donc aussi bien du traitement que de l'utilisation mémoire. L'implémentation de ce modèle de fouille de données Cloud basé sur *Map*/*Reduce* n'est pas pris en compte dans ce document, il fera l'objet de nos futurs travaux de recherche. Ce document est organisé comme suit. Le paragraphe 2 présente la vision globale de l'approche. Le paragraphe 3 expose l'architecture du système. La projection de l'architecture Cloud sur le modèle *Map/Reduce* est abordée dans le paragraphe 4. On conclut dans le paragraphe 5 avec nos suggestions et recommandations.

2 Vision globale

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'applications Grossman 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 data ming 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 et donc elle est 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 Figure 1).

Figure 1. Services de la pile Cloud

La figure ci-dessous 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 par des Tries ou sub-Tries (arbres lexicographiques). Le détail de cette architecture et l'algorithme la supportant sont décrits dans les paragraphes suivants.

2.1 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 Figure 1. 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 génération 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 subdivisé en 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 Figure 2.

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 entier de construction du treillis de Galois. Ce Trie représente en fait la mémoire globale du processus de fouille de données et sa taille impacte son déroulement énormément 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). Pour remédier à cette contrainte, on a opté pour une distribution de cette mémoire (Trie) selon une technique bien précise. Les nœuds constituant cette mémoire sont regroupés dans une couche de mémoire comme schématisé dans la Figure 2.

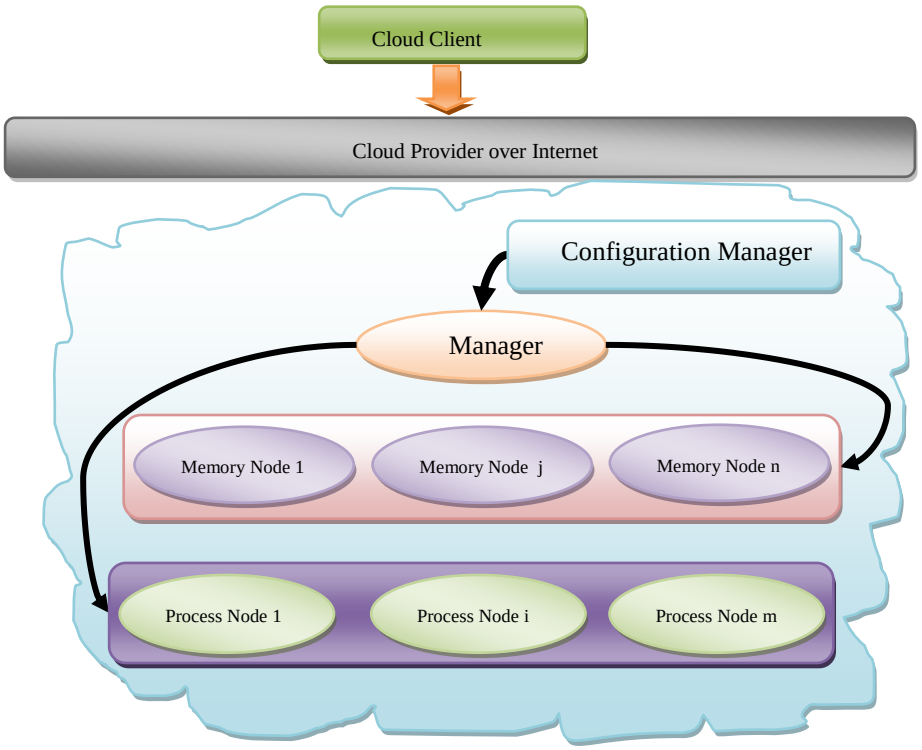


Figure 2. L'infrastructure globale de l'approche Fouille de données Cloud

3 Une architecture parallèle distribuée pour la fouille de données

Dans ce paragraphe, on décrit le fonctionnement de notre architecture CORBA pour faciliter le rapprochement avec le modèle Cloud computing qu'on cible dans ce contexte. On détaille dans ce qui suit les deux volets « traitement » et « mémoires ».

3.1 Aspect traitement

L'architecture proposée dans le schéma ci-dessous est constituée de trois composantes principales dans l'objectif d'améliorer la performance de l'algorithme séquentiel :

- Le Manager : celui-ci utilise d'une part un Trie global pour gérer les concepts constituant le treillis de Galois. D'autre part, le Manager repose sur deux modules pour assurer la communication avec les agents, notamment le Dispatcher et le Collecteur.
  - Dispatcher : distribue les tâches aux Agents (génération des concepts enfants).
  - Collecteur : collecte les résultats et les transmet au Manager. Un résultat est constitué d'une liste de concepts.
- Les Agent : l'Agent est responsable de la génération des concepts enfants en utilisant un Trie local.
- La communication : elle est assurée par le biais d'une infrastructure CORBA basée sur un routeur AMI (*Asynchronous Method Invocation*).

Motivation.

La démarche globale adoptée pour la conception de cette architecture est décrite dans ce qui suit. La première phase a été consacrée à identifier les actions indépendantes de l'algorithme qui peuvent participer à la réduction du temps d'exécution et l'optimisation de l'espace. Dans la deuxième phase, on doit vérifier si ces actions sont dissociables sans impacter le volet communication. Finalement, il reste à étudier les possibilités d'implémentation de l'architecture. En analysant des algorithmes existants Bordat (1986), Choi (2006) et Ganter et Reuter (1991), on a pu distinguer les actions suivantes :

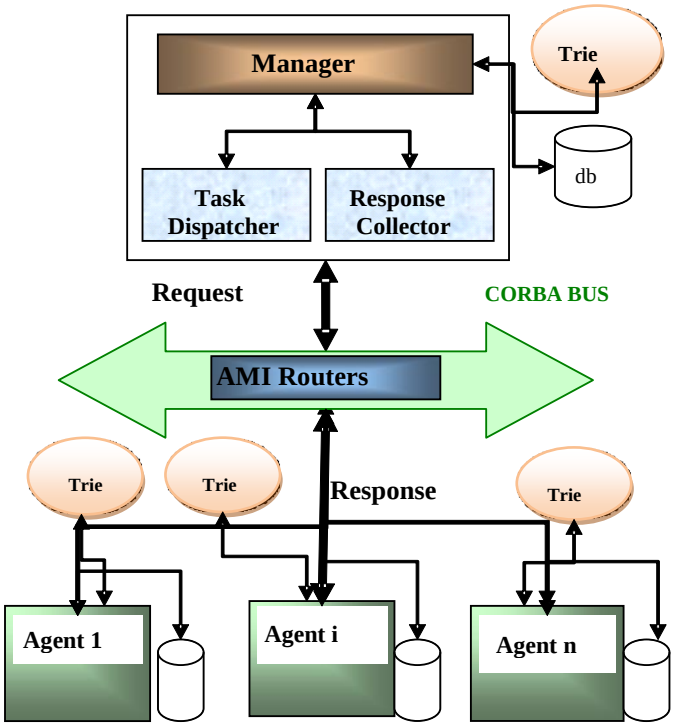


Figure 3. Architecture du système et Algorithme

- La génération des enfants d'un concept
- Le contrôle de fermeture d'un ensemble
- Le contrôle d'existence d'un concept

Le choix a été fait sur le modèle Manager/Agent puisqu'il garantit la scalabilité (le multiplexage des agents implique directement la réduction de la charge d'exécution) et la distribution des services.

La génération des enfants d'un concept est un processus complexe et utilise un arbre local. Cette tâche peut être déléguée aux Agents puisqu'elle est indépendante.

De même, le contrôle de fermeture d'une intention ou une extension peut être aisément délégué aux Agents.

Par contre, L'existence d'un concept est réalisée à l'aide d'un arbre Trie qui comporte au fur et à mesure l'ensemble des concepts générés et donc il doit être partagé par les Agents pour pouvoir tester l'existence. C'est le Manager qui prend en charge la gestion de l'arbre.

On a choisi CORBA pour la communication entre tous les acteurs de cette architecture. L'utilisation de CORBA nous permet d'une part de cacher la complexité des structures de données utilisées dans l'algorithme. D'autre part, CORBA offre des mécanismes de programmation évolués tel que la gestion des événements distribués, le support de la communication asynchrone (AMI) et la programmation orienté objet.

Les services offerts par le Manager et les Agents sont listés ci-dessous.
Manager :

- Gestion de l'arbre (insertion d'un concept, contrôle de l'existence d'un concept)
- Gestion des tâches (distribution, collection, synchronisation)

Agent :

- Génération des concepts enfants
- Contrôle de fermeture de l'intention ou l'extension d'un concept

3.2 Aspect mémoire (Trie global)

Comme schématisé dans la Figure 3, la mémoire mise en jeux dans le processus de génération du treillis de Galois est représentée par le Trie et celle-ci peut atteindre facilement une taille énorme qui impacte le déroulement de ce processus. Pour palier à cette contrainte on a pensé à distribuer la mémoire et donc le Trie similairement à la distribution du traitement. La distribution du traitement est une procédure statique qui s'effectue lors de la configuration de la plateforme d'exécution, alors que la distribution de la mémoire est un phénomène dynamique qui devrait se déclencher sur la base de certains critères dépendant du contexte d'exécution. Notre vision consiste donc à définir une stratégie pour gérer la distribution du Trie. Après analyse, il s'est avéré que la taille du Trie qui est constitué d'un ensemble de nœuds peut être réduite en séparant ses sous-hiérarchies à partir d'un certain nœud tout en gardant l'information liant l'ensemble de ces structures. On aura formé ainsi un réseau de sous-Tries qui nécessite un système d'adressage dynamique pour accéder à leur contenu similairement à l'indexation ou à la pagination. Il reste alors à définir le critère de subdivision qui permettra de passer du Trie aux sous-Tries. On a donc défini la notion de poids d'un nœud qui est le nombre de nœuds se trouvant au dessous de ce nœud. Une fois que ce poids atteigne un certain seuil, le processus de subdivision se déclenche. Comme les branches du Trie sont uniques, l'adresse du nouveau sous-Trie sera constituée simplement de la clé formé de la racine du trie jusqu'au nœud subdivisé. Les nouveaux sous-Tries résultants de ce mécanisme peuvent être hébergés sur différentes machines (sites) accessibles par le biais d'une table d'adressage (table de hachage ou Trie). Ce processus est décrit dans la Figure 5. Selon l'exemple, la subdivision du nœud « 3 » implique la migration de l'hiérarchie au dessous de ce nœud vers le « host i » et une ligne s'insère dans la table d'adressage pour indiquer que toute requête intégrant le nœud « 3 » devrait être redirigée vers le « host i ». De cette manière, on aurait une distribution intégrale du processus de génération du treillis de Galois comme illustré dans la Figure 4.

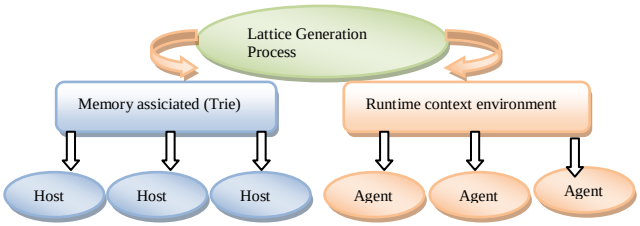


Figure 4. Vue abstraite du processus de génération du treillis de Galois

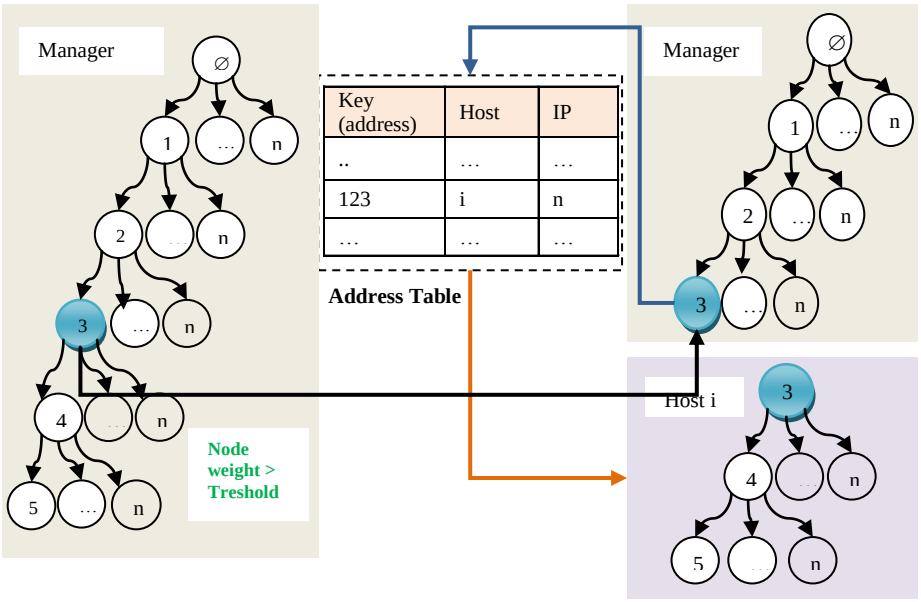


Figure 5. Distribution du Trie (mémoire)

4 Projection de l'architecture Could proposé sur le modèle Map/Reduce

4.1 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 parallèle Dean et 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 formule 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 et de consolider les valeurs intermédiaires ayant une clé commune.
  - **Reduce (Ki, List(Vi)) → List (Vo)**

4.2 Fonctionnement global du modèle Map / Reduce

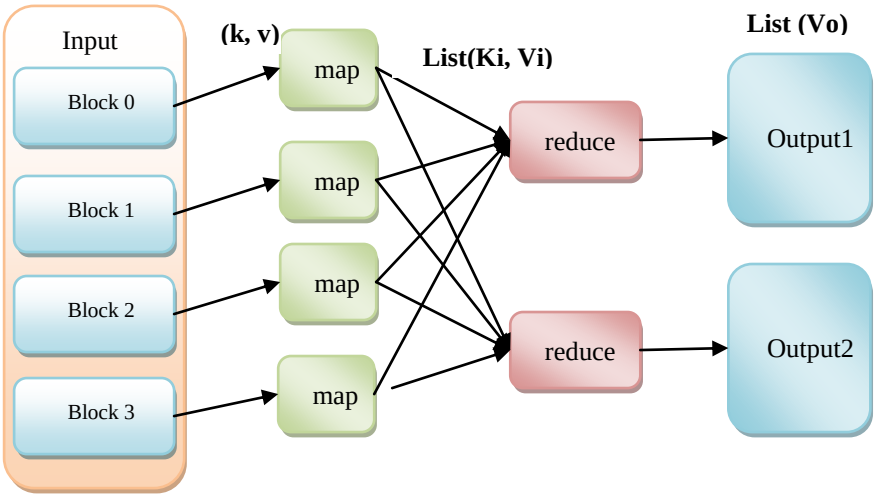


Figure 6 Fonctionnement du modèle Map / Reduce

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. 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.

4.3 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 du moment que ce modèle cible la même stratégie 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
- Il peut répartir la mémoire locale (Trie) en cas de besoin

Reduce : ce processus effectue :

- La consolidation des résultats intermédiaires obtenus par le mapper et les intègre au niveau de la mémoire globale (Trie global)
- Répartition en cas de besoin de la mémoire globale
- 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.

5 Conclusions et perspectives

La distribution et le parallélisme de l'algorithme de construction de treillis de Galois nous a permis :

- la scalabilité de la fouille de données et l'amélioration de sa performance. Cette approche est devenue ainsi intégrable dans un contexte Cloud Computing pour offrir un service fouille de données Cloud exigeant intensivement des ressources en termes de traitement et de mémoires.
- De projeter l'architecture initiale basée sur une infrastructure CORBA vers une architecture Cloud basée sur le modèle Map/Reduce
- de générer la totalité des concepts et par conséquent tous les motifs fermés fréquents et par conséquent les règles d'association sur la base d'un treillis de Galois qui est construit dans une phase préliminaire.
- d'offrir un moyen pour réutiliser les méthodes existantes surtout celles basées sur le treillis de Galois et de construire des infrastructures d'exploitation et de test.

Perspectives :

- Implémentation de la nouvelle architecture basée sur le modèle Map/Reduce en gardant les mêmes principes de l'architecture basée sur CORBA et en utilisant la solution Hadoop à titre indicatif.
- Algorithmes hybrides : En généralisant cette distribution sur plusieurs algorithmes on peut combiner des Agents et des Managers de différents algorithmes et choisir par conséquent les plus performants d'entre eux. Ceci générera des algorithmes hybrides mais sûrement plus robustes que les originaux.
- Optimisation du contrôle de fermeture : les opérations d'inclusions et d'intersection pénalisent le processeur, de ce fait on peut faire recours à la technique diffset de Zaki pour simplifier le calcul.

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 International Conference 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 optimisée de distribution intégrale de la fouille de données basée sur une infrastructure CORBA, ASD (2012)

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)

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 integration issues of a CORBA based parallel distributed architecture into Cloud model to process large datasets easily and transparently. 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. In previous work, we discussed the parallel distribution of the data mining process focusing more on the tasks aspect, in this paper, the integral parallel distributed infrastructure used to implement the Cloud Computing data mining model. Both the distribution of the data mining process and the memory component are discussed in this paper. Our focus is the projection of our CORBA based architecture to a *MapReduce* architecture.

Using Data Mining Techniques for Representing a Course Structure as Weighted Conceptual Maps

Mohammad AlSarem[1, 2], Mostafa Bellafkih[2], and Mohammad Ramdani[1]

[1]Faculty of Sciences and Technique, Mohammadia, Morocco
[2]National Institute of Posts and Telecommunications, Rabat, Morocco

Abstract. With rapid development of the web technologies, many scholars have attempted to adopt e-learning systems to support students during their learning activities. One of tasks that need to solve is representing domain model in an understandable form. This study presents an innovation approach that is used to represent course structure. The conceptual map was selected as a tool for representing domain model. To construct such map, the prerequisite relationships have to mine. Thus, the proposed approach uses data mining techniques namely association rules to extract the relationships among concepts from the grades of learners that are collected during testing process. Furthermore, to show how a concept in such model, affects on other concept the degree of correlation between domain concepts was calculated. As the result, structure of a course will be represented as weighted conceptual map which in its role can be used as a tool for diagnosis learning problem or to provide guidance during learning process.

1 Introduction

With rapid development of the web technologies, many scholars have attempted to adopt e-learning systems to support students during their learning activities. One task that needs to solve is representing domain model in an understandable form. Analysis e-learning systems shows that these systems have limitations in efficiency and affectivity to some extent: guidance learners during their activities depend on the predefined domain model whereas manually constructing this model requires an expert knowledge of the domain and substantial amount of time and effort. In fact, this represents a bottleneck of most adaptive web-based educational systems (Mihál, and Bieliková, 2011). Furthermore, in the constructed domain model often not clear how a concept affects on other. Therefore it is helpful to automate creation of the domain model; in addition, it will be useful if we have an approach to measure how a concept affects on other. In this paper, we present an approach to facilitate process of representation of a course structure. The weighted conceptual maps were selected as a tool for representing domain model. Thus, to aim this goal, the following strategies are using:

1. Using data mining techniques namely fuzzy association rules to mine the prerequisites relationships among concepts from the testing results;
2. Calculating the degree of correlation among concepts to detect how concepts affect on each other;

The remainder of the paper is organized as follows. Section 2 introduces understanding of domain modeling, description of the proposed structure, and tools used to represent it. The

proposed approach of constructing relationships among concepts is given in Section 3. Further analysis of constructed map is introduced in Section 4. Section 5 concludes the whole paper.

2 Understanding of domain modeling

2.1 Domain Model

Generally, domain modeling can be defined as "the process of capturing and structuring knowledge embedded within a selected domain" (Clark et al, 2011). In the literature, numerous terms such as Concept Map, Conceptual Graph, Semantic Network, Knowledge map, Ontology, Association Graph, etc, are all used to refer to the concept of a domain model. Our proposed model is defined as a set of learning concepts of a course and relations between them.

In real educational course, a course structure is consisted of several units. The course unit also could be divided into several sub-units or concepts which has hierarchical structure. In order to aim this structure, we have devised a hierarchical structure of domain knowledge into two different levels as shown in Figure 1:

1. The lower layer which includes teaching concepts and materials;
2. The Upper layer which represents hierarchy of a course units/chapters, where a course with ***n*** lessons/topics can be associated with ***n*** sub-topic/ sub-unit.

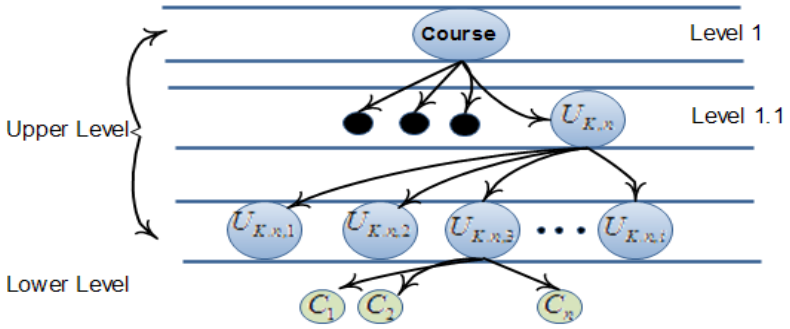


FIG. 1- *An explanation of a course structure*

However, mentioned above structure does not show the relationship among concepts in lower layers and also the cross relationships among chapters/units. To cope with these problems, a concept mapping technique is used to represent domain model.

According to Gupta (2000) a concept mapping technique is a well-defined step to represent the course unites and to develop the domain knowledge. Furthermore, to determine how concept affects on other concept, the weighted conceptual maps (Jong et al., 2004; Novak, 1998) are used.

2.2 Description of the proposed domain model structure

Our domain model consists of the following components:

- ***Concept,*** which presents an actual learning focus of each learning activity. It may also represent an abstract information item from the application domain. Concepts can be either an atomic (primitive) concept that represents single fragment of information or a composite concept that has a sequence of children concepts (sub-concepts). The children of a composite concept are either atomic concepts or composite concepts. In order to minimize author's effort in designing process, we distinguish two classes of domain concepts:
  - ***Internal concepts***, an internal concept can be considered as a container of learning activities having the same learning focus. By other words it contains a set of educational resources having the same learning activity.
  - ***External concepts,*** which have not a directly pedagogical meaning, but have to link to internal concepts, like, assessment tests, quizzes that are used for measuring learners' understanding.
- ***Concept relationship,*** the relationship presents how a concept link to another. Here, we also distinguish between three types of relationships:
  - ***Relationships among internal concepts,*** this type of relationship "prerequisite" meaning, where understanding a concept C_i affects on understanding concept C_j.
  - ***Relationships among external concepts*** can be either "belong to" or "consist of", where a group of concept belong to another one.
  - ***Relationships among external and internal concepts,*** here, we distinguish between relations that are link educational resources to internal concepts and those linking to external concepts. The former relationship type has a "belong to", whilst the latter has an "assessed by" relationship type

To determine how concept affects on other concept, each arrow linking concepts is labeled with a weighted value. Thus, our domain model structure takes form of weighted concept maps.

Definition *The weighted conceptual map is a direct graph* $G = (C, E)$ *,where* $C = \{c_0, c_1, ..., c_n\}$ *is set of vertices,* c_i *- represents concepts,* $E = \{e_0, e_1, ..., e_n\}$ *are set of edges,* e_i *- represents relationship among concepts and all of* e_i *- are signed a conceptual weight* w_{ij} *whose value reveals the effect of concept* C_i *respect to* C_j.

3 Methodology of constructing conceptual maps of a course

The study proposed to use the fuzzy association rules algorithm namely a look ahead fuzzy association rules algorithm ((LFMAlg) (Tseng et al., 2007) to mine the hidden relationships among concepts using the test portfolio of each learner. Constructing conceptual maps of a course includes the following procedures:

3.1 Determining all occurring concepts in test questions and presetting their conceptual weightings by teachers

To minimize complexity of constructing process, we ask experts of domain (teachers/educators) to determine all concepts that can be occurred in test question.

Usually, a test question corresponds to one concept only, but there are situations when a test question may at the same time include two or more than two concepts (this situations can be appeared for concepts that belong to the same unit and also for those units that contain related concepts). Therefore to calculate degree of correlation of concepts, firstly, teacher has to preset the relevance of concepts to questions. For a test question that contains a single concept, the relevance degree will be represented by ‘‘1”. If a test question does not contain any concept, it will be represented by ‘‘0”. If a test question contains more than one concept, the conceptual weight (0–1) distributed to different test questions is presented by the degrees of strong, medium and weak. Using that, the questions-concepts matrix will presented as follows:

$$QC = \begin{array}{c} \\ Q_1 \\ Q_2 \\ \vdots \\ Q_m \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \dots & C_p \end{array} \\ \begin{bmatrix} \mathrm{qc}_{11} & \mathrm{qc}_{12} & \dots & \mathrm{qc}_{1p} \\ \mathrm{qc}_{21} & \mathrm{qc}_{22} & \cdots & \mathrm{qc}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{qc}_{m1} & \mathrm{qc}_{m2} & \dots & \mathrm{qc}_{mp} \end{bmatrix}_{p \times m} \end{array} , where$$

"qc_{ij}" - denotes the weighting or degree of relevance for each concept "C_i” to each question "Q_j".

3.2 Transferring the test portfolio of learners to grades matrix

As learners took a test and their test portfolio transformed into grades matrix, to mine the prerequisite relationships from the numeric testing grades, fuzzy set theory can be useful to transform learners’ grades into symbolic (Al-Sarem et al., 2011a). If the membership functions of each quiz’s grade are known (see Figure 2) and the grades matrix are fuzzificated, then the Look Ahead Fuzzy Association Rule Mining Algorithm (LFMAlg) can be applied to mine rules types “$\mathrm{Q_i, L \rightarrow Q_j, L}$”, “$\mathrm{Q_i, H \rightarrow Q_j, L}$” and “$\mathrm{Q_i, H \rightarrow Q_j, H}$”.

In the fuzzification result, “LOW”, “MIDDLE” and “HIGH” denote “a learner "$\mathrm{S_i}$" has a low grade in question “$\mathrm{Q_j}$”, a learner "$\mathrm{S_i}$" has a middle grade in question $\mathrm{Q_j}$ and a learner "$\mathrm{S_i}$" has a high grade in question $\mathrm{Q_j}$ respectively.

$$G = \begin{array}{c} \\ \mathrm{Q}_1 \\ \mathrm{Q}_2 \\ \vdots \\ \mathrm{Q}_m \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{S}_1 & \mathrm{S}_2 & \dots & \mathrm{S}_n \end{array} \\ \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1n} \\ g_{21} & g_{22} & \cdots & g_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ g_{m1} & g_{m2} & \cdots & g_{mn} \end{bmatrix} \end{array} , where$$

g_{ij} −denotes the score of question $Q_j$of the learner S_i, $g_{ij} \in [0, M]$, and "M" is the maximum value that can obtain by learners.

Later, we will show how the confidences of mind rules use to calculate degree of correlation between concepts.

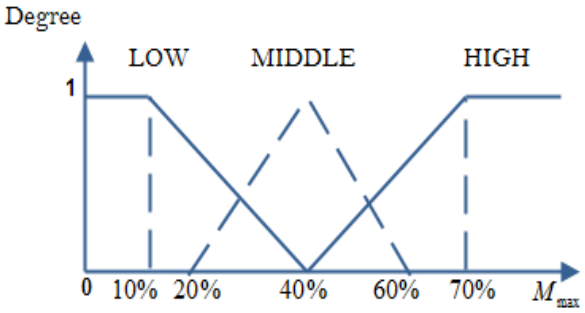


FIG. 2- *The given membership functions of each quiz's grade*

Rule	Description of relationships
$Q_i, L \rightarrow Q_j, L$	It is means that the concepts in question " Q_i" are the prerequisite of those in "Q_j" and explain why getting low grade in question "Q_j" might imply getting low grade on " Q_i"
$Q_i, L \rightarrow Q_j, H$	It is means that the concepts in question "Qj" are the prerequisite of those in"Qi" because "Qi" may be not learned well resulting from "Qj"
$Q_i, H \rightarrow Q_j, L$	It is means that the concepts in question "Qi" are the prerequisite of concepts in "Qj"
$Q_i, H \rightarrow Q_j, H$	It is means that the concepts in question "Qi" are the prerequisite of concepts in "Qj"

TAB. 1- *The explanations of rule types (Al-Sarem et al., 2011a)*

3.3 Calculating degree of correlation among concepts

Armed with the confidence of mind rules by LFMAlg algorithm and the conceptual weight relationships in questions-concepts matrix, we can calculate degree of correlation among concepts as follows:

$$w(c_i \rightarrow c_j) = Max\left(qc_{Q_x c_i} \times qc_{Q_y c_j} \times conf(Q_x \rightarrow Q_y)\right) \quad (1)$$

Where "$w(c_i \rightarrow c)$" denotes the degree of correlation of the relationship "$C_i \rightarrow C_j$" , "C_i" denotes a concept appearing in the question "Q_x", "C_j" denotes a concept appearing in the question "Q_y" ," $qc_{Q_x c_i}$" denotes the weight of the concept "C_i" in the question

"Q_x","$qc_{Q_yC_j}$" denotes the weight of the concept "C_j" in the question "Q_y" (both $qc_{Q_xC_i}$ and $qc_{Q_yC_j}$ can be obtained from QC matrix).

From pedagogical view of point, Eq. (1) can be interpreted this way: the confidence level of test question association rules $conf(Q_x \rightarrow Q_y)$ is the concept of conditional probability: In one hand, it implies that under the condition that a learner obtains, e.g., a low score g_i for answering Question Q_x , there is a probability for the learner to obtain a low score g_j for answering Question Q_y, too.

On other hand, concept in question Q_x is more probably the prerequisite knowledge of those in Q_y (see Table1).The Cartesian product of sets $qc_{Q_xC_i}$ and $qc_{Q_yC_j}$ produces n possible ordered pairs with relevance strength of the priority. Therefore, we propose to choose the largest relevance degree to be the degree of correlation between concept c_i and concept c_j. For example (Al-Sarem et al., 2011a), let the conceptual weights of concept C_i respect to test questions are $qc_{Q_xC_i} = \{1,0.2,0,0,0.6\}$, and the conceptual weights of concept C_j respect to test questions are $qc_{Q_yC_j} = \{0,0.7,0,1,0\}$. Let also the confidence of mind rules are$conf(Q_1 \rightarrow Q_4) = 0{,}83$,$conf(Q_2 \rightarrow Q_4) = 1$, $conf(Q_5 \rightarrow Q_2) = 0{,}92$, $conf(Q_5 \rightarrow Q_4) = 0{,}89$ and $conf(Q_x \rightarrow Q_y) = 0$ for all other rules, then the degree of correlation of the relationship $C_1 \rightarrow C_2$ is $w(c_1 \rightarrow c_2) = \text{Max}(0.83,0.2,0.386,0.534) = 0.83$.

3.4 Constructing the weighted conceptual maps

As mentioned above, the concepts map is selected as a tool for organizing and representing domain model. Thus, after converting the test question association rules into the relationship between concept and concept, concept will represent as circle and arrow direction of each line represents the priority order of learning concepts as shown in Figure 3.

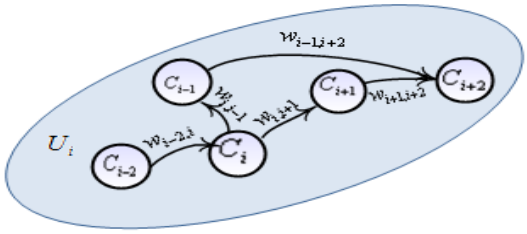


FIG. 3- *Representation concepts of a unit as weighted conceptual map*

4 Description of Experiment

To validate our approach, we applied it using data gathered from the midterm exam of "OOP java course" given to 74 in the third level of faculty of Science in Rabat. The students are divided into four groups with 19 students in each group.

Subject	Information
Education Degree	3ed level of faculty of sciences Rabat
Course Name	Object Oriented Concepts of Java Course

Number of groups	4
Number of Students	74
Average number of students in groups $\bar{\mathbf{x}}$	19
The maximum grade that can be obtained in each item	4
Average Score of Exam	7.9043783
Standard Deviation of Scores	6.683
Number of Test Items	5
Number of Concepts	15

TAB. 2 – *The Related Statistics of Testing Results in Physics Course.*

The exam was given on 7th of January 2011. It consisted of five questions that examined 15 concepts of object oriented Java programming course. In the first questionQ_1, we ask students to write a class and inherit from it two different child classes. This question mainly examines the students' understanding of inheritance structure, and it notes how they pass parameters to the constructors and methods in parent and child classes.

The questions $Q_2$and Q_3cover most of the course concepts. They require the student to write and design complete code. The last two questions $Q_4$and Q_5examine how students would declare class with its behavior and structure. The related concepts of testing paper and Question-concept mapping matrix are shown in Tables 2 and Table 3 respectively.

Concept	Learning Concept	Concept	Learning Concept
C_1	Variable declaration	C_8	Overloading method
C_2	Method declaration	C_9	Encapsulation
C_3	Reusability	C_{10}	Class declaration
C_4	Class structure	C_{11}	Overriding
C_5	Setter and getter method	C_{12}	Object declaration
C_6	Modifier	C_{13}	Inheritance relationship
C_7	Constructor	C_{14}	Aggregation relationship

TAB. 3 – *Concepts List of Testing Paper*

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}
Q_1	0	0	0	0	0	0	0.385	0	0	0	0	0	0.385	0	0.231
Q_2	0.1389	0	0.3335	0.1113	0	0.083	0.139	0.028	0.0556	0	0.028	0	0.083	0	0
Q_3	0	0.158	0	0	0.079	0	0.158	0	0.079	0.342	0.079	0	0.105	0	0
Q_4	0.15	0	0	0	0.3	0	0.25	0	0	0	0	0	0	0.3	0
Q_5	0	0	0	0	0	0	0	0	0	0	0	0.571	0	0	0.429

TAB. 4 - *The Question-Concept Mapping Matrix*

Analysis and Results

After applying our approach with $minsup = 20\%$ and $minconf = 80\%$, the complated wieghted concept maps will be represented as shown in Figure 4.

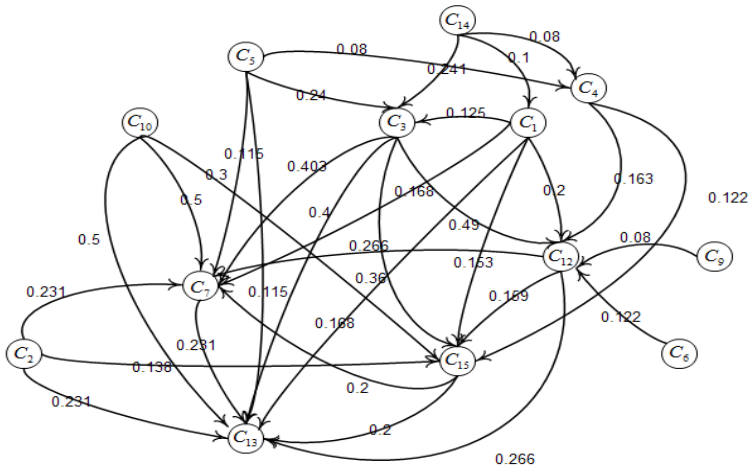


FIG. 4 --*The Constructed Weighted Concept Maps*

Analyzing the above WCMs, we easily can see absents concept C_8 and C_{11} from the final maps. This can be explained regarding the QC-matrix as conceptual weightings of C_8 and C_{11} are very small to influence strongly on calculation the relevance degree of relationships. Hence, in one hand, authors (teachers or course designers) have to write more than one test to get good WCMs. On other hand, we can either update the weightings of the completed WCMs or add new concepts to the existing maps each time new evidence is gotten. Because of this, in the next section, we present an approach to update and merge the constructed WCMs.

5 Conclusions

Constructing knowledge model is a fundamental pillar of any adaptive learning system. However, representing course structure in understandable form is quite time consuming. Therefore, in this study, we present an approach to represent structure of a course as weighted conceptual maps using data mining techniques.

Related to educational technology areas, we found that, the proposed approach can be helpful for the following reasons:

- It represents domain model as weighted concept maps.
- The direction of arrow between concepts reflects the sequence order of concepts
- The weight value labeled with link between concepts C_i and C_j shows how a concept affects on the other.

In addition, the proposed approach can be applied to most computer-based courses which contain explicit <test item, concept> relationships.

Generally, we think that, this study is significant since it provides e-learning curriculum developers with a new algorithm for designing a course management system, moreover for representing a course in understandable form which in its role can be used in diagnostic of

the learning problems and providing learning suggestions for individual student (Al-Sarem et al.,2011b).

References:

V. Mihál, and M. Bieliková (2011). Domain Model Relations Discovering in Educational Texts based on User Created Annotations. *14th International Conference on Interactive Collaborative Learning (ICL2011)- 2011 14th International Conference on. IEEE*, 542-547.

M. Clark, Y. Kim, U. Kruschwitz, D. Songa, D. Albakour, S. Dignum, U. C. Beresi, M. Fasli, and A. De Roeck (2011). Automatically Structuring Domain Knowledge from Text: an Overview of Current Research. *Information Processing & Management 48.3 (2012): 552-568.*

Gupta, K. C. (2000). Concept Maps and Modules for Microwave Education. IEEE Microwave Magazine, 1(3), 56-63.

B. S. Jong, T. W. Lin, Y. L. Wu and T. Chan (2004). Diagnostic and remedial learning strategy based on conceptual graphs. *Journal of Computer Assisted Learning* 20, 377–386.

J. D. Novak (1998), Learning, creating, and using knowledge: Concept maps as facilitative tools in schools and corporations. In Lawrence Erlbaum Associates.

S.S. Tseng, .S. P.C. Sue, J.M. Su, J.F. Weng, and W.N. Tsai (2007), "A new approach for constructing the concept map", *Computers & Education* 49, 691–707

M., Al-Sarem, M.,Bellafkih, M., Ramdani (2011a), "Mining Concepts' Relationship Based on Numeric Grades.", IJCSI I*nternational Journal of Computer Science Issues* 8, 4(2), 136-142.

M., Al-Sarem, M.,Bellafkih, M., Ramdani (2011b), "Improving Learning Performance by Providing Intellectual Suggestions.", *International Journal of Computer Applications* 31(4), 38-42.

Web Usage Mining : Prétraitement des traces pour une analyse multi-vue sur Moodle

Nawal Sael*, Abdelaziz Marzak *
Hicham Behja**
* Faculté des Sciences Ben M'sik, Casablanca, Maroc
saelnawal@hotmail.com
** Ecole Nationale Supérieure d'Arts et Métiers, Meknès, Maroc
h_behja@yahoo.com

Résumé.
Ce travail de recherche vise l'application des techniques du Web Usage Mining sur les environnements e-learning.
Dans des travaux antérieurs, nous avons proposé une nouvelle méthodologie de prétraitement des fichiers de logs de la plate-forme Moodle, en se basant sur la structure du contenu pédagogique respectant le standard SCORM. Ensuite, nous avons développé un outil de prétraitement afin d'automatiser cette phase et de présenter les premiers résultats obtenus.
Dans cet article, nous appliquons une analyse statistique et des techniques de visualisation plus approfondies. Nous définissons de nouvelles variables selon l'arborescence de contenu SCORM et nous présentons des graphiques multi-dimensionnelles qui permettent de mieux comprendre le déroulement de l'apprentissage. Ces variables agrégées fournissent de la connaissance utile sur l'interaction des apprenants avec le contenu pédagogique.

1 Introduction

Au cours de ces dernières années, de nombreuses recherches ont été menées sur l'utilisation des TIC (Technologie de l'Information et de Communication) dans l'université marocaine dans le but d'intégrer l'e-learning dans la formation des étudiants. Pour y contribuer, nous avons mené une expérience d'intégration d'une plate-forme en e-learning au sein de l'ENSAM Meknès, afin de permettre aux étudiants de suivre les cours en-ligne et de bénéficier des apports des technologies de l'information et de communication.

Cependant, dans les environnements classiques d'enseignement, les enseignants sont en mesure d'avoir de l'information sur le déroulement des sessions de formation et l'appréciation des apprenants sur le contenu pédagogique offert (De Ketele et Roegiers, 2009), ce qui permet l'évaluation des programmes d'enseignement (Sheard et al, 2003). Contrairement aux environnements électroniques d'apprentissage, qui manquent d'interaction enseignant-apprenant (Zorrilla et al, 2005) et par conséquent, manquent d'information sur l'interaction des apprenants avec les cours et sur leur suivi, ainsi que d'outils appropriés pour superviser les activités des apprenants (Hijon et Velazquez, 2006).

L'accompagnement du processus d'apprentissage, le suivi et l'évaluation des sessions de formation sont devenus des tâches délicates pour un enseignant ou un tuteur. Ils se trouvent dans l'obligation de chercher d'autres moyens pour la prise de décision. Aider ces enseignants/tuteurs à comprendre le déroulement des apprentissages demeure une nécessité voire, une condition pour la réussite de ce mode de formation.

L'utilisation des technologies de l'information comme support de réalisation des environnements e-learning permet d'enregistrer la trace de toutes les actions utilisateurs, d'avoir de l'information sur le déroulement de l'apprentissage ainsi que sur les profils des utilisateurs (Zorrilla et al, 2007) et (Romero et Ventura, 2007). Quoique, ces informations sont abondantes, non structurées et nécessitent un traitement particulier.

Dans ce travail, nous nous intéressons à l'application des techniques du Web Usage Mining aux plates-formes e-learning afin d'assister l'enseignant/tuteur lors de la prise de décision sur l'efficacité et la pertinence des cours offerts, pour lui offrir de la connaissance sur le déroulement des apprentissages ainsi qu'aider les apprenants à mieux suivre les formations disponibles. Nous nous focalisons sur le prétraitement de logs de la plate-forme Moodle.
La section suivante présente le contexte de travail et l'état de l'art. Nous présentons ensuite notre approche de prétraitement des logs de Moodle. Puis l'expérimentation effectuée et les résultats obtenus. La conclusion situe le travail en cours et évoque les perspectives de recherche.

2 Le Data Mining en Education (EDM)

2.1 Web Mining

Le web mining s'intéresse à l'application des techniques du data mining aux données de navigation sur le web (Cooley, 2003) et (Bing, 2007). Ses techniques couvrent trois disciplines qui se complémentent:

- le WCM : Web Content Mining ou web text mining s'intéresse à l'analyse du contenu des pages web,
- le WSM : Web Structure Mining s'intéresse à la structure d'un site web,
- le WUM : Web Usage Méninge(WUM) est un processus complexe, utilisé pour extraire de la connaissance sur la caractérisation des internautes fréquentant un site Web, et l'identification de leurs motifs de navigation. Il est constitué de trois phases qui se complètent à savoir : le prétraitement, la fouille de données et le post-traitement ou l'analyse des résultats de la fouille

2.2 Data Mining en Education, Educational Data Mining (EDM)

Dans son site web officiel, la communauté internationale de l'EDM [1] (fouille de données éducatives) le définit comme : une discipline émergente, qui s'occupe de développer des méthodes, des techniques et des outils permettant d'explorer les données provenant des milieux éducatifs, pour mieux comprendre les apprenants, leur perception de l'environnement de formation ainsi que leur interaction avec ces environnements.

Baker et Yacef (2009), Baker (2010) estiment que l'EDM s'intéresse principalement à : améliorer le mode et l'environnement d'apprentissage, étudier les formes de soutien pédagogique fournis par les outils d'apprentissage et développer la recherche scientifique vis-à-vis de l'apprentissage et des apprenants. Néanmoins, l'évolution des techniques de data mining en éducation est relativement plus lente que celle en e-commerce. Bien que, cette situation

commence à changer, dans le sens où il y a actuellement un intérêt très croissant pour l'application de ces techniques (Romero et al, 2010).

2.3 État de l'art sur l'EDM

De nombreuses études ont utilisé les techniques d'EDM pour fournir aux tuteurs, aux enseignants et aux propriétaires des plates-formes, des connaissances utiles sur l'interaction des apprenants avec les cours et les formations offertes (Romero et al, 2010). Dans ce qui suit, nous présentons brièvement une analyse de quelques uns de ces travaux.

Dans Machado et Becker (2003) les auteurs ont étudié le potentiel de l'EDM en tant qu'outil de validation de la structure et de la conception d'un environnement d'apprentissage en-ligne. Leur objectif était d'exploiter le WUM pour suggérer des modèles qui peuvent contribuer à l'évaluation de l'utilisation de cet environnement. Marquardt et al (2004) ont développé un outil de prétraitement pour un environnement e-learning. Il permet d'automatiser les tâches de sélection, du nettoyage, de transformation et d'enrichissement généralement effectuées dans la phase de prétraitement. Merceron et Yacef (2005), Zorrilla et al (2007), García et al (2011) ont tenté de surmonter les handicapes liés à l'absence d'un feedback direct entre les enseignants et les apprenants, en offrant aux enseignants/tuteurs des outils appropriés pour suivre et évaluer le progrès de ces apprenants ainsi que leurs processus d'apprentissage.

D'autres travaux utilisent ces techniques pour personnaliser l'accès aux cours (Khribi et al, 2008) (Marquardt et al, 2004) et (Romero et al, 2009). Sfenrianto et al 2012 proposent un système de recommandation aux apprenants en fonction de leur historique de navigation.

Romero et Ventura (2007) ont conduit une large revue de la littérature sur l'EDM entre 1995 et 2005. Ils offrent une vision globale sur le domaine et examinent 81 articles. Ce travail a été étendu en 2009 lorsqu'ils ont mené une étude qui traite environ 306 articles et travaux de recherche, elle regroupe ces travaux en 11 catégories selon leurs objectifs.

Backer (2010) regroupe les méthodes et techniques utilisées dans ce domaine(EDM) suivant la catégorisation suivante: distillation of data for human judgment (statistique, visualisation), prédiction (classification, régression, estimation de densité), clustering, relation data mining (règles d'association, corrélation data mining, extraction de motifs séquentiels, causal data mining) et la découverte de modèle.

2.4 Le prétraitement en EDM

Le prétraitement est une étape fastidieuse dans le processus du web usage mining. De ce fait, de nombreux travaux ont proposé des méthodologies de prétraitement surtout pour le domaine du e-commerce, voir (Cooley, 2003), (Tanasa et Trousse, 2004) et (Sael et al, 2009).

En e-learning, la phase de prétraitement n'a pas reçue suffisamment d'efforts d'analyse et peu de travaux l'ont ciblé dans leurs axes de recherches. De plus, la majorité des recherches sur l'application des techniques de data mining en e-learning, se basent sur les méthodologies de prétraitement des fichiers de logs proposées dans le domaine du e-commerce. Cependant, le contexte du e-learning est très particulier et diffère de celui du e-commerce, qu'il soit au niveau de la structure, de la nature du contenu, des intervenants ainsi que de l'objectif d'analyse.

Dans ce qui suit, nous allons discuter plusieurs travaux qui traitent cette phase en e-learning et discutent ses particularités.

Koutri et al (2005) proposent d'adapter la méthodologie de prétraitement initiée par Cooley et al (2003). Alors que Marquardt et al (2004) ajoutent d'autres contraintes à cette méthodologie pour l'adapter à ce domaine : (le login pour identifier l'utilisateur, la session est l'ensemble des clics utilisateurs lui permettant de réaliser une activité donnée, les transactions ou les épisodes sont identifiés à travers la classification des pages web en « page de contenu, page auxiliaire et page de ressource »). Cependant, cette classification ne paraît pas suffisante, surtout avec la diversité des contenus offerts et la création de nouveaux cours à chaque fois.

Pour Zorrilla et al (2005) une nouvelle session commence à chaque fois que l'apprenant change de cours ou s'il fait une rupture d'utilisation du site de plus de 30 mn. Néanmoins, dans un même cours, l'apprenant peut changer d'activité.

Ba-omar et al (2007) ont adopté un intervalle maximum de 30 minutes entre deux sessions.

Tenant compte de ces différentes recherches, nous concluons qu'il reste encore des particularités dans ce domaine qui ne sont pas exploitées. En réalité, ces étapes de prétraitement permettent d'analyser la réalisation des activités d'un cours donné, mais n'offrent pas d'information sur le déroulement d'une activité particulière et spécialement par rapport au suivi du contenu pédagogique sous format SCORM. Cet aspect sera abordé dans nos travaux de recherche exposés dans la suite.

3 Notre Approche pour le Web Usage Mining

Alors que les travaux réalisés sur le prétraitement des données se limitent à l'analysent des accès soit aux cours globalement, ou à certaines activités telles que le chat et le forum, Nous nous intéressons au prétraitement des informations générées lors de l'interaction des apprenants avec les cours, de la réalisation des activités et du suivi du contenu pédagogique sous format SCORM.

Nous proposons de profiter des apports due ce standard SCORM en tant que support de création du contenu pédagogique. Les principales contributions de ce travail sont :

1. Prétraitement des traces d'accès au contenu pédagogique : La première contribution de ce travail de recherche est la proposition d'une méthodologie de prétraitement des traces d'usage de la plate forme Moodle en général et du contenu pédagogique SCORM d'une façon plus approfondie. Son objectif est de nettoyer et préparer ces traces et les transformer en une BD relationnelle. En plus, nous proposons de redéfinir certains éléments de la terminologie des variables généralement utilisées dans cette phase. Afin d'avoir plus de détail sur les actions utilisateurs ainsi qu'une analyse enrichie de ces traces.
2. Statistique et visualisation : La deuxième contribution de cette recherche pour le WUM concerne l'analyse statistique et la visualisation dans le but d'offrir aux différents utilisateurs (enseignants, tuteurs et apprenants) des informations sur le déroulement des apprentissages via le calcul de variables et l'élaboration de graphique permettant de décrire les accès des apprenants aux contenu pédagogique SCROM et d'avoir une analyse préliminaire de l'avancement de l'apprentissage par rapport aux différentes parties de ce contenu.

3. Data mining : la troisième contribution concerne l'application du clustering et des règles d'association pour obtenir des modèles qui seront évalués et analysés de façon à offrir aux enseignants/tuteurs et aux apprenants des recommandations sur le processus d'apprentissage.

3.1 Analyse des logs de Moodle

Mdl_log est une Table non structurée qui enregistre toute action utilisateur sur Moodle. L'analyse des informations disponibles sur cette table ainsi que des autres tables qui décrivent les cours et les activités, nous permet de dire qu'une action sur Moodle est pour un utilisateur particulier (login), dans un cours donné et une activité ou une ressource spécifique.

Si l'action de l'utilisateur se fait sur un contenu SCORM, elle est faite sur une partie définie de ce contenu. La table mdl_scorm_scoes permet de décrire le contenu pédagogique offert en utilisant l'arborescence du contenu SCORM. Cette arborescence décrit l'organisation de ce contenu en différents niveaux: chapitres, séances ou séquences du cours.

Le système proposé aidera à déterminer les sessions, les visites, les activités et les épisodes parmi les traces enregistrées dans mdl_log pour une période déterminée. Si l'activité est le contenu SCORM, les épisodes sont détectés selon les niveaux de l'organisation de ce contenu. Ainsi, la définition de l'épisode peut être l'accès à un chapitre, à une séance ou à une séquence. Sael et al (2010) offrent plus de détail sur la conception de ce système.

3.2 Terminologie

Dans sael et al (2010), nous avons proposé une restructuration de la table Moodle mdl_log en se basant sur la terminologie spécifiée ci-dessous :

- Utilisateur : Login, défini dans la table user et localisé par son id.
- Session : l'ensemble des accès utilisateur entre une connexion (login) et une déconnexion (logout)
- Visite : pour une session particulière, en fait une visite est l'ensemble des accès successifs à un même cours.
- Activité : décrit les accès successifs à une activité ou à une ressource spécifique.
- Episode : dans le cas où l'utilisateur accède à un contenu pédagogique sous format SCORM, nous proposons une définition de l'épisode qui peut être l'accès à un chapitre du cours, à une séance dans ce chapitre ou à une séquence particulière de cette séance.

3.3 Base de données de prétraitement des traces

Le système proposé, permet le prétraitement des traces d'utilisation de la plate-forme Moodle et leur migration vers une base de données relationnelle qui permet de structurer la table mdl_log tout en lui ajoutant les autres tables définies ultérieurement (session, visite, activité et épisode) ainsi que d'autres tables qui décrivent le cours cible et les activités qui y sont rattachées. Ces objets sont liés directement à l'action de l'utilisateur et trace son interaction avec le contenu pédagogique.

3.4 Collecte des données

Dans ce travail, nous avons recueilli les traces des apprenants à partir de la plate forme FOAD-ENSAM. Les sources de données recueillies étaient les journaux de la plate forme concernant le cours UML qui y' a était suivi par des étudiants de la licence professionnelle Java/C++ de la faculté des sciences ben m'sik. Ce cours respecte la structure suivante : Chapitre → Séance → séquences → Ateliers, et contient 4 chapitres, 14 séances et 42 séquences de travaux.

Il contient également une évaluation à la fin de chaque chapitre ainsi qu'une étude de cas à la fin du cours. Les étudiants ont suivi ce cours en ligne durant les mois de Mai et Juin pendant plus de 4 semaines.

4 Statistique et visualisation

Le prétraitement multi-niveaux des traces d'interactions des apprenants avec le cours permet d'avoir plus de détail sur les profils utilisateur et d'offrir aux enseignants des informations leur permettant d'améliorer la conception et l'organisation du contenu offert. Dans ce qui suit, nous proposons d'exploiter les techniques d'analyse statistiques et de la visualisation pour atteindre ces objectifs et proposer aux enseignants, concepteurs et tuteurs des indicateurs statistiques et graphiques pour les aider à observer et à évaluer une session d'apprentissage.

4.1 Analyse Statistique

Les statistiques sont souvent le point de départ de l'évaluation dans un système e-learning (Zaïane et al, 1998). Elles peuvent être extraites à l'aide d'outils standard conçus pour analyser les logs du serveur web comme AccessWatch, analogique, Gwstat et WebStat. Toutefois, il existe des outils statistiques spécifiques aux données provenant du domaine éducatif tel que Synergo / Colat, (Avouris et al, 2005). Cependant, ces outils ne nous permettent pas d'analyser certaines caractéristiques de notre environnement e-learning (contenu SCORM), pour cette raison, nous avons développé notre propre outil. Ces statistiques sont des simples mesures ou variables calculés.

Nous présentons dans ce qui suit, certains variables qui sont liées directement à la session d'un apprenant :

1. Le nombre de visite dans une session.
2. La durée de la session (différence entre la date de la première et de la dernière action dans une session) en secondes.
3. La durée moyenne des actions d'une session utilisateur.
4. Le nombre d'activité visité par session

D'autres variables qui calculent pour une session, le pourcentage d'accès aux différents niveaux du contenu pédagogique SCORM, à savoir :

5. Le pourcentage d'accès aux différents chapitres,

6. Le pourcentage d'accès aux différentes séances d'un chapitre donné,

Ces variables sont des valeurs agrégées obtenues à travers le prétraitement des traces d'accès au contenu pédagogique en exploitant SCORM. Dans ce qui suit, nous présentons des exemples de variables calculées (Fig. 1,2 et 3).

Pour garantir l'anonymat des informations, les utilisateurs présentés sont représentés par (admin).

Admin a passé une durée de 08:08:15 dans la plateforme
Avec une dureé moyenne de 01:37:39 par session.

FIG. 1 – *Exemple de calcul de variable simple, durée globale et moyenne de session utilisateur*

Le pourcentage de visite des chapitres pour l'utilisateur admin par session

======

La session 1
======
Le chapitre 1 : 15
Le chapitre 2 : 76
Le chapitre 3 : 7

FIG. 2 – *Exemple de calcul de variable, % d'accès aux différents chapitres du contenu SCORM pour un apprenant donné*

Nom	Nombre de visite	Durée globale des sessions	moyenne des session	Pace	% visite du chapitre 1	% visite du chapitre 2	% visite du chapitre 3
User 1	29	08:08:15	01:37:39	7.30769	15	76	7
User	20	08:08:15	01:37:39	0.0	7	84	7

FIG. 3 – *Exemple de calcul de multiple variables, pour deux apprenants*

4.2 Visualisation

La visualisation est une discipline de l'informatique ou nous exploitons les apports des interfaces graphiques pour permettre aux utilisateurs (apprenants, enseignant / tuteur, adminis-

trateur…) de comprendre et analyser facilement de grandes quantités d'informations. Elle permet de rendre des informations assez complexes, plus lisibles via les graphiques multidimensionnelles (Romero et al, 2010).

L'analyse des sessions d'apprentissage et du suivi des apprenants du contenu pédagogique proposé, nous permet de comprendre : Comment ce contenu a été utilisé ? Comment les apprenants interagissent avec ce contenu au fil du temps ? Et d'avoir des suggestions sur la façon dont il a été conçu et structuré.

Pour cela, nous donnons sur la Figure 4 la distribution de l'exploitation des différentes parties du contenu pédagogique dans le temps par chapitre et par apprenant. Figure 5 décrit l'exploitation globale du contenu pédagogique par apprenant comparée à celle du chapitre 2.

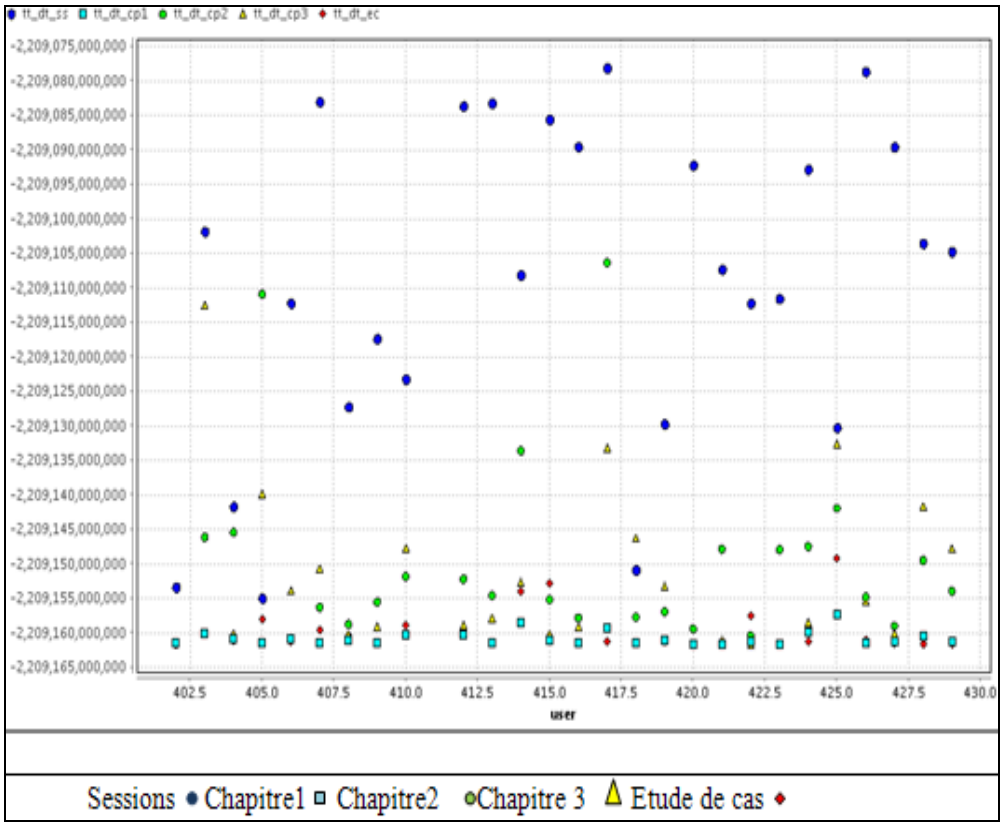


FIG. 4 – Graphique La durée totale d'accès aux différents chapitres

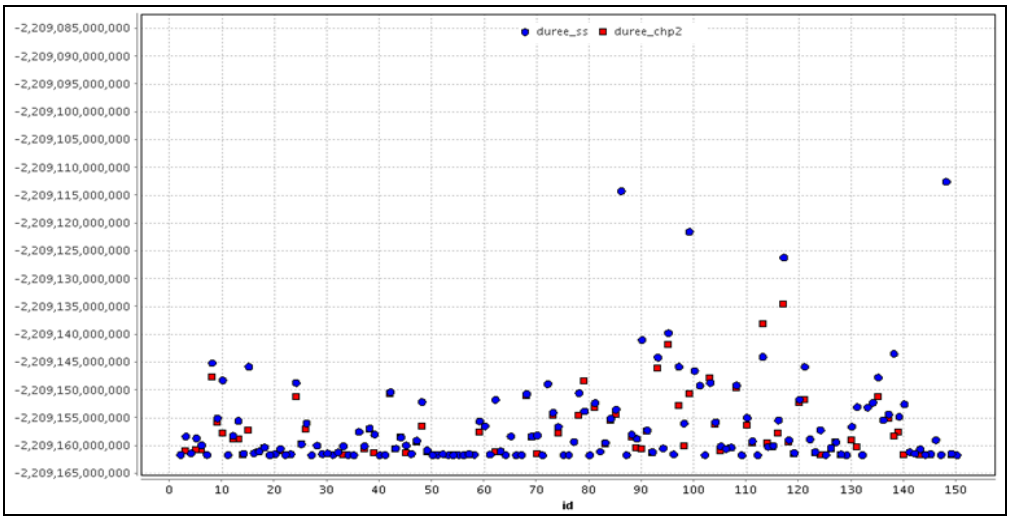


Fig. 5 *Durée global des sessions par rapport à la durée d'accès au deuxième chapitre*

Un phénomène que nous avons remarqué, c'est que : le chapitre 2 est le chapitre qui a pris plus de temps d'enseignement aux apprenants (durée total la plus élevée). Par rapport au chapitre 1 et à l'étude de cas qui n'ont pas été bien suivis ou réalisés par les apprenants.

Le deuxième chapitre est peut être trop long. Une analyse itérative, de la même façon, des interactions des apprenants avec les différentes séances de ce chapitre, pourra nous donner plus de précision sur les parties du contenu qui peuvent éventuellement être plus difficile ou à mieux structurer.

Combinée avec le score des apprenants pendant la réalisation des évaluations liées au contenu de chaque chapitre, nous pouvons valider nos propos sur l'efficacité de la structure actuelle du contenu pédagogique suivi.

5 Conclusion et travaux futurs

Ce travail de recherche vise à aider dans un premier temps les enseignants/tuteurs et de faciliter leurs tâches, en automatisant les étapes de prétraitement et en donnant plus de détails sur le comportement des apprenants et leurs interactions avec le contenu offert.

Dans cet article, nous avons présenté notre travail de recherche sur le WUM pour La plate forme e-learning Moodle FOAD-ENSAM. Il est divisé en quatre étapes, dont deux entre elles ont été exposées tout au long de cet article ainsi que les résultats obtenus.

Dans les étapes prochaines, nous envisageons tout d'abord d'analyser le comportement des apprenants dans le but de mieux comprendre la structure du groupe et analyser le processus de suivi des cours par rapport aux scores finales des apprenants. Pour y arriver, nous projetons d'appliquer le clustering pour analyser la structure du groupe et les règles d'association pour évaluer l'efficacité et l'organisation du contenu pédagogique offert. Ensuite, à la lumière des résultats obtenus, nous proposerons certaines recommandations aux enseignant/tuteurs et aux apprenants.

En perspective à ce travail, nous pouvons généraliser cette recherche pour analyser tous les cours de la plate-forme et offrir un outil de WUM permettant d'analyser d'une façon plus approfondie d'autres activités.

Références

Baker, R. et Yacef, K. (2009) The state of educational data mining in 2009: A review and future visions, J. Educ. DataMining, vol. 1, no. 1, pp. 3–17.

Baker, R. (2010) Data mining for education, in International Encyclopedia of Education, B.McGaw, P. Peterson, and E. Baker, Eds., 3rd ed. Oxford U.K.: Elsevier.

Ba-Omar, H., Petrounias, I., Anwar, F. (2007) A Framework for Using Web Usage Mining to Personalise E-learning. ICALT 2007: 937-938.

Cooley, R. (2003) The Use of Web Structure and Content to Identify Subjectively Interesting Web Usage Patterns. ACM Transactions on Internet Technology, 3(2):93-116.

De Ketele, J.-M. et Roegiers, X., (2009) Méthodologie du recueil d'informations, fondements des méthodes d'observation de questionnaires d'interviews et d'étude de documents.

García, E., Romero, E., Ventura, S. et Castro, C. (2011) A Collaborative Educational Association Rule Mining Tool. The Internet and Higher Education Journal (Special Issue on Web Mining and Higher Education), 14(2), 77-88.

Hijon, R. Velazquez, A. (2006) E-learning platforms analysis and development of students tracking functionality. In Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2006.

Khribi, M.-K., Jemni, M., Nasraoui, O. (2008) Automatic Recommendations for E-Learning Personalization Based on Web Usage Mining Techniques and Information Retrieval. ICALT 2008: 241-245

Koutri, M., Avouris, N., Daskalaki, S. (2005) Chapter 7: A survey on web usage mining techniques for web-based adaptive hypermedia systems , in S. Y. Chen and G. D. Magoulas (ed), Adaptable and Adaptive Hypermedia Systems, IRM Press, pp. 125-149, Hershey.

Machado, L.-D, Becker, K. (2003) Distance Education: A Web Usage Mining Case Study for the Evaluation of Learning Sites. ICALT : 360-361.

Marquardt, C.-G., Becker, K., Dubugras, D., Ruiz, A. (2004) A Pre-Processing Tool for Web Usage Mining in the Distance Education Domain. IDEAS 2004: 78-87

Merceron, A. et Yacef, K. (2005) Educational Data Mining: a Case Study. AIED 467-474.

Romero, C. et Ventura, S. (2007) Educational Data Mining: a Survey from 1995 to 2005. Expert Systems with Applications, 33(1), p.135-146.

Romero, C., et Ventura, S. (2010) Educational data mining: A review of the state of the art. IEEE Transactions on Systems Man and Cybernetics Part C. Applications and Reviews, 40(6), 601-618.

Romero, C., Ventura, S., Zafra, A., De Bra, P. (2009) Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems. Computers & Education 53(3): 828-840.

Sael, N., Marzak, A. et Behja, A. (2009) "Web Usage Mining, Proposition d'une démarche pour le prétraitement des fichiers logs". WOTIC.

Sael, N., Marzak, A. et Behja, A. (2010) "Prétraitement avancé des fichiers logs pour une plate forme d'enseignement à distance" NGN2010.

Sheard, J., Ramakrishnan, S. et Miller, J., (2003) Modelling Learner and Educator Interactions in an Electronic Learning Community, Australian Journal of Educational Technology, 19(2), 211-226.

Sfenrianto, Suhartanto, H., Hasibuan, Z-A. (2012) A dynamic personalization in e-learning process based on triple-factor architecture," Computing Technology and Information Management (ICCM), 2012 8th International Conference on , vol.1, no., pp.69-75, 24-26.

Tanasa, D. et Trousse, B. (2004) Advanced Data Preprocessing for Intersites Web Usage Mining. IEEE Intelligent Systems, 19(2):59,65.

Zaiane, Osmar R., Xin, M., Han, J. (1998) Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs, in Proc. ADL'98 (Advances in Digital Libraries), Santa Barbara,

Zorrilla, M., Mill´an, S., Menasalvas, E. (2005) Data webhouse to support web intelligence in e-learning Environments. In: Proc. of the IEEE International Conference on Granular Computing, Beijing, China.

Zorrilla, M.-E., Mar´ın, D., Alvarez, E. (2007) Towards virtual course evaluation using Web Intelligence. In: Moreno D´ıaz, R., Pichler, F., Quesada Arencibia, A. (eds.) EUROCAST 2007. LNCS,vol. 4739, pp. 392–399. Springer, Heidelberg.

Summary

This research illustrates the use of Web Usage Mining on e-learning domain.

We proposed new data preprocessing methods applied to Moodle logs based on SCORM content structure. Next we developed our preprocessing tool to implement these new methods and here we present the first_hand knowledge gained from it.

In this article, we apply more statistical and visualization techniques and define new statistical variables according to the SCORM content tree. In addition, we present multidimensional graphics in order to understand users' transactions with this e-learning environment. These aggregated variables provide interesting knowledge about student learning process to teachers and tutors.

Planification basée sur la classification par arbre de décision

Sofia Benbelkacem, Baghdad Atmani et Mohamed Benamina

Equipe de recherche SIF « Simulation, Intégration et Fouille de données »
Laboratoire d'Informatique d'Oran - LIO
Université d'Oran
BP 1524, El M'Naouer, Es Senia, 31 000 Oran, Algérie
{sofia.benbelkacem, atmani.baghdad, benamina.mohamed}@gmail.com

Résumé. En Intelligence Artificielle, le mot planification désigne un domaine de recherche qui se propose de concevoir des systèmes pouvant générer automatiquement, au travers d'une procédure formalisée, un résultat défini, sous la forme d'un système intégré de décisions appelé plan. Au lieu d'avoir recours aux algorithmes de planification pour générer des plans, on propose d'exploiter l'apprentissage automatique par arbre de décision pour optimiser le temps.

Dans cet article, nous proposons de générer un modèle de classification par arbre de décision à partir d'un échantillon d'apprentissage contenant des plans auxquels est associé un ensemble de descripteurs dont les valeurs changent en fonction de chaque plan. Ce modèle sera ensuite exploité pour classifier de nouveaux cas en attribuant le plan adéquat.

1 Introduction

La planification suscite actuellement beaucoup d'intérêt étant donné qu'elle combine deux domaines majeurs de l'Intelligence Artificielle, l'exploration et la logique. Le croisement de ces deux domaines a permis d'améliorer les performances au cours des vingt dernières années (Bibai, 2010). Un plan se présente généralement sous la forme d'une collection organisé de descriptions d'opérations (Régnier, 2005).

Généralement, les problèmes de la planification sont résolus à l'aide d'algorithmes de planification. Mais si les algorithmes sont trop longs alors leur exécution peut consommer du temps. Ainsi, au lieu d'utiliser à chaque fois des algorithmes de planification, ce qui pourrait être coûteux en temps de calcul, on propose de faire appel à l'apprentissage automatique, particulièrement l'arbre de décision (Breiman et al., 1984). L'arbre de décision est une méthode issue de la fouille de données (data mining), c'est une structure simple récursive qui permet d'exprimer un processus de classification (Belacel, 1999). Le processus de la classification consiste à affecter une classe à des objets à l'aide d'un modèle entraîné sur un autre ensemble d'objets. Pour ce fait, une correspondance est établie entre un objet décrit par un ensemble de caractéristiques (attributs), et un ensemble de classes disjointes (Belacel, 1999).

Nous proposons d'exploiter le principe de classification par arbre de décision pour la planification. Il s'agit de générer un modèle de classification dont l'utilité est la classification de

nouvelles données. L'idée est d'utiliser l'arbre de décision pour générer un modèle de classification à partir d'un ensemble d'observations ou d'instances. Chaque observation correspond aux valeurs de descripteurs et de classes. La particularité de cette approche est que les classes du modèle sont représentées sous forme de plans.

L'article est organisé comme suit. Dans la section 2, nous citons quelques travaux qui ont impliqué le data mining dans la planification et en particulier les arbres de décision. Ensuite dans la section 3, nous expliquons la démarche que nous avons adoptée qui comporte la génération des plans et la classification par arbre de décision. La section 4 présente quelques résultats de l'expérimentation. Enfin, la section 5 est consacrée à la conclusion du présent travail.

2 Etat de l'art

Nous nous sommes intéressés aux travaux de planification ayant recours aux méthodes de la fouille de données (data mining), particulièrement les arbres de décision. Nous présentons l'état de l'art en deux temps. On commence par des travaux antérieurs qui ont utilisé la planification pour le data mining. Ensuite, nous présentons quelques travaux liés à la planification guidée par arbre de décision.

Kaufman et Michalski (1998) proposent une approche qui consiste en l'intégration de diverses procédures d'apprentissage et d'inférence dans un système permettant de rechercher automatiquement différentes tâches de data mining selon un plan de haut niveau développé par un utilisateur. Ce plan est spécifié dans un langage de production de connaissances, appelé KGL (Knowledge Generation Language).

Kalousis et al. (2008) proposent un système qui combine la planification et le méta-apprentissage pour fournir un appui aux utilisateurs d'un laboratoire virtuel de data mining. L'ajout du méta-apprentissage à la planification fondée sur l'assistant de data mining permet au planificateur de s'adapter aux changements des données et améliorer la décision du planificateur au fil du temps. Le planificateur fondé sur la connaissance s'appuie sur une ontologie de data mining pour planifier le workflow de découverte de connaissances et déterminer l'ensemble des opérateurs valides pour chaque étape de ce workflow.

Zakova et al. (2008) ont proposé une méthodologie qui définit une conceptualisation formelle des types de connaissances et d'algorithmes de data mining ainsi qu'un algorithme de planification qui extrait les contraintes de cette conceptualisation selon les exigences données par l'utilisateur. La tâche de construction du workflow automatique comprend les étapes suivantes : conversion de la tâche de découverte de connaissances en un problème de planification, génération du plan à l'aide d'un algorithme de planification, mémorisation du workflow abstrait généré sous la forme d'une annotation sémantique, instanciation du workflow abstrait avec des configurations spécifiques aux algorithmes nécessaires et stockage du workflow généré.

Fernandez et al. (2009) ont présenté un outil basé sur la planification automatisée qui permet aux utilisateurs, pas nécessairement experts en data mining, d'effectuer des tâches de data mining. L'entrée du système représente une définition des tâches de data mining à réaliser et le résultat est donné sous forme d'un ensemble de plans. Ces plans sont exécutés avec l'outil de data mining WEKA (Witten et Frank, 2005) afin d'obtenir un ensemble de modèles et des statistiques. D'abord, les tâches de data mining sont décrites dans PMML (Predictive Model Markup Language). Ensuite à partir du fichier PMML, une description du problème de plani-

fication est générée dans PDDL (the standard language in the planning community). Enfin, le plan est exécuté dans WEKA (Waikato Environment for Knowledge Analysis).

Miah (2011) présente une étude bibliographique sur l'utilisation des méthodes de data mining pour la planification, plus particulièrement pour la planification de l'évacuation d'urgence. Il fourni également des orientations futures de la recherche.

Crais et Roberts (1991) ont utilisé une série d'arbres de décision pour apporter une aide dans l'évaluation et la planification des interventions auprès de jeunes enfants handicapés. Les arbres de décision sont constitués d'une série de questions d'évaluation conduisant à des suggestions d'intervention.

Wan (1995) a développé une méthodologie de planification pour la conduite d'un jeu de guerre. La méthodologie proposée utilise l'arbre de décision comme un outil analytique pour comparer les plans d'action et trouver la meilleure façon d'accomplir la mission.

Majlender (2003) a représenté des problèmes de planification stratégique par des arbres de décision dynamiques où les nœuds correspondent à des projets dans le but d'aider à l'évaluation des activités d'investissement de plusieurs types. L'analyse de l'investissement fondée sur cette théorie consiste à définir un concept et une méthodologie pour la planification et l'évaluation d'investissements importants.

De la Rosa et al. (2011) ont présenté une approche qui utilise les arbres de décision pour résoudre les problèmes de planification. Cette approche a été implémentée dans un système appelé ROLLER. Cette approche consiste à utiliser les arbres de décision pour sélectionner les actions adéquates dans différents contextes de planification.

Ghoseiri et al. (2012) ont utilisé les arbres de décision dans la planification de la production. Les règles extraites à partir des arbres de décision permettent d'identifier les problèmes des défaillances imprévues dans le programme de production. Cette approche permet aux experts d'enquêter sur les problèmes les plus importants dans le domaine de la production et de proposer des solutions à ces problèmes.

3 Approche proposée

L'objectif de l'approche proposée est double : d'abord, on commence par la construction de l'échantillon d'apprentissage à base de plans ; ensuite, on procède à la classification par arbre de décision.

3.1 Construction de l'échantillon d'apprentissage à base de plans

Un planificateur dispose, en entrée, d'un problème et d'un domaine de planification. Un problème de planification consiste en une description de l'état initial et du but à atteindre. Un domaine de planification est décrit par un ensemble d'actions qui vont permettre des transitions entre les états (Baki et Bouzid, 2006). Une solution au problème de planification est un plan qui permet d'atteindre le but en partant de l'état initial.

On appelle projet l'ensemble des actions à entreprendre afin de répondre à un besoin défini dans des délais fixés. L'organisation et l'enchainement des tâches se présentent généralement sous la forme de tableaux ou de graphes. D'abord, nous décrivons le projet en représentant l'enchaînement des tâches (actions) sous forme d'un tableau afin de générer le graphe ET/OU (Benbelkacem et al., 2012). Un graphe ET/OU est un graphe dont les nœuds représentent des

tâches et les arcs représentent les relations entre les tâches. Une tâche représente l'action réalisée pendant une durée de temps et les relations entre les tâches sont les contraintes à satisfaire (Baki, 2006). Après avoir construit le graphe ET/OU, nous appliquons des algorithmes de planification pour déterminer les plans possibles. Nous utilisons un algorithme de Baki (Baki, 2006) pour générer les plans. Cet algorithme est basé sur un parcours du graphe ET/OU en chaînage arrière. Il consiste à rechercher les chemins possibles entre deux nœuds du graphe ET/OU ; il recherche les chemins qui commencent par un nœud initial et se terminent par un nœud final en utilisant la méthode de recherche en arrière dans le graphe ET/OU. L'algorithme s'arrête lorsque le nœud initial cherché est trouvé. Les plans obtenus à partir de l'algorithme de planification sont tous des chemins du graphe ET/OU qui mènent de l'état initial vers l'état final. Enfin, des descripteurs spécifiques au domaine d'application sont associés aux plans afin de construire l'échantillon d'apprentissage.

Algorithme de planification

Entrées :
G : un graphe ET/OU
$from$: nœud final de G
to : nœud initial de G
$chemin$: le chemin exploré jusqu'à présent
Sorties :
$tousLesChemins$: la liste des chemins recherchés
Début
$tousLesChemins \leftarrow liste\_vide$
$chemin \leftarrow liste\_vide$
RECHERCHECHEMINS$(from, to, chemin, tousLesChemins)$
retourner$(tousLesChemins)$
Fin
procédure RECHERCHECHEMINS$(from, to, chemin, tousLesChemins)$
Ajouter $from$ à $chemin$
si $to = from$ **alors**
ajouter $chemin$ à $tousLesChemins$ et s'arrêter
sinon
pour tout fils prédécesseur($from$) **faire**
$rechercheChemins(fils, to, chemin, tousLesChemins)$
fin pour
Retirer le dernier élément de $chemin$
fin si
fin procédure

3.2 Classification par arbre de décision

L'arbre de décision représente un ensemble de règles pour la classification des données (Belacel, 1999).

Soit $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$ l'échantillon l'apprentissage, c'est l'ensemble d'objets ou de cas qui va être utilisé pour la construction de l'arbre de décision. Chaque cas ω_i est décrit par une série de variables $X_1, X_2, ..., X_p$ dites variables descriptives. A chaque cas ω_i est associé un attribut cible ou une classe notée Y qui prend ses valeurs dans l'ensemble des classes $C = \{c_1, c_2, ..., c_m\}$ (Atmani et Beldjilali, 2007b).

Supposons que l'échantillon d'apprentissage Ω_A issu du domaine Blocksworld [1] comporte plusieurs cas ω_i décrits par trois variables descriptives X_1, X_2, X_3 et auxquels est associée une classe Y qui correspond à un plan.

X_1 : *problem*, représente le nom du problème ;

X_2 : *time*, représente le temps de CPU ;

X_3 : *steps*, représente le nombre d'étapes du plan.

Le tableau 1 illustre quelques cas issu de la base Blocksworld. Dans cet exemple, Y appar-

ω	$X_1(\omega)$	$X_2(\omega)$	$X_3(\omega)$	$Y(\omega)$
ω_1	blocks-4	0.032237	6	P_1
ω_2	blocks-7	0.281196	6	P_5
ω_3	blocks-6	0.147917	10	P_2
ω_4	blocks-5	0.092918	12	P_3
ω_5	blocks-4	0.032703	6	P_1
ω_6	blocks-6	0.154913	12	P_3
ω_7	blocks-5	0.086448	10	P_4
ω_8	blocks-6	0.218894	6	P_1
ω_9	blocks-4	0.041694	10	P_2
ω_{10}	blocks-7	0.782671	10	P_4
ω_{11}	blocks-5	0.116359	6	P_5

TAB. 1 – *Extrait de l'échantillon d'apprentissage Ω_A issu de Blocksworld.*

tient à l'ensemble des classes $C = \{P_1, P_2, P_3, P_4, P_5\}$ où P_1, P_2, P_3, P_4, P_5 correspondent à des plans.

P_1 : (pick-up b)→(stack b a)→(pick-up c)→(stack c b)→(pick-up d)→(stack d c)

P_2 : (unstack b c)→(put-down b)→(unstack c a)→(put-down c)→(unstack a d)→ (stack a b)→(pick-up c)→(stack c a)→(pick-up d)→(stack d c)

P_3 : (unstack c e)→(put-down c)→(pick-up d)→(stack d c)→(unstack e b)→ (put-down e)→(unstack b a)→(stack b d)→(pick-up e)→(stack e b)→(pick-up a)→ (stack a e)

P_4 : (unstack a f)→(stack a d)→(pick-up b)→(stack b a)→(pick-up c)→ (stack c b)→(pick-up f)→(stack f c)→(pick-up e)→(stack e f)

P_5 : (unstack c b)→(stack c d)→(pick-up b)→(stack b c)→(pick-up a)→(stack a b)

Le processus de planification consiste à rechercher une séquence d'opérations permettant de passer de l'état initial à l'état final souhaité. Classiquement, un planificateur dispose d'un problème et d'un domaine de planification. Ce dernier est décrit par un ensemble d'actions permettant des transitions entre les états (Baki et Bouzid, 2006). Les actions utilisées dans les plans ci-dessus sont : *pick-up, stack, unstack* et *put-down*.

1. http ://www.plg.inf.uc3m.es/ipc2011-learning/Domains

Nous utilisons la plateforme Weka (Garner, 1995) pour la construction du modèle de classification à base de plans. Le modèle de classification est constitué de l'arbre de décision et des règles de classification. Un extrait de l'arbre de décision généré à partir de l'échantillon d'apprentissage Ω_A est donné dans la figure 1.

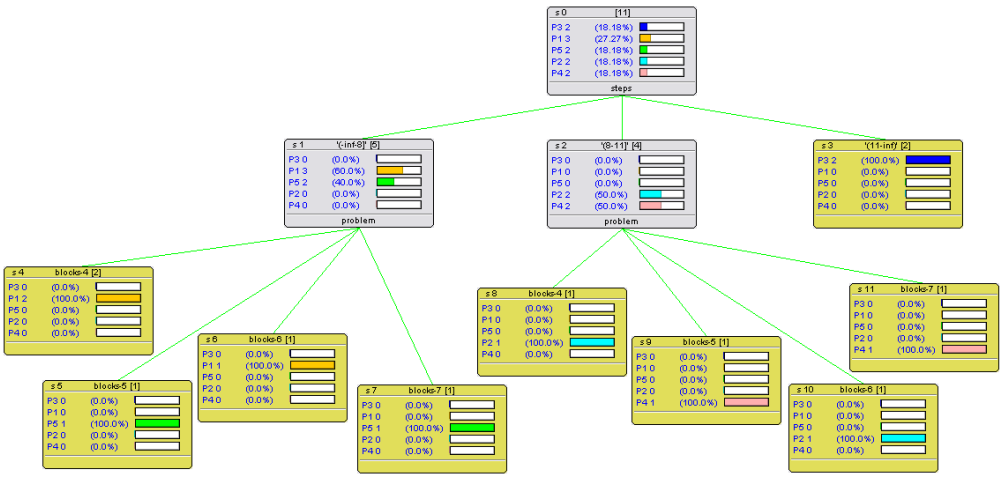


FIG. 1 – *Extrait de l'arbre de décision.*

Les attributs de l'échantillon d'apprentissage peuvent être nominaux ou numériques. Les attributs numériques (continus) nécessitent une procédure particulière, la discrétisation. Discrétiser un attribut numérique consiste à découper son domaine de valeurs en un nombre fini d'intervalles. La discrétisation des données est une phase cruciale car c'est du choix des points de coupure des variables continues que va dépendre la mise au point des modèles de prédiction. Cependant, un choix inapproprié des points de discrétisation des variables peut faire échouer l'opération (Atmani et Beldjilali, 2007a).

Nous remarquons que l'échantillon d'apprentissage Ω_A comporte deux attributs numériques $X_2(\omega)$ et $X_3(\omega)$, ce sont des attributs continus qu'il faut discrétiser. Nous procédons à la discrétisation après avoir défini les points de coupures pour chaque attribut numérique. Les points de coupure sont des intervalles auxquels est attribué un code. Nous utilisons l'outil Weka pour l'étape de discrétisation.

Le premier sommet de l'arbre $s0$ correspond à la racine. La variable X_3 qui correspond à *steps* est la première variable de segmentation utilisée ; elle génère trois sommets fils $s1, s2, s3$ où $s3$ est une feuille dont la classe majoritaire est P_3. La deuxième variable utilisée est X_1 qui correspond à *problem* ; elle produit quatre sommets fils $s4, s5, s6, s7$ qui sont des feuilles. Ce processus est réitéré sur chaque sommet de l'arbre jusqu'à l'obtention de feuilles.

L'objectif de ce modèle de classification est d'attribuer un plan à chaque nouveau cas donné en entrée. Ainsi, au lieu d'appliquer un algorithme de planification pour trouver un plan, nous utilisons la classification par arbre de décision afin de tirer profit de l'expérience acquise.

4 Résultats de l'expérimentation

Pour évaluer l'efficacité de notre approche, nous l'avons testée sur un domaine issu de IPC-2 (The Second International Planning Competition [2]), il s'agit du domaine Blocksworld. Ce domaine est constitué d'un ensemble de blocs et son objectif est de trouver un plan permettant de se déplacer d'une configuration de blocs à une autre. Plusieurs techniques de planification ont été appliquées dans le domaine Blocksworld, parmi lesquelles on trouve BlackBox (Kautz et Selman, 1999), MIPS (Edelkamp et Helmert, 2000), FF (Hoffmann, 2000), HSP2 (Bonet et Geffner, 2001), IPP (Koehler et Hoffmann, 2000), PropPlan (Fourman, 2000), etc.

Les attributs numériques donnés dans l'échantillon d'apprentissage Blocksworld nécessitent une étape de discrétisation. Nous traitons cette étape à l'aide de l'outil Weka qui propose deux modes de discrétisation supervisée et non supervisée. Nous appliquons chacune des ces méthodes de discrétisation sur l'échantillon d'apprentissage et nous obtenons des résultats différents pour chaque attribut discrétisé. Par exemple, le mode de discrétisation supervisée propose deux points de coupures pour l'attribut X_3 et le mode de discrétisation non supervisée propose 10 points de coupures pour le même attribut X_3.

Nous utilisons différentes méthodes (J48, REPTree, IBk) implémentées dans l'outil Weka pour la construction du modèle de classification. J48 et REPTree sont des méthodes utilisées pour la construction de l'arbre de décision alors que IBk représente les k plus proches voisins. Les k plus proches voisins est une méthode couramment utilisée pour la remémoration. Nous proposons de comparer notre approche basée sur l'arbre de décision avec une autre méthode fondée sur les k plus proches voisins.

Pour chaque méthode, nous calculons le taux de succès (%) qui représente le taux d'instances bien classifiées. Les résultats de notre expérimentation sont présentés dans le tableau 2.

Méthode	Mode supervisé	Mode non supervisé
J48	65.02	62.78
REPTree	66.36	65.47
IBk	63.22	50.22

TAB. 2 – *Résultats de l'expérimentation.*

A partir des résultats obtenus, nous remarquons que le taux de succès varie d'une méthode à une autre mais il s'est avéré meilleur avec le mode de discrétisation supervisée. Par ailleurs, les modèles de classification construits avec J48 et REPTree ont donné de meilleurs résultats comparés aux k plus proches voisins (IBk).

Ainsi, le taux d'instances bien classifiées avec les arbres de décision est plus élevé qu'avec les k-plus proches voisins. Par conséquent, nous pouvons constater que nous avons obtenu de meilleurs résultats pour la planification guidée par la classification fondée sur les arbres de décision en comparant avec les k plus proches voisins.

2. http ://idm-lab.org/wiki/icaps/index.php/Main/Competitions

5 Conclusion

Nous avons proposé une approche de planification fondée sur la classification par arbre de décision. Tout d'abord nous avons défini les étapes pour la génération de plans à partir d'une description d'un projet de planification. Ensuite, nous avons expliqué les étapes que nous avons suivies pour la construction du modèle de classification. Nous avons utilisé l'outil Weka pour la construction du modèle de classification à partir de l'échantillon d'apprentissage. Enfin, face à une nouvelle donnée le système se charge de classer cette donnée en lui associant une classe qui correspond à un plan. L'évaluation de notre approche dans le domaine Blocksworld sur plusieurs méthodes a permis de montrer l'efficacité de notre approche.

Comme perspective future de ce travail, nous proposons d'évaluer notre approche dans d'autres domaines et avec d'autres méthodes.

Références

Baki, B. (2006). *Planification et ordonnancement probabilistes sous contraintes temporelles*. Thèse de doctorat, Université de CAEN.

Baki, B. et M. Bouzid (2006). Planification et ordonnancement probabilistes sous contraintes temporelles. *Actes du 15e congrès francophone de Reconnaissance des Formes et Intelligence Artificielle RFIA'2006*, 99–107.

Belacel, N. (1999). *Méthodes de classification multicritère méthodologie et applications à l'aide au diagnostic médical*. Thèse de doctorat, Université Libre de Bruxelles.

Benbelkacem, S., B. Atmani, et A. Mansoul (2012). Planification guidée par raisonnement à base de cas et datamining : Remémoration des cas par arbre de décision. *aIde à la Décision à tous les Etages (Aide@EGC2012)*, 62–72.

Bibai, J. (2010). *Segmentation et évolution pour la planification : le système Divide-And-Evolve*. Thèse de doctorat, Université de Paris-sud XI Orsay.

Bonet, B. et H. Geffner (2001). Planning as heuristic search. *Artificial Intelligence 129*, 5–33.

Breiman, L., J. H. Friedman, R. A. Olshen, et C. J. Stone (1984). *Classification And Regression Trees*. New York: Chapman and Hall.

Crais, E. R. et J. Roberts (1991). Decision making in assessment and early intervention planning. *Language, Speech, and Hearing Services in Schools 22*, 19–30.

De la Rosa, T., S. Jiménez, R. Fuentetaja, et D. Borrajo (2011). Scaling up heuristic planning with relational decision trees. *Journal of Artificial Intelligence Research 40*, 767–813.

Edelkamp, S. et M. Helmert (2000). On the implementation of mips. *Artificial Intelligence Planning and Scheduling, Workshop on Model-Theoretic Approaches to Planning*, 18–25.

Fernandez, S., T. D. la Rosa, F. Fernandez, R. Suarez, J. Ortiz, D. Borrajo, et D. Manzano (2009). Using automated planning for improving data mining processes. *The Knowledge Engineering Review*.

Fourman, M. (2000). Propositional planning. *AIPS-Workshop on Model-Theoretic Approaches to Planning*, 10–17.

Garner, S. (1995). Weka: The waikato environment for knowledge analysis. *Proc. of the New Zealand Computer Science Research Students Conference*, 57–64.

Ghoseiri, K., H. Mazinan, M. Hoseinzadeh, M. Davoodi, et E. khaji (2012). Generating rules to increase production using decision tree. *The 8th International Conference on Data Mining DMIN'12*.

Hoffmann, J. (2000). A heuristic for domain independent planning and its use in an enforced hill-climbing algorithm. *12th International Symposium on Methodologies for intelligent Systems*, 216227.

Kalousis, A., A. Bernstein, et M. Hilario (2008). Meta-learning with kernels and similarity functions for planning of data mining workflows. *ICML/COLT/UAI 2008, Planning to Learn Workshop (PlanLearn)*, 23–28.

Kaufman, K. et R. Michalski (1998). Discovery planning: Multistrategy learning in data mining. *Fourth International Workshop on Multistrategy Learning*, 14–20.

Kautz, H. et B. Selman (1999). Unifying sat-based and graph-based planning. *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, 318–325.

Koehler, J. et J. Hoffmann (2000). On reasonable and forced goal orderings and their use in an agenda-driven planning algorithm. *Journal of Artificial Intelligence Research 12*, 339–386.

Majlender, P. (2003). Strategic investment planning by using dynamic decision trees. *Proceedings of the 36th Hawaii International Conference on System Sciences*.

Miah, M. (2011). Survey of data mining methods in emergency evacuation planning. *Conference for Information Systems Applied Research, 2011 CONISAR Proceedings*.

Régnier, P. (2005). *Algorithmes pour la planification*. Habilitation à diriger les recherches, Université Paul Sabatier, Toulouse.

Wan, N. K. (1995). *Using Decision Trees to Direct the Planning Thought-Process: An Enhancement to the Planning Methodology*. Thèse de master, Fort Leavenworth, Kansas.

Witten, H. et E. Frank (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco: Morgan Kaufmann.

Zakova, M., P. Kremen, F. Zelezny, et N. Lavrac (2008). Using ontological reasoning and planning for data mining workflow composition. *Proc. of ECML/PKDD workshop on Third Generation Data Mining: Towards Service-oriented*, 35–41.

Summary

In Artificial Intelligence, planning refers to an area of research that proposes to develop systems that can automatically generate a result set, in the form of an integrated decision-making system through a formal procedure, known as plan. Instead of resorting to the scheduling algorithms to generate plans, it is proposed to operate the automatic learning by decision tree to optimize time.

In this paper, we propose to build a classification model by decision tree from a learning sample containing plans that have an associated set of descriptors whose values change depending on each plan. This model will then operate for classifying new cases by assigning the appropriate plan.

Fusion de classifieurs SMVs pour la détection d'opinion: Application aux commentaires des journaux en langue arabe

Azizi Nabiha*, Ziani Amel, Tlili-Guiassa Yamina**

*Labged : laboratoire de gestion electronique des documents
nabiha.azizi@univ-annaba.org
**Lri : Laboratoire de recherche en Informatique
guiyam@yahoo.fr
Département d'informatique, Université Badji Mokhtar Annaba, BP 12, Annaba, 23000, Algérie
Z_amel1911@live.fr

Résumé. Cet article s'intéresse à la classification des textes communautaires par apprentissage supervisé. L'opinion peut être exprimée de manière très variée et subtile et donc il est difficile de la déterminer. Une étude approfondie de cette richesse d'information permettrait une meilleure connaissance des utilisateurs, de leurs attentes, de leurs besoins. Pour y parvenir, une étape nécessaire est la classification automatique d'opinion. Les approches basées sur l'apprentissage machine consistent à attribuer des données à un classifieur pour l'apprentissage. Notant que la présence d'une masse textuelle sous forme des commentaires d'articles de journaux en langue arabe en format électronique impose une technique d'exploration particulière. Le but de notre travail est la détection de polarité des commentaires en langue arabe. En effet le système proposé comprend trois phases: la construction et le prétraitement manuel du corpus, l'extraction des caractéristiques et l'apprentissage des classifieurs. Pour la deuxième phase, nous utilisons vingt caractéristiques dont les principales sont l'émotivité, la réflexivité, l'adressage et la polarité. La phase de classification représente dans notre travail la combinaison des plusieurs classifieurs SVMs (Machine à Vecteur de Support), car ils sont les mieux adaptés dans le domaine de fouille d'opinions. Nous avons analysés les deux stratégies des SVMs multi classes qui sont : « un contre tous » et « un contre un » afin de comparer les résultats et améliorer la performance du système global. Nous sommes donc, à notre connaissance, les premiers à mener de telles expériences sur les commentaires et les forums des journaux arabes.

1 Introduction

L'internet social a récemment fait exploser la disponibilité de documents textuels exprimant des opinions ou des sentiments, par exemple dans les groupes de discussions, les blogs, forums et autres sites spécialisés. Les opinions disponibles sur l'internet ont un impact considérable sur les internautes.
L'opinion peut être exprimée de manière très variée et subtile et donc il est souvent difficile de la déterminer exactement. On assiste, ces dernières années, à une prise de conscience de

l'importance de l'opinion sur le web, ce qui explique les nombreux et récents travaux dans ce domaine.
La nécessité de traiter automatiquement les opinions se fait donc fortement ressentir. L'analyse automatique des opinions, aussi appelée fouille d'opinions, concerne l'extraction d'un sentiment dans une source telle qu'un texte sans structure prédéfinie.
Le domaine de la fouille d'opinion peut être divisé en trois sous-domaines:

- L'identification des textes d'opinion qui peut consister à identifier dans une collection textuelle les textes porteurs d'opinion.
- Le résumé d'opinion qui consiste à rendre l'information rapidement et facilement accessible en mettant en avant les opinions exprimées et les cibles de ces opinions présentes dans un texte.
- La classification d'opinion qui a pour but d'attribuer une étiquette au texte selon l'opinion qu'il exprime. On considère généralement les classes positive et négative, ou encore positive, négative et neutre.

De ce fait, plusieurs travaux de recherche se sont intéressés à ce problème. En règle générale, il existe deux types d'approches pour la détection d'opinions et l'analyse du sentiment, celles qui sont basées sur le lexique et celles qui sont basées sur l'apprentissage machine. Les approches basées sur le lexique utilisent un dictionnaire de mots subjectifs. Les approches basées sur l'apprentissage machine consistent à attribuer des données à un classifieur pour l'apprentissage. Ce dernier génère un modèle qui est utilisé pour la partie test.
Nous nous intéresserons ici uniquement aux approches basées sur l'apprentissage machine. Tout système de classification utilisant l'apprentissage automatique se compose de plusieurs phases dont les plus importantes sont: extraction de caractéristiques et le type d'algorithme de classification utilisé.
Ainsi, si aucune méthode de classification ne peut satisfaire entièrement aux exigences d'une application envisagée, l'utilisation simultanée de plusieurs méthodes en même temps peut éventuellement permettre d'en cumuler les avantages sans en cumuler les inconvénients. En effet, le comportement de chaque classifieur vis-à-vis de commentaire à classifier, est déterminé à partir des informations différentes représentants les caractéristiques extraites. L'exploitation des différents résultats générés par les classifieurs utilisant une des méthodes de combinaison de classifieurs, aboutit généralement à une augmentation du taux de clasifiaction. Même si le classifieur est moins performant, la connaissance de son comportement apporte une certaine information utilisable à propos de la vraie classe pendant la combinaison. Donc, le but de la combinaison de classifieurs vise à réduire l'erreur et augmenter la fiabilité de la classification.
Nous nous intéressons dans ce travail à la détection de polarité d'opinion utilisant une des méthodes de classification basée sur l'apprentissage automatique. Toutefois, les méthodes les plus présentes dans la littérature, et qui semblent également être les plus performantes dans le domaine de fouille d'opinion sont les machines à vecteurs de supports (SVM). Comme notre système doit générer 5 classes, il est considéré comme un problème de classification multi classes; pour cela nous avons adopté les deux stratégies « un-contre-tous » et « un contre un » pour décider laquelle entre ces deux est meilleure dans le domaine de fouille d'opinions. Les SVMs sont jugés être très sensibles au paramètres internes tels que la fonction noyau ; en effet, nous avons décidé d'analyser le changement de la fonction noyau en générant plusieurs classifieurs SVMs multi-classes associé chacun avec une fonction noyau différente. Le résultat final sera la combinaison de ces classifieurs afin d'assurer la complémentarité exis-

tante entre les différentes fonctions. Pour cela on a utilisé les deux méthodes de fusion : Vote Majoritaire et Vote Pondéré.

2 Etat de l'art sur la classification d'opinion

Le terme « fouille d'opinions » est utilisé pour évoquer le traitement automatique des opinions, des sentiments et de la subjectivité dans les textes. Ce domaine est connu sous les noms de : opinion mining (Pang et Lee (2008)), sentiment analysis (Liu (2010)), ou encore subjectivity analysis et est souvent associé à un problème de classification sur des textes évaluatifs comme ceux disponibles sur Amazon ou Epinions.
Afin de décider de l'orientation d'un document (Turney, 2002), (Wilson et al., 2004) ou de la valeur positive/négative/neutre d'une opinion dans un document (Hatzivassiloglou & McKeown, 1997), (Yu & Hatzivassiloglou, 2003), (Kim & Hovy, 2004).
Le travail de (Maurel et Dini 2009) été caractérisé par une utilisation mixte d'une technologie symbolique fondée sur des règles et d'une technologie statistique reposant sur l'extraction d'une ontologie du domaine, approche dans laquelle la méthode symbolique a un poids plus important (Dini, 2002), (Dini & Mazzini, 2002), (Maurel et al., 2007), (Maurel et al., 2008), (Bosca & Dini, 2009).
Des travaux allant au-delà ont mis l'accent sur la force d'une opinion exprimée où chaque proposition dans une phrase peut avoir un fond neutre, faible, moyen ou élevé (Wilson et al., 2004). Des catégories grammaticales ont été utilisées pour l'analyse de sentiments dans (Bethard et al., 2004) où des syntagmes adjectivaux comme trop riche ont été utilisés afin d'extraire des opinions véhiculant des sentiments. (Bethard et al., 2004) utilisent une évaluation basée sur la somme des scores des adjectifs et des adverbes classés manuellement, tandis que (Chklovski, 2006) utilise des méthodes fondées sur un modèle pour représenter des expressions adverbiales de degré telles que parfois, beaucoup, assez ou très fort.

En 2002, Turney, Pang et coll. encouragent la recherche dans le domaine de sentiment analysis en classant des critiques de cinéma. (Mukherjee et al., 2010) proposent d'étiqueter un ensemble de structures grammaticales en se basant sur un étiquetage. Part-Of-Speech (POS). (Ding et al., 2007) analysent les cooccurrences de mots à l'intérieur d'une phrase puis les cooccurrences entre les phrases. Ils combinent des règles issues des deux échelles pour obtenir un meilleur taux de bonne classification en classification de sentiments. Dans le même genre. Wilson et al. (2004) ajoutent à la classification selon la polarité, la force de l'opinion exprimée.

3 Le processus général de classification de textes d'opinions

Puisque nous avons besoin d'associer des notes à des textes, nous nous intéressons ici uniquement à la classification d'opinions. Deux grands types de méthodes sont utilisés pour cette tâche. Il y a tout d'abord les approches plutôt linguistiques qui consistent à répertorier le vocabulaire porteur d'opinions, puis à établir des règles de classification selon la présence ou l'absence des mots appartenant à ce vocabulaire.
Il existe également les approches mettant en œuvre des outils issus du domaine de l'apprentissage automatique. Dans notre cas, un ou plusieurs classifieurs sont utilisés pour

apprendre la polarité des opinions traités, puis ils seront responsables de classifier les opinions inconnues.
Nous nous intéressons dans ce travail à l'application de la deuxième approche avec un module d'extraction des caractéristiques pertinentes. Les méthodes utilisées dans ce cadre sont issues de la classification dite supervisée où un classifieur est appris à l'aide d'exemples de données (ici de textes de commentaires) dont on connaît déjà la classe. Les mots des textes sont alors généralement considérés comme des données indépendantes et équivalentes les unes aux autres, leur sémantique n'étant pas explicitement prise en compte.

La Structure générale du système COJA (Classification d'Opinions dans les Journaux Arabes) basée sur la classification supervisée.

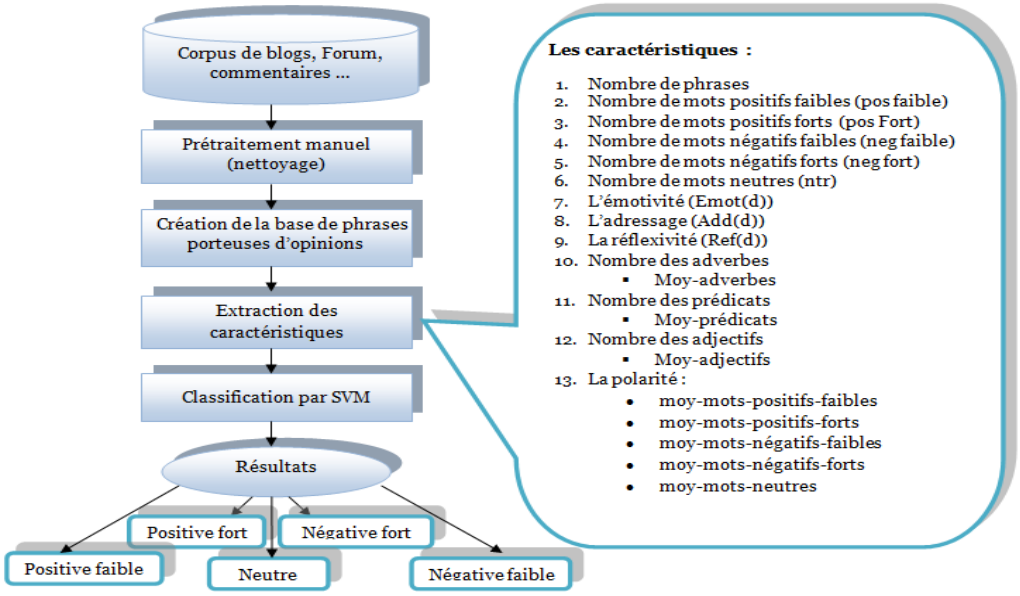


FIG. 1 – Le processus de classification d'opinion par l'approche COJA

3.1 Le corpus d'apprentissage

La classification supervisée nécessite des exemples (données étiquetées) afin de construire le «corpus d'apprentissage». Ce corpus ayant un impact direct sur l'apprentissage des règles, et par conséquent sur la classification, il est nécessaire que les exemples soient représentatifs de l'ensemble des données. Cette hypothèse est généralement difficile à vérifier. En classification d'opinions, ou plus généralement en classification de textes, les corpus étudiés sont assez restreints car l'étiquetage des exemples est souvent effectué à la main. Ceci entraîne un coût élevé et ne permet donc pas l'obtention d'un gros corpus d'apprentissage. [Poirier et al, 2011]

Pendant notre travail, nous avons utilisé un corpus de cent cinquante commentaires sur des articles, recueillis à partir des journaux arabes algériens disponibles sur le net (Eshorouk الشروق, Akher saa آخر ساعة…etc.). L'ensemble touchait plusieurs thèmes différents (économiques, politiques,…etc.). Notre but était d'effectuer une classification supervisée pour pouvoir déterminer la polarité des commentaires: positive fort, positive faible, neutre, négative fort ou négative faible.
Une phase de prétraitements manuels sur ce corpus a été effectué représenté par: la suppression des mots vides, les symboles, les virgules, les points d'exclamations et la traduction des dialectes arabes et autres langues …etc.

Nous présentons un exemple de chaque type d'opinion en langue arabe à classifier. Nous les avons traduits en français pour permettre la compréhension d'un plus grand nombre de lecteurs.

	Polarité	**Les commentaires en français**	**Les commentaires en arabe**
1	Positif faible	Merci pour ce que tu as écris Mr amine zaoui celui qui dirige ministère de la culture en Algérie est celui que le dépense sans dignité.	شكرا على ما كتبت سيد الأمين الزاوي إن وزارة الثقافة بالجزائر يستلمها من يصرف المال العام بغير وجه حق.
2	Positif fort	Félicitation a nous tous, vous avez fait le bon choix. La journaliste leila bouzidi est réellement compétente et d'une personnalité professionnelle lui permettant de bien gérer sa carrière que dieu soit avec elle.	مبروك علينا وعليكم لقد أحسنتم الاختيار. الصحفية القديرة ليلى بوزيدي حقا هي متمكنة وذات شخصية مهنية متحكمة في إدارة مهنتها أعانها الله ووفقها .
4	Négatif faible	Franchement, c'est Article non professionnel de la part d'un expert connu voulant la réussite de l'équipe national .	مقال غير احترافي صراحة من إعلامي كبير يحرص على نجاح المنتخب وقاسي كذلك.
6	Négatif fort	Franchement, j'ai pas aimé le style de l'écrivain, et aussi son point de vus envers ce sujet et j'ai haï ses moqueries pour l'art.	صراحة لم أحب أبدا أسلوب الكاتب ولا رأيه في الموضوع وكرهت استهزاءه بالفن.
8	Neutre	Non, c'est son point de vue personnel	كلا انه رأيه الشخصي.

TAB.1 – Exemples de commentaire a classifier

3.2 L'extraction des phrases porteuses d'opinions

Tout système de classification d'opinions nécessite une étape qui résume l'opinion, en ne gardant que les phrases subjectives.
Dans notre travail, on s'intéresse en priorité au module de classification et vu l'obligation de passer par l'étape d'extraction des phrases porteuses d'opinions qui sont construite à partir du corpus initial. Le résultat de cette étape sera un corpus prétraité contient seulement les phrases subjectives en langue arabe. Donc, on a noté manuellement 150 commentaires pour la phase d'apprentissage du classifieurs et la phase d'évaluation des résultats.

3.3 Extraction des caractéristiques

Nous commençons la phase d'extraction de caractéristiques par la construction des tables des marqueurs d'opinions du corpus arabe utilisé définies comme suit :

- La table des Marqueurs :

La table des marqueurs contient tous les prédicats, les adjectifs et les adverbes construits à partir du corpus avec leurs polarités et intensités.

Exemples: Prédicat (أحبaimer, كره détester, ظن penser. -**Adjectif:** جميلة bien fait, رائعة magnifique, ركيكة lâche. -**Adverbe:** غنية riche, مضجرة fatigante, مفيدةintéressante. -**Intensité:** كثيرا beaucoup,جدا très, مئة بالمئةcent pour cent. -**Négation:** لا (non, ni, pas), ليس لم لن pas. - **Adressage:** أنت tu, يا سيدي الكاتب Mr l'écrivain. **Réflexivité:** أنا moi, رأيي mon opinion.

3.4 Extraction de caractéristiques

Nous avons adopté un ensemble des caractéristiques inspiré des différents travaux dans la littérature et qui ont montrés leur efficacité dans la représentation d'un commentaire. On peut les résumer comme suit !

- **Adverbe :** *Total-adverbes :* $\mathrm{tot(adv)}$= Nombre total des adverbes du document

Moy-adverbes$=\mathrm{tot(adv)/(tot(adj)+tot(adv)+tot(pred)}$ (1)

- **Adjectif:** *Total-adjectives : tot(adj)* = Nombre total des adjectifs du document

Moy-adjectives$=\mathrm{tot(adj)/(tot(adj)+tot(adv)+tot(pred)}$ (2)

-Émotivité: Les chercheurs ont exploité la présence des adverbes et des adjectifs dans un document comme un indicateur qui permet de déterminer les opinions. Nous calculons l'émotivité d'un document en comptant le nombre des adverbes et des adjectifs dans ce document.

$$Emot(d) = \frac{|\{\omega \in d \setminus type(\omega) \in \{adjectif, adverbe\}\}|}{|\{\omega \in d \setminus type(\omega) \in \{predicat\}\}|} \quad (3)$$

-**Adressage:** La plupart des phrases trouvées dans les blogs et les forums contiennent des mots comme suit «أنت, انتم, نفسك, أنفسكم, هو, هي, هم, نفسه, نفسها, نفسهم, أنفسهم» car les utilisateurs écrivent des commentaires sur un sujet, en s'adressant aux autres. De ce fait l'utilisation de ces pronoms d'adressage est très fréquente. Par conséquent, nous considérons que la composante d'adressage dans le cadre de notre détection d'opinions, est comme suit :

$$Add(d) = \frac{|\{\omega \cap \omega' \setminus \omega \in d, \omega' \in A\}|}{|A|+|R|} \quad (4)$$

- **Réflexivité:** Les blogueurs utilisent beaucoup de pronoms réflexifs comme « أنا, أنا شخصيا » « moi, moi-même» lors de l'écriture. Par exemple, l'utilisation de «ي» dans « رأيي »« Je pense que », « من وجهة نظري », « mon point de vu est que », etc. Toutes ces phrases font référence à une opinion d'opinion, et par conséquent, nous incluons la mesure de la réflexivité. L'idée est que tout document avec un plus grand nombre de ces mots sera plus subjectif par rapport à celui qui a moins de nombre de ces mots. Cette mesure est exprimée par la réflexivité. $Ref(d) = \frac{|\{\omega \cap \omega' \setminus \omega \in d, \omega' \in R\}|}{|R|+|A|}$ (5)

- **La négation:** Pour calculer le nombre des mots positifs et négatifs forts il faut prendre en considération la négation:

✓ Marqueur de négation + mot positif fort → un mot négatif fort. Ex: لا أحب.

- **L'intensité:** Si un mot est de polarité positif faible et se situer avant ou après un marqueur d'intensité, il doit être calculé comme un mot positif fort. *Exemple :* أعجبني أكثر فائدة ,جدا

Mais s'il est de polarité négatif faible il doit être calculé comme un négatif fort.*Exemple :*

قاسية كثيرا, أكبر كاذب

En plus il y a un ensemble de caractéristiques que nous avons proposé décrit par le tableau suivant :

Phrase	Nombre de phrases	∑ phrases
Positif fort	Nombre des mots positifs forts	∑ mot (posFo)
Positif faible	Nombre des mots positifs faibles	∑ mot (posFa)
Négatif fort	Nombre de mots négatifs forts	∑ mot (negFo)
Négatif faible	Nombre de mots négatifs faibles	∑ mot (negFa)
Neutre	Nombre de mots neutres	∑ mot (ntr)
Prédicat	Total-prédicats	Nbr(pred)=Nombre total des prédicats du document
	Moy- prédicats	$Nbr(pred)/(tot(adj) + tot(adv) + tot(pred)$
Polarité	Somme polarité	SomPolarite= $\sum mot(posFo) + \sum mot(posFa) + \sum mot(negFo) + \sum mot(ne$ $\sum mot(ntr)$
	Mots positifs forts	$\sum mot(posFo)/SomPolarite$
	Mots positifs faibles	$\sum mot(posFa)/SomPolarite$
	Mots négatifs forts	$\sum mot(negFo)/SomPolarite$
	Mots négatifs faibles	$\sum mot(negFa)/SomPolarite$
	Mots neutres	$\sum mot(ntr)/SomPolarite$

TAB 2 – Les mesures proposées

- L'adressage et la réflexivité: Le compteur des mots d'adressage augmente si le mot est un marqueur d'adressage ou bien il se termine par un marqueur d'adressage. Par exemple :أنت, كلامك, مقالتك, أهانتهم .

3.5 La classification

Ils existent plusieurs méthodes de classification supervisée et beaucoup d'entre elles ont été testées pour la classification d'opinions. On peut citer les arbres de décision, les réseaux de neurones, la régression logistique, ainsi que des méthodes combinant différents classifieurs comme les systèmes de votes ou les algorithmes de Boosting. Toutefois, les méthodes les plus présentes dans la littérature, et qui semblent également être les plus performantes sur les textes, sont les machines à support de vecteurs (Pang et Lee, 2004; Wilson et al., 2004 ; Trinh, 2007 ; Crestan et al., 2007 ; Plantié et al., 2008; Kaiser et al., 2010 ; Poirier 2011 ;).
Les machines à vecteurs de support (SVM) sont une famille d'algorithmes de classification et de régression développées au cours des années 90 par Vapnik, qui sont aujourd'hui considérées comme une des méthodes les plus performantes sur de nombreux problèmes réels, notamment pour les problèmes en grande dimension. Considérons un espace multidimensionnel où chaque trait est une dimension.

- **Un-Contre-Tous (One Versus All):**

La solution la plus simple pour résoudre un problème multi-classes à l'aide des SVMs consiste à le décomposer en un ensemble de sous-problèmes binaires et à construire indépendamment un SVM pour chacun d'entre eux. Ainsi, Cortes et Vapnik proposent d'utiliser une stratégie de décomposition très intuitive et facile à mettre en place. Cette stratégie, communément nommée « un contre tous », consiste à construire autant de SVM qu'il y a de classes. Chaque SVM est alors entraîné à séparer les données d'une classe qui seront étiquetées +1, de celles de toutes les autres classes qui seront étiquetées −1. Chaque SVM est ainsi associé à une classe et sa sortie avant seuillage peut être considérée comme une mesure d'appartenance à la classe. La règle de décision généralement utilisée consiste donc à attribuer la donnée inconnue à la classe correspondant au SVM ayant la plus grande valeur de sortie.

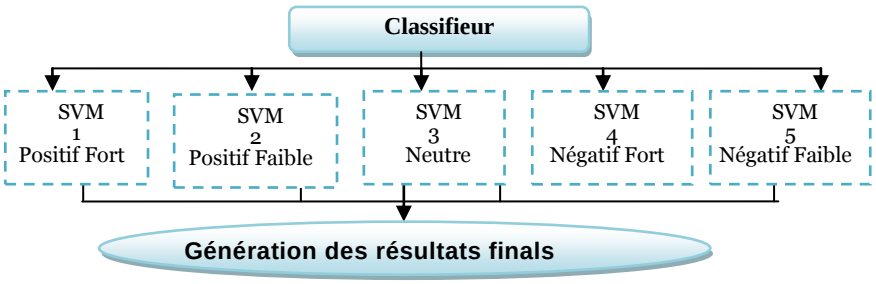


FIG. 2 – Schéma d'un SVM multi classe selon la stratégie UN CONTRE TOUS

Ce classifieur peut donner un seul résultat ou plusieurs. Dans le cas de plusieurs résultats la phase de prise de décision du résultat final doit être générée aléatoirement.

- **La stratégie (un contre un) :**

Une autre stratégie classique consiste à construire un SVM pour chaque paire de classes, soit c(c-1)/2 SVM pour un problème à c classes. Chaque classifieur est donc entraîné à séparer les données d'une classe de celles d'une autre classe. Nous parlerons alors de stratégie « un contre un ».

3.6 La combinaison parallèle des classifieurs

Plutôt que de chercher à optimiser un seul classifieur en choisissant les meilleures caractéristiques pour un problème donné, les chercheurs ont trouvé plus intéressant de combiner plusieurs méthodes de classification. Notre objectif est d'analyser le comportement d'un système combinant plusieurs classifieurs afin d'augmenter les performances de classification. Pour construire un tel système, il y a une méthode qui se base sur l'utilisation d'un même classifieur en modifiant à chaque fois ses paramètres internes; ce qui va générer des classifieurs différents. Cette différence sera traduite par une complémentarité durant le processus de classification d'un commentaire inconnu.
En effet, Vu que chaque classifieurs SVM selon la fonction noyau génère un résultat différent, et vu que dans la littérature, on n'a pas pu prouver la supériorité d'une fonction par rapport aux autres dans le cas général, on a utilisé quatre classifieurs, chacun d'eux à une fonction noyau différente qui sont les suivantes (linéaire, polynomiale, gaussienne et tangente). En combinant ces classifieurs, chacun d'eux va assurer un certain niveau de complé-

mentarité au système global. Le système généré sera construit de 4 classifieurs, dont chacun d'eux offre 5 sorties. Afin de fusionner les résultats des 4 classifieurs, nous avons appliqué deux méthodes de combinaison les plus connus (le vote majoritaire et le vote pondéré).

4 Résultats expérimentaux

4.1. Extraction de caractéristiques

Un exemple de vecteur de caractéristiques généré après l'exécution de la procédure d'extraction de caractéristiques de notre application est illustré par la figure suivante (Fig. 3).

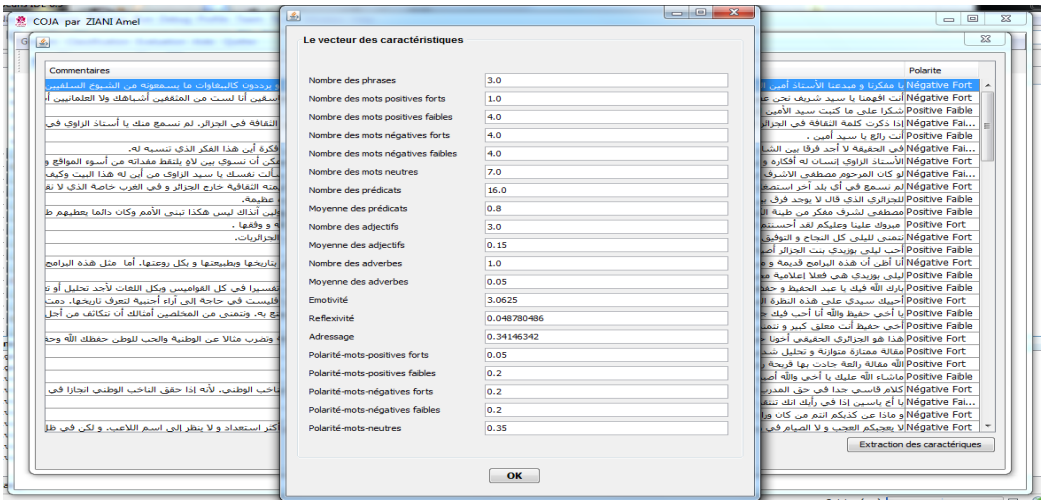


FIG. 3 – phase d'extraction de caractéristiques de l'application COJA

4.2 La phase de classification

Au niveau du test on applique notre système de classification sur un nouvel ensemble différent de celui de l'apprentissage, se composant de 90 commentaires étiquetés manuellement.

- **Pour l'approche un contre tous**

	Linéaire	Polyno-miale	Gaussienne	Tan-gente	Vote	Vote pondéré
Neutre	0.19	0.27	0.37	0.09	0.39	0.39
Positif fort	0.69	0.75	0.84	0.29	0.82	0.92
Positif faible	0.47	0.49	0.55	0.17	0.55	0.59
Négatif fort	0.89	0.91	1.00	0.33	1.00	1.00
Négatif faible	0.45	0.56	0.59	0.27	0.59	0.64

TAB. 2 – Variation du taux de classification de chaque classe pour les quatre fonctions noyaux et avec les deux méthodes de fusion.

A partir du tableau ci-dessus on peut constater que les classifieurs SVMs conçus avec les fonctions noyaux polynomiale et gaussienne donne le meilleur taux de classification par comparaison aux autres SVMs.

On constate notamment que la classe de type « négatif fort » offre le meilleur taux de classification ; en revanche la classe neutre est celle ayant le taux le plus faible. Cela est du à notre avis aux poids des marqueurs représentant ces deux classes.

On remarque que les résultats de la combinaison (soit pour la méthode de vote ou de vote pondéré) sont meilleurs qu'en utilisant un seul classifieur. En effet, malgré que la fonction tangente génère un taux de reconnaissance très faible mais sa présence dans le processus de combinaison enrichie la classification; d'où l'intérêt majeur de la combinaison de classifieurs. Les schémas suivants (Fig. 4 et Fig .5) résument le tableau ci dessus pour chaque fonction noyaux ainsi que pour l'utilisation de la combinaison pour les deux méthodes de fusion Vote et vote pondéré.

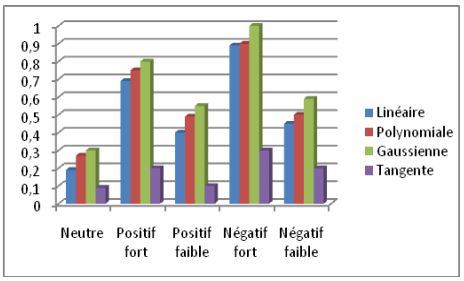


FIG. 4 – Variation du taux de classification de chaque classe pour les quatre fonctions noyaux.

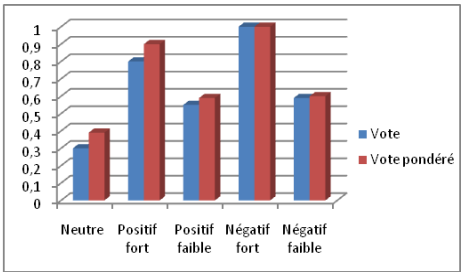


FIG.5 – Variation du taux de classification de chaque classe pour les deux méthodes vote majoritaire et vote pondéré.

- **La stratégie un contre un**

	Linéaire	Polynomiale	Gaussienne	Tangente	Vote	Vote pondéré
Neutre	0.51	0.57	0.64	0.29	0.61	0.65
Positif fort	0.83	0.92	1.00	0.32	1.00	1.00
Positif faible	0.88	0.88	0.88	1.00	0.88	0.93
Négatif fort	0.88	0.98	1.00	0.22	1.00	1.00
Négatif faible	0.49	0.59	0.69	0.19	0.61	0.68

TAB. 3 – Variation du taux de classification de chaque classe pour les quatre fonctions noyaux et avec les deux méthodes de fusion.

En général, les deux méthodes de fusion produisent des bons résultats, mais c'est la méthode de vote pondéré qui a le meilleur taux pour toutes les classes.

Pour mieux voir la différence de comportement des différents SVMs selon la fonction noyau utilisée, on a représenté graphiquement les données du tableau ci-dessus ; la première concernant les classifieurs SVMs pris individuellement et la deuxième pour illustrer le système multi classifieurs utilisant les méthodes de « vote » et de « vote pondéré ».

Nous concluons alors, que pour améliorer le taux de classification, il est préférable d'utiliser la méthode de vote pondéré.

Après la mise en œuvre des deux stratégies, il s'avère que les résultats obtenus par la stratégie « un contre un » sont meilleurs que ceux de la stratégie « un contre tous ». En effet comme le montre le tableau précédant (Tab 4.5), le taux de toutes les classes est augmenté pour tous les fonctions, ce qui signifie que la classification est excellente.

5 Conclusion

Au terme de ce travail, nous avons développé une application qui permet la détection de polarité d'opinions dans les forums en langue arabe par combinaison de plusieurs SVM. Ce système à pour rôle d'extraire les caractéristiques représentant les commentaires du corpus et de les classifier en catégories par la coopération de plusieurs classifieurs SVMs. Notre système opère en trois phases, la première consiste à la construction et le prétraitement manuel du corpus recueillit à partir des journaux arabes algériens. La seconde phase est une extraction des caractéristiques afin de détecter et représenter les commentaires. Le choix de ces caractéristiques est fait par une recherche approfondie sur les plus importantes caractéristiques pouvant représenter de mieux le commentaire tout en évitant la redondance et la confusion des données d'entrée. Enfin, la troisième phase consistant à réaliser le module de classification et dans le but de bénéficier des avantages des systèmes multi-classifieurs, on a proposé un système combinant 04 classifieurs SVMs multi-classes représentant chacun par une fonction noyau différente adoptant les deux stratégies (un contre tous et un contre un) des classifieurs SVMs. Les résultats issus de la combinaison sont très encourageants et nous ont permis de mieux s'investir dans cet axe tout en analysant le rôle que peut jouer chaque paramètre des caractéristiques.

Les tests sur les commentaires en langue arabe avec les deux stratégies « un contre tous » et « un contre un », nous ont permis de prouver que la stratégie un contre un donne de meilleurs résultats avec les commentaires des journaux en langue arabe.

Cette expérience, nous a permis de faire une première exploration du vaste domaine qui est l'opinion mining, et nous incite à aller plus loin dans ce domaine dans le cadre de travaux futurs, car l'opinion mining représente un axe de recherche très prometteur et très passionnant, ce qui explique l'intérêt recrudescent que portent les chercheurs pour ce domaine.

Références

Ardjani, F., Sadouni, K. et Benyettou. A. (2012). Optimisation des SVM Multi-Classe Par Essaim Particulaire (PSO-SVM).

Bethard S., Yu H., Thornton A., Hatzivassiloglou V. Jurafsky D. (2004), « Automatic extraction of opinion propositions and their holders », Actes d' AAAI'04.

BOUGHANEM, M. et BELBACHIR, F. éditeurs (2010), Expérimentation de fonctions pour la détection d'opinions dans les blogs (mémoire de Master).

Chklovski T. (2006), Deriving quantitative overviews of free text assessments on the web, Actes d' IUI'06, p. 155-162.

Dini L. (2002), Compréhension multilingue et extraction de l'information, in , F. Segond (ed.), Multilinguisme et traitement de l'information (Traité des sciences et techniques de l'information), Editions Hermes Science..

Dini L., Mazzini G. (2002), Opinion classification through information extraction , in , A. Zanasi, C. A. Brebbia, , N. F. F. Ebecken, , P. Melli (eds), Data Mining III, WIT Press.

Hatzivassiloglou V., McKeown K. R. (1997), Predicting the semantic orientation of adjectives , Actes d' ACL'97.

Kim S.-M., Hovy E. (2004), Determining the sentiment of opinions », Actes de COLING'04

Liu B (2010), Sentiment Analysis , Invited talk at the 5th Annual Text Analytics Summit.

Maurel S., Curtoni P. et Dini L (2007), L'analyse des sentiments dans les forums.

Pang, B et Lee, I (2008). Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr., vol. 2, n° 1-2, p. 1-135.*

POIRIER, D., FESSANT, F., BOTHOREL, C., GUIMIER DE NEEF, E. et BOULLE, M (2009). Ap proches Statistique et Linguistique Pour la Classification de Textes d'Opinion Portant sur les Films.

POIRIER, D. (2010). La Classification d'Opinion comme préambule à la Recommandation Automatique de Contenus. *In COnférence en Recherche d'Information et Applications 2010*, Tunisie.

POIRIER, D., FESSANT, F. et TELLIER, I. (2011). De la classification d'opinions à la recom mandation : l'apport des textes communautaires. *In TALN 2011 (Traitement automatique des langues naturelles)*, revue semestrielle de l'ATALA 51, 3 (2010).

Turney P (2002), Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, Proceedings of the Association for Computational Linguistics, ACL'02.

Yu H., Hatzivassiloglou V. (2003), Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences, Actes d' EMNLP'03.

.

Summary

This article is interested in the classification of communal texts by supervised training. The opinion can be expressed in a very varied and subtle manner and therefore it is difficult to determine it. A deepened survey of this wealth of information would permit a better knowledge of the users, of their waiting, of their needs. To arrive there, a necessary stage is the automatic classification of opinion. The approaches based on the training plots consist to assign some data to a classifier for the training. Noting that the presence of a textual mass as the commentaries of articles of newspapers in Arabian language in electronic format imposes a particular technique of exploration. The goal of our work is the detection of polarity of the commentaries in Arabian language. Indeed the proposed system consists of three phases: the construction and the manual prétraitement of the corpus, the extraction of the features and the training of the classifiers. For the second phase, we use twenty features of which the main are emotionalism, the reflexivity, the adressage and the polarity. The phase of classification represents in our work the combination of the several SVMs classifiers (Support to Vector Machine), because apparently they adapt better to the domain of opinion mining. We analyzed the two strategies of the SVMs multi classes that are: "one against all "and" one against one" in order to compare the results and to improve the performance of the global system. We are therefore, to our knowledge, the first to lead such experiences on the commentaries and the forums of the Arabian newspapers.

The Adoption of FP-Growth Algorithm to Mine Multilevel Association Rules

Faraj A. El-Mouadib*, Ilham A. El-Areibi**
University of Benghazi
Faculty of Information Technology
Benghazi, Libya
*elmouadib@yahoo.com
**Ilham.elareibi@yahoo.com

Abstract. One of the most researched data mining functionality is the association analysis (discovery of association rules). Many successful applications implemented and developed for single level association rules to produce general knowledge that could be vague or ambiguous. On the other hand, finding fine more precise knowledge has posed some challenges due to the sparsely of data in multidimensional space and the limited number of algorithms. Mining knowledge at multiple levels of abstraction leads to refined knowledge. The objective of this paper is to adopt an association rule mining algorithm (FP-growth) for single level association rules to be used in mining multilevel association rules. This adoption implemented in a system called "ML-FP-growth". The system is tested with a number of data sets and the results are compared with the results obtained from running a system that compares two well known algorithms for multi level association rules (ML-T2 and ML-T2+).

1. Introduction

Since the emergence of the PCs in the early 80's, it became very easy to collect and generate terabytes of data in all real-life fields. Implicit knowledge is buried in such massive amounts of data that can't be discovered via traditional data analysis tools. The urgent need to transform the tombs of data into valuable knowledge has called for the invention of new intelligent set of tools and techniques known as Knowledge Discovery in Databases (KDD). The field of KDD is one of the most attractive subjects in the field of computer science research. In many publications, KDD is treated as a synonym to Data Mining (DM), while others consider it as one step in wider iterative process called KDD Berry and Linoff (2004), Brin et al. (1997). DM is the most essential step in the KDD process because it constitutes the algorithm by which patterns or regularities can be extracted Chen et al. (1996), Devedzic (2002), Han (1995). In general, DM tasks can be classified as: descriptive (unsupervised) and predictive (supervised). One of the well known descriptive mining tasks is association analysis. Association analysis is the process of finding interesting relationship among items in a given transactional database and presents them

in the form of rules (association rules). Association Rule Mining (ARM) was first introduced in 1993, Agrawal et al. (1993) with the first application in the area of market basket analysis. There are three distinct types of association rules: the first depends on the types of values handled in the rule such as; Boolean or quantitative association rules. The second type depends on the data dimensionality such as; single or multidimensional association rules. The third type depends on the levels of abstraction such as; single of multi level association rules.

In this paper, we are concerned with multilevel association rules because the discovered knowledge is more precise. The main aim of this work is to adopt the well-known FP-growth algorithm, which was designed for single level association rules, to be used in the discovery of multilevel association rules.

2. Association Rule Mining (ARM)

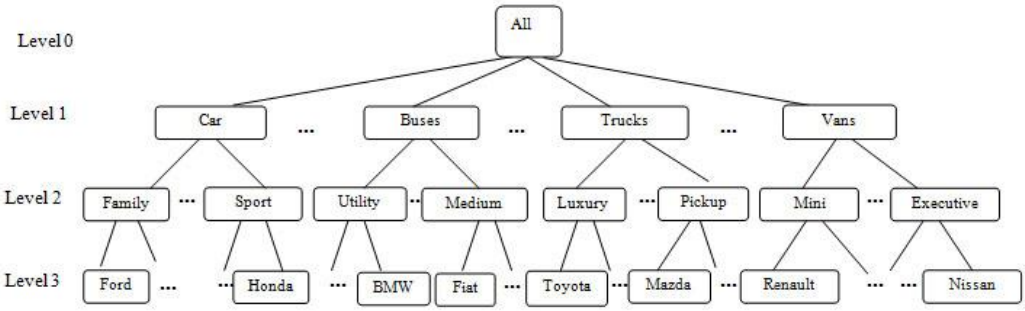
Most of the ARM algorithms had been developed for mining single level association rules. AIS algorithm, Agrawal et al. (1993), was the first algorithm proposed to generate all large itemsets in a transactional database. After the AIS algorithm, many other algorithms were presented i.e. SETM, Houtsma and Swami (1993) and Apriori algorithms Agrawal and Srikant (1994), where the later became the land mark for the ARM algorithms. Many improvements to the Apriori algorithm had been suggested and implemented in search to improve the efficiency of the original algorithm. Such improved algorithms are: AprioriTid and AprioriHybrid Agrawal et al. (1993), Direct Hashing and Pruning (DHP) Fayyad et al. (1996), Sampling, Fayyad et al. (1996), and Dynamic Itemset Counting (DIC), Brin et al. (1997). Such algorithms are Apriori-based because they adopt the same methodology in the generation of the candidate set. Such approaches suffer from two major bottlenecks which are: the high cost to handle huge number of candidate sets and the needed of multiple scans over the database. In response to overcome the Apriori-based algorithms drawbacks, the Frequent Pattern growth (FP-growth) algorithm, Houtsma and Swami (1993), was proposed.

There had been an extension to the Apriori-based single level association rule mining to deal with the concept of multilevel association rules mining. Such algorithms are being ML-T2, ML-T1, ML-Tmax, and ML-T2+, Han and Fu (1995).

3. Concept hierarchy

The efficient mining of multilevel association rules necessitates the use of concept hierarchies. As stated in El-Majressi, (2009), a concept hierarchy organizes concepts (attribute values) in a tree like form (taxonomy) where low-level concepts are very specific and concepts in higher levels are more general. Figure-1 presents vehicle types in a concept hierarchy. This concept hierarchy has four levels of abstraction, which referenced as; 0, 1, 2, and 3. Level 0 represents “All” types of vehicles, which is the most general concept. Level 1 presents some specific types of vehicles such as Cars, Vans, Buses, etc… Level 2 presents the models of vehicles i.e. Family and Sports

for cars, Medium and Mini for Buses and Luxury and Pickup for trucks. Level 3 presents the names of the manufacturing companies.



4. Multilevel association rules

Since early 1990's, the ARM for single level association rule mining has been extensively studied and well developed Srikant (1994), Agrawal et al. (1993), Agrawal et al. (1993), Dunham et al. (2000), Fayyad et al. (1996), Fayyad et al. (1996). Furthermore, finding association rules at single level (high level of abstraction) may already be viewed as common sense knowledge rather than novel one and it represent difficulties to find the desired more specific (fine) knowledge in databases. On the other hand, many applications within ARM require the mining to be carried out at multiple levels of abstraction Chen et al. (1996). Mining association rules at low levels of abstraction is very difficult because patterns at such levels tend to be scattered and lack substantial supports, Han and Fu (1995), Han (1995). Multilevel association rules discovery requires different minimum support and minimum confidence for each level of except for level 0. The process of mining multilevel association rules is performed by the following two steps:

First step: Generate all frequent or large itemsets at each level, except level 0, whose support is above the minimum support for the specific level. Progressively find and collects all frequent itemset at each level starting from level 1 and work down in the hierarchy until no more frequent itemsets can be generated at any level, Han and Fu (1995), Han (1995).

Second step: Generate all association rules from the frequent itemset based on the minimum confidence threshold at each level.

5. FP-Growth algorithm

The two bottlenecks of the Apriori-based algorithm were the main reason for searching for more improved ARM algorithms. The answer came from Han et al. (2000) by introducing an

interesting algorithm called Frequent-Pattern growth (FP-growth) for efficient mining of frequent patterns in large database. The interestingness and efficiency of FP-Tree algorithm can be summarized as follows:

1. The transactional database is compressed into frequent-pattern tree (FP-tree) with the itemset count information.
2. It applies the divide-and-conquer principle to divide the compressed database into a set of conditional databases each of which is associated with one frequent item. The mining is applied on each conditional database separately.

The FP-growth algorithm generates frequent itemsets by only two passes over the database. Like Apriori-based algorithms, the FP-growth algorithm generates the set of frequent 1-itemsets in the first scan of the database. All of the infrequent itemsets are removed by turning the transactional database into FP-tree. According to Han et al. (2000), the FP-tree is defined as follows:

1. The tree consists of a top node called the root, a set of item prefix sub-trees as the branches (internal nodes) and a frequent-item header table.
2. Each item prefix sub-tree has three fields: item-name, count, and node-link. The item-name represents the item in the node. The count field is the number of transactions in reaching the node. The node-link is the links to the next node in the FP-tree carrying the same item-name. The node-link is null if the node is a leaf.
3. The entries in the frequent-item header table have two fields: the item-name and the head of node-link in the FP-tree.

More details on the FP-growth algorithm can be found in Han et al. (2000).

6. Design of ML-FP-Growth system

Our ML-FP-growth system is designed to mine multilevel association rules even though it is based on the FP-growth algorithm, which was intended for mining single level association rules. As it has been stated in Dunham et al. (2000), the frequent patterns for multilevel association rules are calculated at each level (starting at level 1) of the used concept hierarchy and working down to the lower level until it is not possible to generate any frequent patterns. In other words the frequent patterns are discovered at each level. Our system consists of three main processes: data generation process (optional), finding frequent patterns process and multilevel association rules generation process.

- The data generation process is to generate synthetic data given some parameters to scrutinize the data to be specific. The parameters are: the number of items in the database, the total number of transactions in the database, the number of levels in representing the items in the database and the number of items in a transaction.

- The purpose of the frequent patterns generation process is to produce a complete set of frequent patterns from the provided transactional database and under the control of the provided minimum support threshold for each of the levels of the concept hierarchy.
- The multilevel association rules can be generated after the generation of the frequent patterns for all levels providing that the minimum confidence threshold for each level is met.

7. Testing and experiments

All of the experiments are conducted on a laptop computer with Core(TM) 2 Duo CPU, T8300@ 2.40GHz, 1.99 GB of RAM and running under XP professional operating system platform. Our system is tested with a number of multilevel synthetic transactional databases of different sizes with the following characteristics:

1. The number of levels present in an itemset code (S) is 3.
2. The maximum number of items in a transaction (T) is 5.

In generating the synthetic transactional databases, we used synthetic generation program similar to the one described in Srikant (1994). The sizes of synthetic data ranged from 1.23 MB to 81.8 MB with range of 30 to 50,000 transactions and the number of items in the databases ranged from 90 to 729. For the sake of comparison, we used the same databases as the ones used in El-Majressi, (2009).

The results of our system, ML-FP-growth, are compared with the results obtained from the system in El-Majressi, (2009), which is an implementation of two algorithms ML-T2 and ML-T2+ for mining multilevel association rules. The comparison is based on the relative performance and efficiency of the two systems. Even though both systems are implemented in Visual Basic6.0, there some major differences between them as follow:

1. The ML-FP-growth system is based on FP-growth algorithm and the other is based on the concept of Apriori-based algorithm.
2. The FP-growth algorithm was developed for single level association rules while the ML-T2 and ML-T2+ algorithms are developed for multilevel association rules.
3. The ML-FP-growth system doesn't use the concept of candidate itemsets generation while the Apriori-based algorithms do.

7.1 First experiment

The synthetic database for this experiment is about 1.21 MB in size with 30 transactions and 90 different items. This experiment had been conducted nine times each with different minimum support values as depicted in table 1.

An illustration plot of the CPU time verses the different minimum support thresholds for the first experiment is shown in figure 1. The obtained results are interpreted as follows:

1. In general, as the minimum support threshold increased, the CPU time decreased for the two systems.
2. Even though the number of frequent itemsets is the same, the ML-FP-growth system took less time.
3. The overall improvements in performance of the ML-FP-growth system for all values of the minimum support threshold ranged from 44.27% to 80.68%. On the average the improvement was about 64.69%.

Generally, the ML-FP-growth system out performed ML-T2 and ML-T2+ subsystems (i.e. in the case of the ML-T2+, the average the improvement was about 36.84%) for this experiment.

Min. Support value	CPU time(seconds)			Number of Frequent Itemsets			Improvements of	
	System	Sub systems		System	Subsystems			
	ML-FP-growth	ML-T2	ML-T2+	ML-FP-growth	ML-T2	ML-T2+	ML-FP-growth over ML-T2	ML-FP-growth over ML-T2+
1.00	9.749	50.577	321.89	793	793	793	5.2	33.0
1.50	4.860	12.582	6.813	92	92	92	2.6	1.4
1.75	4.510	11.922	6.314	92	92	92	2.6	1.4
2.00	4.345	11.875	6.578	92	92	92	2.7	1.5
2.50	1.751	11.141	2.735	37	37	37	6.4	1.6
2.75	1.718	11.094	2.705	37	37	37	6.5	1.6
3.00	1.703	11.063	2.672	37	37	37	6.5	1.6
4.00	1.297	10.874	5.562	20	20	20	8.4	4.3
5.00	1.156	10.812	3.980	14	14	14	9.4	3.4

TAB. 1– *The experimental results of the first experiment (Minimum support values vs. CPU time,number of frequent itemsets and CPU time improvements).*

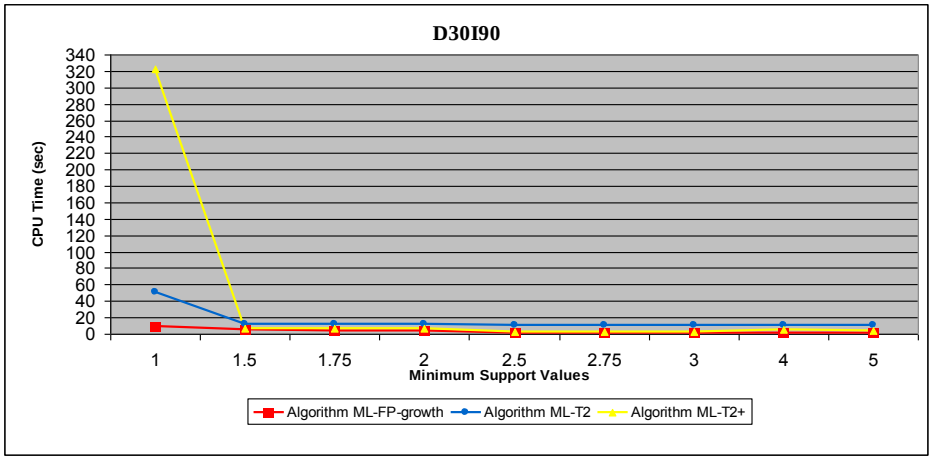


FIG. 1– *An illustration plot of the CPU time vs. Minimum support for the first experiment.*

7.2 Second experiment

The database for this experiment is about 38.30 MB in size and contains 5000 transactions from 620 different items. The minimum support thresholds used here are: 10, 15, 20, 25, 30, 35, 40, 45, and 50. Table 2, depicts the final results of this experiment with the improvements of the ML-FP-growth system over the ML-T2 and ML-T2+ subsystems.

Figure 2 is an illustration plot of the CPU time of the ML-T2 and ML-T2+ subsystems and the ML-FP-growth system verses the different minimum support thresholds. The results are interpreted as follows:

1. When ever the minimum support increased, the CPU time decreased for both systems.
2. Even though the number of frequent itemsets is the same, the ML-FP-growth took less time.
3. The ML-FP-growth system had an average improvement of 69.29% over ML-T2 and an average improvement of 65.66% over ML-T2+.

Min. Support Values	CPU time(seconds)			Number of Frequent Itemsets			Improvements of	
	System	Sub systems		System	Sub systems			
	ML-FP-growth	ML-T2	ML-T2+	ML-FP-growth	ML-T2	ML-T2+	ML-FP-growth over ML-T2	ML-FP-growth Over ML-T2+
10	540.450	2923.384	10248.941	1528	1528	1528	5.4	19.0
15	407.440	2193.973	518.634	911	911	911	5.4	1.3
20	319.928	1625.322	2604.560	644	644	644	5.1	8.1
25	233.629	1273.765	3878.356	426	426	426	5.5	16.6
30	214.803	1379.830	1828.886	336	336	336	6.4	8.5
35	209.182	1229.501	1214.097	293	293	293	5.9	5.8
40	197.034	1057835	858.563	249	249	249	5.4	4.4
45	194.369	1025.778	687.024	218	218	218	5.3	3.5
50	183.194	992.600	613.568	186	186	186	5.4	3.3

TAB. 2. *The experimental results of the second experiment (Minimum support values vs. CPU time, number of frequent itemsets and CPU time improvements).*

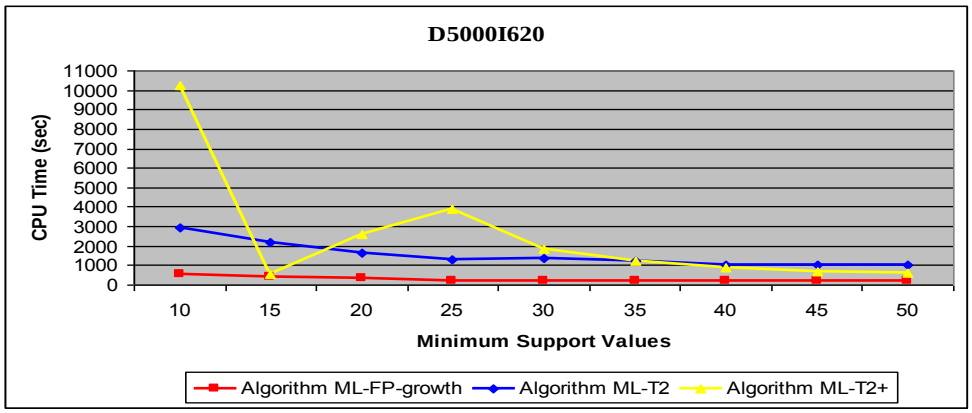


FIG. 2– *An illustration plot of the CPU time vs. Minimum support for the second experiment.*

7.3 Third experiment

The database for this experiment consists of 50,000 transactions that contain 729 different items and the size is about 81.8 MB. In this experiment the minimum supports are: 60, 70, 80, 90, and 100. Table 3, depicts the obtained final results for the CPU time and the improvements of the ML-FP-growth system.

The illustration plot of figure 3 presents the execution time (CPU time) of the ML-T2 subsystem, ML-T2+ subsystem and the ML-FP-growth system verses the different values of the support threshold. The obtained results of table 3 are interpreted as follows:

1. The CPU time decreased for both systems whenever the minimum support increased.
2. The CPU time was better for the ML-FP-growth system than ML-T2 and ML-T2+ subsystems despite they generate the same number of the frequent itemsets.
3. The average improvement of the ML-FP-growth system over the ML-T2 subsystem is about 91.54% and 90.58% over the ML-T2+ subsystem.

Min. Support Values	CPU time(seconds)			Number of Frequent Itemsets			Improvements of	
	System ML-FP-growth	Sub systems		System	Sub systems		ML-FP-growth over ML-T2	ML-FP-growth over ML-T2+
		ML-T2	ML-T2+	ML-FP-growth	ML-T2	ML-T2+		
60	480.737	9212.023	7892.869	2446	2446	2446	19.2	16.4
70	462.265	8112.02	7790.307	1323	1323	1323	17.5	16.9
80	309.104	7928.183	6087.903	1086	1086	1086	25.6	19.7

90	246.281	7227.22	6930.38	1065	1065	1065	29.35	28.14
100	244.891	6301.72	5957.5	1063	1063	1063	25.73	24.33

TAB. 3– *The experimental results of the third experiment (Minimum support values vs. CPU time, number of frequent itemsets and CPU time improvements).*

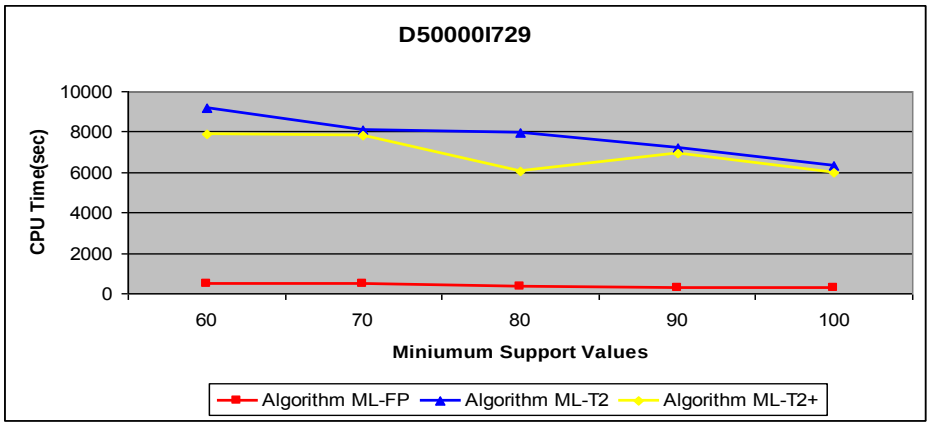


FIG. 3– *An illustration plot of the CPU time vs. Minimum support for the second experiment.*

8. Conclusion and results

The objective of this work was met by extending the FP-growth algorithm from mining single level association rules to mine multilevel association rules. A comparison of the proposed adoption of the ML-FP-growth algorithm with ML-T2 and ML-T2+ algorithms was conducted via a number of experiments to study the performance accuracy and efficiency. From the empirical results, we can conclude the following:

1. The ML-FP-growth had better performance when the minimum support was high.
2. In all of the experiments, the ML-FP-growth system out performed both subsystems ML-T2 and ML-T2+.
3. In all of the experiments, all of the algorithms have produced exactly the same numbers of frequent itemsets nevertheless ML-FP-growth has taken less time to produce such itemsets.
4. The advantages of the ML-FP-growth algorithm are due to:
    a. The number of scans is reduced by the construction of a compact FP-tree that is a representation of the database.
    b. It uses a pattern growth method to avoid the generation of a large number of candidate itemsets.
    c. It used the divide-and-conquer method to reduce the search space by decomposing the data mining task into a set of smaller tasks.

9. Futur work

From the obtained results of our experiments, the authors would like to suggest the following directions for research in the future:

1. To conduct more experiments to evaluate the efficiency of our system with large data sets.
2. To find other evaluating criteria other than Support and Confidence for the multilevel association rules.
3. To extend the ML-FP-growth algorithm in a parallel fashion.

References

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

Agrawal R., Imielinski T. and Swami A., 1993. Data Mining: A performance Perspective. IEEE Transactions on Knowledge and Data Engineering, 914 – 925.

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

Berry M. and Linoff G., 2004. Data Mining Techniques for marketing, sales, and customer relationship management. Second Edition. Wiley, Inc., Indianapolis, Indiana. 44 – 63,

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

Chen M., Han J. and Yu P., 1996. Data Mining: An Overview from a Database Perspective. IEEE Transactions on Knowledge and Data Engineering, 8(6), 866 – 883.

Devedzic V., 2002. Knowledge Discovery and Data Mining in Databases. World Scientific Pub. Co. 10 – 34.

Dunham M., Xiao Y., Gruenwald L. and Hossain Z., 2000. A SURVEY OF ASSOCIATION RULES. Technical report, Southern Methodist University, Department of Computer Science, Technical Report T R00-CSE-8.

El-Majressi A., 2009. An applicative study of two algorithms for multilevel association Rule mining of transactional database. Master Thesis, Department of Computer Science, University of Benghazi, Benghazi, Libya.

Fayyad U., Piatetsky-Shapiro G. and Smyth P., 1996. From Data Mining to Knowledge Discovery in Databases. In Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, USA, 37 – 54.

Fayyad U., Piatetsky-Shapiro G. and Smyth P., 1996. The KDD Process for Extracting Useful Knowledge from Volumes of Data, Communications of the ACM, 27 – 34.

Han J. and Fu Y., 1995. Discovery of Multiple-Level Association Rules from Large Databases. In Proceedings of the 21st International Conference on Very Large Databases. Zurich, Switzerland, 420 – 431.

Han J., 1995. Mining knowledge at multiple concept levels. In CIKM, 19 – 24, 1995.

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

Houtsma M. and Swami A., 1993. Set-Oriented Mining of Association Rules. Research Report RJ 9567, IBM Almaden Research Center, San Jose, California, USA.

Une approche basée-concept pour le routage d'appels

Halima Bahi

Laboratoire LabGED, Département Informatique
Université Badji Mokhtar – Annaba
BP.12, 23000 Annaba
Algérie
halima.bahi@univ-annaba.dz
http://www.labged.net

Résumé. Nous considérons une grande compagnie où on doit transférer des appels téléphoniques (principalement de clients) de sorte que les appels similaires soient transférés vers le même service. Les appels reçus sont transcrits sous forme de documents textes mais les appels se situent dans un contexte multi-langues. Ainsi, pour aborder le problème de routage automatique des appels téléphoniques, il faut d'abord résoudre le problème de représentation du document et compte tenu du multilinguisme, nous proposons de représenter un document par un vecteur de concepts ; lui-même obtenu en faisant la projection de l'appel sur une ontologie du domaine. Pour évaluer cette approche, nous la testons dans le cadre d'une société de télécommunication, qui reçoit des appels de doléances de ces clients et tente de les orienter vers le service concerné.

1 Introduction

La tâche de routage d'appel consiste à diriger un appel de client vers une destination appropriée dans un centre d'appel ou à fournir directement quelques informations simples. Dans les systèmes actuels, une telle interaction est typiquement effectuée par un menu vocal prédéterminé et rigide. Les inconvénients majeurs des menus de navigation pour les utilisateurs sont le temps qu'ils prennent pour écouter toutes les options et la difficulté d'assortir leurs buts à celles-ci.

Pour prendre en charge cette tâche, nous allons supposer qu'il s'agit de classer les messages des clients dans une catégorie ou une autre parmi la liste des destinations (classes ou catégories) possibles.

D'autre part, pour réaliser cette classification et pour prendre en charge de plus d'aspect sémantique du message oral, qui par définition est plus riche en terme d'expressivité qu'un texte, nous utilisons une ontologie car elle permet de gérer les aspects linguistiques et sémantiques au travers de la notion de concept qui dépasse celle de terme. La classification d'un document selon une ontologie revient à l'associer à un, voire plusieurs concepts de cette

ontologie. Ainsi, dans un premier temps, un message reçu est transcris grâce à un module de reconnaissance automatique de la parole (RAP), ensuite il est modélisé sous forme d'un vecteur de mots clés ; ces mots clés représentent les lexicalisations présentes dans l'ontologie. En phase de classification, ce vecteur est projeté sur l'ontologie pour déterminer un ensemble de concepts invoqués dans le message, et finalement, le message sera associé à un concept représentant une destination finale (figure 1).

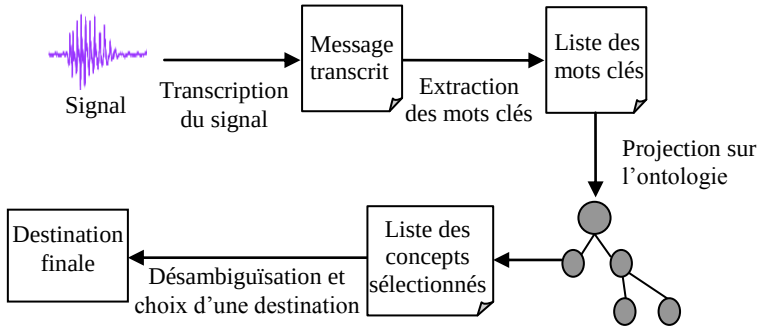


FIG. 1- *Principales étapes du système de routage.*

Pour évaluer cette approche, nous la testons dans le cadre d'une société de télécommunication, qui reçoit des appels de doléances de ces clients et tente de les orienter vers le service concerné.

Le papier est structuré comme suit: dans la deuxième section, nous allons introduire le routage automatique des appels téléphoniques. Dans la troisième section, nous présentons l'ontologie utilisée et les différents éléments conceptuels y afférents. Dans la section 4, nous présentons les étapes suivies lors de la classification avec une évaluation de l'approche. A la fin, nous donnons notre conclusion et nous présentons quelques perspectives à ce travail.

2 Routage des appels téléphoniques

Dans la littérature peu de travaux se sont intéressés au routage des appels téléphoniques par un système automatique, ceci est certainement dû au fait qu'aussi bien l'aspect acoustique que sémantique de la parole spontanée ne sont pas d'un abord facile. En effet, un système de routage d'appels téléphonique se doit d'intégrer deux composantes :

- Un module de reconnaissance vocale pour transcrire ce que dit l'appelant;
- Un module de classification, qui tient compte des énoncés oraux et prédit l'action correcte pour diriger correctement l'appelant.

Dans ce contexte, deux tendances se dessinent: la première consiste en la construction de modèles de langage pour participer à la segmentation du flux de parole en entrée, et la seconde en la représentation basée vecteur de la requête pour permettre une classification de cette dernière.

Beaucoup d'auteurs se sont intéressés à la construction de modèles de langage adéquats pour permettre la segmentation du flux en entrée, nous mentionnons ici deux tendances, la première se base sur des corpus d'appels tels que les travaux de Gorin et al. (1996) qui furent pionniers dans le contexte, et la seconde se base sur des expertises du domaine de l'application (Lee et Chang, 2002).

Dans l'approche basée-vecteur, l'accent est plus mis sur l'aspect classification, où on assume que le la requête en parole est représentée par un « vecteur » de mots et des techniques de reconnaissances de formes sont utilisées pour « router » cet appel (Chu-Caroll et Carpenter, 1999).

3 Une ontologie pour la classification des appels

Pour la construction de l'ontologie *telecom*, nous procédons à un développement de haut en bas. Nous commençons ainsi par le développement des concepts les plus généraux du domaine et on poursuit par la spécialisation des concepts. Nous commençons par créer la classe du concept général *telecom* **(**notre domaine). Puis nous spécialisons la classe *telecom* en créant quelques-unes de ses sous-classes : *c_concepts, c_lexicalisations.*

La classe *c_concepts* englobe toutes les classes de l'application en termes de structure, tandis que la classe *c_lexicatisations* englobe les lexicalisations relatives aux différents concepts. Chaque terme est une entité distincte dans chaque langue qui peut être liée à des concepts ou à d'autres termes et à d'autres variantes du même terme.

3.1 Structure centrale hiérarchique

La classe *c_concepts* est la classe principale de la hiérarchie générale du problème de classification, elle contient deux sous-classe (adsl, telephone), ces deux classes englobent les différents problèmes que peut signaler un client.

Les feuilles de l'arborescence représentent les destinations de notre application, c'est à dire la réclamation pour laquelle un client a appelé (voir figure 2).

3.2 Les lexicalisations

Alors que la structure centrale de l'ontologie de domaine est modélisée sous *c_concept*, les lexicalisations de ces concepts apparaîtront comme des instances de la classe *c_lexicalization*. La modélisation des lexicalisations comme des concepts distincts permettra d'établir des relations entre les différentes lexicalisations qui décrivent un concept. Elle fournira ainsi une sémantique plus riche.

La classe *c_lexicalisation* est modélisée comme une superclasse des *c_noms* et *c_verbes*, ces deux classes contiennent les mots clés qui sont utilisés pour la classification et nous permet d'avoir les différentes déclinaisons d'un mot ou ses synonymes.

Ainsi, à chaque classe de c_concepts peut être associée une ou plusieurs lexicalisations, en d'autre terme un ou plusieurs mots clés. D'autre part, une lexicalisation peut avoir plusieurs termes similaires ou des traductions. On retiendra en particulier, que nous considérons la langue Arabe, la langue française et les dialectes locaux.

4 Le routage d'un appel

Dans tout problème de classification, il est primordial de bien représenter les objets à classer. Dans le cadre de la classification de documents, le plus souvent les documents sont représentés par des vecteurs de caractéristiques. Ces caractéristiques peuvent être des mots qui existent dans le document, mais vu les différentes langues qui coexistent dans un appel téléphonique, nous avons choisi de représenter un document par un vecteur de concepts. Ces concepts sont obtenus en faisant une projection de l'appel sur l'ontologie *telecom*.

4.1 Représentation d'un appel

Il s'agit d'abord d'extraire les informations pertinentes des appels reçus ; cette première étape peut être effectuée par un système de reconnaissance de la parole ou encore par un système de détection de mots clés, où seuls des mots ou des groupes de mots significatifs seront pris en considération (Bahi et Bendib, 2008)(Bahi et Benati, 2009).

Exemple de transcription : Allo j'ai mon numéro de téléphone en dérangement, oui rahou mayemchich

Après la transcription, il, y'a l'extraction du vecteur de caractéristiques ce vecteur contient les mots clés que nous allons utiliser pour la classification. La liste des mots clés a été définie sur la base d'une expertise du domaine. On remarque qu'un mot clé peut être un seul mot ou une séquence de mots, en reconnaissance de la parole, on les appellera plus volontiers des séquences de phonèmes, tandis que dans le domaine la recherche d'information, il s'agit de termes. Pour l'exemple précédent, on obtient le vecteur de termes suivant :

téléphone	dérangement	mayemchich

4.2 Projection sur l'ontologie

Une fois, le vecteur de caractéristiques obtenu, nous allons le projeter sur l'ontologie, c'est-à-dire que nous allons parcourir l'ontologie et associer à chaque terme issu de l'appel le ou les concepts qu'ils caractérisent. Ainsi, pour le vecteur de l'exemple précédent, on obtient le vecteur de concepts suivant :

telephone	derangement	derangement coupure absence_de_connexion

4.3 Le routage

Reconsidérons l'exemple précédent avec le vecteur de concepts produit, les concepts étant : telephone, dérangement, dérangement, coupure et absence_de_connexion.

Le routage se fait sur la base des concepts destinations, dans ce cas nous avons les destinations suivantes : dérangement, coupure et absence_de_connexion. Nous devons aussi tenir

compte du nombre d'apparition d'un concept dans l'appel, ainsi les destinations précédentes seront ainsi pondérées :

- Dérangement (2)
- Coupure (1)
- Absence de connexion (1)

Et de ce fait l'appel sera dirigé vers le concept ayant eu le plus grand poids, en l'occurrence : le dérangement.

4.4 La désambigüisation

Dans les appels réels, il y'a souvent une égalité du poids des destinations candidates, pour cela nous allons faire appel à un module de désambigüisation. Si ce dernier échoue, l'appel est dirigé vers un opérateur humain. Pour illustrer cela, considérons l'exemple suivant : Allo, mon téléphone mayemchich mais rahou khales. Le vecteur mot clé est :

telephone	mayemchich	khales

Le vecteur concepts associé après projection sur l'ontologie est :

telephone	derangement coupure absence_de_connexion	derangement absence_de_connexion

Les concepts destinations deviennent :

- Dérangement (2)
- Coupure (1)
- Absence de connexion (2)

Nous nous trouvons alors dans un cas ambigu, où doit-on diriger cet appel. Pour cela, il existe plusieurs méthodes de désambigüisation en recherche d'information (Baaziz et al., 2005). Pour résoudre cela nous nous sommes inspirés de la méthode de Khan (2000).

Khan (2000) propose une méthode de désambigüisation dans laquelle il préconise que des mots-clés utilisés dans un même contexte désignent ensemble un même concept, même si chaque mot est isolément ambigu. En prolongeant et formalisant l'idée du contexte afin de réaliser la désambiguïsation des concepts, Khan propose un algorithme de désambiguïsation basé sur deux principes: la co-occurrence et la proximité sémantique.

Cet algorithme effectue d'abord la désambiguïsation à travers plusieurs régions de l'ontologie en utilisant le premier principe, et désambiguïse ensuite dans une région particulière en utilisant le second. Une région de l'ontologie détermine un ensemble de concepts se trouvant dans une même zone (voisinage). Dans un premier temps, un ensemble de régions représentant les différents concepts est défini.

Les concepts tels qu'ils apparaissent dans une région sont mutuellement disjoints des autres concepts dans d'autres régions. Une fois les mots-clés appariés avec les concepts de l'ontologie, la région où un plus grand nombre de concepts est sélectionné est identifiée. Cette région sera alors utilisée pour associer des documents à des requêtes utilisateurs.

Nous avons fait le choix de suivre cette méthode de voisinage car nous sommes dans une application où l'ontologie n'est pas très grande et où deux grandes branches se distinguent : le téléphone et l'ADSL.

Reconsidérons l'exemple précédent, et représentons-le sous forme d'arbre de concepts, où les concepts invoqués sont en jaune et ceux candidats sont en bleu. Pour départager les deux destinations candidates (en bleu), l'algorithme de désambigüisation recense les ancêtres des deux concepts et choisira une branche ou l'autre selon le nombre d'ancêtres activés (dont des mots clés ont été détectés dans l'appel). Et de ce fait pour notre exemple, c'est le *dérangement* qui l'emporte.

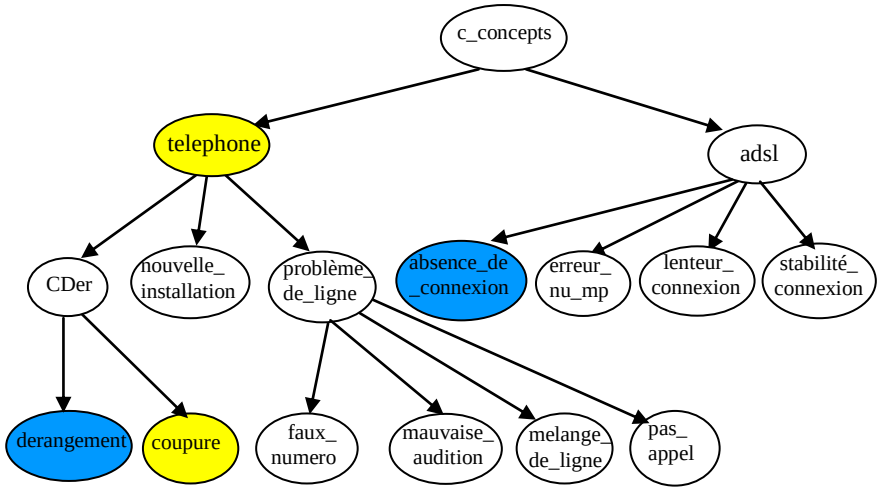


FIG. 2- *Exemple de projection d'un appel sur l'ontologie*e pas indiquer la numérotatioa-lement en minuscules.

4.5 Evaluation

Nous avons évalué notre système de routage sur un ensemble de fichiers transcris, dont voici un exemple :

Documents	Transcriptions
D1	mon téléphone est en dérangement.
D2	la connexion faible ne9der ndire instalation jdida w nbadel le numéro de mon téléphone.
D3	kayen la connexion mai nsite le mot de passe et le nom d'utilisateur.

TAB. 1- *Exemple de documents transcrits*

L'évaluation des systèmes de recherche consiste à vérifier si le système est capable de satisfaire les besoins des utilisateurs réels et potentiels non seulement d'une manière individuelle, mais aussi collective. Ce type d'évaluation est basé généralement sur les mesures du taux de rappel et de précision. Les taux de précision et de rappel sont donnés par les formulations suivantes : Précision = a / b et Rappel = a / c. Où,

a : est le nombre de documents bien routés vers cette destination par le système.

b : est le nombre de documents routés vers cette destination par le système.

c : désigne le nombre de documents d'une destination dans toute la collection.

Classes	Rappel	Précision
dérangement	0,667	0,4
coupure	0,25	0,5
pas_appel	0,5	0,166
mauvaise_audition	1	0,5
mélange_de_ligne	0,333	1
faux_numero	0,333	1
nouvelle_installation	1	0,5
absence_de_connexion	0,2	1
erreur_nu_mp	1	0,666
lenteur_de_la_connexion	0,5	0,5
stabilité_de_la_connexion	0,4	0,5

TAB. 2- *Résultats expérimentaux.*

De ces résultats, on déduit un rappel du système de l'ordre de 0,56 et une précision de : 0,61. On remarque que certaines destinations ont un rappel de 1 ce qui est excellent, ceci est certainement dû au fait que peu d'exemples sont disponibles pour ces réclamations. Effectivement, on appelle rarement pour une nouvelle installation mais aussi les mots clés attachés à ce concept sont très discriminants. La classe *dérangement* nous semble une bonne base d'évaluation car c'est une raison de réclamation très fréquente mais aussi les mots clés qui lui sont associés sont très variés et parfois ils sont communs à plusieurs concepts.

5 Conclusion

Actuellement, la classification des appels téléphonique se fait sur la base d'une correspondance entre des mots clés stockés et les mots clés formulés par les utilisateurs. Le processus de classification des appels devrait être capable de prendre en compte que l'appel téléphonique est dans un contexte spontané et de plus multi-langues. C'est là que la notion d'ontologie intervient, en organisant sous forme de graphe un ensemble de concepts par des relations sémantiques.

Notre proposition est le résultat de la compilation des travaux qui ont posé ce problème. D'abord, nous nous sommes inspirés du travail de (Lee et Chang, 2002) à la DGT, où les mots clés ou termes et les destinations ont été issus d'un travail d'expertise auprès des opérateurs. D'autre part, comme nos prédécesseurs nous avons utilisé l'approche basée-vecteur pour la représentation d'un vecteur. Mais en phase de classification notre apport consiste en

le fait d'utiliser une ontologie. La réalisation permet d'envisager une autre manière d'appréhender ce problème fort intéressant.

Par ailleurs une extension de ce travail consiste en l'utilisation de l'ontologie comme outil de désambiguïsation lors de l'analyse acoustique.

Références

Baaziz M., Boughanem M., Aussenac-Gilles N. et Chrisment C. (2005). Semantic Cores for Representing Documents in IR, *Actes de 20th ACM Symposium on Applied Computing*, Santa fe, New mexico.

Bahi H., Bendib I. (2008). Presentation of password verification system in Arabic, *Actes de la conference maghrébine MCEAI'08* , Oran, Algérie.

Bahi H., Benati N. (2009). A new keyword spotting approach, *International Conference on Multimedia Computing and Systems ICMCS'09*, Ouarzazate, Maroc.

Chu-Carroll J., Carpenter B. (1999). Vector-based Natural Language Call Routing, *Computational Linguistics*, 25(3), p.361-388.

Gorin A.L., Parker B.A., Sachs R.M. et Wilpon J.G. (1996). How may I help you ?, *Actes de IVTTA*, Basking, Ridge.

Khan L. R. (2000). *Ontology-based Information Selection*, Thèse PhD, Université de la Californie du Sud.

Lee C.-J. et Chang J. S. (2002) An Operator Assisted Call Routing System, *Actes de 16th Pacific Asia Conference on Language, Information and Computation*, Jeju, Corée.

Summary

We consider the problem of routing telephone calls in a big company, where similar calls should be routed to the same destination. Received calls are transcribed as textual documents but they have the specificity to be multilingual. To solve the problem of Multilanguage to allow their thematic indexation, the transcript documents must be considered as concept vectors achieved by a projection on a domain ontology. To evaluate this approach, we consider a company of telecommunications, which receives calls from its customers and routes them to the appropriate service.

Etat de l'art des méthodes de construction d'ontologies à partir d'un corpus de texte

Anis Assas*

*Institut Supérieur des Etudes Technologiques de Djerba. Département Technologies de l'Informatique
ISET Djerba, Route Houmt Souk, Midoun 4116 Djerba, Tunisie
assas_anis@yahoo.fr

Résumé. Dans le domaine des sciences informatiques et des technologies de l'information, les ontologies suscitent un grand intérêt pour la communauté scientifique couvrant plusieurs axes de recherche. Les ontologies ont pour objectif de représenter des connaissances au sujet d'un domaine particulier sous un formalisme rationnel pour qu'elles soient traduites dans un langage de spécification interprétable par la machine.
La construction d'une ontologie peut se faire à partir de plusieurs sources d'information tels que les textes, les dictionnaires, les bases de connaissances, les données semi structurées, les schémas relationnels, les bases de données...
Dans cet article, nous nous intéressons particulièrement à la construction d'ontologies à partir de texte. Nous allons commencer alors par définir une ontologie, présenter ses différents types, ses composants ainsi que les différents modèles, outils et environnements de représentation. Nous ferons également un tour d'horizon sur les différentes approches de construction d'une ontologie à partir d'un corpus de texte.

1 Introduction

La représentation des connaissances consiste à trouver les outils et les procédés destinés à représenter et à organiser le savoir humain pour pouvoir l'utiliser et le partager. C'est dans ce cadre que s'inscrivent les ontologies qui constituent une théorie de la connaissance permettant de formaliser et de synthétiser les connaissances issues d'un domaine particulier (Chandrasekaran et al.1999).

Une ontologie peut être alors considérée comme un outil d'ingénierie, composé d'un certain vocabulaire afin de décrire d'une manière explicite une certaine réalité ou encore un outil de communication et de partage efficace permettant la réutilisation et la mise à niveau du savoir.

Durant les deux dernières décennies, l'analyse des textes, en ingénierie des connaissances, a pris un grand essor avec le déploiement des ontologies. Porteurs de connaissances stabilisées et consensuelles, les documents textuels, focalisés sur un champ d'application

bien déterminé, constituent ce qu'on appelle corpus qui peut être très utile pour la construction d'ontologies.

Cependant, l'élaboration de telles ontologies reste une tâche fastidieuse allant de la spécification des besoins et du contexte de l'ontologie à sa maintenance en passant par l'extraction des éléments importants disséminés dans le texte, la conceptualisation et l'opérationnalisation de l'ontologie. Pour ce faire, plusieurs approches et méthodologies ont été proposées tirant profit des méthodes terminologiques, des outils de TAL (Traitement Automatique de la Langue), des méthodes statistiques et d'apprentissage.

2 Les ontologies

2.1 Définitions

Au sens philosophique, une ontologie est une partie de la métaphysique qui s'intéresse à la notion d'existence, aux catégories fondamentales de l'existant et étudie les propriétés les plus générales de l'être (Aristote – 384-322 av J-C).

Dans le domaine de l'intelligence artificielle, on définit une ontologie comme étant une spécification explicite, formelle d'une conceptualisation partagée (Gruber, 1993).

L'ontologie formalise les connaissances issues d'un domaine du savoir et qui sont validées par une communauté scientifique établie.

Qu'elles soient sous forme de thésaurus, de vocabulaire contrôlé, de réseau sémantique, de taxonomie ou de modèle conceptuel, les ontologies servent à la représentation des connaissances, la recherche, l'extraction, l'intégration et l'interopérabilité entre les sources d'information et de connaissances.

Sur le plan pratique, une ontologie est formée de concepts, relations, contraintes et règles d'inférence qu'une base de connaissance peut utiliser (Maedche et Staab, 2000).

2.2 Types d'ontologies

Dans la littérature, on distingue différentes classifications d'ontologies (Mizoguchi et al., 1995), (Van Heijst et al., 1997), (Guarino, 1997), (Gómez-Pérez et al., 2003). D'une manière générale et en se référant au niveau d'abstraction de leur usage, on spécifie, comme le montre la figure 1, les principaux types d'ontologies suivants :

Ontologies génériques : Ces ontologies, appelées aussi, ontologies globales, haut niveau, top level ou upper level, décrivent des concepts généraux, indépendants d'un domaine ou d'un problème particulier comme celui de l'espace, du temps, de la matière ou de l'évènement, etc. Elles sont conçues pour être utilisées dans des emplois divers et pour servir une large communauté.

Exemples : Sowa, Cyc, Dolce, BFO, GFO, Proton, Sumo…

Ontologies de domaine : Elles permettent de spécifier un point de vue sur un domaine particulier. Elles consistent à fournir le vocabulaire des concepts en relation avec un domaine générique comme la médecine, la physique, la géographie ou l'écologie... Les concepts d'une ontologie de domaine sont souvent définis comme une spécialisation des concepts des ontologies génériques.

Exemple : L'ontologie « Enterprise Ontology » qui décrit le domaine de l'entreprise.

Ontologies de tâche ou de méthode : Ces ontologies permettent de décrire un vocabulaire en relation avec une tâche ou une activité générique (conception, diagnostic, vente, planification, construction, évaluation…). Elles offrent un champ lexical standardisé de termes utilisés pour résoudre les problèmes associés à des tâches particulières. Au niveau de ce type d'ontologies, on spécifie une formalisation de la réalisation d'une tâche particulière d'une manière indépendante d'un domaine ou d'un traitement donné.

Ontologies d'application : Elles permettent de définir la structure des connaissances nécessaires à la réalisation d'une tâche particulière. Elles définissent les concepts qui dépendent à la fois d'un domaine particulier et d'une tâche particulière. Ces ontologies peuvent être alors perçues comme une spécialisation à la fois d'une ontologie du domaine et d'une ontologie de tâche ou de méthode.

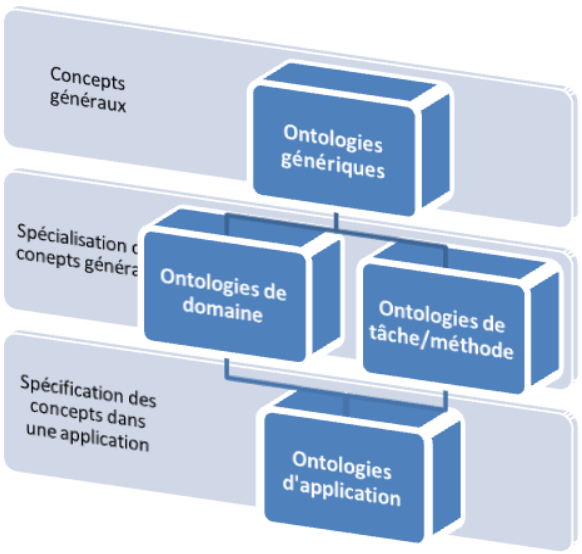


FIG. 1 – *Types d'ontologies*.

On appelle ontologies spécialisées celles de domaine, de tâche ou de méthode et d'application.

(Van Heijst et al. 1997) ajoute aux différentes ontologies ainsi présentées un autre type telles que les ontologies de représentation permettant de spécifier un formalisme de description qui fournit une structure de représentation et des primitives pour décrire les concepts des ontologies de domaine et des ontologies génériques.

Exemple : La Frame Ontology d'Ontolingua.

On distingue également les ontologies légères et les ontologies riches ou lourdes (Studer et al. 1998).

Ontologie légère : elle comprend des concepts, des types atomiques, une hiérarchie IS-A entre les concepts et des relations entre les concepts mais elle n'inclut pas d'axiomes.

Ontologie riche : c'est une ontologie légère qui comprend en plus des contraintes de cardinalité, une taxonomie de relations, des axiomes/héritages sémantiques et elle suppose l'existence d'un système d'inférence.

2.3 Modèles de représentation des ontologies

On distingue deux grands modèles de représentation des ontologies à savoir les modèles conceptuels et les modèles logiques.

Modèles conceptuels : basés sur l'approche BD (Bases de Données), ces modèles exploitent les propriétés des schémas « Entités-Relations » ainsi que les fondements d'UML.

Modèles logiques : ils se reposent sur l'approche IA (Intelligence Artificielle) et les modèles de représentation des connaissances tels que les réseaux sémantiques, les graphes conceptuels, les frames, les logiques de description,…

Parmi ces modèles, les logiques de description ou encore les logiques descriptives (LD) s'avèrent un moyen efficace pour représenter les ontologies. Il s'agit en fait d'une famille de langages de représentation de connaissances, caractérisée par une organisation hiérarchique incluant une boîte terminologique ou conceptuelle (T-Box) pour la description des concepts et des rôles et une boîte assertionnelle (A-Box) qui contient les individus du monde réel (instances). On dispose d'une syntaxe et une sémantique bien définies offrant des procédures d'inférences et de vérification décidables ainsi que des implémentations qui sont disponibles en tant que raisonneurs sur les logiques descriptives tels que FaCT, Racer, KAON2,…

2.4 Outils et environnements des ontologies

Plusieurs langages peuvent être utilisés afin de représenter une ontologie. Parmi les principaux langages, on cite :

- GML (Geography Markup Language)
- RDF (Ressource Description Framework)
- RDFs (RDF Schema)
- DAML+OIL (DARPA Agent Markup Language + Ontology Interchange Language)
- OWL (Web Ontology Language)
- KIF (Knowledge Interchange Format)
- OWL/DL (OWL Descriptive Logic)

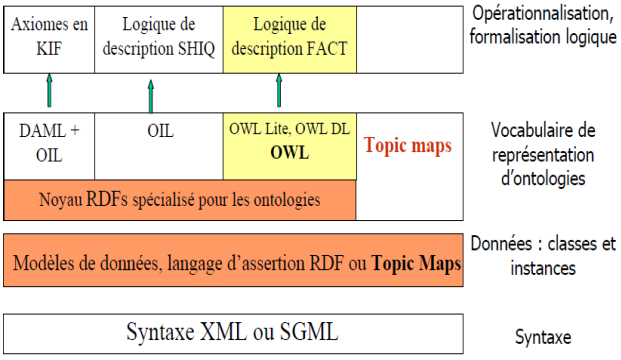


FIG. 2 – Langages de modélisation d'ontologies.

Quant aux environnements et outils de représentation d'ontologies, ils peuvent être regroupés comme suit :

- Outils de construction d'ontologies dépendants de formalisme de représentation
  - Ontolingua (Farquhar et al., 1997) : c'est un serveur d'édition d'ontologies. Il est défini également comme étant un langage de formalisme d'ontologies en tant qu'une extension du langage KIF (Knowledge Interchange Format).
  - OntoSaurus (Swartout et al., 1997) : c'est un outil développé à l'université de Southern California utilisant des formulaires HTML pour éditer les ontologies ainsi que le langage LOOM pour la représentation des connaissances.
  - WebOnto (Domingue, 1998) : c'est une application web conçue à l' « Open University » pour pouvoir développer d'une manière collaborative.des ontologies basées sur le langage OCML (Operational Conceptual Modelling Language).
  - OilEd (Bechhofer et al., 2001) : il s'agit d'un éditeur développé à l'université de Manchester pour la construction de petites ontologies basées sur OIL et DAML+OIL et qui peuvent être testées par la suite à l'aide de moteur d'inférence FaCT.
- Outils de construction d'ontologies indépendants de formalisme de représentation
  - Protégé (Eriksson et al., 99) : cet outil, développé à l'Université de Stanford et considéré comme l'environnement le plus utilisé, est basé sur le modèle de frames et il permet la construction d'ontologies qui peuvent être exportées à plusieurs formats tels que RDF(S), OWL, XML Schema,…
  - ODE et WebODE (Blazquez et al., 1999) : l'outil ODE (Ontology Design Environment) permet la construction d'ontologies à base de frames. Quant à WebODE, il s'agit de l'adaptation de l'ODE pour le web.
  - OntoEdit (Sure et al., 2002) : développé à l'université de Karlsruhe, cet éditeur offre une interface graphique et un ensemble de plugins permettant le codage et l'évolution des ontologies.

2.5 Cycle de vie d'une ontologie

En se référant aux différentes méthodologies trouvées dans la littérature telles que la méthode de Enterprise Ontology de (Uschold et King's 1995), celle de KACTUS de (Bernaras et al, 1996), SENSUS (Swartout et al., 1997), Methontology (Gómez-Pérez et al., 2003) ou TOVE (TOrento Virtual Enterprise) (Grüninger et Fox, 1995), la conception d'une ontologie passe généralement, comme la montre la figure 3, par les principales phases suivantes :

- Spécification des besoins et identification des objectifs et du contexte de l'ontologie.
- Conception et construction de l'ontologie : cette phase consiste à repérer et organiser les termes en utilisant les méta-catégories (concepts, relations, attributs, …). Elle inclut également le codage et l'intégration d'ontologies existantes.
- Utilisation et raffinement : il s'agit de la mise à jour de la connaissance avec les experts du domaine et l'étape de la formalisation.
- Evaluation de l'ontologie : cette étape revient à tester si l'ontologie satisfait les spécifications dans son cadre d'environnement d'application.
- Evolution et maintenance de l'ontologie.

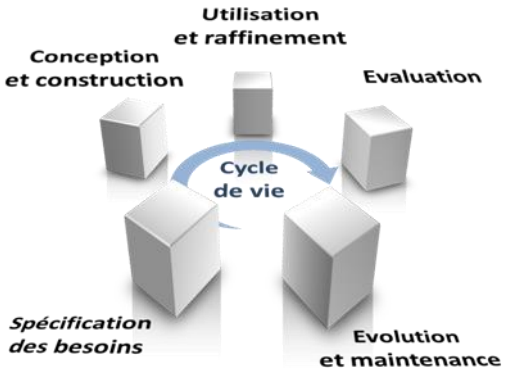


FIG. 3 – Cycle de vie d'une ontologie.

3 Construction d'ontologies à partir de texte

3.1 Corpus de texte

Un corpus est défini comme étant une collection de données langagières sélectionnées et organisées selon des critères linguistiques et extra linguistiques explicites pour servir d'échantillon d'emplois déterminés d'une langue (Habert et al. 1997). Qu'il soit sous forme d'articles scientifiques, documents techniques ou didactiques, retranscriptions d'entretiens, récits de voyages ou autres, un corpus est un recueil de textes ou de paroles, liés à un domaine applicatif bien déterminé et qui sont sélectionnés de la part de l'ingénieur de la connaissance pour un ultime objectif précis.

La légitimité du recours aux corpus de textes en vue de construire les ontologies peut être justifiée par le fait que les textes soient d'une part disponibles, faciles à acquérir et à y accéder d'une manière rapide et d'autre part parce qu'ils sont porteurs de connaissances stabilisées, exprimées dans une langue naturelle qui peuvent être exploitées et partagées par des larges communautés de pratiques.

C'est dans cette optique que la génération d'une ontologie à partir d'une base textuelle permettra de chercher, extraire et traiter au mieux l'information utile et pertinente (filtrage par mots clés, occurrences, classification de documents, lisibilité, maintenance et enrichissement des modèles, …).

Le passage du texte à l'ontologie commence par en extraire les éléments importants.

3.2 Composants de l'ontologie

La construction d'une ontologie à partir d'un corpus de textes repose essentiellement sur les éléments importants suivants :

Candidats termes : Il s'agit d'un mot ou d'une suite de mots susceptible d'être retenue comme entrée (terme, concept) dans une ressource terminologique. L'extraction terminologique est le repérage du vocabulaire conceptuel qui repose sur des méthodes statistiques, linguistiques ou mixtes que nous détaillerons dans la suite de l'article.

Entités nommées : Ce sont des expressions textuelles qui s'apparentent à des noms propres pour désigner des entités référentielles. Elles peuvent avoir aussi la forme d'expressions numériques, url, … On a eu recours aux entités nommées pour peupler les ontologies avec des instances de concepts et pouvoir repérer en corpus les unités textuelles correspondant à des entités pertinentes et les associer à un concept.

Exemples : noms de personnes, de lieux, de compagnies, de médicaments …

Relations : Les relations peuvent avoir deux formes à savoir lexicales et spécialisées. Les relations lexicales ou linguistiques sont formalisées par les experts du domaine telle que l'hyperonymie (relation de subsomption), la synonymie (relation d'équivalence), la méronymie (relation de « partie à tout ») ou l'antonymie (relation pour exprimer les contraires). Quant aux relations spécialisées, elles sont dépendantes d'un domaine particulier.

Exemple : la relation « enseigner à » ayant pour domaine « professeurs » et pour co-domaine « école ».

Classes sémantiques : Il s'agit d'un groupement de mots sémantiquement proches en étudiant leurs distributions au sein du corpus. Les classes sémantiques peuvent être interprétées parfois comme étant des concepts.

Exemples : fenêtre de mots, contexte syntaxique, …

Axiomes ou règles : Il s'agit de modéliser les connaissances en tant qu'axiomes ou règles dans une ontologie en ajoutant des contraintes sur les relations.

3.3 Principales étapes pour la construction d'ontologies à partir de texte

Les étapes communes à la plupart des méthodes de construction d'ontologies peuvent être représentées suivant quatre couches :

Couche corpus : cette phase consiste essentiellement à recueillir et constituer un corpus des documents nécessaires faisant l'objet d'étude.

Couche terminologique : c'est l'étape de l'analyse linguistique du corpus qui consiste à identifier les termes et établir des relations entre eux.

Couche termino-conceptuelle : c'est la phase de conceptualisation qui a pour but la représentation des termes et des relations ainsi que le passage au modèle conceptuel.

Couche ontologique : à ce niveau, on s'intéresse à la formalisation de l'ontologie via les langages de modélisation adéquats (RDF, OWL, …).

3.4 Approches et outils pour la construction d'ontologies à partir de texte

Selon la typologie de (Bourigault & Aussenac-Gilles 2003), on distingue deux grandes étapes pour la modélisation ontologique à savoir l'extraction des candidats termes et leur structuration.

Extraction des candidats termes : Plusieurs outils ont été proposés afin d'extraire les candidats termes à partir d'un texte en faisant recours à une analyse linguistique. Parmi ces outils, on peut citer : Nomino, (anc. Termino) (David & Plante, 1990), Acabit (Daille 1994), OntoLearn (Velardi et Al. 2001), Lexter (Bourigault, 1994), Syntex (Bourigault & Fabre, 2000), FASTR (Jacquemin, 1997), etc.

L'extraction des candidats termes peut se faire également en se basant sur des méthodes statistiques. Il s'agit des méthodes qui s'appuient sur des mesures de similarités pour proposer des associations lexicales récurrentes telle que l'information mutuelle ou le coefficient de Dice pour les mesures d'associations binaires ou encore le système SENTA pour les associations N-aires (calcul de probabilités conditionnelles, indice de cohésion lexicale entre différents mots). Les méthodes statistiques font aussi recours aux égalités approximatives entre mots et observation des répétitions des patrons.

Structuration de termes et regroupement conceptuel : cette étape s'appuie sur des approches structurelles au moyen d'informations syntaxiques permettant d'établir des relations hiérarchiques sous forme d'un réseau terminologique ou encore au moyen d'informations morphologiques, ou par le biais d'analyse de la structure lexicale des termes ou via des connaissances sémantiques (se servir de synonymes et des connaissances sémantiques pour inférer des relations sémantiques entre termes).

La structuration des termes peut se baser également sur des approches contextuelles distributionnelles caractérisées par des dépendances syntaxiques récurrentes ou des fenêtres de mots comme contexte en utilisant des approches statistiques ou des algorithmes de classification (K plus proches voisins, clustering hiérarchique, …). Le groupement des termes peut être fait aussi via des approches contextuelles par patrons lexico-syntaxiques basées sur les définitions des relations sémantiques et sur l'observation de séquences en corpus véhiculant les relations ainsi que la schématisation du contexte lexical et syntaxique des unités lexicales en relation.

Plusieurs autres méthodologies ont été conçues pour la construction d'ontologies à partir de textes. (Mondary et Al. 2008) les répartit en deux grandes familles : les approches terminologiques et les approches non terminologiques.

- Approches terminologiques

Il s'agit d'approches qui s'appuient sur les outils de TAL pour l'extraction des éléments remarquables du texte ainsi qu'à la formalisation de l'ontologie par l'utilisateur.

Parmi les méthodologies appartenant à cette famille, on cite :

Terminae (Aussenac-Gilles et Al. 2008) **:** Il s'agit d'une approche manuelle assistée qui commence par extraire les éléments importants du texte via un extracteur de termes, un concordancier, un analyseur syntaxique, etc. On attribue par la suite une fiche terminologique identifiant pour chaque terme les informations lexicales issues du corpus et en y associant un ou plusieurs termino-concepts qui représentent chacun un sens. Ces derniers dotés de fiches termino-conceptuelles sont structurés manuellement en un réseau termino-ontologique qui est formalisé par la suite en ontologie.

TextToOnto (Maedche et Staab 2000) **:** Il s'agit d'une extension de la plateforme KAON (Karlsruhe Ontology) dédiée à la construction d'ontologie à partir de texte, constituée d'un ensemble de modules pour extraire des candidats termes, des instances de concepts, des règles d'association, des relations entre concepts et une hiérarchie de concepts. Cette approche consiste à enrichir une ontologie existante et comparer deux ontologies entre elles. L'extraction des relations de subsomption entre les candidats-concepts se repose sur la méthode d'analyse des concepts formels ou sur des approches à base de connaissances (patrons d'extraction et wordnet).

Upery (Bourigault 2002) **:** Il s'agit d'un prolongement de l'analyseur syntaxique Syntex. C'est un outil d'analyse distributionnelle exploitant l'ensemble des données présentes dans le réseau de mots et syntagmes construits par Syntex et se basant sur un rapprochement des

syntagmes réduits (mots simples) et maximaux (mots composés les plus longs possibles) via des mesures de proximités.

- Approches non terminologiques

Parmi ces approches, qui se basent plutôt sur des méthodes d'apprentissage rendant la phase de conceptualisation à partir de texte automatique ou semi automatique, on cite :

Text2Onto (Cimiano & Völker, 2005) **:** Il s'agit d'une approche automatique qui utilise l'architecture GATE pour prétraiter le texte. Cette approche, composée de modules pour extraire des candidats termes, des relations entre concepts et des instances de concepts, offre une boîte à outils où l'ontologue peut choisir lui-même les algorithmes d'extraction à utiliser (TF.IDF, fréquence,…). En sortie, on illustre une mesure de confiance (entre 0 et 1) appelée POM (Probabilistic Ontology Model). Le système KASO est couplé à Text2Onto pour affiner le POM à l'aide des méthodes d'acquisition des connaissances (mise en échelle, …).

OntoGen (Fortuna et al. 2006) **:** Il s'agit de chercher les mots clés les plus représentatifs et leur associer automatiquement des instances (documents). Cette approche exploite des algorithmes de fouille de textes non supervisés (k-means, LSI) ou supervisés (svm active learning). C'est une approche itérative, semi automatique où l'expert est guidé dans une démarche descendante pour construire les concepts et choisit quelles zones de l'ontologie suppose affiner.

Asium (Faure et al. 1998) **:** Il s'agit d'une méthode d'apprentissage interactive non supervisée pour la construction de hiérarchies ascendantes. Elle utilise une analyse syntaxique pour extraire les classes de base qu'elle regroupe par niveau en utilisant un algorithme interactif basé sur la distance Asium (prend en compte la fréquence des éléments partagés entre 2 classes, le nombre d'éléments en commun et la cardinalité des classes). On effectue une généralisation des classes en ajoutant des connaissances obtenues par induction qui doivent être validés par l'utilisateur.

4 Conclusion

En résumé, les ontologies présentent un moyen efficace pour la modélisation des connaissances relatives à un domaine particulier. Consensuelles, cohérentes, partageables et réutilisables, les ontologies ont pu être un artefact d'ingénierie et un excellent outil de communication et de partage de la compréhension entre les personnes d'une communauté d'un domaine donné ou des agents logiciels.

A l'issue du survol des méthodes de construction d'ontologies à partir de textes présentées, il semble clair que la plupart des approches proposées (qu'elles soient linguistiques, statistiques ou mixtes) permettent une analyse syntaxico-sémantique et une compréhension superficielles du texte (où on établit parfois des relations sémantiques invalides). Elles portent essentiellement sur la distribution des mots dans le corpus, leur co-présence et leur structuration sans tenir compte de la dynamique conceptuelle. D'où, la propriété statique du produit ontologique. L'attention peut être alors portée sur la génération d'une ontologie à partir d'un corpus de texte permettant de :

- Dépasser les frontières terminologiques pour une conceptualisation des connaissances basée plutôt sur des fondements sémantiques.
- Chercher l'aspect dynamique dans le texte : construire une ontologie offrant une couverture sémantique étendue.

- Modéliser des connaissances qui concernent plutôt les phénomènes, les évènements, les processus dotés d'une dynamique spatiale et temporelle.
- Extraire de la sémantique tout en tirant profit de l'évolution des concepts en fonction de l'espace et/ou du temps. Ceci revient à extraire des relations spatio-temporelles entre les concepts.

Références

Aussenac-Gilles N., S. Despres et S. Szulman (2008). The Terminae method and platform for ontology engineering from texts. Dans Proceeding of the 2008 conference on Ontology Learning and Population. Bridging the Gap between Text and Knowledge, pages 199–223. IOS press, 2008.

Bechhofer S., I. Horrocks, C. Goble, R. Stevens. (2001). OilEd : a Reasonable Ontology Editor for the Semantic Web. Proceedings of KI2001, Joint German/Austrian conference on Artificial Intelligence, September 19-21, Vienna. Springer-Verlag LNAI Vol. 2174, pp 396-408. 2001.

Bernaras A., I. Laresgoiti, J. Corera (1996). Building and Reusing Ontologies for Electrical Network Applications. Proceedings of the European Conference on Artificial Intelligence (ECAI'96). ECAI 96. 1996.

Biebow B. et S. Szulman (1999). TERMINAE. A linguistics-based tool for the building of a domain ontology. Knowledge Acquisition, Modeling and Management, 1621:49–66, 1999.

Blazquez M., M. Fernandez-lopez, J.M. Garcia-Pinar, A. Gomez-Pérez (1998). Building Ontologies at the Knowledge Level using the Ontology Design Environment. In Proceedings of the Workshop on Knowledge Acquisition, Modelling and Management: KAW'98, Banff, Canada.

Bourigault D. et N. Aussenac-Gilles (2003). Construction d'ontologies à partir de textes, Dans. Actes de la conférence TALN 2003, Batz sur-Mer

Bourigault D. et C. Fabre (2000). Approche linguistique pour l'analyse syntaxique de corpus. Cahiers de grammaire, 25:131–151, 2000.

Bourigault D. (2002). Upery. un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. Dans Actes des 9emes journées sur le Traitement Automatique des Langues Naturelles, pages 75–84, 2002.

Bourigault D. (1999). Lexter, Un logiciel d'Extraction de Terminologie. Application à l'acquisition de connaissances à partir de textes. Thèse d'informatique, Ecole des Hautes Etudes en Sciences Sociales, Paris, 1994.

Chandrasekaran B., J.R Josephson, V.R. Benjamin (1999): What are Ontologies, and why do we need them ? Intelligent Systems and their Application, Volume 14, Issue 1, p.20-26. 1999.

Cimiano P. and J. VÄolker. (2005). Text2onto - a framework for ontology learning and data-driven change discovery. In A. Montoyo, R. Munoz, and E. Metais, editors, Proceedings

of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB), volume 3513 of Lecture Notes in Computer Science, pages 227{238, Alicante, Spain, JUN 2005. Springer.

Daille (1994). "Study and Implementation of Combined Techniques for Automatic Extraction of Terminology." In: Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL 1994).

David S. et P. Plante (1990). « De la nécessité d'une approche morpho-syntaxique dans l'analyse des textes », Intelligence artificielle et sciences cognitives au Québec 3(3), pp. 140-154

Domingue J. (1998). Tadzebao and WebOnto: Discussing, Browsing, and Editing Ontologies on the Web. Proceedings of the 11th Workshop on Knowledge Acquisition, Modelling and Management: KAW'98, Banff, Canada.

Eriksson H., R. Fergerson, Y. Shahar and M.A. Musen (1999). Automatic Generation of Ontology Editors. Proceedings of the Workshop on Knowledge Acquisition, Modelling and Management: KAW'99, Banff, Canada.

Farquhar A., R. Fikes, R. and J. Rice (1997). The Ontolingua Server: Tool for Collaborative Ontology Construction. International Journal of Human Computer Studies, 46(6), pp. 707-728.

Faure D., C. Nedellec, and C. Rouveirol. (1998). Acquisition of semantic knowledge using machine learning methods: The system "asium". Technical report. Université Paris Sud, 1998.

Fortuna B., Marko Grobelnik et Dunja Mladenic(2006). Semi-automatic data driven ontology construction system. Dans Proceedings of the 9th International multiconference Information Society IS-2006, Ljubljana, Slovenia, 2006.

Gómez-Pérez A., M. Fernández-López, and O. Corcho, (2003). Ontological Engineering: with Examples from the Areas of Knowledge Management, E-commerce and the Semantic Web, Springer-Verlag, London. 2003

Gruber T.R. (1993). "A translation approach to portable ontology specifications". Knowledge acquisition, 5:199, 1993.

Grüninger M. and M.S. Fox (1995). Methodology for the Design and Evaluation of Ontologies. Proceedings of the IJCAI-95 Worshop on Basic Ontological Issues in Knowledge Sharing.

Guarino N. (1997). Understanding, building and using ontologies. International Journalof Human Computer Studies, 46:293–310, 1997.

Habert B., A.Nazarenko et A. Salem (1997). Les linguistiques de corpus, Armand Collin, 1997.

Jacquemin C. (1997). Variation terminologique. Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus. Mémoire d'habilitation à diriger des recherches en informatique, Université de Nantes, 1997.

Maedche A. et Steffen Staab (2000). "Discovering conceptual relations from text." Dans ECAI 2000, Proceedings of the 14th European Conference on Artificial Intelligence, 2000. IOS Press, Amsterdam, 2000.

Mizoguchi R., J. Vanwelkenhuysen, M. Ikeda. (1995). Task Ontology for reuse of problem solving knowledge. In: Mars N (ed) Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing (KBKS'95). University of Twente, Enschede, The Netherlands. IOS Press, Amsterdam, The Netherlands, pp 46–57,1995.

Mondary T., S. Després, A. Nazarenko et S. Szulman (2008). « Construction d'ontologies à partir de textes. la phase de conceptualisation ». Dans Actes des 19èmes Journées francophones d'Ingénierie des Connaissances (IC2008), Pages 87–98, Nancy, France, juin 2008.

Studer R., V.R. Benjamins, et D. Fense1 (1998). Knowledge engineering: Principles and methods. Data Knowledge Engineering, 25(1-2).161-197, 1998.

Sure Y., J. Angele and S. Staab (2002). OntoEdit: Guiding Ontology Development by Methodology and Inferencing. In Proceeding On the Move to Meaningful Internet Systems, Confederated International Conferences DOA, CoopIS and ODBASE. Pages 1205-1222. Springer-Verlag London 2002.

Swartout B., R. Patil, K. Knight and T. Russ (1997). Towards Distributed Use of Large-Scale Ontologies. Spring Symposium Series on Ontological Engineering, Stanford University, CA, p. 138-148.

Uschold M., M. King (1995). Towards a methodology for building ontologies, in Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI'95, 1995.

Van Heijst G., Guss Schreiber et Bob J. Wielinga (1997). Using explicit ontologies in KBS development. International Journal of Human Computer Studies, 46:183–292, 1997.

Velardi P., Michele Missikoff, and Roberto Basili. (2001). "Identification of Relevant Terms to Support the Construction of Domain Ontologies." In: Proceedings of the workshop on Human Language Technology and Knowledge Management (HLTKM 2001)

Summary

In the field of computer sciences and information technology, ontologies are of great interest in the scientific community regarding several areas of research. Ontologies are intended to represent knowledge about a particular field in a rational formalism and are translated into a specification language interpretable by machine.

The construction of an ontology derive from several sources such as texts, dictionaries, knowledge bases, semi-structured data, relational schemas, data bases...

In this article, we are particularly interested in the construction of ontologies from text. In this context, we will start with defining an ontology, present types, its components and the different models, tools and environments for representing ontologies. We will also shed light on the different methods and approaches for building ontologies from a text corpus.

Une méthode pour la construction des ontologies multi-points de vue en logique de descriptions

Mounir Hemam*, Zizette Boufaida**

* Département d'informatique,
Université de Khenchela, Algérie
mounir.hemam@gmail.com
** Laboratoire LIRE,
Département Technologie des Logiciels et Systèmes d'Information,
Université Constantine 2, Algérie
zboufaida@gmail.com

Résumé. Dans cet article, nous nous intéressons au problème de développement d'une ontologie dans une organisation hétérogène en prenant en compte différents points de vue et terminologies des communautés au sein de cette organisation. Une telle ontologie, que nous appelons ontologie multi-points de vue, confère à un même univers de discours plusieurs descriptions partielles, chacune est réservée à une tâche, une application ou un groupe de personnes particulier. De plus, les différentes descriptions partielles partagent à un niveau global des éléments ontologiques constituant un consensus entre les différents points de vue.
Pour fournir des éléments de réponse à cette problématique nous définissons un modèle de connaissances multi-points de vue fondé sur les notions de point de vue et d'ontologie. Ce modèle de connaissances multi-points de vue est utilisé pour la formalisation de l'ontologie multi-points de vue, en logique de descriptions.

1 Introduction

Une ontologie contient un vocabulaire formalisé regroupant pour une discipline donnée, l'ensemble des concepts, et de leur relations. Son développement croissant en Intelligence Artificielle (IA) vient de son intérêt pour associer du sens à des ressources textuelles, pour localiser et gérer des connaissances dans diverses applications. Depuis la naissance de l'IA, plusieurs formalismes de représentation de connaissances ont été développés, permettant de formaliser les connaissances d'un domaine puis de mettre en œuvre des raisonnements sur ces représentations. Parmi ces formalismes, les logiques de descriptions offrant un potentiel d'expressivité, sont utilisées avec succès pour représenter les ontologies dans plusieurs domaines, en particulier le projet Web Sémantique **(**Baget et al., 2004).

Puisqu'il existe généralement plusieurs façons d'appréhender les connaissances d'un domaine, la construction des ontologies n'est donc pas une tâche facile. Ceci est dû principalement, à la difficulté de trouver des définitions consensuelles des concepts d'un domaine

satisfaisant les définitions propres à chaque utilisateur, qui traduisent son point de vue sur le domaine.

La difficulté de représenter des ontologies est liée principalement à l'existence de plusieurs communautés d'utilisateurs qui peuvent s'intéresser au même domaine mais avec des points de vue différents. Ces communautés, évoluant dans un environnement pluridisciplinaire, coexistent et collaborent entre elles. Chaque communauté a ses intérêts propres et perçoit différemment les entités conceptuelles du même univers de connaissances à représenter.

La plupart des méthodes et méthodologies de constructions des ontologies négligent cette variété de perceptions et elles offrent des outils et des directives pour créer un modèle unique du monde, conçu pour une seule vision du monde. L'approche par point de vue, à laquelle nous nous intéressons, s'oppose à cette approche mono-point de vue et permet donc de modéliser une même réalité selon des points de vue différents. Un point de vue dans son sens large, est la perception qu'une personne ou un groupe de personnes en fait d'un monde observé.

Dans cet article, nous nous intéressons au problème de développement d'une ontologie dans une organisation hétérogène en prenant en compte différents points de vue et terminologies des communautés au sein de cette organisation. Une telle ontologie, que nous appelons *ontologie multi-points de vue* (Hemam et Boufaida, 2009, Hemam et Boufaida, 2011) confère à un même univers de discours plusieurs descriptions partielles telles que chacune soit relative à un point de vue. De plus, les différentes descriptions partielles partagent à un niveau global des éléments ontologiques consensuels et des passerelles. Ces dernières établissent les communications entre les points de vue et représentent ainsi la collaboration interdisciplinaires.

L'objectif visé par ce travail est de proposer une approche de construction des ontologies qui tiennent compte des différents points de vue des utilisateurs. Ainsi, pour atteindre cette objectif, notre démarche est la suivante :

- Au niveau conceptuel, nous proposons un modèle de représentation des connaissances en prenant en compte la notion de point de vue : *modèle de connaissances multi-points de vue*. Le modèle proposé est fondé sur les notions de *point de vue* et d'*ontologie*. Cette dernière représente les connaissances du domaine partagées par plusieurs acteurs et le point de vue représente les connaissances du domaine qui sont pertinentes et visibles selon la perception d'un seul acteur.
- Au niveau formel, le modèle de connaissances multi-points de vue est appliqué pour formaliser l'ontologie multi-points de vue en logique de descriptions. Pour cela, nous utilisons un sous-langage de la logique de descriptions de type $\mathcal{SHOQ}(\mathcal{D})$ (Baader et al., 2003) pour exprimer les différentes notions inhérentes aux points de vue telles que les concepts globaux et locaux, les passerelles, les estampilles ….

La suite de cet article est structurée comme suit. Section 2 présente le paradigme de la multi-représentation basé sur le mécanisme d'estampillage. Dans la section 3, nous nous intéressons à la notion de point de vue. Nous présentons dans la section 4, un processus pour la construction d'ontologies multi-points de vue, en logique de descriptions. Enfin, la section 5 conclut l'article et donne des perspectives pour les travaux futurs.

2 Multi-représentation et le mécanisme d'estampillage

Différentes représentations d'une même entité du monde réel existent en raison des points de vue, du niveau de détail, et de l'intérêt des utilisateurs pour cette entité, d'où la notion de multi-représentation.

Dans (Balley *et al*, 2004) les auteurs ont étudié le problème de la multi-représentation dans les bases de données spatiales afin de permettre d'associer à une même réalité spatiale, plusieurs modélisations alternatives. Dans leurs travaux, ils proposent une extension du modèle conceptuel entité-association. Un mécanisme d'estampillage est proposé dans ce cadre afin de permettre aux concepteurs d'associer à chaque entité, association ou propriété du domaine spatial, plusieurs descriptions selon la résolution sémantique ou spatiale. Une technique d'estampillage pour les logiques de descriptions a été étudiée dans (Benslimane et al., 2006), pour permettre la représentation multiple des concepts dans une même ontologie. Le mécanisme d'estampillage permet aux attributs d'avoir des définitions différentes, i.e. différentes cardinalités ou différents domaines de valeurs (selon des contextes différents).

Cette proposition peut être illustrée à travers un exemple simple. Soit à considérer deux représentations, d'un monde réel, identifiées par les estampilles S1 et S2. La description d'un concept estampillé *Homme_Marié*, dans deux cultures différentes (S1 et S2), est comme suit:

Type Homme_Marié (s1, s2)
s1, s2: Nom: string
s1: Epouse (1,1): Femme
s2: Epouse (1, 4): Femme

Dans le contexte *s*1 un homme marié est un homme ayant exactement une seule épouse. Par ailleurs, dans le contexte *S2* un homme marié peut avoir au maximum 4 épouses. Dans les deux contextes (*S1* et *S2*) le concept *Homme_Marié* est décrit par l'attribut Nom.

3 Notion de point de vue

Plusieurs travaux se sont intéressés à la représentation explicite de points de vue dans différents formalismes de représentation des connaissances. On peut citer par exemple : *KRL* (Bobrow et al., 1977) et *TROPES* (Mariño, 1993) en représentation des connaissances par objets et le travail développé dans (Ribière, 1999) qui introduit les points de vue dans le formalisme des graphes conceptuels. Dans (Bouquet et al., 2004), les auteurs définissent la notion de contexte, qui est très proche de la notion de point de vue, pour désigner une ontologie dont le contenu est décrit dans un contexte particulier et mis en relation avec le contenu d'autres ontologies au travers d'appariements.

Généralement, l'approche par point de vue est fondée sur la conjonction acteur/information. Dans (Benchikha, 2007) le terme de point de vue est défini comme "une position conceptuelle mettant en liaison d'une part un acteur qui observe et d'autre part un monde qui est observé". Les acteurs peuvent observer un même univers de discours produisant des points de vue qui peuvent être considérés de différentes manières :

Points de vue uniformes: dans ce cas, tous les acteurs ont la même vision de l'univers de discours et produisent des représentations équivalentes.

Points de vue complémentaire: dans ce cas, chaque acteur en voit une partie du monde observé. Chaque point de vue est une représentation partielle et cohérente du monde. Les différentes représentations qui découlent des différents acteurs sont alors complémentaires.
Points de vue comparables: dans ce cas les acteurs produisent des représentations comparables au sens plus général/spécifique.

Dans la suite, nous adoptons le terme *d'ontologie multi-points de vue* afin de mettre l'accent sur l'importance de la notion de point de vue pour **1)** résoudre le problème de la représentation multiple **2)** avoir un meilleur accès et une meilleure visibilité des éléments ontologiques (concepts, rôles, individus) **3)** tirer profit de la représentation multi-points de vue des connaissances pour permettre leur évolution. Par ailleurs, pour prendre en compte la notion de point de vue, nous supposons que les différents points de vue sur un même univers de discours sont des visions partielles mais complémentaires. Leur union est une représentation complète et cohérente du monde.

4 Description de la méthode proposée

Notre objectif est de proposer une méthode permettant la construction d'ontologies, en prenant en compte la notion de point de vue.

La méthode que nous proposons est complète, dans la mesure où, partant de données brutes elle permet d'arriver à une ontologie multi-points de vue représentée en logique de descriptions. Pour ce faire, trois principales étapes sont suivies afin d'expliciter et de guider la construction de l'ontologie multi-points de vue.

4.1 Étape de spécification des besoins

Le but visé par cette étape est d'établir une fiche signalétique de l'ontologie. Cette fiche permet de décrire l'ontologie multi-points de vue à construire à travers les quatre aspects suivants :

- **Le domaine de connaissance**. Cet aspect consiste à délimiter aussi précisément que possible le domaine que va couvrir l'ontologie.
- **Les points de vue**. Il s'agit ici de déterminer quels sont les points de vue à représenter. En effet, lorsqu'un domaine est suffisamment vaste et complexe, il est souvent organisé selon plusieurs services, plusieurs tâches, plusieurs groupes de travail ou encore plusieurs communautés. Cette organisation apporte une division *a priori* du domaine en points de vue. Par exemple, dans le domaine « Immobilier » on peut distinguer les points de vue : « *Finance* », « *Taille* » et « *Localisation* ».
- **Les experts du domaine.** Cet aspect consiste à déterminer parmi les experts du domaine, ceux qui sont le mieux à même de modéliser les connaissances relatives à chacun des points de vue, selon leurs spécialités.
- **La portée de l'ontologie.** Cet aspect consiste à déterminer à priori la liste des termes globaux les plus importants désignant les entités du domaine de connaissances à représenter. Dans l'exemple du domaine « Immobilier » on peut déterminer les termes globaux suivants: {Habitat, Appartement, Locataire, Agence}.

4.2 Étape de conceptualisation

La conceptualisation mérite une attention particulière parce qu'elle détermine le reste de la construction de l'ontologie. L'objectif est d'organiser et de structurer la connaissance, en utilisant des représentations semi formelles (tables et graphes) qui sont indépendantes des paradigmes de la représentation de connaissances dans lesquels l'ontologie va être formalisé. Durant cette étape, nous construisons, pour chaque point de vue PVi, une représentation locale selon la perception des experts par rapport au point de vue considéré, ensuite les différentes représentations locales seront connectées par des liens intermédiaires. Pour ce faire, nous distinguons les principales activités suivantes :

4.2.1 Construction d'un glossaire de termes locaux

Un terme peut être la représentation d'une entité pertinente du domaine appelée *concept*, ou d'une *relation binaire* qui lie deux concepts. Cette activité consiste à construire un glossaire local de termes. Ce dernier recueille les termes du domaine qui sont intéressants dans le cadre du point de vue visé et associe à chaque terme identifié une description en langage naturel.

Exemple 1 : Selon le point de vue *Taille* nous pouvons recueillir les termes suivants : {Petit_Appartement, Grand_Appartement, Studio, F1, F2, F3...}.

Exemple 2 : Selon le point de vue *Finance* nous pouvons recueillir les termes suivants: {Appartement_Cher, Appartement_PasCher, HLM, Appartement_LoyerMoyen, Appartement_LoyerHaut, Locataire_Riche…}.

4.2.2 Construction de la hiérarchie locale de concepts

Une hiérarchie locale de concepts (HLC) organise un groupe de concepts entre eux sous forme d'une taxonomie en utilisant la relation de généralisation (i.*e. classe/sous-classe*). Dans une HLC, la relation de généralisation est représentée par un arc (le sens d'un arc est dirigé vers le concept général). Voir l'exemple de la figure 1.

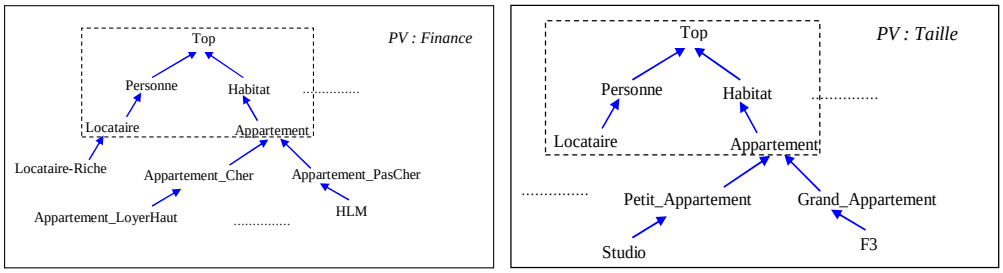


FIG 1– *Hiérarchies locales de concepts selon le point de vue Taille et selon le points de vue Finance*

4.2.3 Construction du dictionnaire de concepts

Le dictionnaire de concepts consiste à décrire tous les concepts représentés dans la hiérarchie de concepts, en représentant pour chaque concept ses attributs qui sont visibles par rapport au point de vue considéré. Un attribut marqué par * est un attribut visible dans tous les points vue. Par ailleurs, l'ensemble des attributs marqués par * constitue ce qu'on appelle la clé du concept. Cet ensemble permet de distinguer une instance de toutes les autres instances de son concept.

Par exemple, pour le concept *Appartement*, le point de vue *Taille* contient, en plus des attributs clés (i.e. numéro, adresse et étage), les attributs qui concernent la conception, l'architecture et la distribution de l'appartement, tandis que le point de vue *Finance* s'occupe des dépenses de location de l'appartement (*Cf.* TAB. 1, TAB. 2).

PV	Concept	Attributs
Taille	Appartement	Numéro *
		Adresse *
		Etage *
		Surface
		Nbr_pièces
		.....

TAB. 1 – *Le concept Appartement est décrit du point de vue Taille comme ne possédant que les attributs qui sont pertinents pour ce point de vue : Surface et Nbr_pièces,...*

PV	Concept	Attributs
Finance	Appartement	Numéro *
		Adresse *
		Etage *
		Loyer
		Charges
		....

TAB. 2 – *Du point de vue Finance, le concept Appartement contient les attributs loyer, charges...*

4.2.4 Constructions de la table des attributs

La description complète d'un attribut pour un concept est donnée en terme des descripteurs de contraintes de type, domaine et cardinalité. Ainsi, pour chaque attribut identifié dans le dictionnaire de concepts, la table des attributs consiste à définir le nom de cet attribut, son concept, son type, sa cardinalité (nombre min ou max de valeurs) et son domaine de valeurs (*Cf.* TAB. 3).

PV	Nom de l'attribut	Concept	Type	Cardinalité (Min/Max)	Domaine des valeurs
Taille	Nbr_pièces	Petit_Appartement	Entier	1..1	{1, 2}
		F1	Entier	1..1	1

TAB. 3 – *Description de l'attribut Nbr_pièces pour les concepts Petit_Appartement et F1*

4.2.5 Construction de la table des instances

Lorsque l'on regarde une instance d'un point de vue particulier, on ne voit que les attributs de l'instance qui sont pertinents pour (connus par) ce point de vue ; on a une vision partielle de l'instance (*Cf.* TAB. 4).

PV	Nom de l'instance	Concept	Attributs	Valeurs
Taille	Chez_hemam	F1	Numéro * Adresse * Etage * Surface Nbr_pièces	112 6 rue benbadis 25000 4 55 1

TAB. 4 – *L'instance « chez_hemam » est considéré comme un F1 d'après le point de vue « Taille » et ne possédant que les attributs qui sont pertinents pour ce point de vue: surface, nbr_pièces*

4.2.6 Liaison des représentations locales

Cette dernière activité consiste à lier les différentes représentations locales des différents points de vue par des liens intermédiaires ; pour cela deux types de liens sont distinguées : *Passerelle* et *Relation globale*.

Passerelles. Une passerelle décrit une règle entre un concept source (ou un ensemble de concepts sources) et un concept cible de deux (ou plusieurs) points de vue différents. Pour cela on distingue les quatre types de passerelles suivants :

- **Passerelle d'inclusion unidirectionnelle.** Exprime l'inclusion ensembliste entre l'extension d'un concept d'un point de vue, source de la passerelle, et celle d'un concept d'un autre point de vue, destination de la passerelle. En termes de logique, la passerelle peut être vue comme une implication: être instance du concept source implique être instance du concept destination.
- **Passerelle d'inclusion avec plusieurs sources.** Dans certains domaines d'application, l'instance doit appartenir à plusieurs concepts de différents points de vue pour que l'on puisse déduire son appartenance à un concept destination d'un autre point de vue. Dans ce cas, qui généralise le cas précédent, une passerelle est décrite par la liste de ses concepts sources des différents points de vue et le concept destination d'un autre point de vue.
- **Passerelle d'inclusion bidirectionnelle**. Exprime l'égalité ensembliste entre deux concepts C et D de deux points de vue différents.
- **Passerelle d'exclusion bidirectionnelle.** Exprime un lien entre deux concepts, de deux points de vue différents, pour lesquels il ne sera pas possible, pour une même instance, d'appartenir en même temps aux deux ensembles d'instances correspondant à ces deux concepts.

La description des différentes passerelles se fait à travers une table d'axiomes logiques. Chaque axiome comporte, la liste des concepts sources et le concept destination sur lesquels porte l'axiome, une définition en langage naturel, et une expression logique (*Cf.* TAB. 5).

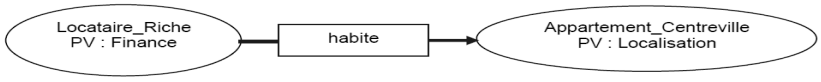
Concept & PV source	Concept & PV Cible	Description	Expression logique
HLM (*PV: Financier*)	Appartement_banlieue *(PV : Localisation)*	Tous les HLM sont dans la banlieue et tous les appartements de banlieue sont des HLM	∀ X, HLM(X) ⇔ Appartement_banlieue(X)

TAB. 5 – *Exemple d'une passerelle d'inclusion bidirectionnelle*

Relations globales. Une relation globale **R** est une relation lexicale, qui permet de relier les sous-concepts hiérarchisés différemment selon des points de vue différents et permet d'exprimer un fait général à propos des membres des concepts qui participent à cette relation. Elle est définie par un concept origine **C** appelé le domaine de la relation **R** et un concept destination **D** appelé le co-domaine de la relation **R**. Cela correspond à l'assertion suivante:

$\forall$ X $\in$ PV_{source}:C, $\exists$ Y$\in$ $PV_{destination}$:D, tel que l'instance X est liée à l'instance Y par la relation R

Exemple : un locataire riche est une personne qui ***habite*** dans un appartement au centre-ville



4.3 Étape de formalisation

L'ontologie conceptuelle obtenue dans l'étape précédente doit alors être formalisée. Le formalisme de représentation utilisé à cette étape est la logique de descriptions. Une ontologie exprimée en LD contient la description des concepts, des rôles et des individus. Le concept représente un ensemble d'objets. L'individu est utilisé pour représenter un objet du domaine. Le rôle correspond à un attribut ou bien à une relation binaire entre deux concepts.

Pour les besoins de la formalisation des ontologies multi-points de vue, nous introduisons, dans la logique de descriptions, les notions suivantes :

1. **Ontologie multi-points de vue**: est une description multiple du même univers de discours selon différents points de vue. Elle est définie comme un quadruplet O = <C^G, R^G, Vp, M>, où $\boldsymbol{C^G}$ l'ensemble des concepts globaux, $\boldsymbol{R^G}$ l'ensemble des rôles globaux, ***Vp*** l'ensemble des points de vue et ***M*** l'ensemble des passerelles.
2. **Point de vue:** est une description partielle d'un univers de discours selon une perception particulière. Un point de vue est défini comme un triplet VP_K= <C^L, R^L, A^L>, où $\boldsymbol{C^L}$ l'ensemble des concepts locaux, $\boldsymbol{R^L}$ l'ensemble des rôles locaux et $\boldsymbol{A^L}$ l'ensemble des individus locaux.
3. **Concept global :** est un concept vu à partir de deux ou plusieurs points de vue avec certaines caractéristiques communes (i.e. attribut et/ou relations).
4. **Concept Local:** c'est un concept qui est vu et décrit localement selon un point de vue donné.

5. **Rôle Global:** c'est une relation entre deux concepts locaux définis dans deux points de vue différents.
6. **Rôle Local:** c'est une relation entre deux concepts locaux définis dans le même point de vue.
7. **Estampille:** nous adaptons le mécanisme d'estampillage défini dans (Benslimane et al, 2006) pour permettre la multi représentation des concepts. Dans notre approche, une estampille (i.e. label) permet de reconnaître pour chaque élément ontologique (i.e. concept, rôle, individu) le point de vue auquel il appartient.
8. **Hiérarchie locale:** Sous un point de vue l'ensemble des concepts locaux sont liés par la relation de *subsomption* (ou de généralité). Cette dernière, permet de les organiser en hiérarchie locale propre au point de vue. Par ailleurs, chaque concept racine (i.e. sommet de la hiérarchie locale) est associé à (subsumé par) un type de concept global.
9. **Passerelle:** l'une des particularités de la représentation multi-points de vue, est l'existence d'un canal de communication entre les différents points de vue. Ce canal de communication, appelé passerelle, permet de représenter des liens consensuels entre les concepts locaux de différents points de vue.
10. **Instanciation multiple:** le mécanisme de l'instanciation multiple permet à un individu d'être instance directe d'un ou plusieurs concepts. Dans le contexte de notre travail, un individu possède la propriété suivante:

    Propriété: *un individu est une instance d'un concept global et instance d'un ou plusieurs concepts locaux définis dans un ou plusieurs points de vue.*

Définition 1 (**Syntaxe d'un concept global**)**.** Soit S= $\{vp_1,\ldots,vp_k,\ldots,vp_m\}$ l'ensemble de noms des points de vue. Un concept global noté par $\mathbf{C^{\hat{0}}}$, peut être formé en utilisant les constructeurs booléens (conjonction, disjonction) et les constructeurs de restriction globaux suivants:

Constructeur de restriction global	Description
$\forall_{vp1,\ldots,vp\,k}$ R.C	Définit un nouveau concept dont toutes ses instances sont reliées, via le rôle *R*, seulement aux instances du concept *C* dans les points de vue vp_1 à vp_k
$\exists_{vp1,\ldots,vpk}$ R.C	Définit un nouveau concept dont toutes ses instances sont reliées, via le rôle R, à au moins une instance du concept C et seulement dans les points de vue vp_1 à vp_k
$\leq_{vp1,\ldots,vpk}$ n R	Spécifie la cardinalité minimale ou maximale du rôle R dans les points de vue vp_1 à vp_k
$\geq_{vp1,\ldots,vpk}$ n R	

Exemple:

Soit les trois points de vue *Taille*, *Finance* et *Localisation* désignés respectivement par les estampilles vp_1, vp_2 et vp_3. L'expression, ci-dessous, définit un concept global avec un attribut *Nbr_pièces* selon vp_1, un attribut *Loyer* selon vp_2, un attribut *Surface* selon vp_1 et vp_2 et un attribut *Adresse* selon les trois points de vue vp_1, vp_2 et vp_3.

$$\text{Appartement}^{\hat{0}} \equiv (\forall_{vp1}\ \text{Nbr\_pièces.Number}) \sqcap (\forall_{vp2}\ \text{Loyer.Number}) \sqcap (\forall_{vp1,vp2}\ \text{Surface.Number})$$
$$(\forall_{vp1,\,vp2,\,vp3}\ \text{Adresse.String}) \sqcap (\geq_{vp1,\,vp2,\,vp3}\ 1\ \text{Adresse}) \sqcap (\leq_{vp1,\,vp2,\,vp3}\ 1\ \text{Adresse})$$

Définition 2 (**Syntaxe d'un concept local**). Soit $vp_i \in$ S. Un concept local, noté **vp_i: C,** peut être défini au moyen de la syntaxe suivante:

vp_i: C → ($C^{\hat{o}}$) | ($\neg$ *C*) | (*C* $\sqcap$ *C*) | (*C* $\sqcup$ *C*) | ($\exists$ *R. C*) | ($\forall$ *R. C*) | ($\geq n$ *R*) | ($\leq n$ *R*)| *R*.{a,b...}

Exemple:

vp_1: Petit_Appartement $\equiv$ Appartement $^{\hat{o}}$ $\sqcap$ (Nbr_pièces. {1, 2})

Définit un concept local, dans le point de vue vp_1, comme étant un appartement et dont la valeur de l'attribut Nbr_pièces *est dans l'ensemble* {1, 2}.

Définition 3 (**Syntaxe d'un rôle local**). Un rôle local, noté **vp_i: R,** peut être défini selon la forme suivante:

vp_i: R (C, D) *où R est un nom de rôle local défini dans le point de vue* ***vp_i****, C et D sont deux concepts locaux définis dans le même point de vue* ***vp_i***

Exemple:

vp_2: habite_par (Appartement_Cher, Locataire_Riche)

Définit un rôle local entre deux concepts locaux définis dans le même point de vue vp_2

Définition 4 (**Syntaxe d'un rôle global**). Un rôle global, note par **$R^{\hat{o}}$**, peut être défini selon la forme suivante:

$R^{\hat{o}}$ (vp_i: C, vp_j: D) *où R est un nom de rôle global, C et D sont deux concepts locaux définis dans deux points de vue différents.*

Exemple:

habite $^{\hat{o}}$ (vp_2: Locataire_Riche, vp_3: Appartement_CentreVille)

Définit un rôle global entre deux concepts locaux définis dans deux PV différents (vp_2 et vp_3)

Définition 5 (**Syntaxe d'une relation de subsomption**). *Sous un point de vue* ***VP_i****, une hiérarchie locale, notée* ***$vp_i/\mathcal{H}$****, est définie par le triplet ($C^{\mathcal{L}}$, ∂, $\sqsubseteq$) où* : $\mathcal{C}^{\mathcal{L}}$ est l'ensemble des concepts locaux, ∂ est une fonction de $C^{\mathcal{L}}$ dans $C^{\mathcal{G}}$ qui associe chaque concept racine (*i.e.* le plus général) S de $C^{\mathcal{L}}$ à un seul concept global $C^{\hat{o}}$ de $C^{\mathcal{G}}$ et $\sqsubseteq$ est la relation de subsomption utilisée pour exprimer explicitement un lien d'ordre direct selon les deux formes suivantes :

vpi: D $\sqsubseteq$ vpi: C *où C et D sont deux concepts locaux définis dans le même point de vue vp_i,*

vp_i: S $\sqsubseteq$ $C^{\hat{o}}$ *où S est le concept le plus général défini dans le point de vue* ***vp_i*** *et $C^{\hat{o}}$ est un nom de concept global.*

Exemple:

vp_2 : HLM $\sqsubseteq$ $vp_{2:}$ Appartement_PasCher

Exprime un lien de subsomption entre deux concepts locaux définis dans le même point de vue. En effet, sous le point de vue vp_2 *, tous les HLM sont des appartements pas cher.*

$vp_{2:}$ Appartement_PasCher $\sqsubseteq$ Appartement $^{\hat{o}}$

Exprime un lien de subsomption entre le concept local Appartement_PasCher, défini sous *le point de vue* $vp_{2,}$ *et le concept global* Appartement [ô]

Définition 6 (**Syntaxe d'une passerelle).** Une passerelle en logique de descriptions s'exprime de quatre manières :

vpi: X $\xrightarrow{\sqsubseteq}$ vpj: Y (Passerelle d'inclusion) (1)

vp1: $X_1 \sqcap ... \sqcap$ vpk: $X_k \xrightarrow{\sqsubseteq}$ vpj: Y (Passerelle d'inclusion avec plusieurs sources) (2)

vpi: X $\xleftrightarrow{\equiv}$ vpj: Y (Passerelle d'inclusion bidirectionnelle) (3)

vpi: X $\xleftrightarrow{\perp}$ vpj: Y (Passerelle d'exclusion bidirectionnelle) (4)

Exemple:

vp_2 : HLM $\xleftrightarrow{\equiv}$ vp_3: Appartement_Banlieue

Exprime que les deux concepts locaux, définis dans deux points de vue différents, sont équivalents. En effet, tous les HLM sont dans la banlieue et tous les appartements de banlieue sont des HLM

$vp_{1:}$ Plus_TroisPièce $\sqcap$ $vp_{3:}$ Appartement_CentreVille $\xrightarrow{\sqsubseteq}$ $vp_{2:}$ Appartement_Cher

Signifie que tous les appartements de plus de trois pièces qui se trouvent au centre-ville sont des appartements chers

Définition 7 (**Syntaxe d'un individu local).** Un individu local est une instance d'un concept local défini sous un point de vue donné. Chaque individu local est décrit selon la forme suivante:

vp_i: C (a) *où* ***C*** *est un concept local défini dans le point de vue* $\boldsymbol{vp_i}$ *et* ***a*** *est un nom d'individu.*

Exemple:

vp_1: Petit_Appartement (chez-Benali) vp_3: Appartement_Banlieue (chez-Benali)

Indiquent que l'individu chez-Benali est une instance de Petit_Appartement sous le point de vue vp_1 *et aussi une instance de Appartement_Banlieue sous le point de vue* vp_3

5 Conclusion et perspectives

Dans ce article, nous avons présenté une approche pour la construction d'une ontologie avec des points de vue différents. L'élément clé de notre approche est de permettre la description de ce type d'ontologies, sans éliminer l'hétérogénéité mais en faisant cohabiter l'hétérogénéité (au niveau local) et le consensus (au niveau global). A chaque point de vue correspond une représentation locale. Les différents points de vue partagent à un niveau global, des éléments ontologiques (i.e. concepts et rôles globaux) et des passerelles. Ces dernières, permettent de lier différents concepts locaux provenant de différents points de vue. Par ailleurs, dans le modèle multi-points de vue proposé, une instance représente un individu particulier d'un concept global, qui est rattaché, dans chaque point de vue, à son concept local d'appartenance. De ce fait, un individu peut être manipulé à partir d'un seul point de vue sans se préoccuper des autres, ce qui réduit le nombre d'attributs et de concepts à regarder et simplifie ainsi la hiérarchie de spécialisation (subsomption) de concepts. Enfin, une

définition formelle de la syntaxe, en logique de descriptions, des différents concepts introduits pour le support des points de vue a été présentée.

Dans le modèle multi-points de vue proposé, nous n'avons pas introduit de relations entre les points de vue. Par exemple, un point de vue ne peut pas être défini comme sous-point de vue d'un autre point de vue. Cette capacité pourrait être affectée en considérant qu'un point de vue correspond à un ensemble de critères qui caractérisent le contexte défini par le point de vue, et que l'ajout d'autres critères (caractéristiques) à cet ensemble créera un autre point de vue, qui sera un sous-point de vue du point de vue en question.

Références

Baader, F., Horrocks, I., & Sattlerl, U., Description Logics as Ontology Languages for the Semantic Web. In Festschrift in honor of Jorg Siekmann, LNAI. Springer, 2003.

Baget., J.-F., Canaud., E., Euzenat., J. et Hacid., M.-S. (2004). "Les langages du Web Sémantique", In Le Web sémantique, CHARLET J., LAUBLET P. & REYNAUD C. (Ed.), Revue Information - Interaction - Intelligence, Vol.4, N°.1, pp. 21-43.

Balley S., Parent C., et Spaccapieta S. (2004)., Modeling Geographic Data with Multiple Representation. In journal Geographical Information science, Vol 18, p. 327-352.

Benchikha, F., Intégration des points de vue dans les bases de données à objets : Le modèle Multi-Viewpoint DataBase, Thèse de Doctorat, Université de Constantine, 2007.

Benslimane, D., Arara, A., Falquet, G., Maamar, Z., Thiran, P., & Gargouri, F., Contextual Ontologies: Motivations, Challenges, and Solutions. In Fourth Biennial International Conference on Advances in Information Systems ADVIS 2006 (Springer), 2006.

Bobrow, D.G., & Winograd, T., An overview of KRL, a knowledge representation language. Cognitive Science, Vol. 1, 1977

Bouquet, P., Giunchiglia, F., Van Harmelen, F., Serafini, L., & Stuckenschmidt, H., Contextualizing Ontologies. In Journal of Web Semantics, 1(4): pp 325--343, 2004.

Hemam, M., Boufaida, Z.: Représentation d'ontologies multi-points de vue : une approche basée sur la logique de descriptions. Short paper presented at the 20es Journées Francophones d'Ingénierie des Connaissances, Tunisia (2009)

Hemam, M., Boufaida, Z.: MVp-OWL: A Multi-Viewpoints Ontology Language for the Semantic Web. Int. J. Reasoning-based Intelligent Systems, Inderscience Publishers, Vol. 3, N° 3/4 (2011) 147–155

Mariño, O., Raisonnement classificatoire dans une représentation à objets multi-points de vue. Thèse d'informatique, université Joseph Fourier, Grenoble, 1993.

Ribière, M., Représentation et gestion de multiples points de vue dans le formalisme des graphes conceptuels. Thèse de doctorat en informatique, Nice-Sophia Antipolis,1999

Ribière, M., & Dieng, R., Introduction of viewpoints in conceptual graph formalism. In 5th International Conference on Conceptual Structures, ICCS'97, Washington, USA, LNAI 1257, Springer Verlag, pp. 168-182, 1997.

Construction d'ontologies à partir des besoins métier d'un Système d'Information Décisionnel

Aziza Sabri *, Laila Kjiri**
ENSIAS, Université Mohammed-V-Souissi, Rue Mohammed Ben Abdellah Regragui,
B.P. 713 Agdal, Madinat Al Irfane – Rabat, Maroc
*azizasabri@yahoo.com
**kjiri@ensias.ma
http://www.ensias.ma

Résumé. Au niveau conceptuel d'un SID, l'étape d'expression des besoins soulève plusieurs problèmes majeurs qui sont issus de la diversité des perspectives, des points de vue, des contextes et des profils acteurs, de l'incohérence et de l'ambiguïté sémantique. Pour remédier à cela, nous avons défini les besoins d'une organisation sous forme de buts. Chaque but est formulé par un verbe, une section paramètres_faits et une section paramètres_dimensions. Ensuite, nous avons défini une interprétation des buts informationnels qui aide à extraire les données décisionnelles à analyser. Cette interprétation est réalisée à l'aide d'un méta-modèle de buts et à l'aide de l'ontologie. Ainsi, nous allons détailler le processus de construction des ontologies à partir des besoins métier d'un SID. Nous visons à construire un vocabulaire métier qui sera ainsi partagé et utilisé par les concepteurs dans l'expression des buts métier. Le processus de construction des ontologies s'appuie sur un ensemble des étapes explicitées dans le corps de l'article.

1 Introduction

Généralement la conception d'un SID passe par les trois niveaux de conception habituels: conceptuel, logique et physique. Pour le niveau conceptuel, la phase d'expression des besoins pour la conception d'un SID consiste à identifier les données spécifiques requises par les acteurs de l'entreprise pour l'exercice de leur métier (marketing, qualité, finance, etc.). Chaque acteur de SID peut formuler ses besoins en langage naturel dans lequel il exprime une requête avec les buts à atteindre. Cependant, cette démarche soulève plusieurs problèmes majeurs qui sont issus de la diversité des perspectives, des points de vue, des contextes et des profils acteurs, de l'incohérence et de l'ambiguïté sémantique. Pour remédier à cela, dans un travail antérieur (sabri et kjiri, 2011), nous avons défini les besoins d'une organisation sous forme de buts élaborés à partir des anciens rapports d'analyse ou en interviewant les décideurs. Les buts sont classifiés en trois types à savoir les buts tactiques, les buts stratégiques et les buts informationnels. Ces buts sont exprimés à l'aide d'un modèle facilitant la définition des besoins et permettant de guider le concepteur dans la définition ultérieure d'un schéma multidimensionnel. Chaque but est formulé par un verbe, une section paramètres_faits et une section paramètres_dimensions. Ensuite, nous avons défini une interprétation des buts informationnels qui aide à extraire les données décisionnelles à analyser. Cette interprétation est réalisée à l'aide d'un méta-modèle de buts et à l'aide de l'ontologie.

Dans ce présent travail, nous avons facilité la tâche au concepteur afin de créer les concepts ontologiques à partir des données décisionnelles extraites des buts informationnels. L'objectif de notre proposition est d'assurer un partage de la connaissance entre l'ensemble des décideurs et de résoudre les problèmes d'ambigüité sémantique. Pour cela, nous avons utilisé deux types d'ontologies: L'ontologie de paramètres_faits qui définit un vocabulaire sémantique et syntaxique commun pour l'ensemble des concepts définis dans le but informationnel. Cette ontologie est utilisée en particulier pour spécifier les concepts ontologiques des faits et ceux des mesures mentionnées par les décideurs lors de la formulation des buts. Et l'ontologie de paramètres_dimensions qui définit également un vocabulaire sémantique et syntaxique commun pour l'ensemble des concepts définis dans le but informationnel. Cette ontologie est utilisée en particulier pour spécifier les concepts ontologiques des dimensions qui sont liées à chaque table de fait et ceux des attributs des dimensions. Ensuite, nous allons détailler le processus de construction des deux ontologies (parramètres_faits et paramètres_dimensions) à partir des besoins métier d'un SID ainsi que les relations qui les relient. Nous visons à construire un vocabulaire métier qui dépend des activités de chaque organisation et qui sera ainsi utilisé par les concepteurs dans l'expression des buts métier. Le processus de construction des ontologies s'appuie sur un ensemble des étapes qui seront explicitées en détail dans le corps de l'article.

L'article est organisé de la façon suivante. Tout d'abord, nous présentons la démarche d'expression des besoins métier d'un SID sous forme de buts. Dans la section 3, nous présentons le processus de traitement et de formalisation des besoins métier d'un SID. Dans la section 4, nous détaillant l'activité de construction des ontologies de résolution des buts informationnels formalisés. Enfin, nous terminons notre article par une conclusion.

2 Expression des besoins métier d'un SID sous forme de buts

Généralement, les décideurs expriment leurs besoins en termes de buts et d'objectifs. Dans cette partie nous allons définir le concept but et la classification des buts dans la littérature avant d'entamer nos propositions pour l'expression des besoins métier d'une organisation.

2.1 Le concept but

Plusieurs études s'intéressant au concept de but l'ont défini comme un objectif que le futur système doit satisfaire (Anton, 1996) (Rolland, 1999) (van Lamsweerde, 2001) (Kavakli04). Le but permet de capturer le besoin concret, réel et aide à la découverte et à l'opérationnalisation des besoins d'une organisation. Il permet, ainsi, d'établir le lien entre les fonctionnalités attendues du système d'information décisionnel et les objectifs de l'organisation. Dans la littérature (Annoni, 2007) (Elgolli, 2008), les buts sont classifiés selon trois types (stratégique, tactique et informationnel) qui seront décrits dans la section suivante.

2.2 Classification des buts collectés d'une organisation

La structure organisationnelle se compose des décideurs stratégiques (les managers, les stratèges, la direction générale), qui expriment des buts stratégiques liés à la gouvernance de l'organisation tout en ayant une vision synthétique et globale des buts stratégiques de

l'organisation. Egalement, elle comporte des décideurs tactiques exprimant les buts tactiques (directeur marketing, directeur financier, directeur informatique) et ayant une perspective relative à un secteur d'activité en particulier, un groupe de métiers conjointement liés ou une classe d'utilisateurs d'une organisation.

Un but stratégique indique un résultat souhaité à long terme de l'organisation, il est lié à la stratégie adoptée par l'organisation. Il peut représenter une intention d'un décideur ou de toute l'organisation. Un but tactique est une perspective particulière relative à un secteur d'activité en particulier. Un ensemble des buts tactiques requiert un ensemble d'informations afin d'apporter une réponse et aider à la prise de décision. Un but tactique est opérationnalisé avec un ensemble de buts informationnels.

Dans notre travail, nous définissons un but informationnel comme un besoin formulé par des décideurs en termes d'informations recherchées. Ce type de buts est lié à des métiers spécifiques, exprimant une vision verticale de la prise de décision au sein de l'organisation. Ils sont exprimés sous forme de requêtes des exigences analytiques. Ils sont porteurs des indicateurs qui sont fondamentaux dans la construction de schéma multidimensionnel.

Ainsi, dans la suite de cet article, nous allons nous focaliser sur les buts informationnels. L'idée est de permettre aux décideurs d'exprimer leurs buts informationnels en langage naturel suivant une structure explicite représentée par un méta-modèle sémantique.

2.4 Modèle de but informationnel : Proposition structurelle

Afin de faciliter la tâche à l'analyste de SID pour détecter systématiquement les données décisionnelles, nous avons défini une nouvelle version de modèle de buts informationnels (Sabri et kjiri, 2011). La formulation de ces buts, suivant cette définition, va permettre de déterminer les dimensions et les faits à inclure dans le modèle multidimensionnel. En effet, le choix des faits, des tables de faits, des dimensions et des attributs d'une dimension est contextuel. Leur extraction sera donc directement faite à partir des buts formulés par les différents acteurs de SID. Ainsi, la nouvelle structure de but est la suivante :

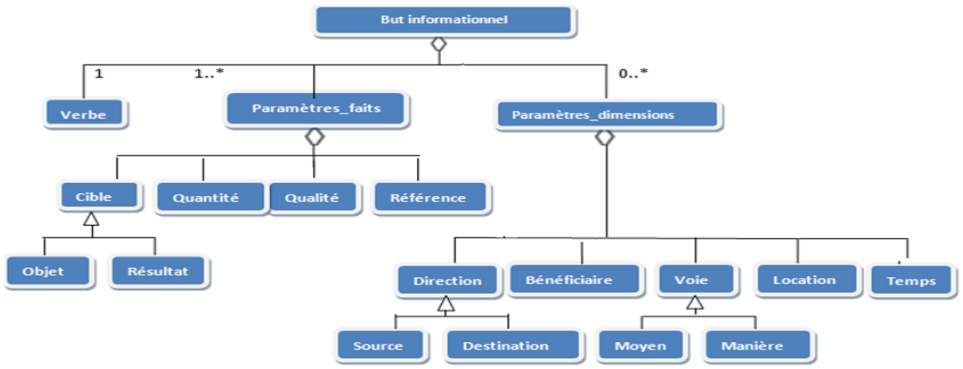


FIG. 1 - *Modèle sémantique proposé pour représenter un but informationnel.*

Chaque but informationnel sera formulé avec un «verbe», une section nommée « Paramètres_faits » et une autre nommée « Paramètres_dimensions » :

- Section « Paramètres_faits» : contient le nom et les indicateurs des tables de faits. Ainsi, la table de faits contiendra la cible, la référence, la qualité et la quantité :
  - **Cible**: La cible concerne les entités affectées par le but. Il y a deux types de cibles: l'objet et le résultat. L'objet existe avant la réalisation du but et peut éventuellement être modifié ou supprimé par celui-ci, alors que le résultat est l'entité qui résulte de la réalisation du but désigné par le verbe.
  - **Référence** : Elle est l'entité par rapport à laquelle une action est effectuée ou un état est atteint ou maintenu.
  - **Qualité**: C'est une propriété qui doit être atteinte ou préservée.
  - **Quantité**: Elle mesure la quantité qui devrait se produire.
- Section « Paramètres_dimensions» : contient les noms des tables de dimensions. Ainsi, les paramètres direction (source ou destination), location, voie (moyen ou manière), temps et bénéficiaire peuvent désigner des tables de dimensions :
  - **Direction**: Les deux types de direction sont appelés la source et la destination, et identifient respectivement l'endroit initial et l'endroit final de l'objet. La source est le point de départ du but (source d'information ou lieu physique).
  - **Voie**: spécialisée par les deux paramètres : la manière qui spécifie la façon d'atteindre le but et le moyen qui est l'outil par lequel le but est atteint.
  - **Bénéficiaire**: La personne ou le groupe en faveur de qui le but doit être atteint.
  - **Location**: Elle situe l'intention dans l'espace.
  - **Temps**: Il situe l'intention dans le temps.

  Pour aider le concepteur à organiser les buts collectés et atteindre les données décisionnelles pour établir les schémas multidimensionnels, nous avons proposé un ensemble de modèles qui seront présentés dans la partie suivante.

3 Modèles de traitement et formalisation des besoins métier

Dans (Sabri et kjiri, 2012a), nous avons proposé une démarche qui peut se faire en trois principaux processus. Le premier consiste en la délimitation de l'environnement métier pour définir l'ensemble des contextes qui lui sont attachés ainsi que la structure organisationnelle afin de localiser les acteurs à intervenir dans l'étape suivante. Le second détaille le traitement des besoins recensés sous forme de buts. Enfin, le troisième processus est relatif à la production du schéma en étoile.

Dans un travail antérieur (Sabri et kjiri, 2012b), nous avons détaillé le processus de traitement et de formalisation des besoins. Ce processus est formé des étapes suivantes : l'établissement du modèle diagnostic de l'organisation, le recensement des besoins sous forme de buts à travers le cas d'utilisation «Expression des besoins d'un SID», la classification des besoins collectés (en besoins stratégiques, tactiques et informationnels), la création du «Modèle d'association des buts stratégiques à des buts tactiques», suivie du «Modèle d'association des buts tactiques à des buts informationnels» et, enfin, l'établissement du «Modèle de formalisation des buts informationnels». Dans les sections suivantes, nous allons définir les modèles établis suite à l'exécution de ce processus.

3.1 Modèle « Diagnostic de l'Organisation »

Le contexte représente les informations locales concernant le point de vue d'un acteur d'une application (Rifaieh, 2004). Dans (Sabri et kjiri, 2012a), nous avons défini le contexte

métier d'une organisation en considérant qu'un contexte est un ensemble d'informations qui caractérise une situation. Chaque situation représente un fait qui fait référence à un ou plusieurs événements produits dans l'organisation. Un contexte peut détailler un autre contexte. Nous avons considéré également qu'un métier de l'organisation représente un ensemble d'activités définies dans des contextes différents.

Nous avons proposé le modèle suivant pour systématiser la collecte des informations concernant le métier de l'organisation. Il s'agit du modèle «Diagnostic de l'Organisation» :

Métier de l'organisation :
Activité 1 :
Liste 1 des Contextes : Contexte1, contexte 2, contexte3…
Activité 2 :
Liste 2 des Contextes : Contexte1, contexte 2, contexte3…
…
Activité n :
Liste n des Contextes : Contexte1, contexte 2, contexte3…

FIG. 2 - *Modèle «Diagnostic de l'Organisation»*

Ce modèle sert comme document de diagnostic de l'organisation. Il permet de noter les activités liées au métier de l'organisation ainsi que les contextes rattachés à chacune de ces activités. L'ajout d'un contexte ou d'une activité s'effectue d'une manière itérative.

3.2 Modèle d'association des buts stratégiques à des buts tactiques

Après la classification des buts en trois types (stratégiques, tactiques et informationnels), nous procédons à l'association de chaque but stratégique à l'ensemble des buts tactiques qui lui sont attachés selon un contexte prédéfini. Cette association se réalisée selon le modèle suivant:

Activité			
Contexte			
But Stratégique 1	**But Stratégique 1**	**...**	**But Stratégique n**
Liste 1 des buts tactiques	Liste 2 des buts tactiques		Liste n des buts tactiques

FIG. 3 - *Modèle d'association des buts stratégiques à des buts tactiques*

Ce modèle sert comme base de définition des buts tactiques liés à chaque but stratégique. L'ajout d'un nouveau but tactique se fait d'une manière itérative.

3.4 Modèle d'association des buts tactiques à des buts informationnels

Après l'établissement des modèles d'association des buts stratégiques à des buts tactiques, nous procédons de la même manière en associant chaque but tactique à l'ensemble des buts informationnels qui lui sont attachés selon un contexte prédéfini. Cette association s'effectue selon le tableau suivant:

Activité			
Contexte			
But stratégique			
But tactique1	But tactique 2	...	But tactique n
Liste 1 des buts informationnels	Liste 2 des buts informationnels	…	Liste n des buts informationnels

FIG. 4 - *Modèle d'association des buts tactiques à des buts informationnels*

Ce modèle sert comme base de collecte des buts informationnels associés à chaque but tactique. En effet, l'ajout d'un nouveau but informationnel s'effectue d'une manière itérative.

De plus, les buts informationnels sont formés en langage naturel suivant un modèle linguistique. Ce modèle permet d'extraire les données décisionnelles afin d'établir le modèle en étoile. La section suivante décrit la formalisation des buts informationnels pour l'extraction des faits et des dimensions.

3.5 Formalisation des buts informationnels : extraction des faits et des dimensions

Afin de faciliter la tâche aux analystes, nous proposons le modèle de formalisation suivant qui permet d'extraire facilement les données décisionnelles:

Activité		
Contexte		
But stratégique		
But tactique1		
Verbe	**Paramètres_faits**	**Paramètres_dimensions**
V1	PF1	PD1
V2	PF2	PD2
V3	PF3	PD3

FIG. 5 - *Modèle de formalisation des buts informationnels*

Le modèle ci-dessus explicite clairement la structuration des buts informationnels liés à chaque but tactique. Ce modèle encapsule les faits et les dimensions en sont associées. Ainsi, l'extraction des paramètres de faits et de dimensions est systématique.

Dans la section suivante, nous allons détailler le processus de création des concepts ontologiques des données décisionnelles afin de partager le vocabulaire retenu entre l'ensemble des utilisateurs de SID et de résoudre les problèmes de l'ambigüité sémantique détectés lors de l'expression des besoins.

4 Construction des ontologies de résolution des buts informationnels formalisés

Dans cette partie, nous allons détailler le processus de construction des ontologies à partir des besoins métier d'un SID. Nous visons à construire des ontologies métier qui dépendent des activités de chaque organisation. Ces ontologies se basent sur un vocabulaire standarisé des concepts multidimensionnels extraits à partir des deux sections prédéfinies (paramètres_faits et paramètres_dimensions).

4.1 Définition de l'ontologie :

L'ontologie est un mot qui a fait couler beaucoup d'ancre durant des siècles et qui était une énigme pour les philosophes. Ce terme est apparu en informatique au début des années 1990, et a été défini par (Gruber, 1993), notamment dans le domaine de l'intelligence artificielle, comme « une ontologie est une spécification explicite d'une conceptualisation ». Autrement dit, une ontologie définit le vocabulaire partagé pour aboutir à une compréhension commune d'un domaine donné. Elle contient les définitions des concepts dans le domaine et les relations entre eux. Elle peut contenir des axiomes et des règles qui nous permettent l'inférence des nouvelles connaissances.

4.2 Données décisionnelles :

A partir des buts informationnels de l'organisation, nous pouvons extraire les concepts décisionnels nécessaires pour représenter les schémas multidimensionnels. Chaque schéma multidimensionnel possède une table de fait déterminée à partir de la section Paramètres_faits et une ou plusieurs tables de dimensions déterminée à partir de la section Paramètres_dimensions. Une table de faits contient des mesures sous forme d'attributs représentant les éléments d'analyse et une table de dimension contient un ou plusieurs attributs permettant d'avoir des mesures suivant différentes perspectives d'analyses. Le schéma suivant représente la structure des données décisionnelles.

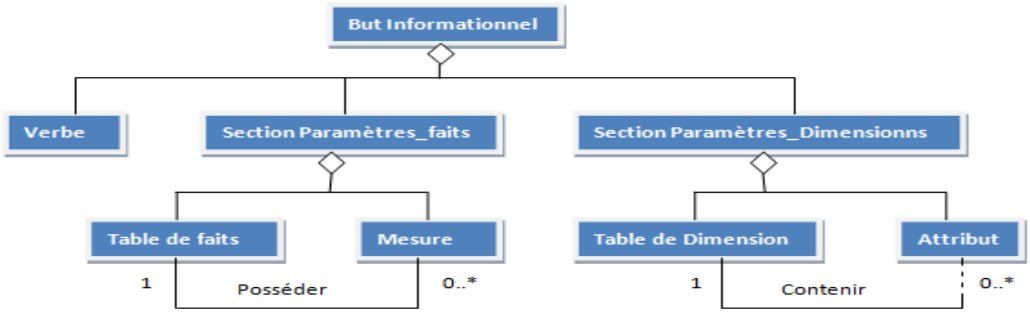


FIG. 6 – *Modèle des données décisionnelles en notation UML*

Nous avons pu définir une interprétation des buts informationnels et en extraire les données décisionnelles à analyser pour un métier donné. A l'issue de cette interprétation nous allons nous servir de l'ontologie pour mettre à la disposition des concepteurs un vocabulaire, voire un référentiel, qui définit les faits et les dimensions pour un métier donné, qui favorise le partage de la connaissance entre l'ensemble des décideurs et qui aide à résoudre les problèmes des ambiguïtés sémantiques et syntaxiques.

4.3 Lien sémantique entre les données décisionnelles

Nous définissons les liens sémantiques (LS) reliant les occurrences des données décisionnelles extraites à partir des buts informationnels dans un métier donné. Les LS

supportées par notre ontologie sont : l'identité, la synonymie, l'équivalence et l'homonymie. Lors de l'enrichissement de l'ontologie par les concepts ontologiques choisis, nous pouvons avoir des redondances. Ainsi, l'optimisation élimine les liens sémantiques redondants et, par conséquent, améliore les résultats des interrogations et favorise un gain de temps d'accès.

4.4 Processus de déduction des ontologies

Le processus de déduction des ontologies représente un ensemble des étapes qui mènent à la construction des ontologies de résolution des buts informationnels prédéfinis suivant un cycle de vie itératif et progressif. Le diagramme d'activités suivant schématise ces étapes.

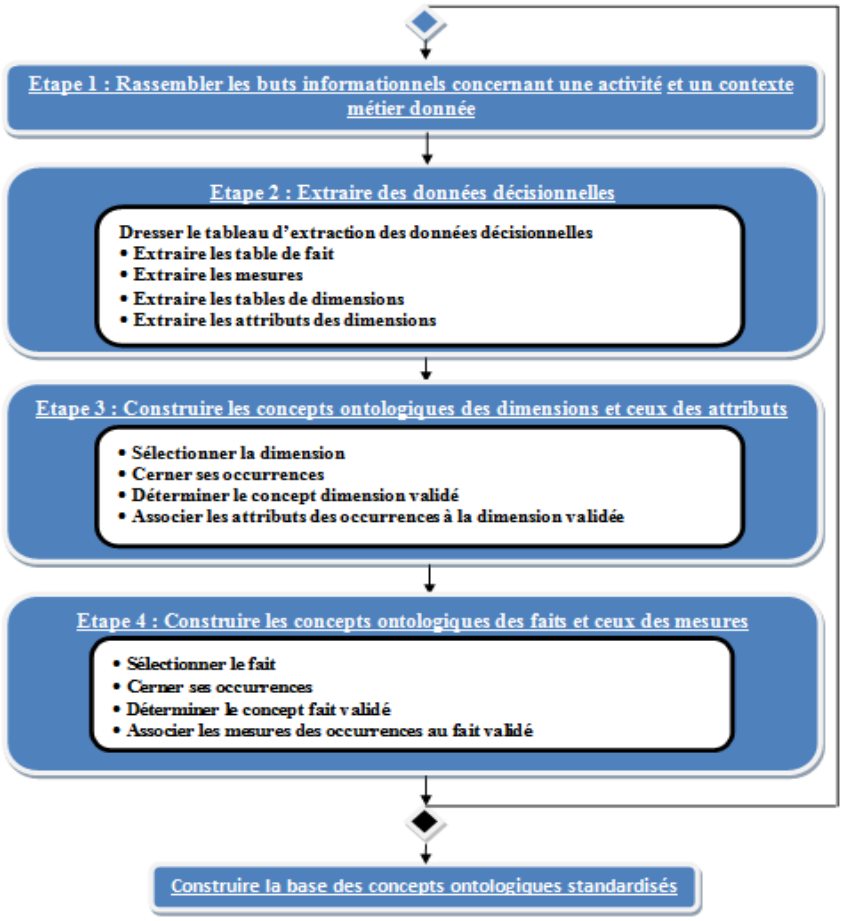


FIG.7 - *Processus de déduction des ontologies*

Les ontologies déduites seront alimentées par les concepts ontologiques des faits, les concepts ontologiques des mesures, les concepts ontologiques des dimensions et les concepts

ontologiques des attributs associées à ces dimensions. La démarche de construction de ces ontologies se compose des étapes suivantes : (1) Rassembler les buts informationnels concernant une activité donnée, (2) Extraire des données décisionnelles, (3) Construire les concepts ontologiques des dimensions et ceux des attributs, (4) Construire les concepts ontologiques des faits et ceux des mesures.

L'étape (1) consiste à délimiter l'activité de l'organisation concernée et le contexte métier pour lequel nous allons construire le vocabulaire standardisé puis collecter l'ensemble des buts informationnels à analyser. Dans l'étape (2), le concepteur est amené à extraire les données décisionnelles (faits, mesures, dimensions, attributs) en dressant le tableau d'extraction des données décisionnelles. Dans l'étape (3), pour construire les concepts ontologiques des dimensions et ceux des attributs, le concepteur doit sélectionner la dimension, cerner ses occurrences, déterminer le concept dimension validé et associer les attributs des occurrences, retenus comme vocabulaire ontologique après validation des concepteurs, à la dimension validée. Et dans l'étape (4), le concepteur doit sélectionner le fait, cerner ses occurrences, déterminer le concept fait validé et associer les mesures, retenues comme vocabulaire ontologique après validation, des occurrences au fait validé afin de construire les concepts ontologiques des faits et ceux des mesures. Lors de la construction des ontologies, l'intervention des concepteurs de SID, tout au long de processus de cette construction, est nécessaire, voire obligatoire. En effet, elle facilite l'approbation des résultats obtenus et permet la résolution des ambiguïtés rencontrées.

4.5 Tableau d'extraction des données décisionnelles

Le tableau d'extraction des données décisionnelles est une extension de modèle de formalisation de buts informationnels dans lequel nous déduisons les données décisionnelles (faits, mesures, dimensions, attributs) supportées par chaque but informationnel lié à un métier donné. Ainsi, les données décisionnelles établies concernent un contexte sélectionné pour une activité donnée.

Activité				
Contexte				
	faits	**Mesures**	**Dimensions**	**attributs**
But informationnel 1				
But informationnel 2				
But informationnel n				

FIG.8 - *Tableau d'extraction des données décisionnelles*

A travers ce tableau, nous mettons à la disposition des concepteurs un vocabulaire métier complet des données décisionnelles afin de leur faciliter la construction des concepts décisionnels ontologiques.

4.5 Tableau de validation des concepts ontologiques des dimensions

Le tableau de validation de concepts ontologiques des dimensions permet de cerner les liens sémantiques entre toutes les dimensions afin de pouvoir valider le choix de concepts ontologiques des dimensions. A l'issue de cette validation, nous pouvons construire les

concepts ontologiques des dimensions qui seront partagé entre l'ensemble des utilisateurs de SID.

Dimension : D1		
Attributs_dimension : A1_D1, A2_D1, A3_D1… An_D1		
Occurrences (O)	**Type d'occurrence**	**Attributs**
O1	Identité/ synonymie, équivalent/ l'homonyme	A1_O1, A2_O1,…
O2	Identité/ synonymie, équivalent/ l'homonyme	A1_O2, A2_O2,…
On	Identité/ synonymie, équivalent/ l'homonyme	A1_On, A2_On,…
Concept ontologique retenu		
Attributs retenus		

FIG.9 - *Tableau de validation de concepts ontologiques des dimensions*

Le tableau récapitule la tache de comparaison et de validation des dimensions extraites à partir des buts informationnels et facilite le choix de concept ontologique des dimensions à standardiser. Les attributs de la dimension retenue comme concept ontologique représentent les attributs initiaux de la dimension D1 comparés et validés avec les attributs des occurrences.

4.6 Tableau de validation des concepts ontologiques des faits

Le tableau de validation de concepts ontologiques des faits permet de cerner les liens sémantiques entre touts les faits afin de pouvoir valider le choix de concepts ontologiques des faits de la part des concepteurs. A l'issue de cette validation, nous pouvons construire les concepts ontologiques des faits qui seront partagé entre l'ensemble des utilisateurs de SID.

Fait : F1				
Mesures associées : M1_F1, M2_F1, M3_F1… Mn_F1				
Dimensions associées : D1_F1, D2_F1, D3_F1… Dn_F1				
Occurrences (O)	**Type d'occurrence**	**Mesures**	**Attributs**	**Dimensions**
O1	Identité/ synonymie, équivalent/ l'homonyme	M1_O1, M2_O1,…	A1_O1, A2_O1,…	D1_O1, D2_O1
O2	Identité/ synonymie, équivalent/ l'homonyme	M1_O2, M2_O2,…	A1_O2, A2_O2,…	D1_O2, D2_O2
On	Identité/ synonymie, équivalent/ l'homonyme	M1_On, M2_On,…	A1_On, A2_On,…	D1_On, D2_On
Concept ontologique retenu				
Mesures retenues				
Attributs retenus				
Dimensions retenues				

FIG.10 - *Tableau de validation de concepts ontologiques des faits*

Le tableau récapitule la tache de comparaison et de validation des faits extraits à partir des buts informationnels et explicite le choix des concepts ontologiques des faits à standardiser. Les mesures de fait sont extraites après validation des mesures associées au fait F1 avec celles

des occurrences afin d'éviter l'ambigüité ou une éventuelle redondance. Ces mesures représentant les concepts ontologiques des mesures standardisées. Egalement, les dimensions associées au fait, retenu comme concept ontologique, représentent les dimensions initiales de fait F1 assemblés avec les autres dimensions associées aux occurrences de ce fait.

5 Conclusion et perspectives

A travers le travail présenté dans cet article, nous avons rappelé notre démarche d'expression des besoins métier d'un SID. Nous avons proposé un modèle sémantique pour formuler les buts informationnels permettant une représentation systématique des besoins des décideurs en adoptant une approche intentionnelle. Les buts sont formulés via un verbe, une section des paramètres_faits et une section des paramètres_dimensions. La formulation de ces buts suivant cette définition va permettre de déterminer les dimensions et tables de faits à inclure dans le modèle multidimensionnel. Dans la deuxième partie, nous avons cité l'ensemble des modèles de traitement et de formalisation des besoins métier d'un SID qui permettent d'associer l'ensemble des buts métier de l'organisation. Ces modèles sont : le modèle « Diagnostic de l'Organisation » qui permet de noter l'ensemble des activités liées au métier de l'organisation ainsi que les contextes rattachés à chacune de ces activités, le modèle d'association des buts stratégiques à des buts tactiques, le modèle d'association des buts tactiques à des buts informationnels, et le modèle de formalisation des buts informationnels. Ensuite, nous avons détaillé le processus de construction des ontologies à partir des besoins métier d'un SID qui représente un ensemble des étapes menant à la construction des ontologies de résolution des buts informationnels prédéfinis suivant un cycle de vie itératif et progressif. Ces ontologies seront alimentées par les concepts ontologiques des faits, les concepts ontologiques des mesures, les concepts ontologiques des dimensions et les concepts ontologiques des attributs associées à ces dimensions. La démarche de construction de ces ontologies se compose des étapes suivantes : (1) Rassembler les buts informationnels concernant une activité donnée, (2) Extraire des données décisionnelles, (3) Construire les concepts ontologiques des dimensions et ceux des attributs, (4) Construire les concepts ontologiques des faits et ceux des mesures.

Dans un futur travail, nous allons nous intéresser à formaliser notre solution. Nous souhaitons développer un environnement à base des ontologies proposées et utilisant notre modèle sémantique de formalisation des buts informationnels pour pouvoir assister les décideurs dans la spécification de leurs besoins décisionnels.

6 Bibliographie

Annoni, E., Eléments méthodologiques pour le développement des systèmes décisionnels dans un contexte de réutilisation, Thèse de Doctorat, Université de Toulouse 1, Toulouse, France, 2007.

Anton A., « Goal based Requirements Analysis », Proc. of the 2 Int. Conf. On Requirements Engineering, 1996.

El Golli, I.G., Ingénierie des Exigences pour les Systèmes d'Information Décisionnels : Concepts, Modèles et Processus (la méthode CADWE), Thèse de Doctorat, Université Paris-Panthéon-Sorbonne, France, 2008.

Gruber T.R., « A translation approach to portable ontology specification. », Knowledge Acquisition 5, 1993, p. 199-220.

Kavakli E., Loucopoulos P., « Goal Modelling in Requirements Engineering: Analysis and Critique of Currents Methods », Information Modeling Methods and Methodologies; Krogstie J., Halpin T. and Keng S., 2004.

Rifaieh, R. D., Utilisation des ontologies contextuelles pour le partage sémantique entre les systèmes d'information dans l'entreprise, Thèse de doctorat, Institut National des Sciences Appliquées, Villeurbanne, France, 2004.

Rolland C., Loucopoulos P., Kavakli V., Nurcan S., « Intention-based Modelling of Organizational Change », Int. Workshop on Evaluation of Modelling Methods in System Analysis and Design, 1999, Heidelberg, Germany.

Sabri A., Kjiri L., (2011).Vers une ontologie pour la formulation des besoins d'un Système d'Information Décisionnel, *International Workshop on Information Technologies and Communication (WOTIC'11)*, ENSEM, Casablanca, Maroc, 13-15 Octobre 2011.

Sabri A., Kjiri L., (2012a). Une démarche d'analyse à base de patrons pour la découverte des besoins métier d'un Système d'Information Décisionnel, *Atelier aIde à la Décision à tous les Etages AIDE, 12e Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances EGC*, 31 janvier - 3 février 2012, Bordeaux, France, 2012.

Sabri A., Kjiri L., (2012b). Processus de traitement et de formalisation des besoins métier d'un Système d'Information Décisionnel, *Inforsid2012 : INFormatique des ORganisations et Systèmes d'Information et de Décision*, Montpellier, France, 29-31 Mai 2012.

Van Lamsweerde A., « Goal-oriented Requirements Engineering : a Guided Tour », Proc.RE'01, 5th IEEE Int. Symposium on Requirements Engineering, Toronto, 2001.

Summary

At the conceptual level of a SID, the step of expressing requirements raises several major problems that come from the diversity of perspectives, points of view, contexts and actor's profiles, inconsistency and ambiguity semantics. To remedy this, we defined the requirements of an organization in the form of goals. Each goal is expressed by a verb, a section paramètres_faits and a section paramètres_dimensions. Then, we define an interpretation of the informational goal which helps to extract business intelligence data for analyzing. This interpretation is made by using a meta-model goals and using ontology. Thus, we will detail the process of building ontologies based on business requirements of the SID. We aim to build a business vocabulary that will be shared and used by designers in the expression of business goals. The process of building ontologies based on a set of steps explained in the article's body.

Relational.OWL2E : Une nouvelle approche de représentation du schéma d'une base de données relationnelle basée sur OWL2

Naïma Souâd Ougouti*, Hafida Belbachir*, Dolière Francis Some*,
Ismael Abraham Ouattara*
* Laboratoire LSSD, USTO-MB BP 1505 El M'Naouer,
Oran, Algeria
{ s_ougouti, h_belbach, some.frncs, ismaelouattara03 }@yahoo.fr

Résumé. Un des domaines de recherche qui intéressent la communauté scientifique de nos jours est le problème de la médiation dans les systèmes pair-à-pair (P2P). Dans ce contexte nous avons proposé un nouveau système de gestion de données hétérogènes et distribuées dans un environnement P2P nommé MedPeer. Parmi les fonctions de ce système, nous nous intéressons dans cet article à la description des bases de données relationnelles en utilisant des ontologies. Nous proposons donc Relational.OWL2E, une nouvelle approche qui génère à partir du schéma relationnel une ontologie basée sur le langage OWL dans sa deuxième version (OWL2).

1 Introduction

On assiste depuis quelques années à l'émergence de nouvelles applications qui ont besoin de partager des informations entre différents systèmes. C'est le cas du e-gouvernement, du e-Learning, du e-commerce, de la bioinformatique ou encore des bibliothèques électroniques. Or, dans ce contexte, les systèmes d'informations, conçus et développés par des organisations différentes, constituent généralement des sources de données hétérogènes et autonomes.

Ainsi, l'interopérabilité est devenue une nécessité pour répondre au besoin d'échange d'informations entre systèmes d'informations hétérogènes. Elle traduit la capacité d'un système d'informations à collaborer avec d'autres systèmes de nature parfois très différente. Elle a pour objectif de développer des architectures et des outils pour le partage, l'échange et le contrôle des données.

Dans ce contexte, nous avons présenté un nouveau système d'intégration des données dans un environnement P2P nommé MedPeer (Ougouti et al., 2011). Ce système repose sur une architecture Super-pair s'appuyant sur un regroupement des pairs selon le type de média (Textes, Images, Bases de données relationnelles, semi-structurées,..). Cette architecture combine l'approche centralisée et celle non structurée prenant ainsi les avantages de la recherche centralisée, de l'autonomie, de la répartition des charges et de la robustesse pour une recherche distribuée.

Chaque super-pair gère les pairs traitant le même type de média qu'il est en charge de représenter, il est choisi en fonction de ses capacités en terme de capacités de calcul et de

Relational.OWL2E

bande passante. Il doit en outre disposer de toutes les informations nécessaires pour pouvoir orienter les requêtes qui lui arrivent vers les pairs pertinents. Les super-pairs forment entre eux un réseau pair-à-pair pur. Lorsque les pairs ont des schémas différents à gérer, une médiation sémantique est nécessaire. Dans cet article, nous nous intéressons à ce dernier problème puisque nous présentons un nouveau format de représentation pour les schémas relationnels basé sur le langage d'ontologies Web OWL dans sa deuxième version nommé Relationnel.OWL2E. En exploitant les différentes opportunités fournies par OWL2 (Golbreich C. et Wallace K, 2009) et notre ontologie, nous sommes capables aujourd'hui de décrire et de partager n'importe quel schéma d'une base de données relationnelle.

Cet article est organisé comme suit :

Dans la section 2, nous présentons un état de l'art des principales approches de description des bases de données relationnelles. En section 3, nous introduisons Relationnal.OWL2E, notre ontologie OWL2 pour la représentation d'un schéma relationnel. Dans la section 4, nous illustrons notre approche par un exemple. Enfin nous terminons par une conclusion.

2 Etat de l'art

En voulant tirer profit des bénéfices qu'apporte le web sémantique, plusieurs travaux dont le but est le passage d'une base de données relationnelle à un format plus récent (XML/RDF/OWL) ont vu le jour. Nous avons choisi de présenter quatre approches (Perez de Laborda et al. , 2005) (Dejing D. et al., 2006) (Intellidimension Company, 2000) (Nguyen T.D.T, 2008), d'autres méthodes plus récentes se trouvent dans (Sequeda J.F., 2012) (Arenas M. et al., 2011) (Cullot, N et al., 2007).

Relational.OWL (Perez de Laborda et al. , 2005) permet de traduire presque tous les concepts du modèle relationnel en ontologies OWL, du schéma relationnel aux données en passant par les contraintes d'intégrité. Ce système définit quatre classes et un ensemble de propriétés qui permettent de relier entre elles.

Dans le tableau1 sont répertoriées les classes prédéfinies et dans le tableau 2 se trouvent les différentes propriétés.

Les préfixes rdf, rdfs, dbs, xsd et owl représentent des espaces de noms utilisés dans l'ontologie Relational.OWL

rdf :ID	rdfs :subClassOf	rdfs :comment
dbs : Database	rdf :Bag	Classe des bases de données
dbs :Table	rdf :Seq	Classe des tables de la base
dbs :Column	rdf :Ressource	Classe des colonnes
dbs :PrimaryKey	rdf:Bag	Clé primaire d'une table

TAB.1-*Classes de Relationnal.OWL* Classe des bases de données

rdf :ID	rdfs :domain	rdfs :range	rdfs :comment
dbs :hasTable	dbs:Database	dbs:Table	Ensemble de tables
dbs :hasColumn	dbs:Table	dbs:Column	Ensemble de colonnes
dbs:isIdentifiedBy	dbs:Table	dbs:PrimaryKey	Clés primaires des tables
dbs:references	dbs:Column	dbs:Column	Clés étrangères
dbs:length	dbs:Column	xsd:nonNegativeInteger	Longueur d'un champ.
dbs:scale	dbs:Column	xsd:nonNegativeInteger	Partie décimale

TAB.2- *Propriétés de Relational.OWL*

OntoGrate (Dejing D. et al., 2006) est un système d'intégration de bases de données relationnelles dans un environnement P2P (Pair-à-Pair). Pour représenter les schémas relationnels en ontologies OWL, les auteurs ont étendu l'expressivité des langages d'ontologies web. Ils ont ainsi introduit un nouveau langage, Web-PDDL, extension de PDDL (*Planning Domain Defintion Language*) basé sur la logique des prédicats du premier ordre. Dans un premier temps, les concepts de la base de données sont traduits en utilisant le langage Web-PDDL. Une fois l'ontologie générée, le système dispose d'un adaptateur de syntaxe, PDDOWL, qui traduit ainsi la première ontologie Web-PDDL en ontologie OWL. Dans l'ontologie finale générée, une table est transformée en une classe, sous classe de la classe sql:Relation (Définie dans le système OntoGrate, comme étant la classe représentant les tables), un attribut est transformé en propriété OWL, une contrainte est vue comme un axiome (règle) et une clé primaire comme une contrainte fonctionnelle OWL (owl:FunctionalProperty).

RDF Gateway (Intellidimension Company, 2000) est un système qui permet de traduire le schéma relationnel d'une base de données en une ontologie RDFS ou OWL, via le paramètre *schema_type*, qui spécifie la sortie par défaut de l'ontologie.

Le SQL Data service est un module du système RDF Gateway qui interroge la base de données et en extrait le schéma relationnel qu'il traduit en ontologies RDFS ou OWL. Dans ce système une table est traduite en classe, un attribut en une propriété *rdfs:property* pour une sortie RDFS ou *owl:DatatypeProperty* pour une sortie OWL, une clé étrangère en une propriété *rdfs:property ou owl:ObjectProperty* et enfin les types de données des attributs sont traduits en type de données du Schéma XML.

OWL_K (K pour *Key* ou clé) (Nguyen T.D.T, 2008) est une extension du langage OWL pour la gestion des contraintes d'identification. Les contraintes d'identification sont l'équivalent des clés primaires du modèle relationnel. Ce travail a été motivé par les difficultés du dialecte OWL DL à capturer toute la sémantique des contraintes d'identification. Le vocabulaire par défaut de OWL a été étendu afin de prendre en compte les contraintes d'identification.

Le système propose :

- La classe *ICAssertion* qui représente la contrainte d'identification.
- La propriété *onClass* qui est la classe (table) sur laquelle porte la contrainte d'identification
- La propriété *byProperty* qui représente une propriété (attribut) participant à la contrainte d'identification.

Le langage de logique descriptive par défaut de OWL a aussi été étendu pour prendre en compte la sémantique des nouveaux concepts introduits.

Dans ce système, les types de données sont traduits en utilisant XML Schéma et les clés étrangères sont traduites en utilisant les contraintes de cardinalités (*owl :minCardinality, owl :cardinality, owl :maxCardinality*).

3 Relational.OWL2E

La solution que nous proposons ici est une extension du système Relational.OWL proposé en (Perez de Laborda et al. , 2005). Notre choix s'est porté sur cette approche à cause de sa particularité à traduire presque tous les concepts du modèle relationnel en ontologies OWL.

Notre apport majeur se situe au niveau de la sémantique que nous avons ajoutée aux différents concepts des bases de données relationnelles, en traduisant entre autres, les attributs par des types de données riches du schéma XML, les clés primaires, les clés uniques, les clés étrangères, et en tenant compte des contraintes NULL et NOT NULL du modèle relationnel.

Nous avons appelé cette ontologie Relational.OWL2E parce qu'elle est basée sur OWL2 et sur Relational.OWL que nous avons étendu (E).

Nous obtenons les informations sur le contenu de la base de données à partir de son dictionnaire de données (catalogue) et nous donnons dans ce qui suit la représentation des tables, attributs (colonnes), types de données avec éventuellement les restrictions nécessaires, les clés primaires, les clés uniques et les clés étrangères.

Nous avons donc défini 5 classes et 7 relations (propriétés) entre elles, elles sont résumées dans les deux tableaux suivants :

Classes	Commentaires
Database	Classe représentant une base de données
Table	Classe représentant une table
Column	Classe représentant une colonne
PrimaryKey	Classe représentant une clé primaire
UniqueKey	Classe représentant une clé unique

TAB3.- *Classes dans Relational.OWL2E*

Propriétés	rdfs :domain	rdfs :range	Commentaires
has	owl :Thing	owl :Thing	Une chose possède une autre chose.
hasTable	Database	Table	Une table fait partie d'une base de données
hasColumn	Table PrimaryKey UniqueKey	Column	Une colonne fait partie d'une table
hasPrimaryKey	Table	PrimaryKey	Une clé primaire identifie une relation
hasUniqueKey	Table	UniqueKey	Une table peut avoir des contraintes d'unicité sur certains attributs
hasForeignKey	Table	Table	Une table référence une autre table dans une relation de clé étrangère.
references	Column	Column	Une colonne référence une autre colonne dans une clé étrangère.

TAB.4-*Propriétés dans Relational.OWL2E*

3.1 Sérialisation de l'ontologie Relational.OWL2E.

Dans ce qui suit nous donnons quelques extraits de sérialisation, dans la syntaxe RDF/XML de l'ontologie Relational.OWL2E dont les classes et les propriétés ont été décrites ci-dessus.

Définition d'une classe

```
<owl:Class rdf:ID="Table">
       <rdfs:subClassOf   rdf:about="http://www.w3.org/1999/02/22-rdf-
       syntax-ns#Bag"/>
       <rdfs:label xml:lang="fr">Table</rdfs:label>
       <rdfs:label xml:lang="en">Table</rdfs:label>
       <rdfs:comment  xml:lang="fr">Classe  des  tables  d'une  base  de
       données.</rdfs:comment>
       <rdfs:comment     xml:lang="en">The     class     of     database
       tables.</rdfs:comment>
</owl:Class>
```

Définition d'une propriété

```
<owl:ObjectProperty rdf:ID="hasTable">
             <rdfs:subPropertyOf rdf:resource="#has" />
             <rdfs:domain rdf:resource="#Database"/>
             <rdfs:range rdf:resource="#Table" />
```

Relational.OWL2E

```
            <rdfs:label xml:lang="fr">aPourTable</rdfs:label>
            <rdfs:label xml:lang="en">hasTable</rdfs:label>
            <rdfs:comment xml:lang="fr">Une base de données
            contient un ensemble de tables </rdfs:comment>
            <rdfs:comment xml:lang="en">A Database has a set of
            Tables.</rdfs:comment>
</owl:ObjectProperty>
```

3.2 Algorithme de passage d'une base de données relationnelle à une ontologie OWL 2 : cas de MySQL

3.2.1 Première partie

On commence par extraire les tables d'une base de données

- Le nom de la base de données représente une classe, de type *Database* de Relational.OWL2E
- Le nom de chaque table sera exprimé comme valeur de la propriété *hasTable* de Relational.OWL2E

Pour chaque table, on extrait la liste des attributs, la clé primaire, les clés uniques et étrangères.

- Le nom de chaque table est une classe de type *Table* de Relational.OWL2E
- Le nom de chaque attribut sera exprimé comme valeur de la propriété *hasColumn*
- La clé primaire sera exprimée par la propriété *hasPrimaryKey* sur la classe *PrimaryKey* (représentant la clé primaire) et contenant la liste des attributs participant à la clé, chaque attribut étant exprimé comme valeur de la propriété *hasColumn*
- La clé unique sera exprimée similairement à la clé primaire, mais avec les propriétés *hasUniqueKey* vers la classe *UniqueKey* contenant la liste des attributs participants à la clé, chaque attribut étant exprimé comme valeur de la propriété *hasColumn*
- La clé étrangère sera exprimée par la propriété *hasForeignKey*. La valeur de cette propriété sera la table référencée par la clé étrangère. Chaque attribut de la clé étrangère sera exprimé comme instance de la classe *Column* et liée à la colonne référencée par la propriété *references* ayant pour valeur l'attribut référencé.

Le résultat de cette première partie de l'algorithme sera une ontologie reprenant les termes du schéma d'une base de données relationnelle dans les termes du schéma de l'ontologie que nous avons définit.

3.2.2 Deuxième Partie

1) Chaque attribut sera exprimé comme une propriété de type de données, dont le domaine (*rdfs :domain*) est le nom de la classe représentant la table contenant l'attribut et l'image (*rdfs :range)* le type de données de l'attribut traduit en un type de données du schéma XML de la façon suivante :
   - Les types de données entier seront exprimés par le type *integer* du schéma XML, avec éventuellement des restrictions sur les intervalles de valeurs du type, grâces aux facettes *maxInclusive, maxExcusive, minInclusive, minExclusive* du schéma XML.

- Les types de données numériques à virgules (décimaux), nous les exprimons par le type *decimal* du schéma XML avec éventuellement des restrictions grâces aux facettes *totalDigits* (nombre total de caractères du type) et *fractionDigits* (nombre de chiffres après la virgule).
- Les types de données textes seront exprimés par le type *string* du schéma XML. Nous utilisons les facettes *minLength* et *maxLength* pour exprimer le nombre de caractère minimal et maximal autorisé dans les chaînes de ce type.
    - Pour la valeur de la facette *minLength*, si l'attribut accepte les valeurs nulles, alors, *minLength* vaudra 0, sinon 1
    - Pour le type set, on le traduit par un type string, dont la valeur de la facette *maxLength* est à extraire du catalogue MySQL.
    - Pour le type de données *enum*, on l'exprimera par la propriété *owl :oneOf* composée des différentes valeurs prises par l'attribut *enum.*
- Les types temporels seront exprimés par l'un des nombreux types temporels du schéma XML.
- Les types de données binaires, sont similaires aux types de données textes, à la seule différence qu'ils sont exprimés par le type *hexBinary* du schéma XML.Comme dans le cas des types textes, la valeur de la facette *minLength* est 0 si les valeurs de l'attribut peuvent être nulles ou 1 sinon.

2) La clé primaire sera exprimée par la propriété *owl :hasKey de OWL2* sur le nom de la classe représentant la table contenant cette clé et ayant pour valeurs la liste des attributs participant à la clé primaire.
3) Chaque clé unique (son nom) sera exprimée comme une sous-classe (de la classe contenant la clé unique) contenant la propriété *owl :hasKey* sur une classe équivalente de la classe représentant la table contenant la clé unique et ayant pour valeurs la liste des attributs participant à la clé unique
4) Les clés étrangères seront exprimées par des restrictions de propriétés (owl:Restriction) sur le nom de chaque attribut participant à la clé (owl:onProperty) vers l'attribut référencé (owl:someValuesFrom).

4 Exemple d'application

Dans cette section, nous donnons quelques extraits de l'ontologie OWL qui décrit la base de données relationnelle *Elevage* sous MYSQL. Elle est constituée de trois tables Espece, Race et Animal.

4.1 Schéma relationnel à décrire

```
CREATE DATABASE Elevage ;

CREATE TABLE Espece (
    id smallint(6) not null auto_increment,
    nom_latin varchar(40) not null,
    primary key(id),
    unique key nom_latin (nom_latin));
```

Relational.OWL2E

```
CREATE TABLE Race (
    id smallint(6) not null auto_increment,
    espece_id smallint(6),
    primary key(id),
    constraint fk_race_espece_id foreign key(espece_id) references Espece(id));

CREATE TABLE Animal(
    id smallint(6) not null auto_increment,
    sexe enum('male', 'femelle'),
    date_naissance datetime not null,
    nom varchar(30),
    espece_id smallint(6) not null,
    race_id smallint(6),
    primary key(id),
    constraint fk_espece_id foreign key (espece_id) references Espece(id),
    constraint fk_race_id foreign key(race_id) references Race(id));
```

4.2 Quelques extraits de l'ontologie Relationnal.OWL2E générée

Description de la base de données

```
<owl:Class rdf:ID="Elevage">
            <rdf:type rdf:resource="#Database">
            <hasTable rdf:resource="#Espece" />
            <hasTable rdf:resource="#Race" />
            <hasTable rdf:resource="#Animal" />
</owl:Class>
```

Description d'une table

```
<owl:Class rdf:ID="Espece">
      <rdf:type rdf:resource="#Table" />
      <hasColumn rdf:resource="#Espece.id" />
      <hasColumn rdf:resource="#Espece.nom_latin" />
      <hasPrimaryKey>
            <PrimaryKey>
            <hasColumn rdf:resource="#Espece.id" />
      </PrimaryKey>
            </hasPrimaryKey>
            <hasUniqueKey>
            <UniqueKey>
            <hasColumn rdf:resource="#nom_latin" />
      </UniqueKey>
            </hasUniqueKey>
</owl:Class>
```

Description d'un attribut

< !--Attribut id de la table Espece--!>

```
<owl:DatatypeProperty rdf:ID="Espece.id">
            <rdf:type rdf:ressource="#Column" />
            <rdfs:domain rdf:ressource="#Animal" />
            <rdfs:range
rdf:ressource="http://www.w3.org/2001/XMLSchema#short" />

   </owl:DatatypeProperty>
```

<!--Attribut nom_latin de la table Espece-->

```
<owl:DatatypeProperty rdf:ID="Espece.nom_latin">
<rdf:type rdf:ressource="#Column" />
<rdfs:domain rdf:ressource="#Espece" />
<rdfs:range>
<rdfs:Datatype>
<owl:onDatatype
rdf:ressource="http://www.w3.org/2001/XMLSchema#string" />
<owl:withRestrictions rdf:parseType="Collection">
<rdf:Description>
<xsd:minLength
rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">1</xsd
:minlength>
</rdf:Description>
<rdf:Description>
<xsd:maxLength
rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">
40</xsd:maxLength>
</rdf:Description>
</owl:withRestrictions>
</rdfs:Datatype>
</rdfs:range>
</owl:DatatypeProperty>
```

Clé primaire et clé unique

<!-- Clé primaire de la table Espece -->

```
      <owl:Class rdf:about="Espece">
      <owl:hasKey rdf:parseType="Collection">
            <owl:DatatypeProperty rdf:resource="#Espece.id" />
      </owl:hasKey>
</owl:Class>
```

<!-- Clé unique de la table Espece -->

```
<owl:Class rdf:about="Espece">
      <owl:equivalentClass>
            <owl:Class>
```

Relational.OWL2E

```
            <owl:hasKey rdf:parseType="Collection">
            <owl:DatatypeProperty
rdf:resource="#Espece.nom_latin" />
            </owl:hasKey>
            </owl:Class>
      </owl:equivalentClass>
</owl:Class>
```

Clé étrangère

```
<!-- Clés étrangères de la table Race  -->
<owl:Class rdf:about="Race">
      <owl:equivalentClass>
      <owl:Restriction>
      <owl:onProperty rdf:ressource="#Race.espece_id" />
      <owl:someValuesFrom rdf:ressource="#Race.id" />
      </owl:Restriction>
      </owl:equivalentClass>
</owl:Class>
```

5 Conclusion

L'intégration des bases de données relationnelles dans le web sémantique passe par une description préalable de celles-ci.

Dans ce travail, nous avons introduit notre nouvelle approche qui a cet objectif en utilisant une ontologie. Nous avons pour cela détaillé les différentes étapes de l'algorithme qui mènent d'un schéma relationnel vers l'ontologie Relationnal.OWL2E écrite avec le langage OWL2, tout en respectant toute la sémantique contenue dans la structure du schéma.

L'algorithme fourni peut être facilement implémenté pour des systèmes de gestion de bases de données relationnelles autres que MySQL.

L'ontologie fournie, peut être améliorée et corrigée pour prendre en charge d'autres concepts spécifiques.

Références

Arenas M., Prud'hommeaux E. and Sequeda J, (2011), Direct mapping of relational data to RDF. W3C Working

Cullot, N., Ghawi, R. and Yetongnon, K. (2007), DB2OWL: A Tool for Automatic Database to Ontology Mapping. In: Proc. of 15th Italian Symposium on Advanced Database Systems (SEBD 2007), Torre Canne, 491-494.

Dejing D., LePendu P., Shiwoong K., and Peishen Q., (2006), Integrating Databases into the Semantic Web through an Ontology-Based Framework. In ICDEW '06: Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06), Washington.

Golbreich C. et Wallace K. (2009), OWL2 Web Ontology Language New Features and Rationale. W3C Recommendation.

Intellidimension Company. RDF Gateway, (2000). Available at http://www.intellidimension.com.

Nguyen T.D.T. ,(2008), A Dl-Based Approch To Integrate Relational Data Sources Into The Semantic Web. Thèse de doctorat, Université de Nice-Sophia Antipolis, France.

Ougouti, N.S., Belbachir, H., Amghar and Y., Benharkat, A.N., (2011), Architecture Of MedPeer : A New P2P-based System for Integration of Heterogeneous Data Sources. In: Proceedings of the International Conference on Knowledge Management and Information Sharing (KMIS 2011), Paris, 351-354.

Perez de Laborda et C., Conrad S., (2005), Relational.OWL – A Data and Schema Representation Format Based on OWL. In: Second Asia-Pacific Conference on Conceptual Modelling (APCCM2005), Newcastle, volume 43 of CRPIT, 89–96.

Sequeda J.F., Tirmizi S.H., Corcho O., and Miranker D.P. (2012), Survey of directly mapping sql databases to the semantic web. Knowledge Eng. Review.

Summary

Nowadays, scientific community is more and more interested by the mediation problem into Peer-to-Peer systems (P2P) and by migration of data sources into semantic web. In this context we have proposed a new heterogeneous and distributed data management system in a P2P environment called MedPeer. Among this system functions, we are interested in this article in relational databases description by using ontologies. We thus propose Relational.OWL2E, a new approach which generates starting from relational schema an OWL2 based ontology.

BRMAP : Un outil d'Alignement des ontologies

Saida Gherbi*
Mohamed Tarek Khadir **, Habiba Belleili ***

*Université Badji Mokhtar. BP 12, 23000 Annaba, Algérie
Gharbi@labged.net
** Université Badji Mokhtar. BP 12, 23000 Annaba, Algérie
khadir@labged.net
*** Université Badji Mokhtar. BP 12, 23000 Annaba, Algérie
Belleili@labged.net

Résumé. Le but de cet article est de proposer un nouveau outil qu'améliore le rendement des techniques d'alignement qui se basent sur la structure ou la richesse du langage de représentation des ontologies, appelé BRMAP (Background Raisoner MAPping**).** Il est fondé sur une approche combinant les techniques d'appariement terminologiques, structurelles et les connaissances complémentaires en intégrant le raisonneur Pellet. Cet outil est implémlenté en trois phases, la première phase d'alignement calcule les valeurs de similarité entre les concepts de l'ontologie source et cible en utilisant les mesures terminologiques et structurelles. La deuxième, la phase d'ancrage apparie chaque concept des deux ontologies non aligné à la phase précédente avec les concepts de l'ontologie complémentaire et donne des ancres. La dernière phase, la dérivation s'appuie sur l'ontologie complémentaire, le raisonneur, les ancres pour définir des appariements.

1 Introduction

La représentation explicite d'une ontologie dépend toujours d'assomptions implicites comme les objectifs des concepteurs, leurs capitaux connaissances, rendant l'objectivité de la représentation un but difficile à concrétiser. Ces connaissances implicites sont à l'origine de différentes formes d'hétérogénéité entre les ontologies, même entre les ontologies décrivant le même domaine. Le choix d'une ontologie particulière ou l'exploitation de plusieurs d'entre elles devient difficile, ce que nécessite une comparaison entre ces ontologies afin de passer de l'une à l'autre ou de les intégrer, en générant le plus automatiquement possible des relations ou appariements entre les concepts de deux ontologies.

L'évolution continuelle des ontologies Euzenat et al. (2007) a engendré plusieurs versions de la même ontologie, ce qui a met les développeurs et les ingénieurs de la connaissance dans la confusion, ne sachant pas ce qui a changé. L'alignement va permettre d'identifier les différences entre deux versions : les entités qui ont été ajoutés, supprimés ou renommés.

La recherche d'alignement entre les concepts des ontologies se base des techniques d'alignement terminologiques, structurelles, extensionnelles et sémantiques, qui proviennent de disciplines variées, telles que la fouille de données, les sciences du langage, les statistiques ou la représentation des connaissances. Ces techniques d'alignement tirent parti des différents aspects des ontologies (leur structure, les noms des différents éléments, les objets, la sémantique du langage) P. Shvaiko et al. (2008), M.ELBYED. (2009), J. Euzenat et al.

(2011), P. Shvaiko et al. (2012). Pour compléter ces méthodes d'appariement qui ne donnent pas des bons résultats quand les ontologies à apparier sont faiblement structurées ou se limitent à de simples hiérarchies de classification. De nombreux travaux portent actuellement sur l'utilisation de connaissances complémentaires, dites de "background" ou de support, représentées le plus souvent sous la forme d'une 3ème ontologie, Reynaud et al. (2006), Aleksovski et al. (2006, 2006b), Sabou et al. (2006), P. Shvaiko et al. (2012).

Le schéma général d'un processus d'alignement utilisant des connaissances complémentaires pour aligner deux ontologies, est devenu aujourd'hui un processus bien connu. Par contre, la façon dont le processus est implémenté et les problèmes posés dépendent en grande partie de la nature des connaissances complémentaires utilisées, dans notre cas il s'agit de connaissances décrivant une turbine à vapeur dans le domaine de l'industrie électrique.

Notre approche se fonde sur l'utilisation d'une ontologie complémentaires O_{BK} et un raisonneur, afin de découvrir automatiquement des correspondances sémantiques entre concepts de deux ontologies O_{Src} et O_{Tar} hétérogènes, représentées dans le langage de description d'ontologie OWL recommandé par le W3C. L'ontologie complémentaire est une description de la turbine à vapeur donnée par un expert, cette description est plus détaillée que les deux ontologies à aligner.

2 L'alignement En Utilisant Une Ontologie Complémentaire

Cet alignement utilise une ontologie complémentaire (Background ontology) O_{BK} Reynaud et al. (2006b), pour aligner les concepts des deux ontologies, source O_{Src}, et cible (target) O_{Tar}. Pour simplifier la présentation générale de ce processus, nous considèrerons que chaque ontologie O ne comprend qu'un ensemble de concepts C et un ensemble de relations R entre ces concepts.

L'approche générale se décompose en 2 phases : l'ancrage et la dérivation, et consiste à identifier l'existence d'un appariement de la forme (X_{Src} relation Y_{Tar}) où X_{Src} Є C_{Src}, Y_{Tar} Є C_{Tar}, et relation Є R, l'ensemble des relations exprimables entre deux concepts appartenant respectivement à l'une et à l'autre des ontologies considérées.

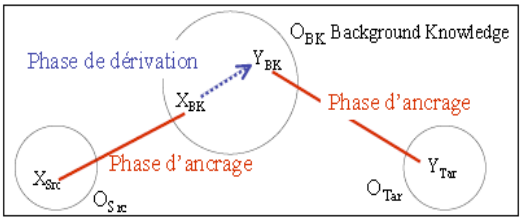


FIG. 1–*Schéma général de dérivation d'un mapping (X_{Src} relation Y_{Tar}).* Safar *et al. (2007).*

L'ancrage consiste tout d'abord à apparier chacun des 2 concepts X_{Src} et Y_{Tar}, pris indépendamment l'un de l'autre, avec un ou des concepts de la 3ème ontologie (O_{BK}), c'est-à-dire, à identifier des mappings de la forme (X_{Src} relation X_{BK}) et (Y_{Tar} relation Y_{BK}) où X_{BK} et Y_{BK} Є C_{BK} et sont appelés des ancres ou points d'ancrage.

La dérivation consiste ensuite à s'appuyer sur la structuration de O_{BK} pour :

- soit rechercher s'il existe des relations entre les différents points d'ancrage X_{BK}, Y_{BK} identifiés, afin d'essayer d'en dériver des relations (des mappings " sémantiques ") entre les éléments des ontologies à apparier,
- soit utiliser une mesure de similarité entre noeuds d'un même graphe, pour identifier pour chaque ancre X_{BK} d'un concept de l'ontologie source, l'ancre Y_{BK} du concept de l'ontologie cible qui lui est le plus similaire.

Notre approche qui sera illustrée dans la section suivante, se différencie de celle-ci par le fait qu'à la phase d'ancrage nous calculons la similarité au niveau linguistique et structurelle, et à la phase de dérivation nous utilisons un raisonneur afin d'inférer les relations existantes entre les différents points d'ancrage identifiés au sein de O_{BK}, puis d'en dériver des relations entre les éléments de O_{Src} et O_{Tar}.

3 Notre Approche

Afin de simplifier la présentation de notre approche, nous considèrerons qu'une ontologie O est un ensemble de concepts C et un ensemble de relations R entre ces concepts. Notre processus d'alignement entre ontologies se décompose en trois phases. La première consiste à mettre en correspondance les concepts de O_{Src} avec les concepts de O_{Tar} , en identifiant un mapping de la forme (X_{Src} relation Y_{Tar}) où X_{Src}ЄC_{Src}, Y_{Tar} Є C_{Tar}, et relation Є R, où R est l'ensemble des relations exprimables entre deux concepts appartenant respectivement à l'une et à l'autre des ontologies considérées. Les concepts non alignées dans la phase précédente vont être l'entrée de la deuxième phase appelée phase d'ancrage, et le résultat de cette dernière sera l'entrée de la phase de dérivation.

3.1 La phase d'alignements

Elle se base sur le calcule des valeurs de similarité entre les classes et les propriétés décrivant les concepts des deux ontologies (O_{Src} et O_{Tar}). Ce calcul utilise les mesures de similarité linguistiques et structurelles appelées les valeurs de similarité partielles entre deux entités, qui vont être stockées dans une base des similarités (Vecteurs). Ces valeurs de similarité partielles sont ensuite agrégées par une mesure de combinaison (la somme pondérée par des poids pour atteindre une seule valeur de similarité finale entre deux entités comprise entre 0 et 1).

- La similarité au niveau linguistique : La similarité linguistique de deux classes est calculée à partir des composantes linguistiques de la classe. En OWL, les composantes linguistiques dans une description d'une classe sont le nom et les étiquettes. Elles sont décrites en utilisant des primitives telles que "rdf :id" pour les noms, et "rdfs :label" pour les étiquettes. Dans notre algorithme on s'intéresse uniquement à la mesure de similarité construite à partir du nom de la classe parce qu'elle reflète son identification unique dans toute l'ontologie. Ce nom peut être un mot, un terme, ou une expression (une combinaison des mots). Le calcul de la valeur de similarité de deux noms est effectué dans deux étapes : la normalisation et la comparaison.

  L'étape de normalisation convertit un nom de classe en un ensemble d'unités lexicales appelées des tokens. Un nom est découpé en plusieurs tokens grâce à la

ponctuation, à la casse (majuscule), aux symboles spéciaux, aux chiffres, par exemple, le nom de classe "AirCanada " est découpé en deux tokens "Air " et "Canada". Elle inclut aussi une expansion de token : les abréviations, les acronymes sont élargis, par exemple le token "WS " est élargi à " Web ", "Sémantique". Cette expansion est effectuée grâce à un dictionnaire externe, dont chaque entrée est une paire composée d'un token (abréviation ou acronyme) et d'un ensemble de mots qui correspondent au token. Ce dictionnaire est soit construit spécialement pour le domaine des ontologies à aligner, soit il s'agit d'un dictionnaire général contenant des termes communs et ce dernier est le dictionnaire opté pour notre approche. Les tokens dans l'ensemble de tokens sont enfin rendus minuscules pour être comparés après.
La comparaison de la similarité de deux noms de deux classes est la comparaison entre deux ensembles de tokens correspondant à ces noms. La similarité de deux tokens, qui sont actuellement des chaînes de caractères courtes, est calculée en employant la métrique n-gram qui apporte un avantage très important en permettant de contrôler la taille du lexique et de la maintenir à un seuil raisonnable pour de très grands noms.

- La similarité au niveau structurelle : Consiste à calculer la similarité entre les descriptions contextuelles de deux concepts c, c',où la description contextuelle d'un concept c est représentée par un vecteur dspcont=(dsp1,dsp2,...,dspn) où $\forall i \in (1,...,n)$, dspi=(pi,ri) dont pi dénote la iéme propriété du concept et ri la restriction de cette propriété. La restriction d'une propriété est une contrainte de deux types: contrainte de valeur ou de cardinalité.

3.2 La phase d'ancrage

Cette phase dont l'algorithme sera décrit dans la figure 2, consiste tout d'abord à apparier chacun des deux concepts X_{Src} et Y_{Tar}, pris respectivement de deux ontologies source et cible non alignées dans la première phase, avec un ou des concepts de la 3ème ontologie (O_{BK}), c'est-à-dire, à identifier des mappings de la forme (X_{Src} relation X_{BK}) et (Y_{Tar} relation Y_{BK}) en se basant sur les techniques utilisées dans la phase précédente, où X_{BK} et Y_{BK} appartiennent à C_{BK} et sont appelés des ancres ou points d'ancrage.

Algorithme ancrage**(**x_{Src}, x_{Bk}**) ;**
Entrée les deux concepts $x_{Src} \in X_{Src}$ et $x_{Bk} \in X_{Bk}$, **.**
Sortie ancre
Début Algorithme
Def $Dsp_{cont}(x_{Src})$=[ ], $Dsp_{cont}(x_{Bk})$=[ ] : *les vecteurs de description contextuelles de propriété de* x_{Src} et x_{Bk} *respectivement ;*
Def $Ancre(x_{Bk})$=[] est un vecteur contenant l'ensemble des ancres ;
Def $VMap(x_{Src}, x_{Bk},s)$=[] est un vecteur contenant x_{Src}, x_{Bk} et s le degré de similarité entre eux ;
Def $w_{LA} \in [0,1]$: le poids associé à la similarité linguistique, $(1-w_{LA})$ le poids associé à la similarité structurelle.
Def x=0, y=0, s=0

pour chaque concept $x_{Src} \in X_{Src}$
$x=S_{Linguistique}(x_{Src}, x_{Bk})$;
$y=S_{Structurelle}(x_{Src}, x_{Bk})$;
$s=w_{LA}*x+(1-w_{LA})*y$;
Ajouter $[x_{Src}, x_{Bk}, s]$ dans $VMap(x_{Src}, x_{Bk}, s)$
pour chaque valeur de $VMap(x_{Src}, x_{Bk}, s)$
$Ancre(x_{Bk,} x_{Src})=[x_{Bk}, x_{Src}]$;
retourner $Ancre(x_{Bk}, x_{Src})$;
Fin.

FIG. 2 - *L'Algorithme d'ancrage*

3.3 La phase de dérivation

Cette phase s'appuie sur la structuration de l'ontologie O_{BK} et le raisonneur Pellet Sirin et al (2004). Ce raisonneur inclut des méthodes sémantiques permettant d'effectuer des déductions sur les connaissances d'O_{Bk} par l'application de techniques de raisonnement, ces méthodes s'appuient sur des formalismes de représentation des connaissances fondées sur des logiques formelles, et les techniques citées au dessous appliquent les algorithmes de subsomption pour déduire les correspondances entre les ancres;

- subsomption : ($X_{BK} \leq Y_{BK}$) ou ($X_{BK} \geq Y_{BK}$)
- équivalence : ($X_{BK} \leq Y_{BK}$ et $Y_{BK} \leq Y_{BK}$) signifie que ($X_{BK} \equiv Y_{BK}$)

Afin d'avoir des mappings de la forme (X_{Src} relation Y_{Tar}), l'ensemble R des relations utilisées est l'ensemble $\{\leq, \geq, \equiv\}$ où $X \leq Y$ peut se lire, suivant les cas, «X isA Y», «X part-of Y» ou plus généralement «X narrower-than Y». Les mappings cherchés sont dérivés en exploitant des règles de la forme :

- Si ($X_{Src} \leq X_{BK}$) et ($X_{BK} \leq Y_{BK}$) et ($Y_{BK} \leq Y_{Tar}$) alors ($X_{Src} \leq Y_{Tar}$)
- Si ($X_{Src} \geq X_{BK}$) et ($X_{BK} \geq Y_{BK}$) et ($Y_{BK} \geq Y_{Tar}$) alors ($X_{Src} \geq Y_{Tar}$).

Ces règles utilisent aussi la relation d'équivalence, ≡, en considérant que l'existence d'une relation de type $A \equiv B$ permet de rajouter les deux relations $A \leq B$ et $A \geq B$ et qu'inversement, le fait d'avoir pu dériver les deux relations $X_{Src} \leq Y_{Tar}$ et $X_{Src} \geq Y_{Tar}$ permet de dériver la relation $X_{Src} \equiv Y_{Tar}$.

Ces différentes règles permettent ainsi de dériver des mappings « sémantiques », i.e. des mappings reliant deux concepts par un lien de type isA ou isEq dont la sémantique est bien définie et qui peuvent être justifiés et prouvés par des mécanismes d'inférences Sabou et al (2006). Cette phase est réalisée selon l'algorithme de la figure suivante :

Entrée Les deux Vecteurs $VMap(x_{Src}, x_{Bk},s)$ et $VMap(y_{Tar},y_{Bk},s)$;
Sortie Relation entre x_{Src}, x_{Bk}
Début Algorithme
Def $Reasoner(x_{Bk},y_{Bk})$ fonction utilise le raisonneur et retourne $r \in R=\{\leq, \geq, \equiv\}$;
Def $DérivBack(x_{Bk},y_{Bk},r)$ =[] vecteur contenant des ancres et la relation r entre eux;

pour chaque concept $x_{Src} \in X_{Src}$
pour chaque concept $x_{Tar} \in X_{Tar}$
r=$Reasoner(x_{Bk},y_{Bk})$;
$DérivBack(x_{Bk},y_{Bk},r)$=[x_{Bk},y_{Bk},r];

pour chaque élément de DérivBack(x_{Bk},y_{Bk},r)
si ($x_{src} \leq x_{bk}$) et ($x_{bk} \leq y_{bk}$) et ($y_{bk} \leq y_{tar}$) **alors** ($x_{src} \leq y_{tar}$)
Retourner x_{src} *isA* y_{tar} ;
si ($x_{src} \geq x_{bk}$) et ($x_{bk} \geq y_{bk}$) et ($y_{bk} \geq y_{tar}$) **alors** ($x_{src} \geq y_{tar}$).
Retourner y_{tar} *isA* x_{src} ;
si ($x_{src} \leq y_{tar}$) et ($x_{src} \geq y_{tar}$) **alors**
Retourner x_{src} *isEq* y_{tar} ;
Fin.

FIG. 3 *: L'algorithme de dérivation*

4 EXPÉRIMENTATION

Notre travail de recherche a été guidé par son application sur les turbines à vapeur. Nous présentons le contexte d'utilisation de notre outil BRMap dans une première sous-section. Ensuite nous présentons et discutons les résultats obtenus dans les expérimentations réalisées.

4.1. Contexte d'utilisation

Notre outil a été conçu pour augmenter le nombre d'appariement découverts dans le domaine de l'ingénierie ontologique où les ontologies sont hétérogènes et en continuelle évolution, ce qui engendre différentes versions, cette hétérogénéité met les développeurs et les ingénieurs de la connaissance dans la confusion, ne sachant pas ce qui a changé, l'alignement va permettre d'identifier les différences entre deux versions : les entités qui ont été ajoutées, supprimées ou renommées. Afin de tester notre outil nous avons besoin de trois ontologies une source, l'autre cible et une troisième complémentaire. La construction de ces ontologies est faite à l'aide d'un expert en utilisant Protege 2000.

4.2. La construction des ontologies

Pour construire les ontologies utilisées dans le test de BRMap, nous avons choisi l'outil d'édition des ontologies "Protege2000"[1] qui permet de décrire les ontologies en OWL et spécialement OWL DL. Cet outil intègre le plugin DataMaster spécialisé dans l'import des structures et des données de bases de données relationnelles, il propose une méthode d'import des tables de la base de données relationnelle comme des concepts OWL. Ce plugin est exploité pour convertir les données originales des ontologies décrivant la turbine à vapeur du sonalgaz[2] par ses experts dans des tables à l'aide du Microsoft Office Access, où chaque table a été convertie en un concept et chaque ligne de la table a été convertie en une instance du concept correspondant. Les valeurs des attributs ont été instanciées avec les valeurs des champs correspondants de la table.

Ce langage est favorisé dans notre cas à cause de son mécanisme de raisonnement basé sur la logique de description et sa syntaxe supportée par XML. Il a ses fondements dans la logique

[1] http ://protege.stanford.edu/index.shtml
[2] www.sonelgaz.dz/

de description comparé à d'autres formalismes Baader et al (2003), Horrock et al (2004) . Il intègre aussi le plugin DataMaster exploité pour construire les trois ontologies suivantes:

- L'ontologie source O_{Src} décrivant les cas d'interventions de l'ingénieur où chaque cas est défini par la cause de panne, défaut, symptôme et remède, extraite de la base de données " diagnostic.mdb ".
- L'ontologie target O_{Tar} décrivant les caractéristiques (code, description, zone, fonction,...) des composants, extraite de la base de données " centraletopo.mdb ".
- L'ontologie complémentaire O_{Bk} décrit les cas d'interventions et les caractéristiques de la turbine à vapeur, extraite des deux bases de données précédentes.

4.3. Résultats

Notre outil a été implémenté en Java, qu'offre plusieurs API permettant de manipuler les ontologies tel que : OWL-API, Jena[3]. De plus, il existe également plusieurs API de calcul de mesures de similarité réalisées en Java : n-gramme, etc. Nous les réutilisons dans notre application avec d'autres librairies comme Pellet, et d'autres algorithmes que nous avons développés.

Afin d'évaluer les résultats d'expérimentation de BRMap et d'autres outils d'alignement comme Coma++ et XMAPDjeddi (2009) , nous avons besoin d'utiliser les mesures de Précision, Rappel et Fallout Do et al (2002) qui sont des métriques largement exploitées pour estimer la qualité des alignements obtenus. L'objectif principal de ces mesures est l'automatisation du processus de comparaison des méthodes d'alignement ainsi que l'évaluation de la qualité des alignements produits. La première phase dans le processus d'évaluation de la qualité de l'alignement consiste à résoudre le problème manuellement. Le résultat obtenu manuellement est considéré comme l'alignement de référence et il sera donné par un expert en industrie.

Ensuite la comparaison du résultat de l'alignement de référence avec celui de l'appariement obtenu par la méthode d'alignement produit trois ensembles : Nfound, Nexpected et Ncorrect. L'ensemble Nfound représente les paires alignées avec la méthode d'alignement. L'ensemble Nexpected désigne l'ensemble des couples appariés dans l'alignement de référence. L'ensemble Ncorrect est l'intersection des deux ensembles Nfound et Nexpected. Il représente l'ensemble des paires appartenant à la fois à l'alignement obtenu et l'alignement de référence. La précision est le rapport du nombre de paires pertinentes trouvées, i.e., "Ncorrect", rapporté au nombre total de paires, i.e., "Nfound". Il renvoie ainsi, la partie des vraies correspondances parmi celles trouvées. Ainsi, la fonction précision (P) est définie par :

$$\text{Précision} = \frac{|N_{correct}|}{|N_{found}|} \qquad (1)$$

[3] www.hpl.hp.com/semweb/jena2.htm

Le *rappel* est le rapport du nombre de paires pertinentes trouvées, "*Ncorrect*" (N_C), rapporté au nombre total de paires pertinentes, "*Nexpected*" (N_e). Il spécifie ainsi, la part des vraies correspondances trouvées. La fonction *rappel (R)* est définit par :

$$\text{Rappel} = \frac{N_{correct}}{N_{expected}} \qquad (2)$$

La mesure *Fallout (F)* permet d'estimer le pourcentage d'erreurs obtenu au cours du processus d'alignement. Elle est définie par le rapport des paires erronées, "(*Nfound −Ncorrect*)", rapporté au nombre total des paires trouvées,"*Nfound*" (N_f),

$$\text{Fallout} = \frac{N_{found} - N_{correct}}{N_{found}} \qquad (3)$$

Enfin, les résultats obtenus par les différents outils d'alignement vont être illustré dans les tableaux suivants :

Paramètres Sélection	**N_e**	**N_f**	**N_C**	**P**	**R**	**F**
Seuil(S) -> (0.1 - 0.5)	60	130	18	0.1384	0.3	0.5384
MaxDelta(D) -> (0.01 - 0.1)		130	18	0.1384	0.3	0.5384
MaxN (N)-> (1 – 5)		130	18	0.1384	0.3	0.5384
S=0 .5 *D-> (0.01 - 0.1)* *N-> (1 – 5)*	60	130	18	0.1384	0.3	0.5384
S-> (0.1 - 0.5) *D = 0.1* *N-> (1 – 5)*	60	130	18	0.1384	0.3	0.5384
S-> (0.1 - 0.5) *D -> (0.01 - 0.1) N= 5*	60	130	18	0.1384	0.3	0.5384

Tab. 1- *Résultats du test d'alignement COMA++: $O_{Src} \rightarrow O_{Tar}$*

Paramètres de Sélection		**Ne**	**Nf**	**NC**	**P**	**R**	**F**
W_{la}	**S**						
0.5	0.1	60	8058	18	0.0022	0.3	0.9977
0.5	0.2		7956	18	0.0023	0.3	0.9973
0.5	0.3		18	18	1	0.3	0.7
0.5	0.4		18	18	1	0.3	0.7
0.5	0.5		18	18	1	0.3	0.7
0.5	0.6		18	18	1	0.3	0.7
0.5	0.7		18	18	1	0.3	0.7
0.5	0.8		18	18	1	0.3	0.7
0.5	0.9		18	18	1	0.3	0.7
0.5	1		18	18	1	0.3	0.7

Tab. 2- *Résultats du test d'alignement COMA++: $O_{Src} \rightarrow O_{Tar}$*

Paramètres de Sélection		N_e	N_f	N_C	P	R	F
W_{la}	S						
>=0.5	>=0.4	60	60	60	1	1	0

Tab. 3- *Résultats du test d'alignement BRMAP : $O_{Src} \rightarrow O_{Tar}$*

Le test du tab 1. est fait suivant la configuration donnée par les concepteurs du système combinatoire COMA++ présentée dans le Tab. 4. Il est nécessaire avant de lancer l'exécution de ce programme, de stabiliser une combinaison des différents paramètres de ce système pour donner les meilleurs résultats possibles pour chaque alignement.

Matchers		Aggregat	Direction	Selection	CombSim
No reuse	5 single		-LargeSmall -SmallLarge -Both	-MaxN(1-4) -Delta(0.01-0.1) -Thr(0.3-1.0) -Thr(0.5)+MaxN(1-4) -Thr(0.5)+Delta(0.01-0.1)	-Average -Dice
	11 combinaisons	-Max -Average -Min			
Reuse	2 single				
	12 combinaisons	-Max -Average -Min			-Average
∑= 16+14		3	3	36	2

Tab. 4- *Les différentes combinaisons de paramètres testées par les concepteurs de COMA Do et al (2002).*

A partir des Tab. (1, 2, 3) ; nous remarquons que :

- P_BRMAP >=PXMAP > PCOMA++ (4)
- R_BRMAP >RXMAP >= RCOMA++ (5)
- F_BRMAP <FXMAP < FCOMA++ (6)

La fonction 4 montre que BRMAP donne des résultats plus précis, mieux pertinents(5) avec peu d'erreur (6); que les deux autres outils utilisés, mais avec un temps d'exécution plus lent.

Comme l'alignement manuel se base sur l'utilisation d'une ontologie de référence du domaine de connaissances des ontologies à aligner, et notre approche consiste à utiliser une ontologie complémentaire et un raisonneur pour automatiser l'alignement des concepts, donc BRMAP est un outil automatisant l'alignement manuel.

5 Conclusion et perspectives

Dans cet article, nous avons présenté le processus d'alignement des ontologies et son importance dans le domaine industriel. Ensuite nous avons implémenté l'idée d'utiliser une ontologie plus détaillée que les deux ontologies à aligner, dites de background ou de support, en proposant l'utilisation d'un raisonneur à la phase de dérivation afin d'automatiser la dé-

couverte de correspondance sémantique entre les différents ancres trouvés lors de la phase d'ancrage qui se base sur les mesures de similarité linguistique et structurelle. Cette implémentation a engendré notre plugin BRMAP, ce dernier a donné des résultats plus satisfaisants que d'autres exploités pendant l'étape de test, mais avec un temps d'exécution considérable, à cause de la capacité volumineuse de l'ontologie complémentaire utilisée.

L'approche que nous avons proposée dans ce papier est appliquée sur des ontologies décrivant une turbine à vapeur dans le domaine industriel bien détaillé. Nous envisageons de poursuivre ce travail en testant encore l'approche sur d'autres domaines d'application. Dans un second temps nous améliorons les performances du plugin implémenté afin de le tester sur OAEI[4] (Ontology Alignment Evaluation Initiative) . Enfin, l'approche sera élargie pour aligner des ontologies non restreintes à des hiérarchies de concepts.

Références

Aleksovski, Z., Klein, M., Ten Kate, W., Van Harmelen F. (2006). "Matching Unstructured Vocabularies using a Background Ontology", Proceedings of the 15th International Conference on Knowledge Engineering and Knowledge Management (EKAW'06), Springer-Verlag.

Aleksovski Z., Klein M., Ten Kate W., Van Harmelen F. (2006b). "Exploiting the Structure of Background Knowledge used in Ontology Matching". ISWC'06 Workshop on Ontology Matching (OM-2006), Athens, Georgia, USA.

Baader F., Horrocks I. et Sattler U. (2003). "Description logics as ontology languages for the semantic web", Dans Hutter D. et Stephan, W. (éditeurs), Festschrift in honor of Jörg Siekmann. Lecture Notes in Articial Intelligence. Springer-Verlag, 2003.

Djeddi W. (2009). Conception et Réalisation d'un algorithme d'Alignement d'ontologies en OWL, mémoire Présentée Au Département d'informatique Université Badji Mokhtar Annaba, Pour l'obtention du diplôme de MASTER2 en Informatique Option : Sciences et Technologies de l'information et de la Communication.

Do H., Melnik S., Rahm E. (2002) " Comparison of schema matching evaluations", Proceedings of the 2nd Int. Workshop on Web Databasess, German Informatics Society, Erfurt, Germany, p. 221-237.

Do, H.H. et Rahm, E. (2002) "Coma – a system for flexible combination of schema matching approaches", Dans Proc. VLDB, pages 610–621.

Euzenat, J. et Shvaiko P. (2007) "*Ontology matching*", Springer, Heidelberg (DE).

Euzenat J., Bach, T.L., Barrasa, J., Bouquet, P., Bo, J.D., Dieng-Kuntz, R., Ehrig, M., Hauswirth, M., Jarrar, M., Lara, R., Maynard, D., Napoli, A., Stamou, G., Stuckenschmidt, H., Shvaiko, P., Tessaris, S., Acker; S.V. et Zaihrayeu, I.(2004). "State of the art on ontology alignment", deliverable 2.2.3, IST Knowledge web NoE, Knowledge web NoE, 80p..

[4] http://oaei.ontologymatching.org

J. Euzenat, C. Meilicke, H. Stuckenschmidt, P. Shvaiko, C. Trojahn: Ontology Alignment Evaluation Initiative: six years of experience Journal on Data Semantics, 2011.

Gruber T. R. (1993). A translation approach to portable ontology specifications, Knowledge Acquisition, 5((2): 199 – 220.

Horrocks i., Patel-schneider p. f., Boley h., Tabet s., Grosof b. (2004). Dean m., "SWRL: A Semantic Web Rule Language Combining OWL and RuleML", W3C Member Submission.

M. ELBYED, ROMIE, une approche d'alignement d'ontologies à base d'instances, Thèse de doctorat de l'institut national des telecommunications dans le cadre de l'école doctorale S&I en co-accréditation avec l'universite d'evry-val d'essonne, Soutenue le 16 Octobre 2009.

P. Shvaiko, J. Euzenat: Ten Challenges for Ontology Matching In Proceedings of ODBASE, 2008.

P. Shvaiko, J. Euzenat: Ontology matching: state of the art and future challenges IEEE Transactions on Knowledge and Data Engineering, 2012.

Reynaud C., Safar B., (2006). "When usual structural alignment techniques don't apply". ISWC '06 Workshop on Ontology Matching (OM-2006), Poster, Athens, Georgia, USA.

Reynaud C., Safar B. (2006b). "Structural Techniques for Alignment of Taxonomies: experiments and evaluation", In TR 1453, LRI, Université Paris-Sud.

Safar Brigitte, Reynaud Chantal, Calvier François. (2007) "Techniques d'alignement d'ontologies basées sur la structure d'une ressource complémentaire", Université Paris-Sud, CNRS (LRI, UMR 8623) & INRIA (Futurs), 91405 Orsay, France.

Sabou, M., D'Aquin M., Motta E. (2006), "Using the Semantic Web as Background Knowledge for Ontology Mapping", ISWC'06 Workshop on Ontology Matching (OM-2006), Athens, Georgia, USA.

Sirin E. et Parsia B. (2004). "Pellet : An owl dl reasoned", Dans Haarslev, V. et Möller, R. (éditeurs), Proceedings of the International Workshop on Description Logics (DL2004).

Summary

The main goal of this paper is to improve the performance of alignment techniques that are based on the structure or the richness of language representation of ontologies in proposing an approach combining matching techniques terminological, structural, and those used the background knowledge in integrating the reasoner Pellet.

Fusion automatique des ontologies : modélisation booléenne

Fawzia Zohra Abdelouhab*, Baghdad Atmani**, Bouziane Beldjilali***

Equipe de recherche Simulation, Intégration et Fouille de données (SIF)
Laboratoire d'Informatique d'Oran (LIO). Université d'Oran.
BP 1524, El M'Naouer, Es Senia, 31 000 Oran, Algérie
*fzabdelouhab@yahoo.fr, **atmani.baghdad@{univ-oran.dz, gmail.com},
***bouzianebeldjilali@yahoo.fr

Résumé. Notre travail s'insère dans le cadre de l'entreposage des données hétérogènes par entrepôt en vue de construire des contextes d'analyse appelés des cubes multidimensionnels pour la fouille de données. Il se focalise, essentiellement sur l'intégration des données hétérogènes tout en gardant leurs qualités sémantiques.
Dans cet article nous présentons les premières phases d'analyse de notre projet qui consiste en une nouvelle approche de fusion des ontologies en utilisant la modélisation booléenne et ceci en quatre étapes : la première étape, qui permet d'obtenir des graphes d'ontologies, consiste à extraire la structure arborescente des différents fichiers OWL représentants les ontologies en entrée. Ensuite, nous construisons pour chaque graphe d'ontologie, l'ensemble des règles de production des concepts, où les concepts de la partie prémisse de la règle subsument les concepts de la partie conclusion. Dans la troisième étape, nous utilisons le moteur d'inférence de la machine cellulaire CASI (Cellular Automata for Symbolic Induction) pour fusionner les ensembles de règles obtenues dans la deuxième étape. La transformation de l'ensemble des règles fusionnées en un arbre est faite dans la quatrième étape. L'arbre obtenu représente l'ontologie globale de fusion.

1 Introduction

Le rôle des systèmes d'intégration de données est de répondre aux besoins des utilisateurs à travers des interfaces d'accès uniformes aux sources contenant ces données (Zerdazi et Lamolle, 2005). Le défi de l'intégration de données est de faire cohabiter les sources hétérogènes, de plus en plus nombreuses, souvent réparties et indépendantes, dans un seul système uniforme, appelé système d'intégration, sans contraindre le comportement ni l'autonomie de chacune d'elles.

Des réflexions sérieuses ont été faites ces dix dernières années sur la normalisation du web et un consensus mondial a été établi à fortiori par les recommandations W3C[1]. Ces réflexions nous placent à la croisée de plusieurs domaines de recherche tels que l'ingénierie des connaissances, l'informatique décisionnelle, le traitement et la classification automatiques des concepts. C'est dans cette connectivité que nous avons structuré notre problématique dont la formulation serait la modélisation et l'intégration des données hétérogènes au

[1] : http://www.w3.org/2003/06/Process-20030618/tr.html#RecsW3C

sein d'un espace commun appelé entrepôt de données via une ontologie du domaine servant de pivot dans cette approche d'intégration.

Notre travail s'insère dans le cadre du projet national PNR[2] intitulé *architecture orientée service pour le programme élargi de vaccination* où il s'agit de l'entreposage par entrepôt des données médicales de vaccination..

2 Etat de l'art

L'intégration des données est un processus qui consiste à rapatrier des données à partir de différentes sources hétérogènes pour, soit les stocker dans une base commune (approche Entrepôt de Données), soit les traiter localement (Approche médiateur). Dans l'une ou dans l'autre des approches, la complexité du problème reste la même dû au fait que les informations sous-jacentes se trouvent dans des sites différents (Boussaid et al. 2006).

Selon Nguyen, (2006), les approches d'intégration peuvent être classifiées suivant trois critères : selon la manière de stocker les données à intégrer ou selon la manière de relier les schémas des sources locales avec le schéma global ou encore sur le degré d'automaticité d'intégration. Avec l'engouement actuel du web d'autres critères peuvent, aussi, rentrer en jeu pour distinguer les approches d'intégration selon qu'elles tiennent compte du critère de scalabilité du web (i.e., l'augmentation des accès concurrents sur les différentes sources du web) et de l'intéropérabilité de sa structure et de sa sémantique (caractéristique inhérente aux ontologies).

Notre travail s'apparente donc d'une part à des travaux sur l'intégration de données du Web plus précisément l'intégration de sources de données autonomes et hétérogènes, et d'autre part, à des travaux sur l'intégration de données guidée par une ontologie qui étudient, quant à eux, comment trouver des correspondances entre les ontologies des sources de données à intégrer et comment les utiliser. Quant au problème de scalabilité du web et de son engouement, les ontologies ont été utilisées dans les systèmes d'intégration pour représenter, justement, le schéma global de médiation et/ou les schémas des sources locales (Dibie, 2009).

Plusieurs systèmes d'intégration à base d'ontologies, ou par ontologies ont vu le jour et ont apporté un plus considérable dans des domaines aussi variés que la médecine, le droit, l'indexation de séquences audiovisuelles, et l'électronique pour n'en citer que cela (Mena et al. 2000).

Selon Bellatreche, (2006), il existe deux catégories de ces systèmes : les premiers utilisent une structure à base d'une ontologie unique comme les Projets OntoBroker, SIMS, COIN, Picsel cités dans (Khouri, 2009) mais ils souffrent de manque d'autonomie au niveau des sources locales. Les deuxièmes sont à base d'ontologies multiples et apportent une meilleure solution tels que les projets ONION et caBIG cités dans (Khouri, 2009).

Dans les approches à base d'ontologies multiples chaque source est décrite sémantiquement par sa propre ontologie, appelée ontologie locale qui est mise en correspondance avec une ontologie partagée modélisant un domaine particulier, qu'on appelle ontologie globale. Dans cette catégorie nous trouvons, entre autre, les travaux de (Diallo, 2006) qui propose une Architecture à Base d'Ontologies pour la Gestion Unifiée de deux types de don-

[2] : http://www.nasr-dz.org/dprep/pnr2/projets-pnr/PNR_a.htm

nées Structurées et non Structurées basée sur une approche de médiation par ontologies. Sa conception repose sur l'utilisation des technologies du Web Sémantique et de plusieurs types d'ontologies pour la caractérisation sémantique des sources non structurées (textuelles). Les ontologies servent d'une part à définir le schéma global d'intégration (ontologie globale) et, d'autres parts, les différentes sources à intégrer. Des correspondances sont établies entre l'ontologie globale et les différentes ontologies locales.

Sais, (2007), propose un système d'Intégration Sémantique de Données structurées représentant des tableaux collectés et extraits à partir du Web. Cette intégration de type entrepôt de données rentre dans le cadre du projet eDot guidée par une Ontologie du domaine. Sa contribution est une méthode générique et automatique d'enrichissement sémantique d'informations structurées représentant des tableaux pour découvrir des relations candidates à l'enrichissement d'un entrepôt et d'une ontologie. Cette étude monte l'intérêt et la faisabilité d'approches complètement automatiques, non supervisées et guidées uniquement par une ontologie.

L'objectif de Zimmerman, (2008), est de modéliser la sémantique d'un ensemble des connaissances produites indépendamment les unes des autres, formant un réseau et mises en correspondances. Dans chaque nœud du réseau se trouve une ontologie, reliée aux autres par des correspondances formant des alignements d'ontologies. Afin de favoriser l'utilisation d'ontologies indépendantes et préexistantes, il définit une sémantique formelle exploitant le principe de médiation. Il a mit en place un formalisme qui exploite au mieux les logiques locales déjà établies, les met en corrélation par un procédé original qui distingue d'un côté la représentation locale, propre à chaque nœud dans le réseau, et la représentation des connaissances inter-ontologies propre au médiateur. Cependant, son travail se trouve confronté à la complexité du raisonnement distribué.

Les travaux de Dibie, (2009), portent sur l'intégration des données guidée par une ontologie à travers la réalisation du projet ONDINE (ONtology-based Data INtEgration). Ceci consiste en l'intégration et l'interrogation d'une BD relationnelle et d'une base des graphes conceptuels. Son système d'intégration repose sur une ontologie de domaine qui est construite à partir des bases locales (à partir de leurs schémas, de leurs attributs et des contraintes associées).

De ces recherches nous remarquons que l'intégration des données passe par les étapes suivantes : elle associe, d'abord, à chaque source son ontologie locale ; Ensuite elle intègre les ontologies des sources en établissant des relations sémantiques (équivalence, subsomption…) entre leurs concepts et, enfin, elle peuple les données dans l'entrepôt en exploitant les correspondances ontologiques établies dans l'étape précédente.

On se place, pour notre travail, dans le cas d'une approche matérialisée (entrepôt de données). On souhaite effectuer une intégration sémantique de manière automatique. On s'intéresse donc aux approches à base d'ontologies conceptuelles. L'avantage d'utiliser un entrepôt de données est que le problème de réécriture des requêtes suivant le schéma des sources des données ne se posera plus comme dans le cas du médiateur. En effet, les données intégrées étant archivées selon le schéma de l'entrepôt, leur interrogation se fera directement. Par contre, le problème de scalabilité peut être limité si le schéma global est défini de manière à couvrir l'ensemble du domaine (ontologie). Pour cela une modélisation et un stockage booléen de l'entrepôt rendraient notre problématique assez pertinente pour ce critère puisqu'elle permet d'optimiser l'espace de stockage et augmenter l'accès aux données.

3 Contribution

Nous avons conçu notre solution en impliquant une autre voie de recherche qui est l'utilisation de la machine cellulaire CASI (Atmani et Beldjilali, 2007) pour ses techniques mathématiques formelles et ses performances confirmées dans bien des travaux de fouilles de données à partir de connaissances; (Barigou et al. 2012), (Brahami et al. 2010), (Mansoul et Atmani, 2009). L'utilisation de la machine CASI nous permet de concevoir un nouveau système automatique d'intégration des données guidée par la fusion booléenne des ontologies.

La modélisation booléenne d'intégration est motivée par la structure du modèle de données qui se présente sous forme de graphe OWL et facilite bien la génération automatique du graphe de l'ontologie. Un avantage certain à cette modélisation est la réduction dans l'espace de stockage de l'entrepôt de données et aussi dans le temps de calcul des requêtes décisionnelles.

Dans ce sens, notre contribution, par rapport aux systèmes d'intégration existants, ne tente pas de les concurrencer mais se limite à appréhender une autre voie de recherche pour apporter un plus au niveau de la modélisation des données, de l'espace de stockage, du temps d'exécution et dans la complexité de programmation.

Le processus d'intégration des données que nous avons développé se décompose en trois phases importantes :

3.1 Phase de pré-intégration

A partir des graphes d'ontologies, une fouille de données est faite pour extraire la structure spécifique des données et la présenter sous forme de règles d'association. Cette phase, réalisée dans le projet BICS-XML (Abdelouhab et Atmani, 2009), consiste à extraire la structure arborescente des données. Dans notre projet, les sources étudiées sont semi-structurées du type XML ou OWL parce qu'elles respectent la structure d'arbre à travers le système des balises.

3.2 Phase de Matching/Mapping des schémas

Le Matching est un processus qui, étant donné deux schémas, effectue des correspondances entre les éléments et les attributs des schémas, et retourne comme résultat les valeurs de similarités sémantiques entre les deux schémas (Zerdazi et Lamolle, 2005).

Dans notre approche les éléments des schémas que nous manipulons sont des concepts constituants les ontologies à intégrer. Selon Diday, (2008), un concept est défini par une "intension" et une "extension". L'intension est un ensemble de propriétés caractéristiques du concept appelées « Attributs » et l'extension est l'ensemble des individus appelés « Instances » du concept qui satisfont ces propriétés. Dans l'exemple de la figure1, le concept Personne de l'Ontologie O1 a pour extension l'ensemble du personnel de la faculté d'Es-Science qu'il soit enseignant, étudiant ou administrateur.

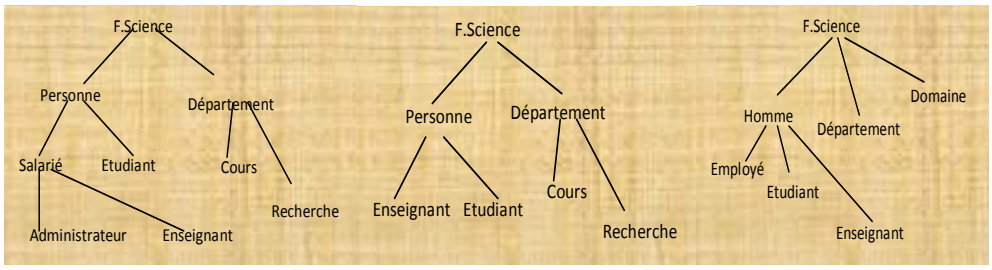
Selon Maiz, (2008), le calcul de la similarité entre deux concepts est basé sur la terminologie du concept, ses propriétés et ses relations avec son voisinage. Seulement, cette similarité n'est pas suffisante pour conclure que les deux concepts sont similaires ou pas. Pour qu'ils le soient complètement il faut que leurs instances le soient également. Ce qui s'appelle le

Mapping ; se sont des expressions décrivant le moyen dont les instances du schéma cible (final) sont dérivées à partir des instances de schéma source (initial). Elles décrivent la correspondance sémantique entre les instances de schémas en complémentarité avec le Matching.

Notre solution par rapport au problème de Matching et de Mappings des schémas des ontologies réside dans le projet CASI-Match. Ce dernier est un algorithme de classification hiérarchique qui permet de classifier les concepts formant les ontologies en entrées en deux classes. La première, est la classe de synonymie et la deuxième est la classe de disjonction. En appliquant CASI-Match sur les ontologies O1 et O2 de la figure1, nous obtenons une similarité entre le concept Personne de O1 et le concept Homme de O3. Cet algorithme n'étant pas encore finalisé, ne sera pas détaillé dans le spectre réduit de cet article.

3.3 Phase de fusion

Pour étayer notre approche nous allons présenter notre démarche à travers un exemple simple et assez expressif de trois ontologies représentant le même domaine mais définies de manières différentes (Maiz, 2008). Il s'agit de la gestion du personnel universitaire, illustrée par la figure1 comme suit :



(a) Ontologie O1 *(b) Ontologie O2* *(c) Ontologie O3*

FIG. 1 – *Exemple de trois Ontologies.*

A partir de chaque graphe, nous en générons une base de connaissance qu'on notera BC correspondante formée d'un ensemble de règles sous la forme : R_i : Si $Prémisse_i$ Alors $Conclusion_i$

Avec une représentation cellulaire selon le principe suivant :

- les concepts de $Prémisse_i$ et $Conclusion_i$ vont constituer les faits : FAITS.
- les R_i vont constituer les règles : REGLES.

Ces règles produites seront intégrées dans la BC de CASI pour exploitation en inférence.

Exemple : la partie du graphe (F.Science→Personne) est représentée par la règle suivante : R1 : Si F.Science Alors Personne.

La partie du graphe (F.Science → Personne → Salarié) est représentée par la règle suivante : R2 : Si F.Science, Personne Alors Salarié.

De cette façon nous obtenons la BC suivante rassemblant l'ensemble de toutes les règles des trois Ontologies.

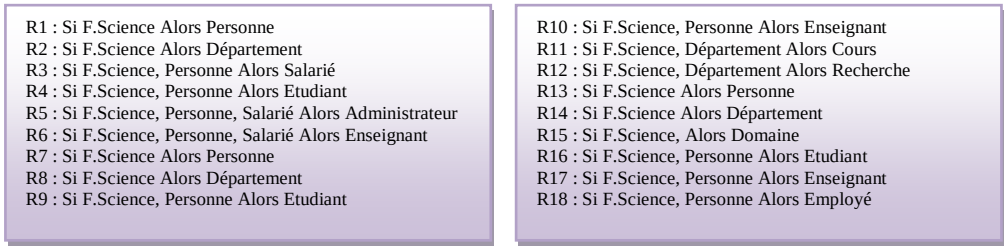


FIG. 2 – *Base de connaissances globale.*

Nous remarquons sur cette BC obtenue que le nom du concept Homme de l'ontologie O3 a été remplacé par le nom Personne puisque l'algorithme CASI-Match aura détecté une similarité entre les deux.

Les règles de la BC ainsi obtenues sont répertoriées selon deux catégories:

***D**es règles incluses formées par l'ensemble de règles redondantes notées R_{IN}(des règles ayant même prémisses) avec des conclusions différentes. Elles expriment une relation de subsomption.*

***D**es règles identiques formées par l'ensemble de règles redondantes ayant les mêmes conclusions notées R_{ID}. Elles expriment une relation d'équivalence.*

La fusion consiste à rassembler toutes les règles redondantes comme suit :

- L'ensemble de toutes les R_{IN} sera remplacé par une seule Règle dont la conclusion sera formée par l'union de toutes les conclusions de l'ensemble des règles qu'elle remplace. Exemple : R1 : Si F-Science Alors Personne et R2 : Si F-Science Alors Département, seront remplacées par Si F-Science Alors Personne, Département.
- L'ensemble de toutes les R_{ID} sera remplacé par une seule Règle.

L'algorithme que nous utilisons simule le fonctionnement d'un automate cellulaire qui est une grille composée de cellules changeant d'état dans des étapes discrètes. Après chaque étape, l'état d'une cellule est modifié selon les états de ses voisines calculés dans l'étape précédente. Les cellules sont mises à jour d'une manière synchrone, et les transitions sont effectuées simultanément. Pour cela nous utilisons deux matrices booléennes exprimant les différentes couches d'automates finis. La première matrice, *CELFACT*, exprimant la base des faits et, la deuxième matrice, *CELRULE*, exprimant la base de règles. Chaque élément de la matrice représente une cellule de l'automate. Chaque inférence du moteur crée une configuration de la machine CASI. A chaque itération nous obtenons une couche de l'automate formée par les états des matrices. A chaque étape, une cellule peut être active (1) ou passive (0), selon qu'elle participe ou pas à l'inférence.

Le principe est simple :

- Toute cellule *i* de *CELFACT* est un fait établi si sa valeur est 1, sinon, il est à établir.
- Toute cellule *j* de *CELRULE* est une règle candidate si sa valeur est 1, sinon, elle ne doit pas participer à l'inférence.

La configuration initiale de la machine est donnée par l'état initial de *CELFACT* et *CELRULE*. Ces états sont représentés par des vecteurs d'Etat Entrée, d'Etat Interne et d'Etat de sortie. Le vecteur IF : indique le rôle du *Fait* dans le graphe : Si IF = 0, le Fait est du type sommet (ie, un nœud complexe : qui fait référence à d'autres nœuds); et Si IF = 1, le Fait est du type *attribut=valeur* (ie, un nœud atomique : qui contient des données simples).

Initialement, toutes les entrées des cellules de *CELFACT* sont passives (*EF* = 0), exceptées celles qui représentent la BF initiale (*EF*(1) = 1). Dans notre cas le nœud racine de la première règle F-Science représente le Fait Initial. A partir de ces notations nous construisons les couches *CELFACT* et *CELRULE* Fig3. En plus de ces deux couches, la machine CASI utilise deux autres matrices d'incidence R_E et R_S représentant la correspondance d'entrée et de sortie des faits par rapport aux règles :

- la relation d'entrée, notée iR_Ej, est formulée comme suit : $\forall i \in \{1, \ldots, l\}\ \forall j \in \{1, \ldots, r\}$ si (le Fait $i \in$ à la *Prémisse* de la règle j) alors $RE(i, j) \leftarrow 1$.
- la relation de sortie, notée iR_Sj, est formulée comme suit : $\forall i \in \{1, \ldots, l\}\ \forall j \in \{1, \ldots, r\}$, si (le Fait $i \in$ à la *Conclusion* de la règle j) alors $RS(i, j) \leftarrow 1$.

La configuration initiale de la machine est générée comme suit:

CELFACT

Fait	**EF**	**IF**	**SF**
F.Science	1	0	0
Personne	0	0	0
Département	0	0	0
Salarié	0	0	0
Etudiant	0	0	0
Cours	0	0	0
Recherches	0	0	0
Administrateur	0	0	0
Enseignant	0	0	0
Domaine	0	0	0
Employé	0	0	0

CELRULE

Règles	**ER**	**IR**	**SR**
R1	1	1	1
R2	0	1	1
R3	0	1	1
R4	0	1	1
R5	0	1	1
R6	0	1	1
R7	0	1	1
R8	0	1	1
R9	0	1	1
R10	0	1	1
…..			
R18	0	1	1

FIG. 3 – *Couches CELFACT et CELRULE.*

$\mathbf{R_E}$	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	…	R18
F.Science	1	1	1	1	1	1	1	1	1	1		1
Personne	0	0	1	1	1	1	0	0	1	1		1
Département	0	0	0	0	0	0	0	0	0	0		0
Salarié	0	0	0	0	1	1	0	0	0	0		0
Etudiant	0	0	0	0	0	0	0	0	0	0		0
Cours	0	0	0	0	0	0	0	0	0	0		0
Recherches	0	0	0	0	0	0	0	0	0	0		0
Administrateur	0	0	0	0	0	0	0	0	0	0		0
Enseignant	0	0	0	0	0	0	0	0	0	0		0
Domaine	0	0	0	0	0	0	0	0	0	0		0
Employé	0	0	0	0	0	0	0	0	0	0	…	0

$\mathbf{R_S}$	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	…	R18
F.Science	0	0	0	0	0	0	0	0	0	0		0
Personne	1	0	0	0	0	0	1	0	0	0		0
Département	0	1	0	0	0	0	0	1	0	0		0
Salarié	0	0	1	0	0	0	0	0	0	0		0
Etudiant	0	0	0	1	0	0	0	0	1	0		0
Cours	0	0	0	0	0	0	0	0	0	0		0
Recherches	0	0	0	0	0	0	0	0	0	0		0
Administrateur	0	0	0	0	1	0	0	0	0	0		0
Enseignant	0	0	0	0	0	1	0	0	0	1		0
Domaine	0	0	0	0	0	0	0	0	0	0		0
Employé	0	0	0	0	0	0	0	0	0	0	…	1

3.3.1 Section de niveau 3

Le moteur d'inférence cellulaire de la machine CASI est organisé en cellules qui passent d'une configuration à une autre en appliquant deux fonctions de transitions **δ*rule* et δ*fact*** sur les couches CELFACT et CELRULE. Une fonction de transition est l'ensemble de règles qui détermine le nouvel état de chaque cellule selon son état précédant, et les états précédents des cellules de son voisinage.

La Fonction de Transition δfact pour effectuer une transition de l'instant t à t+1

(EF, IF, SF, ER, IR, SR) $\xrightarrow{\delta fact}$ **(EF, IF, EF, ER+(R_E^t. EF),IR,SR)**

La Fonction de Transition δrule pour effectuer une transition de l'instant t+1 à t+2:

(EF, IF, SF, ER, IR, SR) $\xrightarrow{\delta rule}$ **(EF + (RS · ER), IF, SF, ER, IR, ^ER)**

où matrice R_E^t désigne la transposée de RE et ^ER désigne la négation du vecteur booléen ER.

Nous considérons G_0 la configuration initiale de notre automate cellulaire (voir la Fig3) et, $\Delta = \delta rule \circ \delta fact$ la fonction de transition globale : $\Delta(G_0) = G_1$ obtenu en deux étapes :

1- On applique la fonction de transition *δfact sur* G_0 nous obtenons G'_0 (Fig5). *δfact* permet de filtrer les règles candidates à l'inférence. Ce sont toutes les règles (dont ER=0) et possédant le même ensemble de prémisse que la première règle sélectionnée par l'automate (ie ER=1). Dans notre exemple il s'agit des règles suivantes : R_2, R_7, R_8, R_{13}, R_{14} et R_{15}. Elles seront marquées en mettant ER à 1 pour sortir de la compétition.

2- On applique, ensuite, la deuxième fonction de transition *δrule sur* G'_0 nous obtenons le graphe G_1(Fig6). *δrule* permet de valider les Faits Conclusion des règles sélectionnées par *δfact* en mettant EF à 1. Ensuite, elle désactive les règles sélectionnées en mettant SF à 1.

Le processus se répète d'une configuration à une autre jusqu'à ce qu'il n'y a plus de règle candidate (dont ER=0) à sélectionner. Les règles se feront désactiver au fur et mesure que l'on valide leurs Faits Conclusions.

CELFACT

Fait	EF	IF	SF
F.Science	1	0	1
Personne	1	0	0
Département	1	0	0
Salarié	0	0	0
Etudiant	0	0	0
Cours	0	0	0
Recherches	0	0	0
Administrateur	0	0	0
Enseignant	0	0	0
Domaine	1	0	0
Employé	0	0	0

CELRULE

Règles	ER	IR	SR
R1	1	1	0
R2	1	1	0
R3	1	1	0
R4	1	1	0
R5	1	1	0
R6	1	1	0
R7	1	1	0
R8	0	1	1
R9	0	1	1
R10	0	1	1
…..			
R18	0	1	1

FIG. 6 – Configuration obtenue avec *δrule*

Supposons que $G = \{G_0, G_1, .., G_q\}$ est l'ensemble des configurations de notre automate cellulaire. L'évolution discrète de l'automate, d'une génération à une autre, est définie par la séquence G_0, G_1,..., G_q, où $G_{i+1} = \Delta(G_i)$. Ce qui représente la configuration finale des deux matrices d'Entrée/Sortie donnée comme suit :

$\mathbf{R_E}$	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	…	R20
F.Science	**1**	**1**	**1**	**1**	0	0	0	0	0	0		0
Personne	0	**1**	0	1	0	0	0	0	0	0		0
Département	0	0	1	0	0	0	0	0	0	0		0
Salarié	0	0	0	1	0	0	0	0	0	0		0
Etudiant	0	0	0	0	0	0	0	0	0	0		0
Cours	0	0	0	0	0	0	0	0	0	0		0
Recherches	0	0	0	0	0	0	0	0	0	0		0
Administrateur	0	0	0	0	0	0	0	0	0	0		0
Enseignant	0	0	0	0	0	0	0	0	0	0		0
Domaine	0	0	0	0	0	0	0	0	0	0		0

$\mathbf{R_S}$	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	…	R20
F.Science	0	0	0	0	0	0	0	0	0	0		0
Personne	1	0	0	0	0	0	0	0	0	0		0
Département	1	0	0	0	0	0	0	0	0	0		0
Salarié	0	1	0	0	0	0	0	0	0	0		0
Etudiant	0	1	0	0	0	0	0	0	0	0		0
Cours	0	0	1	0	0	0	0	0	0	0		0
Recherches	0	0	1	0	0	0	0	0	0	0		0
Administrateur	0	0	0	1	0	0	0	0	0	0		0
Enseignant	0	1	0	1	0	0	0	0	0	0		0
Domaine	1	0	0	0	0	0	0	0	0	0		0

FIG. 7 – *Configuration finale de l'Automate*

A partir de cet état final, nous appliquons le processus inverse de la modélisation booléenne pour retrouver la BC finale à partir des matrices de l'automate. Cette BC optimale représente le résultat de la fusion donnée comme suite :

R1 : Si F.Science Alors Personne, Département, Domaine
R2 : Si F.Science, Personne Alors Salarié, Etudiant, Enseignant
R3 : Si F.Science, Département Alors Cours, Recherche
R4 : Si F.Science, Personne, Salarié Alors Administrateur, Enseignant

Dont le graphe est illustré par la Figure 8 :

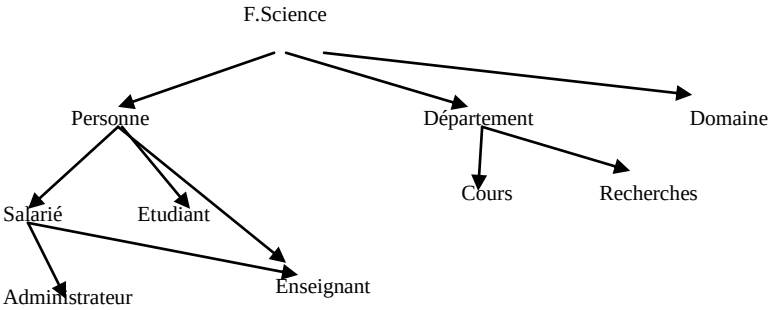


FIG. 8 – *Arbre Final de Fusion.*

4 Conclusion et perspectives

L'originalité de notre travail par rapport à l'état de l'art est que nous avons reconsidéré le problème dans son ensemble en introduisant à partir des couches les plus basses une modélisation booléenne pour garantir, à la fois, une construction booléenne automatique de l'ontologie et une optimisation considérable de l'espace de stockage de l'entrepôt de données. Nous sommes revenus, et cela était nécessaire, sur les différentes étapes du processus d'intégration pour ajuster la méthode de modélisation booléenne et améliorer les résultats obtenus.

Cette étude s'inscrit dans une perspective originale centrée autour des automates cellulaires pour modéliser l'ontologie en BC et implémenter des algorithmes de fusion. Notre apport est de montrer que la modélisation booléenne peut apporter sa contribution à ce problème comme cela a déjà été prouvé pour le problème de la classification des données (Azzag et al. 2004). Les auteurs répartissent les données à classifier sur une grille de cellules 2D et à travers la fonction locale de transition des cellules, ils favorisent la constitution de regroupement des données similaires pour des cellules voisines.

En utilisant la machine cellulaire CASI nous avons développé un algorithme de fusion des ontologies tout en réduisant la quantité de stockage et le temps d'exécution. En effet, cela est due à l'utilisation de la représentation booléenne des matrices R_E et R_S, et à la multiplication booléenne employée par les fonctions de transition *δfait* et *δregle*.

Notre projet est actuellement en cours de réalisation et a pu être validé par des expérimentations dans le cadre de projet de fin d'étude. Une perspective à court terme est le passage à l'échelle avec une expérimentation sur des données réelles issues de la vaccination.

Références

Abdelouhab, F., B. Atmani (2008). Intégration automatique des données semi-structurées dans un entrepôt cellulaire, ASD 2008, ISBN 978-9981-1-3000-1, dépôt légal: 2168/2008, pp 109-120.

Abdelouhab, F., B. Atmani (2009). Extraction de structure d'un document XML : Modélisation Booléenne, ASD 2009, ISBN 978-9961-9913-0-5, dépôt légal: 5226/2009 pp. 67-81.

Atmani, B., B. Beldjilali (2007). Knowledge Discovery in Database: Induction Graph and Cellular Automaton, Computing and Informatics Journal, V.26, N°2 (2007) 171-197.

Azzag H., Picarougne F., Guinot C., Venturini G.,(2004), *Un survol des algorithmes biomimétiques pour la classification*. Classification Et Fouille de Donnée, pages 13-24,

Barigou, F., B. Atmani, B. Beldjilali, (2012). Using a Cellular Automaton to Extract Medical Information from Clinical Reports, JIPS Volume 8, pp.67-84

Bellatreche L., Xuan D., Pierra G. & Dehainsala H. (2006). Contribution of ontology-based data modeling to automatic integration of EC within ED. Computers in Industry Journal.

Boussaid O., R. Ben Messaoud, R. Choquet, S. Anthoard (2006). Conception et construction d'entrepôts en XML. EDA'06 Versaille 19.

Brahami, M., Atmani, B., Mokaddem, M.(2010), CARTOCEL : un outil de cartographie des connaissances guidée par la machine cellulaire CASI. ;In EGC(2010)625-626.

Diallo, G (2006), Une Architecture à Base d'Ontologies pour la Gestion Unifiée des Données Structurées et non Structurées, thèse de doctorat à l'Université Joseph Fourier. Grenoble.

Dibie,B, J, (2009). Intégration de données guidée par une ontologie : Application au domaine du risque alimentaire. Habilitation à Diriger des Recherches de l'Université Paris-Dauphine.

Diday, E. (2008), Principes d'Analyse des données symboliques et application à la détection d'anomalies sur des ouvrages publics. EGC 2008: 211-212.

Khouri, S.(2009), Modélisation conceptuelle à base ontologique d'un entrepôt de données, Mémoire de Magistère, Université Oued-Smar Alger.

Maiz, N., O. Boussaid et F. Bentayeb (2008). Fusion automatique des ontologies par classification hiérarchique pour la conception d'un entrepôt de données. EGC2008.

Mansoul, A., B. Atmani,(2009), Fouille de données biologiques: vers une Représentation Booléenne des Règles d'Association. CIIA2009 .

Mena,E.,Illarramendi,A.,Kashyap,V.etSheth,A.P.(2000).Observer : An approach for query processing in global information systems based on interoperation across preexisting ontologies. 8(2):223-271.

Nguyen X. D,(2006), Intégration de bases de données hétérogènes par articulation à priori d'ontologies: Applications aux CCI,. Thèse de doctorat, Université de Poitiers.

Sais, F. (2007), Intégration Sémantique des Données Guidée par une Ontologie, Thèse de Doctorat. Université Paris-Sud.

Zerdazi A., M Lamolle (2005), HyperSchéma XML: Un modèle d'intégration sémantique par enrichissement de schémas XML, *MajecSTIC 2005 : STIC (2005) 143-150.*

Zimmerman A (2008), Sémantique des réseaux de connaissances : gestion de l'hétérogénéité fondée sur le principe de médiation, thèse de doctorat, Université Joseph Fourier. Grenoble.

Summary

Our work fits within the framework of the storage of the heterogeneous data by warehouse in order to build contexts of analysis called multidimensional cubes for the excavation of data. It is focused, primarily on the integration of the heterogeneous data while keeping their semantic qualities. In this article we present the first phases of analysis of our project which consists of a new approach of fusion of ontologies by using Boolean modeling and this in four stages: the first stage, which makes it possible to obtain graphs of ontologies, consists in extracting the tree structure of various files OWL representatives ontologies as starter. Then we build for each graph of ontology the whole of the rules of production of the concepts, where the concepts of the premise part of the rule subsume the concepts of the conclusion part. In the third stage, we use the inference engine of cellular machine CASI (Cellular linen Automata for Symbolic Induction) to amalgamate the whole of rules obtained in the second phase. The transformation of the whole of the rules amalgamated into a tree is made in the fourth stage. The tree obtained represents the total ontology of fusion.

Raisonnement classificatoire appliqué à la classification d'individus dans une ontologie multi-points de vue

Meriem Djezzar*, Zizette Boufaida**

* Laboratoire LIRE,
Département Technologie des Logiciels et Systèmes d'Information,
Université Constantine 2, Algérie
djezzar.meriem@gmail.com
** Laboratoire LIRE,
Département Technologie des Logiciels et Systèmes d'Information,
Université Constantine 2, Algérie
zboufaida@gmail.com

Résumé. Une ontologie multi-points de vue confère à un même univers de discours, plusieurs descriptions partielles, telles que chacune soit relative à un point de vue. Ces descriptions partielles partagent à un niveau global, des éléments ontologiques (concepts et rôles globaux) et des liens sémantiques (passerelles) constituant un consensus entre les différents points de vue. Dans ce papier, nous proposons, dans un premier temps, la description en logique de descriptions d'une ontologie multi-points de vue. Ensuite, sur la base du modèle de représentation développé, nous proposons un mécanisme de classification d'un individu. Cette classification se déroule selon un ou plusieurs points de vue et permet de retrouver, pour chacun d'eux, les concepts locaux dont un individu est susceptible d'être une instance. La notion de passerelle permet de déduire l'appartenance d'un individu à des concepts locaux d'un point de vue en se basant sur les résultats obtenus sur d'autres points de vue.

1 Introduction

Les ontologies permettent une modélisation formelle d'un domaine, en décrivant ses concepts et les relations qu'ils entretiennent ainsi que les individus (ou instances) qui leurs sont associés. La conception d'une ontologie se fait en deux étapes, d'abord sa conceptualisation (Bendaoud et al., 2007) qui a pour but de faire émerger les concepts et relations représentant le domaine visé, ainsi que les axiomes qui permettront d'ordonner concepts et relations dans l'ontologie et de classer leurs instances, ensuite son peuplement ou instanciation qui consiste à associer des instances à des concepts existants.

Dans de nombreux contextes applicatifs plusieurs modèles conceptuels couvrant un même domaine sont développés indépendamment les uns des autres par des communautés différentes. L'hétérogénéité entre les connaissances exprimées au sein de chacun d'entre eux doit être résolue. La modélisation des connaissances d'un domaine surtout complexe par une ontologie définie comme conceptualisation consensuelle et partageable s'avère dans ce cas une tache ardue. Cette difficulté provient principalement du fait que chaque modèle conceptuel dépend du point de vue du concepteur du monde et de sa terminologie usuelle.

Dans ce travail, nous nous intéressons à problématique de la représentation d'une ontologie dans une organisation hétérogène en prenant en compte différents points de vue et terminologies des communautés au sein de cette organisation. Une telle ontologie, que nous appelons ontologie multi-points de vue, confère à un même univers de discours plusieurs descriptions partielles telles que chacune soit relative à un point de vue. De plus, les différentes descriptions partielles partagent à un niveau global des éléments ontologiques consensuels et des passerelles. Ces dernières établissent les communications entre les points de vue et représentent ainsi la collaboration interdisciplinaire.

Pour définir l'ontologie multi-points de vue (Hemam et Boufaida., 2011), nous nous sommes inspirés de travaux qui traitent de la représentation des connaissances par objets auxquels sont associés des points de vue (Mariño., 1993). Pour cela, nous utilisons un sous-langage de la logique de descriptions (LD) $\mathcal{SHOQ}$ (Baader et al., 2003) pour exprimer les notions inhérentes aux points de vue tels que les concepts globaux et locaux, les passerelles, les estampillages.

Dans cet article, nous nous intéressons au raisonnement par classification basé sur le modèle de représentation des ontologies multi-points de vue proposé.

Le raisonnement par classification est l'un des principaux mécanismes de raisonnement associés aux LD. En effet, la structuration de la connaissance en classes, sous-classes et instances favorise l'utilisation de la classification pour récupérer les connaissances implicites, les relations entre une nouvelle situation et des situations déjà connues (Djezzar et al., 2012). Le terme *classification* supporte deux mécanismes distincts : **(i)** la classification de classes (aussi appelée catégorisation) qui consiste à organiser et à maintenir une hiérarchie de classes, en insérant de nouvelles classes à leur place et **(ii)** la classification d'instances qui consiste à trouver, dans une hiérarchie de classes, la classe d'appartenance la plus appropriée pour une instance.

Intuitivement, le raisonnement par classification d'instances consiste à faire descendre une instance le plus bas possible dans la hiérarchie des classes, en la comparant aux descriptions fournies par les sous-classes.

Dans notre étude, la classification multi-points de vue tire profit des passerelles pour faire descendre l'individu plus rapidement dans un point de vue en utilisant les résultats des classifications obtenues dans d'autres points de vue.

La suite de cet article est structurée comme suit. Dans la section 2, nous introduisons les logiques de descriptions comme formalisme de représentation d'ontologies. La section 3 est consacrée à la notion de points de vue et à son intégration dans le domaine de la représentation des connaissances. Nous rappelons dans la section 4, le modèle de représentation d'ontologies multi-points de vue proposé puis comment l'utiliser dans le processus de classification d'individus. Enfin, la section 5 conclut l'article et donne des perspectives pour les travaux futurs.

2 Les logiques de descriptions

Les Logiques de Descriptions, anciennement appelées logiques terminologiques (Napoli., 1997), forment une famille de langages de représentation de connaissances utilisée pour représenter la connaissance terminologique d'un domaine d'application. Une ontologie exprimée en LD contient la description des concepts, des rôles et des individus. Les *concepts* représentent des ensembles d'objets, possédant des caractéristiques communes. Les *individus*

sont utilisés pour représenter les objets du domaine. Les *rôles* correspondent à des relations binaires entre ces objets. Généralement, les LD permettent de représenter deux types de connaissances. La TBox, ou le niveau *terminologique*, est un ensemble d'axiomes, de la forme $C \leqslant D$ ou de la forme $C \equiv D$ où C et D étant deux concepts. La première forme permet l'introduction d'un concept dit *primitif*, tandis que la deuxième forme permet la définition d'un nouveau concept dit *défini*. La ABox, ou le niveau *factuel*, est un ensemble d'assertions de la forme a (C) ou (a, b) r où a et b sont deux individus, C est un concept et r, un rôle. Le premier type d'assertions correspond à une instanciation de concept, le second, à une instanciation de rôle.

La bonne formalisation des LD tient à leur sémantique. Cette dernière, s'exprime grâce à des notions ensemblistes. Une *interprétation* dans ce cadre est un couple $I = (\Delta^I, \cdot^I)$, où Δ^I est un ensemble non vide, appelé *domaine d'interprétation* et contient l'ensemble des objets du domaine, et $\cdot^I$, une *fonction d'interprétation* associe à un concept $\boldsymbol{C}$ (un ensemble d'objets du domaine, i.e. un sous-ensemble de Δ^I), à un rôle $\boldsymbol{r}$ (un sous-ensemble de $\Delta^I \times \Delta^I$), et à un individu $\boldsymbol{a}$ (un objet du domaine, élément de Δ^I).

Les LD supportent deux relations clés. La relation de subsomption permet d'organiser les concepts d'une ontologie du plus général au plus spécifique. Intuitivement, un concept D est subsumé par un concept C (ce qui est noté $D \sqsubseteq C$) si C est plus général que D (l'ensemble des individus de C contient l'ensemble des individus de D). La relation « être instance de » quant à elle, permet de connaître les individus d'un concept. Ceci engendre deux types d'inférences, liées au raisonnement classificatoire :

- Le *test de subsomption* permet de vérifier qu'un concept C subsume un concept D. Il est à la base du processus de classification de concepts qui consiste à déterminer l'ensemble des ascendants directs d'un concept dans la hiérarchie des concepts, (les subsumants les plus spécifiques), et l'ensemble des descendants immédiats du concept (les subsumés les plus généraux).
- Le *test d'instanciation* vérifie si un individu a est l'instance d'un concept C dans une ontologie. Il est à la base du processus d'*instanciation* (i.e. classification d'instances) qui nous intéresse dans cette étude.

3 Notion de point de vue

La représentation des connaissances est le domaine précurseur dans l'introduction de la notion de point de vue dans des modèles à objets en raison du caractère multi-disciplinaire des applications. En effet, la plupart des applications en intelligence artificielle nécessitent l'intervention de plusieurs experts ayant chacun une connaissance particulière du domaine d'étude qui procure aux objets manipulés une représentation et un type de raisonnement particulier.

KRL (*Knowldge Representation Language*) (Bobrow et al., 1977) est l'un des premiers langages de représentation des connaissances à reconnaître qu'un objet peut être vu de plusieurs façons, selon le point de vue de l'observateur. Dans le formalisme KRL, la notion de point de vue (exprimée par le terme perspective) est représentée au niveau des instances. Un individu (appelée *unité individuelle*) a une première perspective qui est la classe la plus générale à laquelle il appartient, unité de type *Basic* et il peut avoir d'autres perspectives parmi les unités de spécialisation de celle-ci.

Le système TROPES (Mariño., 1993) est un système de représentation des connaissances par objets (RCO) avec multi-points de vue. TROPES emploie le concept[1], la passerelle et le point de vue. Ce dernier a un double rôle. Il permet de voir le concept selon un certain angle et de ce fait, seuls les attributs du concept pertinents pour le point de vue sont visibles. Il permet aussi d'organiser les spécialisations du concept en une hiérarchie de classes significative pour le point de vue. Par ailleurs, la notion de passerelle est proposée afin de mettre en relation les classes d'un concept vu selon différents points de vue.

TROPES a fourni une base de réflexion à d'autres projets. C'est le cas notamment des travaux développés dans (Ribière., 1999) et le système Kasimir (d'Aquin et al., 2004). Dans le premier cas, l'auteur utilise le formalisme des graphes conceptuels pour représenter et faire cohabiter différents points de vue d'experts sur un même sujet. Le système d'aide à la décision en cancérologie Kasimir s'intéresse à la représentation multi-points de vue des connaissances contenues dans les référentiels (sortes de protocoles de décision médicaux) du domaine de cancérologie. Ce travail a d'abord été étudié dans un cadre de la représentation par objets (d'Aquin et al., 2005) puis a été implanté dans le cadre des logiques de descriptions distribuées (Borgida et Serafini., 2003) et du langage C-OWL (Bouquet et al., 2004).

Dans la suite, nous adoptons le terme *d'ontologie multi-points de vue* afin de mettre l'accent sur l'importance de la notion de point de vue pour **1)** résoudre le problème de la représentation multiple **2)** avoir un meilleur accès et une meilleure visibilité des éléments ontologiques (concepts, rôles, individus) **3)** tirer profit de la représentation multi-points de vue des connaissances pour permettre leur évolution. Par ailleurs, pour prendre en compte la notion de point de vue, nous supposons que les différents points de vue sur un même univers de discours sont des visions partielles mais complémentaires. Leur union est une représentation complète et cohérente du monde.

4 Présentation de notre approche

4.1 Représentation d'ontologies multi-points de vue

Pour les besoins de la formalisation des ontologies multi-points de vue, nous introduisons, dans la logique de descriptions, les notions suivantes :

1. *Une ontologie multi-points de vue* est une description multiple d'un même univers de discours selon différents points de vue. Elle est définie par un quadruplet $O = <C^G, R^G, Vp, M>$, où : C^G est l'ensemble des concepts globaux, R^G est l'ensemble des rôles globaux, Vp est l'ensemble des points de vue et M est l'ensemble des passerelles.
2. *Un point de vue* est une description partielle d'un univers de discours selon une perception particulière. Un point de vue est défini par un triplet $VP_K = <C^L, R^L, A^L>$, où : C^L est l'ensemble des concepts locaux, R^L est l'ensemble des rôles locaux et A^L est l'ensemble des individus locaux.

[1] TROEPS ne considère pas, comme il est habituel de le faire en RCO et en LD, les termes de classe et de concept comme des synonymes.

3. *Concept global* est un concept vu par l'ensemble des points de vue avec certaines propriétés[2] communes. Ces dernières, sont visibles par tous les points de vue et constituent ce qu'on appelle la clé du concept global.
4. *Concept local* est un concept qui est vu et décrit localement selon un point de vue donné.
5. *Rôle global* est une relation entre deux concepts locaux définis dans deux points de vue différents.
6. *Rôle local* est une relation entre deux concepts locaux définis dans le même point de vue.
7. *Estampill*e: Nous adaptons le mécanisme d'estampillage[3] utilisé dans (Benslimane et al., 2006) pour permettre la multi représentation des concepts dans le formalisme de la logique de descriptions. Dans notre approche, une estampille (i.e. label) permet de reconnaître pour chaque élément ontologique (i.e. concept, rôle, individu) le point de vue auquel il appartient.
8. *Hiérarchie locale :* Sous un point de vue $\mathcal{VP}_i$, une hiérarchie locale, notée $vp_i/\mathcal{H}$, est définie par le triplet $(\mathcal{C}^{\mathcal{L}}, \partial, \sqsubseteq)$ où : $\mathcal{C}^{\mathcal{L}}$ est l'ensemble des concepts locaux, ∂ est une fonction de $\mathcal{C}^{\mathcal{L}}$ dans $\mathcal{C}^{\mathcal{G}}$ qui associe chaque concept racine (*i.e.* le plus général) $\mathcal{S}$ de $\mathcal{C}^{\mathcal{L}}$ à un seul concept global $\mathcal{C}^{\mathcal{S}}$ de $\mathcal{C}^{\mathcal{G}}$, $\sqsubseteq$ est la relation de subsomption utilisée pour exprimer explicitement un lien d'ordre direct
9. *Passerelle:* L'une des particularités de la représentation multi-points de vue est l'existence d'un canal de communication entre les différents points de vue. Ce canal de communication, appelé passerelle, permet de représenter des liens consensuels entre les concepts locaux de différents points de vue. Une passerelle s'exprime de quatre manières :

$vpi: X \xrightarrow{\sqsubseteq} vpj: Y$ (***Passerelle d'inclusion unidirectionnelle***) **(1)**

Signifie qu'un individu qui est une instance du concept source X sous le point de vue VPi est aussi une instance du concept destination Y sous le point de vue VPj.

$vp1: X_1 \sqcap \ldots \sqcap vpk: X_k \xrightarrow{\sqsubseteq} vpj: Y$ ***(Passerelle d'inclusion avec plusieurs sources)*** **(2)**

Signifie qu'un individu qui est une instance de chacun des concepts sources ($vp_1:X_1 ... vp_k:X_k$) définis sous des points de vue disjoints est aussi une instance du concept destination Y sous le point de vue VPj.

$vpi: X \xleftrightarrow{\equiv} vpj: Y$ (***Passerelle d'inclusion bidirectionnelle***) **(3)**

Exprime l'égalité entre les deux ensembles d'extensions des deux concepts locaux X et Y définis sous deux points de vue différents.

$vpi: X \xleftrightarrow{\perp} vpj: Y$ **(*Passerelle d'exclusion bidirectionnelle***) **(4)**

Signifie que les deux concepts X et Y sont incompatible.

10. *Instanciation multiple:* Le mécanisme d'instanciation multiple permet à un individu d'être une instance directe d'un ou de plusieurs concepts. Dans le contexte de notre travail, un individu possède la propriété suivante:

[2] Le terme propriété est pris au sens large et inclut les relations binaires entre concepts globaux et les relations unaires (attributs).

[3] Les estampilles sont des concepts de modélisation utilisés afin de différencier les représentations multiples d'une même réalité (Balley et al., 2004).

– ***Propriété:*** *un individu est une instance d'un concept global et une instance d'un ou de plusieurs concepts locaux définis dans un ou plusieurs points de vue.*

Un individu possède donc une description de base (i.e. description globale) et peut être décrit partiellement selon un ou plusieurs points de vue.

Concept global
Appartement $^{\hat{0}} \equiv (\forall_{vp1}$ Nbr_pièces.Number) $\sqcap (\forall_{vp2}$ Loyer.Number) $\sqcap (\forall_{vp1,vp2}$ Surface.Number) $(\forall_{vp1, vp2, vp3}$ Adresse.String) $\sqcap (\geq_{vp1, vp2, vp3} 1$ Adresse) $\sqcap (\leq_{vp1, vp2, vp3} 1$ Adresse) *Définit un concept global avec un attribut* Nbr_pièces *selon* vp_1*, un attribut* Loyer *selon* vp_2 *, un attribut* Surface *selon* vp_1 *et* vp_2 *et un attribut* Adresse *selon les trois points de vue* vp_1, vp_2 *et* vp_3
Concept local
vp_1: Petit_Appartement $\equiv$ Appartement $^{\hat{0}} \sqcap$ (Nbr_pièces. {1, 2}) *Définit un concept local, dans le point de vue* vp_1*, comme étant un appartement et dont la valeur de l'attribut* Nbr_pièces *est dans l'ensemble* {1, 2}.
Relation de subsomption
vp_2 : HLM $\sqsubseteq$ $vp_{2:}$ Appartement_PasCher *Exprime un lien de subsomption entre deux concepts locaux définis dans le même point de vue. En effet, sous le point de vue* vp_2 *, tous les HLM sont des appartements pas cher.* $vp_{2:}$ Appartement_PasCher $\sqsubseteq$ Appartement $^{\hat{0}}$ *Exprime un lien de subsomption entre le concept local* Appartement_PasCher, défini sous *le point de vue* $vp_{2,}$*, et le concept global* Appartement $^{\hat{0}}$
Rôle local /global
vp_2: habite_par (Appartement_Cher, Locataire_Riche) *Définit un rôle local entre deux concepts locaux définis dans le même point de vue* vp_2 habite $^{\hat{0}}$ (vp_2: Locataire_Riche, vp_3: Appartement_CentreVille) *Définit un rôle global entre deux concepts locaux définis dans deux PV différents* (vp_2 *et* vp_3)
Passerelle unidirectionnelle/bidirectionnelle
vp_2 : HLM $\xleftrightarrow{\equiv}$ vp_3: Appartement_Banlieue *Exprime que les deux concepts locaux, définis dans deux points de vue différents, sont équivalents. En effet, tous les HLM sont dans la banlieue et tous les appartements de banlieue sont des HLM* $vp_{1:}$ Plus_TroisPièce $\sqcap$ $vp_{3:}$ Appartement_CentreVille $\xrightarrow{\sqsubseteq}$ $vp_{2:}$ Appartement_Cher *Signifie que tous les appartements de plus de trois pièces qui se trouvent au centre-ville sont des appartements chers*
Multi-instanciation
vp_1: Petit_Appartement (chez-Benali) vp_3: Appartement_Banlieue (chez-Benali) *Indiquent que l'individu* chez-Benali *est une instance de* Petit_Appartement *sous le point de vue* vp_1 *et aussi une instance de* Appartement_Banlieue sous le point de vue vp_3

TAB. 1 - *Exemple d'une ontologie multi-points de vue en LD*

4.2 Processus de classification multi-points de vue

Le processus de classification multi-points de vue peut être perçu comme le passage d'un état initial (de l'individu à classer et de son concept global d'appartenance, ainsi que ses valeurs valides pour l'ensemble des propriétés communes qui forment la clé du concept global), à un état final dans lequel l'individu est classé le plus bas possible dans les différentes hiérarchies de subsomption locales des différents points de vue impliqués. Par ailleurs, un utilisateur expert dans un domaine est sollicité pour lancer la classification de l'individu sur un sous-ensemble de points de vue liés à ses centres d'intérêts et ses connaissances. Les autres points de vue peuvent être utilisés pour inférer des connaissances additionnelles à travers l'usage des passerelles. Nous considérons pour cela les deux ensembles des points de vue suivants:

1- **Les points de vue principaux** : ce sont les points de vue supposés connus par l'utilisateur et par rapport auxquels il veut classer un individu.
2- **Les points de vue intermédiaires** : ce sont les autres points de vue non concernés par la classification de l'individu mais qui peuvent être utilisés pour aider et accélérer la descente de l'individu dans les différents points de vue principaux.

La mise en œuvre du processus de classification multi-points de vue se déroule selon l'algorithme illustré dans la figure 1 :

Entrée :

- L'individu **I** à classer
- Le concept global $\mathbf{C^{\hat{0}}}$ auquel **I** appartient
- La liste ordonnée des points de vue principaux [$VPp_1 \ldots VPp_n$]
- La liste des points de vue intermédiaires [$VPi_1 ... VPi_m$]

Sortie :

- Dans chaque $\mathbf{VPp_i}$, l'individu **I** est rattaché à ses concepts d'appartenance les plus spécialisés

Début
Pour tout $VPp_i \in [VPp_1 ... VPp_n]$ **Faire**
Tant_que (Actif (VPp_i)) **Faire**
Appariement ;
Propagation_Marques ;
Mise_A_Jour ;
Fin_Tant_Que
Fin_Pour
/* *tous les points de vue principaux ont été explorés* ******/
Rattachement ;
Fin.

FIG 1- *Algorithme de classification multi-points de vue.*

L'algorithme de classification multi-points de vue commence avec l'individu I à classer, supposé déjà rattaché à son concept global, et la liste des points de vue principaux. L'ordre de parcours des points de vue principaux est donné explicitement par l'utilisateur, ce qui aura pour effet de privilégier certains points de vue par rapport aux autres. L'algorithme de classification se termine lorsque tous les points de vue principaux ont été explorés et l'individu I

est classé le plus bas possible dans chacun de ces points de vue. Par ailleurs, sous un point de vue principal $\mathbf{VPp_i}$, l'espace de recherche pour la classification de l'individu est une hiérarchie de subsomption de concepts. Cette dernière, notée $\mathcal{H}_{\mathbf{C}^{\hat{0}}}$, est constituée de l'ensemble (la famille) des concepts locaux qui sont des sous-concepts (des descendants) du concept global $\mathbf{C}^{\hat{0}}$ auquel appartient l'individu à classer.

Cette classification multi-points de vue entraîne une vision "parallèle" de la descente de l'individu dans les différents points de vue. En effet, à travers l'utilisation des passerelles, plusieurs points de vue peuvent être modifiés "simultanément". L'utilisateur ne voit qu'un seul point de vue à la fois, le point de vue principal actuel, sur lequel se fait la classification à un instant t.

Pour tout point de vue principal actuel ($\mathbf{VPp_i\_Actuel}$), la classification de l'individu repose sur une boucle qui comporte trois étapes principales: 1) Appariement, 2) Propagation_Marques et 3) Mise_A_Jour.

4.2.1 Appariement

Sous un point de vue principal, le parcours de la hiérarchie $\mathcal{H}_{\mathbf{C}^{\hat{0}}}$ s'effectue en profondeur, sur l'ensemble des sous-concepts de $\mathbf{C}^{\hat{0}}$, en partant du premier sous-concept fils de $\mathbf{C}^{\hat{0}}$. La procédure ***Appariement*** est utilisée pour tester, à tout moment de la descente dans la hiérarchie $\mathcal{H}_{\mathbf{C}^{\hat{0}}}$, si l'individu I appartient au concept courant $\mathbf{C_i\_Courant}$, en faisant une comparaison entre les valeurs des propriétés de l'individu I et les caractéristiques de la description du concept courant. Pour faire cette comparaison, la procédure demande à l'utilisateur les valeurs des propriétés de I dont elle a besoin. Ainsi l'individu I est considéré comme une instance de C_i _courant s'il possède les propriétés (i.e. rôles) décrites dans C_i _courant et les valeurs de ces propriétés doivent satisfaire les descriptions[4] (i.e. les contraintes) de ces propriétés dans le concept C_i _courant.

Les résultats fournis par la procédure ***Appariement*** sont au nombre de trois :

1- L'appariement est **sûr** si pour chaque propriété P_J de C_i la valeur de cette propriété P_J dans I satisfait sa description D_J dans C_i.
2- L'appariement est **impossible** s'il existe au moins une valeur d'une propriété P_J de I qui ne satisfait pas sa description D_J dans C_i.
3- L'appariement est **possible** si toutes les valeurs des propriétés P_J de I satisfont leurs descriptions D_J et s'il subsiste des propriétés de C_i non valuées dans I. Ceci veut dire que la connaissance que l'on a de l'individu ne permet ni d'affirmer ni de nier son appartenance au concept C_i.

La procédure d'appariement ne s'applique que sur les concepts non marqués. Une fois le résultat de l'appariement obtenu, le concept C_i _Courant sera marqué par l'une des étiquettes : sûr, impossible ou possible.

4.2.2 Propagation_Marques

A partir des résultats fournis par la procédure ***Appariement***, la procédure ***Propagation_Marques*** procède à une propagation des étiquettes sûr, possible ou impossible sur les

[4] Propriétés et leurs descriptions définissent les conditions nécessaires et suffisantes d'appartenance d'un individu à un concept.

concepts locaux des différents points de vue. Ce marquage, basé sur la sémantique de la relation de subsomption (propagation locale) et des passerelles (propagation globale), permet d'une part de réduire l'espace de recherche et d'autre part d'économiser de futurs appariements. La propagation de ces marques obéit aux règles suivantes :

Propagation locale

1. Si un concept est sûr pour l'individu I, tous ses sur-concepts (ses subsumants) doivent aussi être sûrs.
2. Si un concept est impossible pour l'individu I, tous ses sous-concepts (ses subsumés) deviennent aussi impossibles.
3. Si un concept est marqué possible pour l'individu I, tous ses sous-concepts (ses subsumés) deviennent temporairement possibles.

Propagation globale

1. Cas des passerelles d'inclusions à une seule source : si le concept source est sûr pour l'individu I, alors le concept destination doit aussi l'être. Inversement, si le concept destination est marqué impossible alors le concept source doit aussi l'être.
2. Cas des passerelles d'inclusions avec plusieurs sources : si tous les concepts sources sont marqués sûrs pour l'individu I, alors le concept destination doit aussi l'être. Par ailleurs, si l'un des concepts sources est non marqué (ou marqué possible) et tous les autres sont marqués sûrs et le concept destination est marqué impossible, alors le concept source qui est non marqué (ou qui est marqué possible) devient impossible.
3. Cas des passerelles d'exclusion : si l'un des deux concepts de ce type de passerelle est marqué sûr l'autre doit être marqué impossible.
4. Cas des passerelles d'inclusion bidirectionnelles : si l'un des deux concepts de cette passerelle est marqué sûr (impossible) l'autre doit être aussi marqué sûr (resp. impossible).

De ce fait, à travers les règles décrites ci-dessus, un concept local non marqué ou marqué comme *possible* peut changer de statut et prendre l'étiquette *impossible* ou *sûr*.

4.2.3 Mise_A_Jour

La procédure ***Mise_A_Jour*** réalise quatre principales opérations selon l'ordre suivant :

1- **Mise à jour de l'espace de recherche :** Dans chaque point de vue principal, l'espace de recherche pour la classification est constitué initialement de l'ensemble des descendants du concept global $\mathbf{C^0}$. Ainsi, à partir des résultats obtenus par la procédure ***Propagation_Marques***, cette première opération consiste à réduire les différents espaces de recherche des différents points de vue principaux. Pour ce faire, nous considérons les deux conditions suivantes :
   - Si un concept X est marqué impossible alors la sous-hiérarchie de racine X est élaguée.
   - Si un concept X est marqué sûr alors tous ses ascendants sont éliminés.
2. **Mise à jour des passerelles :** Cette opération consiste à mettre à jour l'ensemble des passerelles, à considérer lors de la procédure *Propagation-Marques*, en éliminant celles qui sont devenues inactives. En toute généralité, une passerelle (uni ou bidirectionnelle) devient inactive, c'est-à-dire ne sert plus à faire des propagations de mar-

ques, si tous ses concepts (sources et destination) ont une marque sûre ou impossible. Trois cas particuliers (*Cf.* Figure 2) sont distingués :

- Une passerelle d'inclusion unidirectionnelle (avec une ou plusieurs sources) est inactive si son concept destination est marqué sûr.
- Une passerelle d'inclusion unidirectionnelle (avec plusieurs sources) est inactive si l'un de ses concepts sources est marqué impossible.
- Une passerelle d'exclusion bidirectionnelle est inactive si l'un de ses concepts est marqué impossible.

De ce fait, aucune nouvelle marque pour un des concepts non marqués (ou marqués comme possible) ne permet de faire de nouvelles inférences.

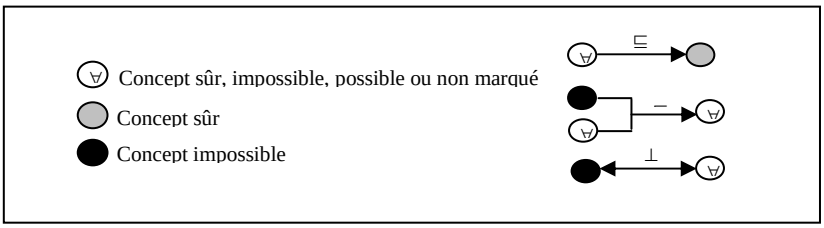


FIG 2- *Passerelles à l'état inactif*

3. **Mise à jour des points de vue intermédiaires :** Cette opération consiste à mettre à jour l'ensemble des points de vue intermédiaires qui vont participer à la descente de l'individu I. Ainsi, un point de vue intermédiaire (**VPi**) passe à l'état inactif, c'est-à-dire n'appartient plus à l'ensemble considéré, lorsque toutes ses passerelles sont devenues inactives.
4. **Mise à jour des concepts les plus spécifiques (CPS) :** Cette opération consiste à mettre à jour, pour chaque point de vue principal, l'ensemble de ses concepts les plus spécifiques (i.e. les plus bas) qui sont marqués comme sûrs (*Cf.* Figure 3).

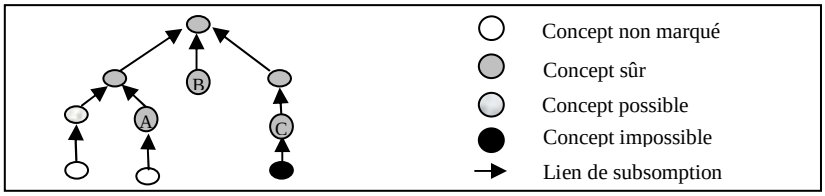


FIG 3- *Mise à jour des CPS : Résultat {A, B, C}*

4.2.4 La fonction Actif (VPp_i)

Le but de la classification est de descendre l'individu le plus bas possible dans les points de vue principaux. Un point de vue principal duquel on a déjà tiré toute l'information possible, n'est plus utile à la classification. A cet effet, la fonction **Actif (VPp_i)** est utilisée pour tester, à tout moment de la classification, si le point de vue courant $\mathbf{VPp_i}$ est actif ou non (i .e. il est encore possible de faire descendre l'individu ou pas). Ainsi :

1. Un point de vue principal est *inactif* si tous ses concepts les plus spécifiques (les plus bas) marqués comme sûrs sont des concepts terminaux. Un concept sûr terminal peut être *ouvert* ou *fermé* :
   - Un concept sûr terminal est *fermé* s'il n'a pas de sous-concepts (i.e. il est une feuille), ou bien si tous ses sous-concepts sont marqués comme des concepts impossibles.
   - Un concept sûr terminal est *ouvert* si au moins l'un de ses sous-concepts est temporairement marqué par l'étiquette possible et les autres sont marqués par impossible.
2. Un point de vue principal est *actif* s'il a des concepts non marqués. De ce fait, il est toujours possible de faire descendre plus bas l'individu vers ces sous-concepts .

Un point de vue actuel, **VPp_i _Actuel**, est un point de vue *actif* si l'algorithme de classification l'utilise pour classer un individu. Dans le cas où le **VPp_i _Actuel** passe à l'état *inactif*, le prochain point de vue principal, à prendre comme actuel doit être choisi parmi ceux qui sont à l'état actif, et cela selon l'ordre établi par l'utilisateur dans la liste des points de vue principaux.

4.2.5 Attachement

Une fois tous les points de vue principaux explorés (i.e. ils sont tous à l'état inactif), la dernière étape consiste, pour chacun d'entre eux, à attacher l'individu I à chaque élément appartenant à CPS_Ter = { $\mathbf{vp_i}$:C_i | marque ($\mathbf{vp_i}$:C_i)= sûr et $\mathbf{vp_i}$:C_i est un concept terminal} .

5 Conclusion et perspectives

Dans cet article, nous avons présenté une approche pour la représentation d'une ontologie avec des points de vue différents. L'élément clé de notre approche est de permettre la description de ce type d'ontologies, sans éliminer l'hétérogénéité mais en faisant cohabiter l'hétérogénéité (au niveau local) et le consensus (au niveau global). A chaque point de vue correspond une représentation locale. Par ailleurs, les différents points de vue partagent à un niveau global, des éléments ontologiques et des passerelles. Ces dernières, permettent de lier différents concepts locaux provenant de différents points de vue et ainsi d'inférer des informations provenant d'un point de vue en fonction de celles connues dans un autre. Le mérite de cette approche est de permettre à des mécanismes de raisonnement de fonctionner localement sur chaque point de vue, ou bien sur des assemblages de ces points de vue. Le mécanisme de raisonnement que nous avons utilisé sur l'ontologie MPV est la classification d'un individu. Cette classification prend en compte les caractéristiques propres au modèle multi-points de vue : le concept global, les points de vue et les passerelles. La classification se déroule en même temps dans les hiérarchies locales des différents points de vue impliqués. Les résultats obtenus dans un point de vue peuvent être utilisés par un autre point de vue grâce aux passerelles.

Le processus de classification multi-points de vue mérite d'être raffiné. Ce dernier, pourra être utilisé, comme module, pour annoter sémantiquement les ressources d'une organisation hétérogène, en prenant en considération l'aspect multi-points de vue. Ceci, en se basant sur l'exploitation et l'instanciation de l'ontologie multi-points de vue.

REFERENCES

Baader, F., Horrocks, I., & Sattlerl, U., Description Logics as Ontology Languages for the Semantic Web. *In Festschrift in honor of Jorg Siekmann, LNAI. Springer,* 2003

Balley, S., Parent, C., & Spaccapieta, S., Modeling Geographic Data with Multiple Representation. *In journal Geographical Information science. Vol 18, N°4 pp. 327-352, ISSN 1365-8816,* 2004

Bendaoud, R., Rouane, M., Hacene, Toussaint, Y., Delecroix, B., & Napoli A., Construction d'une ontologie à partir d'un corpus de textes avec l'acf. *In F. Trichet (Ed.), A tes des 18eme journées francophones d'ingénierie des connaissances,* 2007.

Benslimane, D., Arara, A., Falquet, G., Maamar, Z., Thiran, P., & Gargouri, F., Contextual Ontologies: Motivations, Challenges, and Solutions. *In Fourth Biennial International Conference on Advances in Information Systems,* 2006.

Bobrow, D.G., & Winograd, T., An overview of KRL, a knowledge representation language. *Cognitive Science, Vol. 1*, 1977

Bouquet, P., Giunchiglia, F., Van Harmelen, F., Serafini, L., & Stuckenschmidt, H., Contextualizing Ontologies. *In Journal of Web Semantics, 1(4): pp 325--343,* 2004.

Borgida, A., & Serafini. L., Distributed description logics: Assimilating information from peer sources. *Journal of Data Semantics, 1:153–184,* 2003

d'Aquin, M., Lieber, J., & Napoli, A., Représentation de points de vue pour le raisonnement à partir de cas. *In Langages et Modèles à objets (LMO'04),* pp. 245-258, 2004.

d'Aquin, M., Lieber, J., & Napoli, A., Decentralized Case-Based Reasoning for the Semantic Web. *In Proceedings of the 4th International Semantic Web Conference*, 2005

Djezzar, M., Hemam, M., & Boufaida, Z., Ontological Re-Classification of Individuals: A Multi-Viewpoints Approach", In Second International Conference on Model and Data Engineering (MEDI'2012), 7602, LNCS, Springer, pp 91-102, 2012.

Hemam, M., Boufaida, Z., MVp-OWL: A Multi-Viewpoints Ontology Language for the Semantic Web. Int. J. Reasoning-based Intelligent Systems, Inderscience Publishers**,** Vol. 3, N° 3/4 (2011) 147–155

Napoli, A., Une introduction aux logiques de descriptions. *Rapport de recherche N° 3314, INRIA,* 1997.

Mariño, O., Raisonnement classificatoire dans une représentation à objets multi-points de vue. *Thèse d'informatique, université Joseph Fourier, Grenoble,* 1993.

Ribière, M., Représentation et gestion de multiples points de vue dans le formalisme des graphes conceptuels. *Thèse de doctorat en informatique, Nice-Sophia Antipolis,* 1999.

La Réutilisation des connaissances ontologiques dans le processus d'affaires

Moufida Aouachria *
Dr. Ramdane Maamri **

* Département d'informatique, École Doctorale en S.T. I. C
Université de Batna, Algérie
aouachria.hayet@yahoo.fr
** Département d'informatique, Laboratoire LIRE
Université Mentouri-Constantine, Algérie
rmaamri@yahoo.fr

Résumé. Nous avons remarqué que les approches pour la réutilisation à deux phases développées dans l'ingénierie de domaine et adaptées par Caplinskas sur les processus d'affaires, sont mises en place avec succès ; mais elles ne proposent pas de dynamisme à l'exécution. De plus, elles ne conviennent pas très bien aux environnements dynamiques changeants. D'où, dans cet article, nous proposons une approche à trois phases : la phase d'ingénierie du domaine de processus, la phase de l'ingénierie applicative de processus et la phase de l'exécution de processus. Cette approche basée sur les ontologies pour permettre la réutilisation des connaissances du processus d'affaires durant l'ingénierie de domaine d'application puis dans l'environnement d'exécution approprié. Pour réaliser cette approche, nous devons tout d'abord séparer l'ontologie de processus d'affaires de celle du domaine d'application. Ensuite réutiliser l'ontologie de processus dans les différents domaines d'application pour qu'il soit exécuté dans un environnement d'exécution approprié.

1 Introduction

Récemment les enjeux de l'ingénierie du processus d'affaires sont devenus une partie importante dans les méthodologies avancées de l'ingénierie d'entreprise. En particulier, on recense la Gestion des Processus d'affaires (BPM) (SMITH et FINGAR, 2003) et l'Architecture Orientée Service (SOA) (ERL, 2005), qui facilitent la réutilisation des logiciels (Donatas et Čaplinskas, 2007). Cependant, représenter les connaissances sur les familles des processus d'affaires similaires et la réutilisation de ces connaissances dans des projets de l'ingénierie d'entreprise reste un problème ouvert (Donatas et Čaplinskas, 2007).

Dans ce contexte, un système intégré des ontologies pour soutenir la modélisation d'entreprise a été développé à Toronto Virtual Enterprise (TOVE) projet (Fox, 1992). L'approche TOVE est critiquée puisqu'elle exige trop d'efforts pour instancier le modèle d'une entreprise particulière. Dans le même aspect, nous avons un autre travail concernant le projet d'entreprise (Ontologie D'Entreprise) à l'Université d'Édimbourg (Uschold et al, 1996) pour la modélisation d'entreprise. Les termes de cette ontologie sont exprimés sous une forme restreinte

et structurée de la langue naturelle complétée par quelques axiomes formels utilisant Onto lingua. Par conséquent, il ne supporte pas le raisonnement automatique (Albertas et al, 2000).

Nous avons constaté aussi que les travaux de Caplinskas et ses coéquipiers (Donatas et Čaplinskas, 2007), (Albertas et al, 2000), (Ciuksys et al, 2006) sont très intéressants, dans la mesure où ils ont proposé une approche à base d'ontologies qui permet la réutilisation des connaissances au niveau du processus d'affaires. La réutilisation présentée dans cette approche permet la résolution de variabilité d'une manière statique, c'est-à-dire durant la construction du processus d'affaires localisé dans un domaine d'application choisi, sans tenir compte de la variabilité qui peut avoir lieu durant l'exécution de ce processus. Ils ont ainsi proposé une approche à deux phases : L'ingénierie du domaine de processus et l'ingénierie de processus.

D'après le travail de JIANQI YU (JIANQI, 2010) sur les lignes de production, les approches pour la réutilisation à deux phases développées dans l'ingénierie de domaine et adaptées par Caplinskas sur les processus d'affaires, sont mises en place avec succès ; mais elles ne proposent pas de dynamisme à l'exécution. De plus, elles ne conviennent pas très bien aux environnements dynamiques changeants. De ce fait, nous trouvons qu'il a pris en considération la variabilité qui aura lieu durant l'exécution d'applications, où il a proposé une approche à trois phases : Une phase d'ingénierie domaine, Une phase d'ingénierie applicative et Une phase d'exécution. Pour les raisons citées précédemment, dans cet article, nous allons nous baser sur l'approche proposée par JIANQI YU sur les lignes de production et largement sur celle de Caplinskas et ses coéquipiers sur les processus d'affaires, qui ont été toutes les deux visées à la réutilisation, pour construire notre approche.

Donc notre contribution, consiste à :

Premièrement, distinguer deux parties dans la phase de l'ingénierie de processus proposée par Caplinskas : la phase de l'ingénierie applicative de processus et la phase de l'exécution de processus. Ainsi, l'architecture de l'approche que nous allons proposer comporte trois phases : la phase d'ingénierie du domaine de processus, la phase de l'ingénierie applicative de processus et la phase de l'exécution de processus.

Deuxièmement, pour réaliser cette approche nous devons tout d'abord séparer l'ontologie de processus d'affaires de celle du domaine d'application. Ensuite réutiliser l'ontologie de processus dans les différents domaines d'application pour qu'il soit exécuté dans un environnement d'exécution approprié. C'est pourquoi, nous proposons d'enrichir deux types d'ontologies réutilisables de haut niveau : Ontologie du domaine d'application de haut niveau et Ontologie du processus d'affaires de haut niveau. Les deux ontologies sont basées sur un système de méta-concepts, qui constitue une ontologie de haut niveau, que nous traitons également. La notion de processus d'affaires générique est introduite pour désigner une famille du processus d'affaires similaires. Nous mettons beaucoup plus l'accent sur la conceptualisation de variabilité entre les membres de même famille, où nous proposons un modèle pour résoudre les points de variation statiques et dynamiques, qui est notre contribution principal dans cet article.

Troisièmement, nous présenterons une étude de cas pour illustrer l'approche proposée.

La suite de ce travail est organisée comme suit : la section 2 décrit quelques notions de base. Dans la section 3, nous proposerons l'architecture générale de notre approche. La section 4 montrera les ontologies réutilisables de haut niveau qui ont été utilisées dans l'architecture proposée. Dans la section 5, nous présenterons une étude de cas pour expérimenter l'approche présentée. Enfin, dans la section 6 nous conclurons et donnerons des perspectives à cette recherche.

2 Notions de base

Avant de présenter notre approche, nous devons tout abord rappeler quelques notions de base. Alors, pour bien définir la notion de connaissance, il faut faire la différence entre les trois termes : donnée, information et connaissance. **Les données** ne sont ni vraies, ni fausses, elles sont transmises à un système ou un programme qui les traite, les modifie et les fait évoluer. Les données deviennent **informations** quand elles prennent un sens soit pour le système soit pour l'utilisateur. L'information**,** constituée de données, devient **connaissance** à partir du moment où elle sert de fondement à une inférence, au déclenchement d'un processus (Lame, 2002). L'élément de tout système basé sur la connaissance est une base de connaissances. Une base de connaissances est capable de stocker des données mais est également capable de leur associer une représentation formelle. La représentation formelle de connaissances a pour objectif de rendre celles-ci interprétables aussi bien par une machine que par un être humain. Les logiques de description (Baader et al, 2003) forment une famille de langages de représentation de connaissances d'un domaine d'application d'une façon structurée et d'une sémantique formelle, l'une des raisons qui nous amène à les utiliser pendant la phase de configuration dans l'approche à proposer.

Les logiques de description (LD) (Baader et al, 2003) reposent sur trois notions de base: les concepts représentant des classes (ensemble d'objets), les rôles (relations liant deux objets) et les individus. Pour décrire ces éléments, deux structures sont utilisées : la T-BOX et la A-BOX. La T-BOX (Boîte Terminologique) comprend la description des concepts et des rôles. La A-BOX (Boîte Assertionnelle) est constituée des individus, de leur description et des règles qui leur sont attachées.

La notion de **processus d'affaires** est définie comme un ensemble partiellement ordonné d'activités liées qui créent une valeur en transformant une entrée en une sortie plus précieuse. Les entrées et les sorties peuvent être des artefacts et / ou des informations et la transformation peut être effectuée par des acteurs humains, machines, ou les deux (Donatas et Čaplinskas, 2007). **Un processus d'affaires générique** (appelé aussi domaine de processus) est défini comme étant une abstraction d'une famille de processus d'affaires. Les membres de cette famille sont caractérisés par des points (parties) communs importants, et chaque membre particulier est également spécifié par des points supplémentaires (variabilité) (Donatas et Čaplinskas, 2007). Un processus d'affaires générique est décrit par une sorte de modèle de caractéristiques (KANG et al, 1990) et par une ontologie. Le processus d'affaires générique ne comprend pas les connaissances de contrôle sur la séquence des activités d'affaires. Les connaissances de contrôle sont ajoutées plus tard, réutilisant l'ontologie de processus dans un domaine particulier d'application. Ces processus génériques sont utilisés pour générer des processus particuliers, qui sont alors situés dans le domaine d'application choisi. Dans notre proposition, nous allons utiliser trois techniques de réutilisation : **La réutilisation dans l'ingénierie de connaissances.** Dans notre cas, nous devons séparer les connaissances de processus d'affaires de celles de domaine d'application. Ensuite réutiliser les connaissances de ce processus dans les différents domaines d'application. **La réutilisation dans l'ingénierie de domaine**. L'ingénierie de domaine, est conçue pour modéliser des familles d'application (CzARNECKI et EISENECKERl, 2000). L'ingénierie de domaine englobe les trois activités suivantes : l'analyse de domaine, conception de domaine, et l'implémentation de domaine. Les résultats de l'ingénierie de domaine sont réutilisés lors de l'ingénierie d'application, qui est, le processus de la production de systèmes concrets. On a donc deux activités de développement : Le développement pour la réutilisation (Ingénierie de domaine) et le développement avec la

réutilisation (Ingénierie d'application). **La réutilisation dans l'ingénierie des systèmes basés sur les ontologies**. L'ontologie est une spécification formelle et explicite d'une conceptualisation partagée (Borst, 1997). Ce qui fait partie d'une ontologie de domaine est limité par la capacité de représentation de méta-modèle (le langage utilisé pour construire le modèle). En général, une ontologie se compose de trois parties : Définitions des Concepts, Définitions des Rôles et Définitions de l'Inférence.

La section suivante présente notre proposition.

3 Proposition

L'approche proposée comporte trois phases, la figure "Fig. 1" illustre les activités principales durant chaque phase :

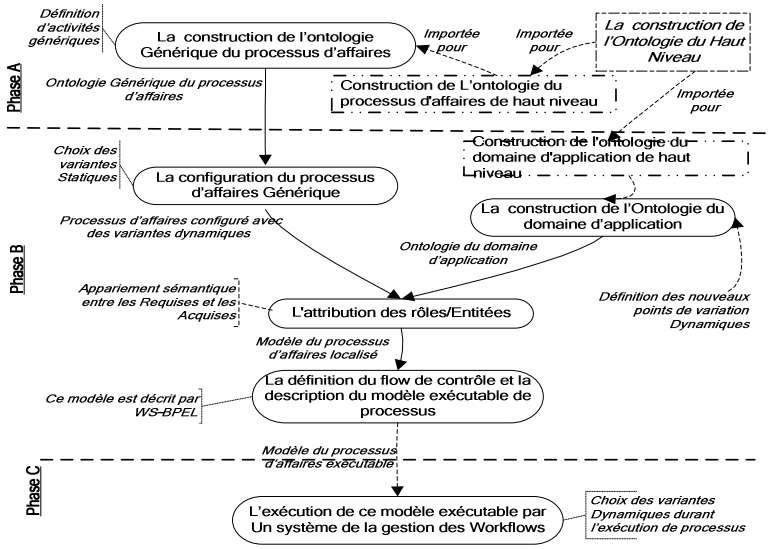


FIG. 1 – *Les activités principales durant chaque phase de l'approche proposée.*

3.1 La phase d'ingénierie du domaine de processus (A)

L'objectif de cette phase est la définition d'un domaine du processus d'affaires particulier. À cet égard, nous avons défini trois sous-activités qui sont : **L'analyse du domaine de processus.** Permet d'identifier les éléments communs ou variables entre les processus de ce domaine. Parmi les méthodes d'analyse du domaine, nous utilisons la méthode FODA (KANG et al, 1990), où le domaine est décrit sous forme d'un modèle de caractéristiques (Feature Model) (Donatas et Čaplinskas, 2007). **La conception du domaine de processus.** Permet de raffiner les termes définis par le modèle des caractéristiques et ajouter aux connaissances l'ontologie épistémique. L'ontologie du processus d'affaires générique (décrit en termes de rôles) produite est basée sur l'ontologie du processus d'affaires de haut niveau (cf. la section 4) (Donatas et Čaplinskas, 2007). **L'implémentation du domaine de processus.** Permet de créer des assets (artefacts) réutilisables à partir du modèle de caractéristiques et l'ontologie de processus. L'on-

tologie de processus comme une ressource réutilisable est représentée à l'aide du langage d'ontologie Web (OWL). Nous choisirons l'outil Protégé-OWL, pour implémenter les différentes ontologies proposées dans cette approche (Donatas et Čaplinskas, 2007).

3.2 La phase d'ingénierie applicative du processus d'affaires (B)

Elle permet la réutilisation des éléments, déjà définis dans la phase précédente. (Générer le processus d'affaires particulier et le placer dans le domaine d'application choisi). Cette phase commence par deux activités parallèles : **L'analyse du domaine d'application.** Nous tirerons profit du modèle de domaine existant et nous décrirons les besoins du client utilisant les caractéristiques à partir du modèle de domaine. Cependant les nouvelles exigences des clients qui ne figurent pas dans le modèle de domaine nécessitent un développement personnalisé. Le résultat de cette étape est une Ontologie du domaine d'application (décrit en termes d'entités). Elle est basée sur l'ontologie du domaine d'application de haut niveau (cf.la section 4). **La configuration du processus d'affaires générique.** Pour résoudre le problème de la configuration de processus, nous suivrons l'approche à base de Logiques de Description (DL) (Baader et al, 2003) décrite dans (Donatas et Čaplinskas, 2007), parce que nous définissons toutes les ontologies proposées utilisant l'Ontology Web Langage DL (OWL DL).

Donc, une solution au problème de configuration peut être proposée pour être un modèle de la base de connaissances donnée de type Logiques de Description (DL). Dans ce cas, l'espace de configuration est défini par le TBox qui décrit la hiérarchie de concepts et la hiérarchie de rôles. Donc, l'espace de configuration (TBox) pour le problème de la configuration *Paiement* dans un processus de commande, par exemple, est présenté dans le Tableau 1 :

Paiement_Partie ⊆ (Payer_par_Bill U Payer_à_la_livraison U Payer_par_Carte_de_crédit) ⊆ =1 Partie_De ∩ Partie_De.Paiement	(01)
Payer_par_Bill⊆¬ (Payer_à_la_livraison∩ Payer_par_Carte_de_crédit)	
Payer_A_la_livraison⊆¬ (Payer_par_Bill∩Payer_par_Carte_de_crédit)	
Payer_par_Carte_de_crédit⊆¬ (Payer_par_Bill∩Payer_A_la_livraison)	
A_Une_Paiement_Partie ⊆ A_Une_Partie	
A_Une_Paiement_par_Bill ⊆ A_Une_Paiement_Partie	(02)
A_Une_Paiement_A_la_livraison ⊆ A_Un_ Paiement_Partie	
A_Une_Paiement_par_Carte_de_crédit⊆A_Une_Paiement_Partie	
T ⊆ ∀ A_Une_Paiement_Partie. Paiement_Partie	
T⊆∀A_Une_Paiement_par_Bill.Payer_par_Bill	(03)
T⊆∀A_Une_Paiement_A_la_livraison. Payer_A_la_livraison	
T⊆∀A_Une_Paiement_par_Carte_de_crédit.Payer_par_Carte_de_crédit	
Paiement ⊆ Partie_Processus_de_commande ∀ A_Une_Partie. Paiement_Partie ∩ ≥ 1 A_Une_Paiement_Partie ∩ ≤ 1 A_Une_Paiement_par_Bill ∩ ≤ 1 A_Une_Paiement_A_la_livraison ∩ ≤ 1A_Une_Paiement_par_Carte_de_crédit.	(04)

TAB. 1 – *L'espace de configuration (TBox) pour un problème de configuration Paiement.*

TBox commence par donner ce qu'on appelle l'axiome de couverture pour le concept *Partie_De_Paiement*, aussi bien qu'exiger de *Partie_De_Paiement* d'être une et une seule partie de concept *Paiement*. Les axiomes additionnels assurent la disjonction de concepts *Payer_par_Bill, Payer_à_la_livraison* et *Payer_par_Carte_de_crédit* (01). Alors TBox présente la hiérarchie de rôle, déclarant que le rôle *a_une_Partie_De_Paiement* est un sous rôle

de rôle *a_une_Partie*, et les trois rôles *a_Une_Paiement_par_Bill*, *a_Une_Paiement_par_Carte_de_crédit* et *a_Une_Paiement_à_la_livraison* sont des sous rôles de rôle *a_une_Partie_De_Paiement* (02). La suite de TBox est une série de restrictions imposantes sur les rôles (03). Enfin des exigences sont énoncées pour le concept *Paiement* (04). Il doit être une partie du processus de commande, toutes les parties qu'il a doivent être des instances de concept *Partie_De_Paiement*, et il doit avoir au moins une *Partie_De_Paiement*, au plus un *Paiement_par_Bill*, au plus un *Paiement_à_la_livraison* et au plus un *Paiement_par_Carte_de_crédit*. ABox initial inclut exclusivement des variantes obligatoires. Dans notre exemple l'ABox initial est très simple :

$$A = \{t : Transaction, p : Paiement, a\_une\_Paiement\ (t\ ;\ p)\}.$$

Utilisant cette approche, le choix des variantes optionnelles ou alternatives nécessite que ABox soit augmenté avec les individus correspondants et qu'un raisonneur OWL DL soit sollicité pour vérifier si ABox est conformé à l'espace de configuration. Donc, ABox deviendra :

$$A = \{t : Transaction, p : Paiement, a\_une\_Paiement\ (t\ ;\ p), ppcc : Paiement\_par\_Carte\_de\_crédit, a\_Une\_Paiement\_par\_Carte\_de\_crédit(p\ ;ppcc)\}.$$

Dans le cas contraire, le choix doit être défait. Il est important que dans notre cas, le problème de processus de configuration soit résolu en mode interactif et que le processus de résolution des problèmes soit itératif. Pour assurer les dépendances entre les variantes, nous avons utilisé les règles de Horn (Motik, 2005). La règle selon laquelle le choix de la variante *Livraison_Électronique* nécessite de choisir les variantes *Facture_Affichée_En_Ligne* et *Paiement_par_Carte_de_crédit*, peut être écrite comme suit :

$$a\_une\_Livraison\_Électronique\ (a;\ le) \rightarrow a\_une\_Facture\_Affichée\_En\_Ligne\ (f;\ fael)\ \&\ a\_une\_Paiement\_par\_Carte\_de\_crédit\ (p;\ ppcc)\ (I)$$

Ici *a, le, f, fael, p, ppcc* sont des individus des concepts : *Accomplissement, Livraison_Électronique, Facture, Facture_Affichée_En_Ligne, Paiement, Paiement_par_Carte_de_crédit*, respectivement. Le processus d'affaires configuré et l'ontologie du domaine d'application sont des entrées pour la prochaine étape, qui est l'attribution des rôles.

L'attribution des rôles. L'attribution des rôles aux entités est réalisée en correspondant les requises de rôles aux acquises d'entités, selon l'algorithme (Tableau 2.) suivant qui est proposé dans (Donatas et Čaplinskas, 2007) et enrichi par les concepts que nous avons ajoutés :

1 Répéter Rôle. Requises ← Capacités et autorités requises ;
2 Répéter Entité. Acquises←Capacités et autorité acquises ;
Si (Entité. Acquises = Rôle. Requises) **alors**
Attribution de rôle, Rôle ← Entité ; aller à **1** ; **FinSi**
Si (Entité. Acquises englobant Rôle. Requises) **alors**
Cette entité doit être reconstruite Et divisée en plusieurs entités spécifiques (la spécification des Acquises) ; aller à **2** ; **FinSi**
Si (Entité. Acquises sont englobées par les Rôle. Requises) **alors**
Ces entités doivent être composées une seule entité, qui est plus grande, entité composite (la généralisation des Acquises) ; aller à **2** ; **FinSi**
Jusqu'à ce que toutes les Entités soient vérifiées ;
Si (aucun candidat d'entité active n'a joué un certain rôle) **alors** Une nouvelle entité doit être créée ; **FinSi**
Jusqu'à ce que tous les rôles soient attribués.

TAB. 2 – *Algorithme de l'attribution des rôles.*

Pendant l'étape de configuration, on peut conserver certains points de variabilité qui seront résolus lors de la phase suivante.

La définition de contrôle du flow et la description de modèle exécutable du processus. Après avoir localisé le processus d'affaires dans un domaine d'application choisi. Nous devons définir, dans cette étape, l'ordre d'exécution des activités de ce processus, par conséquent, nous ajouterons les connaissances de contrôle qui définissent l'ordre d'exécution de ces activités (Donatas et Čaplinskas, 2007). Donc, un modèle exécutable de processus d'affaires est produit. Nous pouvons utiliser, un langage d'exécution de processus tel que WSBPEL (OASIS, 2007), qui va exécuter par un système de gestion de Workfows. Ce système illustre l'exécution d'activités du processus.

3.3 La phase d'exécution du processus d'affaires (C)

La troisième phase de notre approche a comme but de gérer l'exécution de processus d'affaires défini précédemment par un langage tel que WSBPEL. Cela par la résolution des variabilités qui existent encore dans ce modèle exécutable du processus d'affaires. Comme nous l'avons déjà mentionné, selon l'approche proposée dans (Donatas et Čaplinskas, 2007) en cas d'absence d'un service ou d'une ressource pour l'exécution d'une activité pendant l'exécution d'un processus d'affaires, ils n'ont pas proposé un autre choix. En d'autres termes, ils n'ont pas proposé ou montré la résolution des points de variations qui peuvent avoir lieu durant l'exécution de processus. La différence majeure entre notre proposition et celle de (Donatas et Čaplinskas, 2007), c'est que, en cas d'absence des services ou ressources nécessaires pour exécuter une activité, nous devons permettre leurs créations ou bien nous devons exécuter une autre activité avec des services ou des ressources d'exécution disponibles. Pour cette raison, nous pouvons distinguer un autre type de points de variations, qui se compose des points de variation dynamiques reliés à l'environnement d'exécution. Cela veut dire, que certains services ou ressources, dans cet environnement, doivent être disponibles pour qu'une activité s'exécute. En cas d'absence de ces services ou ressources, comme nous avons déjà mentionné, nous avons la possibilité ou le choix d'exécuter une autre activité avec un service ou une ressource d'exécution est disponible ou bien de créer ces services manquants. Dans l'exemple du processus de commande, si nous n'avons pas un protocole ou un service pour faire une *Livraison_Électronique*, nous pouvons faire une *Expédition* durant l'exécution de ce processus dans un domaine d'application et environnement d'exécution bien choisi. Dans ce cas là, nous allons choisir l'exécution de l'activité *Expédition* ou lieu d'exécuter l'activité *Livraison_Électronique.*

La section suivante présente les différentes ontologies de haut niveau utilisées dans cette proposition.

4 Les ontologies de haut niveau utilisées

4.1 L'ontologie de haut niveau

Tant l'ontologie du domaine d'application que l'ontologie de processus, elle devrait être décrite par un certain système commun de méta concepts. Cela signifie qu'une certaine ontologie de niveau plus haut "Fig. 2" est exigée. La différence entre l'ontologie de haut niveau

décrite dans (Donatas et Čaplinskas, 2007) et notre ontologie c'est que nous avons introduit le concept de Temps, que nous considérons important, surtout durant l'exécution du processus d'affaires.

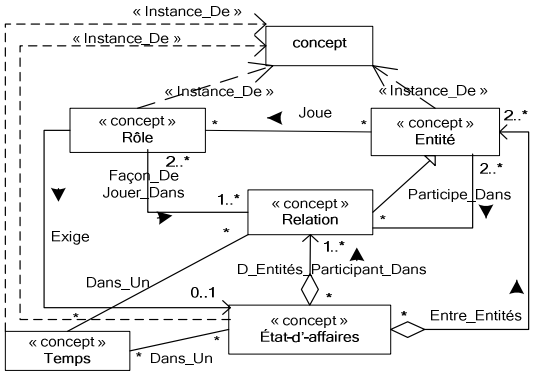


FIG. 2 – *L'ontologie de haut niveau.*

4.2 L'ontologie du domaine d'application de haut niveau

Cette ontologie "Fig. 3" est considérée comme une base pour définir les concepts de l'ontologie d'un domaine d'application particulier. La différence entre notre ontologie et celle proposée par les auteurs de (Donatas et Čaplinskas, 2007) c'est que nous avons ajouté quelques concepts tels que : L'Intervalle de Temps, qui est une instance de concept Temps, représente l'intervalle du temps estimé pour changer l'état d'une Entité passive par une Entité active. Il peut aussi représenter l'Intervalle de temps pour atteindre l'Objectif ou les Sous Objectifs d'une Activité d'affaires. Autorité, dans certain cas les Entités actives ont comme Acquises des Autorités au lieu de Capacités ou bien les deux en même temps, qui peuvent changer l'état d'une Entité passive.

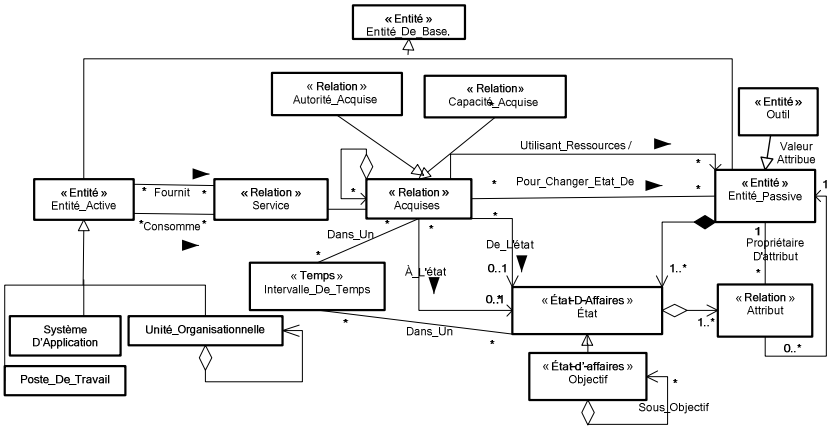


FIG. 3 – *L'ontologie de haut niveau du domaine d'application du processus d'affaires.*

4.3 L'ontologie du processus d'affaires de haut niveau

L'ontologie du processus de haut niveau qui a été proposée par Caplinskas (Donatas et Čaplinskas, 2007) est un sous-ensemble de BPDM (BPDM, 2007). Notre contribution dans cette partie d'ontologie, est d'ajouter aux requises l'autorité pour exécuter une activité ou une sous-activité, dans un Intervalle de temps estimé "Fig. 4". Ces concepts ne sont pas suffisants pour représenter les variabilités fournies par les modèles des caractéristiques. Ainsi, les concepts tels que : Variabilité, Variante, Point de variation, etc. doivent être inclus dans l'ontologie du processus de haut niveau (Donatas et Čaplinskas, 2007).

Notre apport dans la conception de variabilité est d'adapter le modèle général de variabilité dans les lignes de produits, qui est proposé par Becker (Becker et al, 2003), aux processus d'affaires :

La Variabilité. Représente une capacité de changer ou adapter un système (Gurp et al, 2001) ; dans notre contexte, le processus avec variabilités peut être adapté dans divers domaines d'application. Le Point de variation : est une place dans une asset logiciel où la variabilité se produit (Becker et al, 2003). La Variabilité correspond à un ensemble de points de variation, qui définissent l'étendue de variabilité (c'est-à-dire ; plus est élevé le nombre de variantes, plus est élevé le degré d'adaptabilité du système). Pour la raison de simplicité, nous permettons à la variation du processus de se produire seulement dans les activités, et d'appeler ces activités des activités génériques (Donatas et Čaplinskas, 2007).

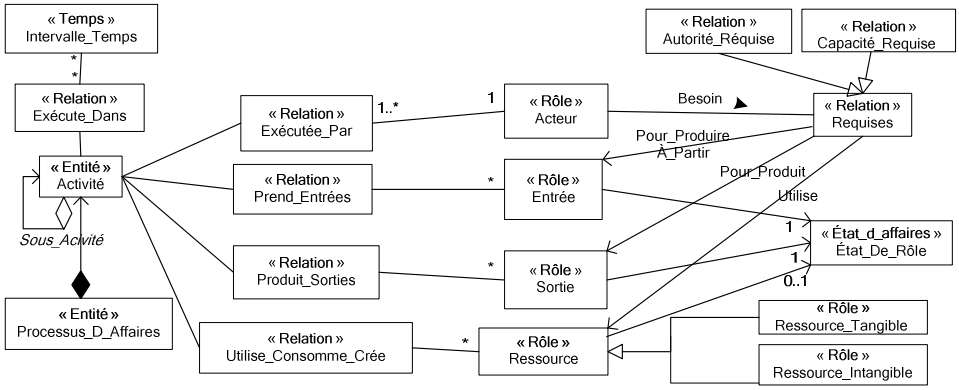


FIG. 4 –*L'ontologie du processus d'affaires de haut niveau (Conceptualisation de processus).*

Dans notre cas, nous distinguons deux types des points de variation : les points de variation statiques qui seront résolus pendant la phase de l'ingénierie applicative du processus d'affaires ; les points de variation dynamiques qui seront résolus durant la phase d'exécution du processus, afin de l'adapter aux changements de contexte. Nous nous sommes basés sur (JIANQI, 2010), pour distinguer deux états pour un point de variation, fermé ou ouvert : Si le point de variation est « ouvert. », dans ce cas, il est possible pendant l'ingénierie applicative d'ajouter une nouvelle variante ou de modifier des variantes existantes. Si le point de variation est « fermé. », dans ce cas, il n'est pas possible de sortir des choix proposés par le point de variation. Le Tableau 3 représente un algorithme qui illustre la résolution de la variabilité au sein d'un processus d'affaires générique.

Modèle de processus d'affaires génériques {contient des points de variation qui doivent être résolus afin de configurer ce processus premièrement dans un domaine d'application et deuxièmement dans un environnement d'exécution bien choisi.}
{La résolution des points de variation statiques :}
Répéter
Si (Points De Variation = Points De Variation Statiques) **alors :**
- Résolution des points de variation (Choix de variantes statiques, les dépendances restreignent ces choix) {c'est une configuration durant la phase d'ingénierie applicative du processus} ;
- Ajouter ces variantes statiques au profil de dérivation statiques ;

FinSi
Jusqu'à ce que tous les Points De Variation Statiques soient résolus;
Modèle de processus d'affaires localisé dans un domaine d'application bien choisi ;
{La résolution des points de variation dynamiques :}
Répéter
Si (Points De Variation = Points De Variation Dynamiques) **alors :**
- Résolution des points de variation (Choix de variantes Dynamiques) {c'est-à-dire la configuration durant la phase d'ingénierie de l'exécution de processus} ;
- Ajouter ces variantes Dynamiques au profil de dérivation Dynamiques ;

FinSi
Jusqu'à ce que tous les Points De Variation Dynamiques soient résolus ;
Modèle de processus d'affaires exécutable {Localisé dans un environnement d'exécution bien choisi.}, qui va être exécuté par un système de gestion des Workflows.

TAB. 3 – *Un algorithme qui illustre la résolution de la variabilité au sein d'un processus d'affaires générique.*

La Variante choisie doit être intégrée dans le système au Point de Variation. Pour réaliser cette intégration, des Mécanismes d'Implémentations (Puhlmann et al, 2005) précisent des techniques qui dépendent du type de variantes et le lieu où l'intégration doit être effectuée. Les moments exacts où les décisions retardées sont prises sont variables. Ces moments, sont généralement appelés "Binding Time" (JIANQI, 2010). Un profil de dérivation comporte un ensemble d'affectations. Chaque affectation représente une décision prise (Choix de variante, que se soit statiques ou dynamiques). Par exemple, La variante (A) a été choisie au point de variation (B) pour la variabilité (C) au BindingTime (D). Si aucune affectation n'est disponible pour une variabilité, donc la Variabilité n'est pas liée au profil de dérivation.

Dans la section suivante, nous présenterons une étude de cas pour illustrer la proposition.

5 Étude de cas

Pour bien comprendre l'approche proposée, nous avons pris l'exemple d'un processus de conférence simplifié qui est considéré comme un processus générique. Cela signifie qu'il peut être réutilisé dans différentes universités, collèges et pour faire une présentation lors des conférences. Dans le cadre de notre travail, nous avons appliqué l'approche proposée sur ce processus générique de Conférence. Premièrement, nous avons donné une conceptualisation détaillée de l'application de cette approche sur le processus de conférence. Cela durant les trois phases proposées : l'ingénierie du domaine de processus, l'ingénierie d'applicative du processus et la phase de l'exécution de processus. Nous avons décrit principalement les points suivants :

- Nous avons construit un modèle des caractéristiques du processus de conférence. Nous avons défini la logique du déroulement des activités de processus Conférence. Nous avons exposé une conceptualisation de toutes les activités de processus conférence (Donner conférence, Déverrouiller la salle, Verrouiller la salle, Poser la question, Répondre à

la question, Commencer La conférence et Fermer La conférence), sous forme d'ontologies du processus à partir de méta-concepts «Ontologie du processus d'affaires de haut niveau» mentionné dans la section 4. Nous avons aussi défini une ontologie, à partir de méta-concepts « Ontologie de domaine d'application de haut niveau » mentionné dans la section 4, elle représente l'Ontologie du domaine d'application du processus de cours dans une université particulière. Pour résoudre le problème de la configuration de processus, nous avons suivi l'approche à base de la Logique de Description (DL) (Baader et al, 2003) décrite dans (Donatas et Čaplinskas, 2007), parce que nous avons défini toutes les ontologies proposées utilisant l'Ontology Web Langage (OWL) à base de la Logique de Description (DL). Nous avons remarqué que pendant cette étape, certains points de variation, peuvent rester non résolus, la résolution ou la prise de décisions sur ces points de variations sera durant la phase de l'exécution de processus. Le processus de conférence configuré et l'ontologie du domaine d'application (correspondant au processus cours) sont des entrées pour la prochaine étape, qui est l'attribution des rôles.

- L'attribution des rôles est réalisée en faisant correspondre les requises aux acquises, selon l'algorithme qui a été proposé dans (Donatas et Čaplinskas, 2007) et enrichi par les concepts que nous avons ajoutés. Nous avons ajouté des entités à l'ontologie de domaine d'application selon l'algorithme d'attribution des rôles, dans ce cas la nous sommes arrivés à construire l'ontologie du domaine d'application finale de processus qui correspond à l'ontologie du processus de conférence configurée. Nous avons défini l'ordre d'exécution des activités de ce processus, nous avons donné un diagramme qui représente cet ordre. À la fin de cette description et durant la phase de l'exécution du processus cours, nous avons donné un modèle des caractéristiques qui présente des points de variation dynamiques qui peuvent être résolus durant l'exécution du processus.

Deuxièmement, nous avons donné un aperçu schématisé sur l'aspect d'implémentation de cette approche. Nous étions devant quatre problèmes principaux à implémenter :

Nous avons utilisé l'outil Protégé pour implémenter toutes les ontologies déjà mentionnées dans cette approche, avec les différentes relations, restrictions et règles de dépendances entre les différents concepts. Le raisonneur Racer a permis de tester la cohérence et la consistance de ces ontologies. L'outil Protégé avec le raisonneur Racer, nous a permis aussi d'utiliser la logique de description pour résoudre le problème de la configuration de processus. Cela est réalisé en utilisant les différentes relations, restrictions et règles de dépendances entre les différents concepts. Notons que le plugin SWRL RULES nous a permis de définir ces règles de dépendances "Fig. 5".

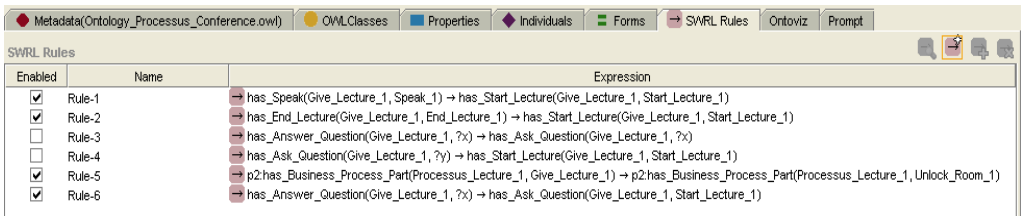


FIG. 5 – *Les règles (01), (02), (03), (04), (05), (06) sont implémentées avec l'onglet SWRL Rules d'outil Protégé*

- Pour l'attribution des rôles, nous avons utilisé la fonction Map de plugin Prompt de l'outil Protégé "Fig. 6". L'implémentation de modèle exécutable de processus cours est

possible via un langage d'exécution des processus tel que WSBPEL, ce dernier doit être exécuté par un système de gestion des Workflow.

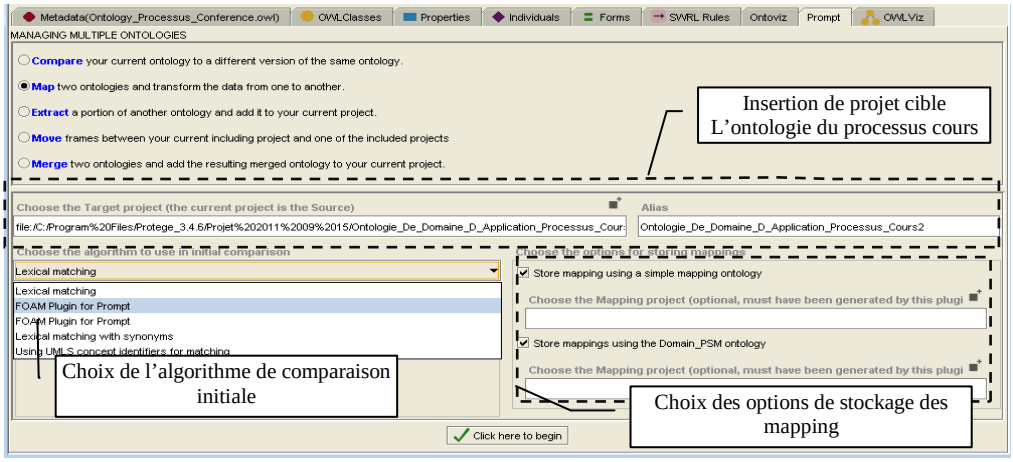


FIG. 6 – *La fonction Map de plugins Prompt dans l'outil Protégé*

Dans ce qui suit nous présentons une conclusion générale de notre travail ainsi que les perspectives entendu pour des travaux future.

6 Conclusion

Dans cet article, nous avons d'abord proposé une architecture à trois phases qui illustre la réutilisation. Nous avons distingué deux phases durant la phase de l'ingénierie de processus, proposée par Caplinskas, ce qui ajoute une certaine dynamique à l'exécution du processus d'affaires ; ces deux phases nommées : Phase d'ingénierie applicative du processus et Phase d'exécution du processus. Ensuite, pour soutenir l'architecture déjà présentée, nous avons enrichi deux types d'ontologies réutilisables de haut niveau : l'une pour la représentation des connaissances d'un processus d'affaires générique, pendant la phase de l'ingénierie de domaine du processus (nous nous sommes beaucoup plus concentrés sur la conceptualisation de la variabilité) et l'autre ontologie est désignée pour la représentation des connaissances d'un domaine d'application particulier, pendant la phase d'ingénierie applicative du processus. Les deux ontologies sont basées sur une ontologie de haut niveau, que nous avons exposé aussi.

Enfin, nous avons procédé à une étude de cas pour illustrer l'approche proposée. Nous avons utilisé l'outil Protégé pour implémenter les trois types d'ontologies déjà mentionnés pour le domaine d'application choisi avec les différentes relations, restrictions et dépendances entre les différents concepts. Nous avons aussi utilisé la logique de description pour résoudre le problème de la configuration de processus, ensuite pour l'attribution des rôles, nous avons utilisé la fonction Map de plugin Prompt d'outil Protégé. Ce qui distingue cette approche des autres (par exemple les approches à base de Enterprise Resource Planning (ERP)) est que les processus d'affaires sont adoptés pour les domaines d'application, et non pas inversement. Parmi nos perspectives, nous comptons procéder à la réalisation de la phase d'exécution du processus pour construire un processus d'affaires exécutable et l'exécuter dans un système Workflow.

Références

Albertas CAPLINSKAS, Audron eLUPEIKIEN, E Olegas VASILECAS (2000). The Role of ontologies in reusing domain and enterprise engineering Assets; INFORMATICA,2003,Vol.14,No.4,455–470 Institute of Mathematics and Informatics, Vilnius.

Baader.F, D. Calvanese, D.L. Mcguinness, D. Nardi and P.F. Patel-Schneider (2003). The description logic handbook; Theory, implementation, and applications. Cambridge University Press. 574pp.

Becker, M. (2003). Towards a general model of variability in product families. In J. van Gurp, J. Bosch (Eds.), Proceedings of the 1st Workshop on Software Variability Management, Groningen, The Netherlands. pp. 19–27

Borst, W. N (1997). Construction of Engineering Ontologies. Center for Telematica and Information Technology, University of Tweenty, Enschede, NL.

Business Process Definition MetaModel (BPDM) (2007). Object Management Group, Inc. Access through Internet http://www.omg.org/cgi-bin/ doc?bmi/ [2007-03-01].

Ciuksys, D., and A. Caplinskas (2006). Modelling of reusable business processes: an ontology-based approach. In A.G. Nilsson, R. Gustas, W. Wojtkowski,W.G.Wojtkowski, S. Wrycza, and J. Zupancic (Eds.), Advances in Information Systems Development, Vol. 1. Springer. pp 71–82.

CzARNECKI, K.; EISENECKER, U (2000). Generative Programming: Methods, Tools, and Applications. Addison-Wesley Professional,.

Donatas Čiukšys, albertas Čaplinskas (2007)., Ontology-based approach to reuse of business process knowledge, ISSN 392-056. INFORMACIJOS MOKSLAI.

ERL, T. (2005). Service-Oriented Architecture : Concepts, Technology, and Design. Upper Saddle River: Prentice Hall PTR.

Fox, M.S (1992). The TOVE project: a common-sense model of the enterprise, industrial and engineering applications of artificial intelligence and expert systems. In F. Belli and F.J. Radermacher (Eds.), Lecture Notes in Artificial Intelligence, 604. Springer–Verlag. pp. 25–34.

JIANQI YU (le 16 juin 2010), Ligne de produits dynamique pour les applications à services, Thèse préparée au sein du Laboratoire d'Informatique de Grenoble (LIG).

KANG, K; COHEN, S.;HESS, J.; NOVAK, W.; PETERSON, A. (1990). Feature-Oriented Domain Analysis (FODA) Feasibility Study. Technical Report CMU/SEI-90-TR-21, SEI. Carnegie Mellon University, Pittsburgh, Pennsylvania.

Lame, G, Construction d'ontologie à partir de texte, une ontologie du droit dédiée à la recherche d'information sur le Web, Thèse de doctorat, Ecole des Mines de Paris, 2002.

Motik B., Sattler U. and Studer R (2005). Query Answering for OWL-DL with Rules.Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 3(1), 41–60.

Puhlmann, F, A.Schnieders, J. Weiland and M. Weske (2005) Variability mechanisms for Process Models.PESOA-Report No.17/2005.

SMITH, H; FINGAR, P (2003). Business Process Management (BPM): The Third Wave.Meghan Kiffer Press.

Uschold, M., M. King, S. Moralee, Y. Zorgios (1996.). The enterprise ontology. Knowledge Engineering Review, 13(1), 31–90.

Van Gurp, J., J. Bosch, M. Svahnberg (2001). On the notion of variability in software product lines. In R. Kazman, P. Kruchten, C. Verhoef, and H. van Vliet (Eds.), Proceedings of the Working IEEE/IFIP Conference onSoftware Architecture. IEEE Computer Society Press. pp. 45–54.

Web Services Business Process Execution Language v2.0 (2007). OASIS standard. Access through Internet: http://docs.oasis-pen.org/wsbpel/2.0/OS/wsbpel-v2.0-S.pdf.

Summary

We noticed that the approaches for reuse with two phases developed in domain engineering and adapted by Caplinskas on business processes, are implemented successfully, but they do not offer dynamic execution. In addition, they do not agree very well with the changing dynamic environments.

Where, in this article, we propose an approach in three phases : the domain process engineering phase, the application process engineering phase and phase of process execution. This approach based on ontology to allow knowledge reuse of business process during the engineering application domain and in the appropriate execution environment.

To achieve this approach, we must first separate business process ontology and application domain ontology. Then reuse the process ontology in different application domains to be run in a suitable execution environment.

Textual Knowledge Modeling By Dynamic Ontology: Application on cancer disease inflammatory and non inflamatory

N.Taleb , B.Tighiouart
Research laboratory in data processing, Annaba, Algeria
Nora.taleb@univ-annaba.org

Abstract – Domain knowledge is evolving over time and thus is should be also expressible at Ontology level. This paper concerns Ontology detecting changes from texts, which is one of the key problems facing Ontology users today. We propose a bottom up methodology consisting in a rigorous analysis of changes, which ensures that the Ontology remains consistent after changes have been applied on the linked corpus.
The methodology is implemented in order to lighten the cognitive load of users, it includes a change detection mechanism that allows generating automatically a detailed overview of changes that have occurred based on a set of change definitions. . The experiment is done on medicals data on cancer disease to detect the parameters of the disease that promotes cancer inflammatory versus non inflammatory

1 Introduction

The Ontology evolution concerns the capacity to update the existing Ontology following the appearance of new changes on the corpus of texts, and to maintain its uniformity and its coherence [1] [2]. This aspect remains poorly studied in the literature in spite of a lot of researches aligned around the construction of static ontologies.

In this perspective, we propose a methodology supporting firstly the dynamics of the Ontology built from a corpus of texts, and secondly to follow the changes made in the studied domain.

The proposed methodology, OntoEvol, like grouping together the principles of the studied methodologies and avoiding the limits encountered in the existing methodologies. The methodology OntoEvol details mainly the following three important phases of the evolution process : the representation of the changes, their semantics and their implementation. These steps are based on the modules MOCO and MOTO; the first module concerns the management of the evolution of corpus and the second concerns the ontological evolution.

The implementation will be detailed by presenting a prototype of a software tool corresponding to OntoEvol. The experiment has been made on a medical corpus of cancer-related texts linked to its Ontology “CHIFA”.

We first make a survey of the main approaches of evolution with their limits. We present the methodologies AIFB and IMSE on which we based to draw our approach.

We then present our proposition for the ontological evolution based on the changes in textual knowledge, its basic principle, and the process of evolution.
The process of evolution will be detailed according to two axes: corpus evolution and Ontology evolution.

We will finish this paper by an experiment in the medical domain of cancer. In this experiment, the developed software tool will be exposed at the end of part.

2 State of the art

Several studies highlight the major importance of Ontology evolution as well as the lack of the approaches to manage this evolution [1]-[2]. The authors propose methodological elements concerning the evolution notion. Heflin and others have developed SHOE[3] [4], a language based on HTML which offers primitives for the management of the multiple versions. Oliver, Shahar, Musen, and Shortliffe have defined a conceptual model, CONCORDIA[5], to manage the changes of a medical terminology. McGuinnes provides theoretical recommendations to guarantee a process of evolution of the Ontology with a minimum of errors by using fusion techniques [6]. Several works as [7], [8], [9] based on texts, evoke the necessity of associating an independent lexicon with an Ontology.

Finally, Stojanovic [10], Klein and Noy[11], Luong [12], Rogozan [13] propose approaches to manage evolution, but all these works take into account only the Ontology and put aside the terminology. In addition a few works are interested in the evolution from a corpus of texts. In [7], [8], [9], the problem concerning the Ontology evolution in relation with textual knowledge is evoked, but no methodology for evolution is proposed. Currently, only two approaches deal with the evolution of an Ontology on the methodological level: (1) the approach developed by the AIFB team at the Karlsruhe University [1], and (2) the approach developed by the team of the IMSE department of the Amsterdam University [1]-[14].

2.1 AIFB methodology to support Ontology evolution

The AIFB methodology consisted of five main steps [2], [12], [15] - [16]:

1. Representation of the changes. This step aims at the edition of the elementary or complex changes [11].
2. Semantics of the changes. The task of this step is to allow the resolution of all the additional changes in a systematic way.
3. Implementation. This step aims at the execution of the changes, once approved by the users.
4. Propagation of the changes. The purpose is to modify automatically the authorities and the dependent artifacts to ensure their consistency with the evolved Ontology.
5. Validation. The users evaluate the evaluation result and repeat the process if necessary.

2.2 IMSE methodology to support Ontology versioning

The IMSE approach considers the use of several versions of an Ontology and the access to artifacts being made by means of these multiple versions. The authors use then the term versioning to describe their approach [1]-[14].

2.3 Limits of the studied approaches

Theauthors of the AIFB methodology do not propose any step of analysis of the effects of the changes on the compatibility relationship between the evolved Ontology and the dependent artifacts, in our case the corpus of texts. This is an important limit since Ontology evolution can be provoked by the evolution of dependent artifacts. Secondly, the authors develop an evolution process which does not take into account the management of versions.
Concerning the IMSE versioning methodology, it does not support the Ontology evolution process, but it ensures the management of the versions of Ontology after its evolution. It offers an analysis model to analyse the relationship between the different versions of the Ontology, but without considering the management of the access to the dependent artifacts (i.e., referenced objects, Ontology, applications, corpus of texts) using Ontology version. Furthermore, the authors have not developed a functional framework to integrate the proposed methodological elements.
The AIFB and IMSE approaches are currently the only ones having methodological indications to support Ontology evolution, but they do supply neither a complete methodology nor a framework for integrating the proposed elements and do not deal with the management of the Ontology evolution from a corpus of texts.
Our objective is to integrate these approaches into a unified methodology. The OntoEvol methodology describes the complete process of Ontology evolution where the pieces of knowledge are extracted from a corpus of text. In this approach we have to preserve the first version of the Ontology because it is validated by domain experts.

3 OntoEvol methodology

The **OntoEvol** methodology manages the evolution around the couple of corpus-Ontology. It respects the global evolution process (EP) of the AIFB methodology and it uses the versioning principle of the IMSE methodology.
OntoEvol is an equipped approach. It includes techniques, tools and algorithms to describe the global evolution process. We will focus on the structural level and the lexical one to represent the Ontology and the corpus of the domain.
In our domain of study which is the cancer disease, the Ontology describes the categories of cancer, the causes, the symptoms, as well as the possible treatments at present.
We are recovered some steps from the evolution process of AIFB methodology in order to present the evolution process of our methodology. These main steps are:

1. Representation of the changes: consists of the identification and the edition of the changes which can intervene on the Ontology further to changes made on the correspondent corpus of texts.
2. Semantics of the changes

3. Implementation

3.1 Representation of the changes and MOCO module

We are in a medical domain where the knowledge expresses them in natural language. The studied corpus is in French language; it is characterized by its continuous evolution in the time further to the knowledge evolution. These changes can be: Addition of a new text, deletion of an existing text, update of the corpus.

The methodology OntoEvol proposes a module of corpus treatment "MOCO" to clearly identify the changes introduced into the studied corpus, and provides to the second phase "Semantics of the changes" two bases: Base of terms, and Base of verbal relations.

The module MOCO is grouping a set of techniques and tools. These techniques are extracted from the linguistic and statistical approaches for the texts treatment. MOCO is decomposed into two big phases: Phase of corpus pretreatment, and Phase of evolution.

For the realization of these two phases we used a statistical linguistic approach LSTAT [17] which we designed in our laboratory LRI[1] in another research project.

3.1.1. The evolution phase

It is the phase of acquisition of the knowledge itself, it is divided into two steps: the acquisition of terms and their structure and the acquisition of the relations.

1) Terms acquisition

This phase consists of the noun phrases construction which can be the terms of the domain. The inputs of this phase are the linguistic filters, the text (to add, to delete, and to modify) and the initial corpus. The outputs are the base of terms and the base of relations between terms.

Several approaches have been developed for the analysis of texts and the extraction of noun phrases susceptible to be afterward the terms representing the domain. There are different categories of these like linguistic approaches, statistical approaches, or hybrid approaches [18];

In our research framework, we chose to use the hybrid approach *LSTAT* for the knowledge acquisition [113,114], developed within the *LRI* laboratory.

2) Terms structuring

We adopt the structuring in network of syntactic dependence (head and expansion) by adding the statistical and semantic associations. Our hypothesis of structuring is:

The terms which share the same head, co-occurring often or are linked together by one of the semantic relations (Synonym, hyponyms, and causal clause) represents the same subject.

This structuring allows to extract the relations inside a terms (relations between expansions), as well as the relations outside terms. The objective is to recover the dependence syntactic, statistical and semantic at the same time, to supply a coherent view on the corpus content.

This structure will provide us with a concepts base from a terms base.

[1] Informatics research laboratory

This principle consists in synthesizing the terms which have the same head in a single tree structure [19],[20],[21], consisting in a single header and of a set of expansions.

3) Relations extraction

This is the second part in the evolution phase of the corpus treatment module "MOCO". It consists of the relations domain extraction . It exploits the obtained terms to acquire and structure the appropriate relations for the domain, it describes:
1. How to identify specific relations with the domain by investigating linguistic contexts between terms?
2. The passage from the relations between terms towards the relations between concepts.

To enrich our model, the relations acquisition process tries at first to identify the relations which have for marker "is a" "est un" in French, explicitly indicated in the corpus, then proceeds to search of the left and right argument of the relation which are terms belonging to terms classes, after that it identifies the concepts expressing these terms (that means looking for the class of a given term), the class of the left term represents the left concept, and the class of the right term represents the right concept.

4) The" is-a" relation

This type of relation can express a definition or a synonymy; a part of the relations expressed by this marker was extracted in the section "structuring of terms".
To enrich our model, we tried to identify all the sentences which contain the relative word ("est un" and its flexions "sont des") with the only constraint that: the left and right arguments are terms and are not sentences.

3. 2. Semantic changes and the MOTO module

It is the second phase in the Evolution Process EP; it consists in determining the changes made on the corpus of texts toward the Ontology.

This phase is based mainly on a second module of Ontology treatment MOTO, which contains a set of algorithms based on text mining techniques.
The role of this module is to manage the Ontology evolution, while respecting the constraints of the ontological model, such as the Ontology must evolve from one consisting state to another consisting state by preserving its coherence.
The Ontology treatment module MOTO uses the results of the module MOCO as input data.
Let us remind that our job does not consist in making changes directly on the Ontology, but the evolution of the Ontology is managed during the time when the corpus of the domain is updated.

Before starting the presentation of this phase we should identify the types of changes that may take on the ontological model.

The ontological model

To formalize our Ontology, we are inspired by the work [22], [23] and the results of the first phase (representation changes).
We consider that the Ontology consists of two levels (structural, and lexicon):

- The Ontology structure is represented by the tuple: *S :={ C, R, H, X}* where:

- *C:* the set of concepts.
- *R*: the set of relations.
- *H:* CxC: a partial order on C it defines the hierarchy of concepts. IF $(C_1,C_2) \in$ H then C_1 is a sub concept of C_2 .
- **X:** R→ CxC: is the signature of an associative or not taxonomic relation.

- The lexicon of the Ontology which is a tuple L :{ Lc, Lr, F, G} where:

- *Lc, lr* : Separated sets containing terms with their frequencies associated with the concepts and relations.
- *F, G*: are two reference relations who allow reaching the terms associated with the concepts or with the relations.

We note that a concept C_i can be identified by several terms (Ci is contained in the C set of the structure S of the Ontology O)
Operationally, we use the software Protégé or the Ontology implementation. This software integrates the representation language OWL, which presents the interest to exploit, on one hand, a description logic to describe the concepts and the relations between concepts,, and on the other hand a language of tags which can be used in a context of information search.

The types of changes and their effects on the Ontology

Each change in the corpus can cause others changes in the Ontology.
The analysis of the consequences of changes allows us to predict the effects on the Ontology elements.
The Ontology treatment module MOTO contains a conversion module of changes in OWL format. This module converts directly the MOCO results and integrates them with the basic Ontology. It has an algorithm to check the ontological constraints before any changes.
The changes may be represented at different levels of granularity [24].
A usual changes classification of the Ontology is quoted in [10]. This classification is based on the level of abstraction such as the elementary changes, and the compound changes.
For each type of change, we describe the syntax, the parameters and its semantics. We also describe the conditions before and after which must be satisfied to make every change. (Table I, II, III, V)

3.2.1 Case 1: Adding

A concept which will be used in the classification algorithm is defined as:
A concept Ci is the elementary entity of the Ontology.
It belongs to the concepts set C of the Ontology O. Its attributes belong to the lexicon set L_C of the Ontology O.
The attributes of the concept C_i are given by Lc (C).
For the concepts classification we will used a *Similarity measure.* It determines if the concepts are independents or equivalents. Its calculation is based on the vector of the concept C.
The calculation of this measure is like comparing the extract term ti with all the identifiers of all the concepts, that's mean with the set Lc.

We define two types of similarity measure: terminological and conceptual. The terminological similarity measure **Simt** is used to identify the similarity between a term ti and the attributes of a concept Cj, denoted by Simtj (ti, Cj).
The similarity of terms is found by using one or more distances between strings and making reference to WorldNet.
If ti is an identifier of Cj then Simtj (ti, Cj) = 1, the terminological similarity of ti with all the concepts C is defined by the following formula:

$$\text{Simt}(ti, C) = \sum_{j=1}^{n} \text{Sim}_j(ti, Cj)$$, n: concept number of C

1) Presentation of changes
To add a text, the expert of the domain has to pass through the section in the corpus summary which address to the corresponding section in the text.
The corpus summary is represented as a tree in which each node indicates a section in the corpus summary.
Three cases can be considered for the addition.

- The extracted term appears as an attribute of a concept Ci, i.e. in the lexicon of Ci (Lc (Ci)). It means that this new term is similar to one of the attributes of the concept Ci. This can be assured by calculating the similarity measure Sim (ti,C) :

1. If Sim (ti, C) <> 0, then there may be an attribute that already exists in the list of attributes of a concept Ci (Lk(C)), this is the case for "A2 Changement" (Table I)
2. If Sim (ti, C) = 0, then it is not similar to any existing attribute in the list of attributes of all concepts Lc (Ci), with i=1 .. n (n number of concept), it is the case of "A1,A3 changements" (Table I).

- The extracted term does not appear in the Lc (Ci) for all concepts, three cases may arise:

1. If this is a new unique identifier of an already existing concept. In this case, the processing can not be automated, since the new term ti must be validated as an identifier of a concept of C by the expert of the domain.
2. If the new term ti is a concept son to Ci or to one of his descendants in Cd concepts (direct or not direct).

TABLE. I. CHANGES ANALYSIS: ADDED CASE

Changement A1	
Syntax	***Semantics***
Add_C (ti)	*Add as a class in the Ontology*
Pre-condition	***Post-condition***
$t_i \in T_{ajouté}$; $t_i \notin$ Lc $C_i \in$ Lc ; Ti $\notin$ Hc $Ci \notin$ Hc; $Sim(t_i, C_i) \geq 1$	$t_i \in C_i$; ti $\in$ Lc(ci) ; ci $\in$ Lc $nboc(t_i,c_i) = 1$ (ti,ci) $\in$ Hc
Changement A2	
Syntax	***Semantics***
INC (ti,ci)	*Increment the number of occurrences of Ti in the class Ci*
Changement A3	
Syntax	***Semantics***
Add (ti,C)	*Add as a new attribute of a given concept*
Pre-condition	***Post-condition***

Ti ∉ Lc(Ci) ; t_i ∈ T_{Add}	t_i ∈ Lc (Ci)
Pre-condition	***Post-condition***
t_i ∈ Lc(C_i) sim (ti,ci)<>0[1]	Nboc (t_i,c_i)=nboc (t_i,c_i)+1
Changement A4	
Syntax	***Semantics***
Add (ti,Lr)	*Add as a new relationship in the database of relationships terminology*
Pre-condition	***Post-condition***
Ti ∉ Lr[2]; t_i ∈ T, (t_i,c_i) ∉ H (ti,rang,dom) ∉ Lr	(t_i,rang,domain) ∈ Lr
Changement A5	
Syntax	***Semantics***
INC (ti)	*Increment the number of occurrence if it already exists as a terminological relation.*
Pre-condition	***Post-condition***
t_i ∈ Lr R ∈ Rang (ti) ; D ∈ Domain(Ti)	Nboc (t_i,c_i)=nboc (t_i)+1

Before concluding the case study of addition, we propose the following addition algorithm which uses a similarity measure. This classification algorithm supports the study made by Table II.

3.2.2. Case 2: The deletion

The deletion in our study may occur in some cases, e.g. : a symptom is no a sign of the disease, an analysis is no necessary, a treatment causes side effects, ...
In these cases, the removal of unnecessary parts of texts is needed so that the corpus is still consistent with the studied domain.
This deletion affects the consistency of the Ontology with actual knowledge domain. The process of deletion consists of three steps:

1. Deletion of the text.
2. Construction of two bases: terms and relations bases.
3. Execution of an algorithm for updating the Ontology.

The first step uses the corpus summary, in order to quickly accessing to the fragment which must be deleted.
The second step uses the module MOCO described above to build the list of terms and of relations to be deleted.

1) Presentation of the changements

The Changes which may be occurred in this case of the evolution are decomposed into two parts:
Remove terms

[1] Lc : the lexicon of the concept Ci
[2] Lr : lexique de la relation, ou la liste des relations terminologique

The deletion of a term in the text can affect the structure or lexicon of the Ontology in question. The term which must be removed is either an instance (attribute of a concept), or a concept. The deletion of a concept leads to the updating of the hierarchy H of the Ontology O and resolve all problems of attributes inheritance.

TABLE III CHANGE ANALYSIS - DELETION CASE.

Changement S1	
Syntax *Dell_indiv(ti)*	***Semantics*** *If it is included in concepts then it is a son*
Precondition $t_i \in T_{Dell}$; Ti $\in$ Lc(ci) $\lvert Lc(ci) \rvert > 1$; Nbocc (ti)=1	***Post-condition*** Ti $\notin$ Lc(ci)
Changement S2	
Syntax *Decr_indiv(ti)*	***Semantics*** *Just decrement the number of occurrence*
Precondition $t_i \in T_{Dell}$, ti $\in$ Lc (ci) ; nbocc (ti)>1	***Post-condition*** $t_i \in$ Lc (ci) nbocc (ti)=nboc(ti)-1
Changement S3	
Syntax Dell_concept(ti)	***Semantics*** Delete the concept
Precondition Ti $\in$ TDell ; Ti $\in$ Lc(ci) Lc(ci)=1;Nboc(ti)=1 ; Ci $\in$ H	***Post-condition*** $t_i \notin$ Lc (ci) *Ci* $\notin$ *H*
Changement S4	
Syntax *Dell_rel(domain,rang,ti)*	***Semantics*** *Remove the relationship as an attribute of a concept if it has the same identifier.*
Precondition $t_i \in T_{Dell}$, (domain, rang, ti) $\in$ Lr	***Post-condition*** $t_i \in$ Lc (ci) *Si* nbocc (ti)= 1 *Sinon* nboc(ti)- 1

2) Presentation of the deletion algorithm

Considering T as the initial corpus presented by the vector T = (t1, ..., tn) where ti: the terms of the corpus.

Considering T2 is the text to be deleted, we propose the following algorithm to manage the Ontology evolution.

The algorithm uses a measure that we called the *confidence measure*, defined as follows:

$C = \frac{Lc \cap Bt}{Bt}$ where :

Lc: the set of all individuals of the Ontology O, i.e. all terms of the initial corpus.

Bt: the set of all extracted terms from deleted text.

If (C = 1), then it is a complete deletion of all the corpus, so all the Ontology.

To manage the deletion of the ontological entities we have used an algorithm of deletion.

4 The architecture of OntoEvol

This architecture is a recursive process, based on a corpus and an initial Ontology. It manages the evolution in order to obtain a different version of the corpus and the Ontology. These versions will be used as the inputs for the next evolution. The architecture of the methodology OntoEvol described in Fig.5 contains:

• Module of corpus processing (MOCO), dedicated to the first phase in the evolution PE (changement representation), its role is to deal with changes on the corpus of text and provide the results of these changements in the form of: term base, and the relationship base.

• Module of the Ontology processing (MOTO), dedicated to the second phase in the evolution PE (semantic of the changes), its role is to deal with changes to be made on the Ontology starting from the results provided by the module (MOCO), by using addition and deletion algorithms.

As regards the phase of the changement propagation, in our case study, it is really limited, because the only artefacts depending to the Ontology is the corpus of the text.

The implementation phase is provided by the tool S-OntoEvol (described below) to automate the management of evolution.

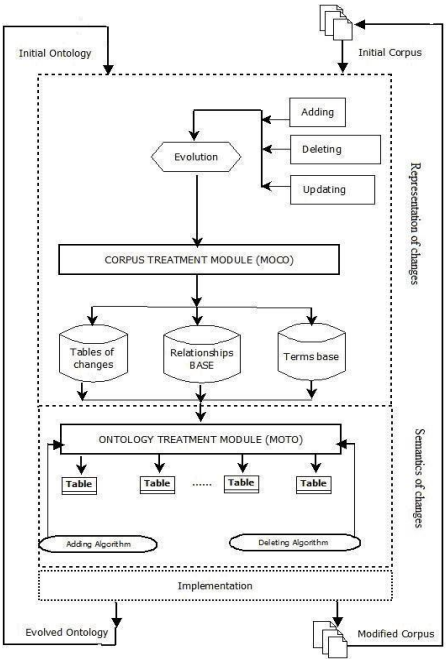


Fig. 5. The general architecture of OntoEvol

5 Conclusion

In this article we have presented our equipped methodology OntoEvol which manages the evolution of an Ontology from a corpus of texts.

After a study of the art state, we didn't find a methodology that supports both textual knowledge and Ontology. We tried to resume the process of evolution described in the methodology of the AIFB, and we relied on the principle of versioning studied in the IMSE methodology to manage only Ontology versions which we worked with: VO (initial) and V (supporting evolution).
This work involves analysis techniques of the texts in the domain of TALN. We proposed a prototype combining the principles of OntoEvol.
Our experiment was done on the medical domain of disease Cancer, describes the corpus and Ontology Cancer "Chifa". We have not treated the case where the Ontology is dependent with other artifacts, as in our case study, it is considered as the corpus of text. This can open multiple perspectives of research.
- How to manage the evolution of a dependent Ontology with other ontologies from a corpus of text.
- Can we manage the evolution of a dependent Ontology from a corpus of texts when they are distributed among multiple users?

References

[1] A. Maedche, S. Staab, N. Stojanovic, R. Studer, and Y. Sure, Semantic Portal – The SEAL approach, *Creating the Semantic Web*. Edition MIT Press, MA, Cambridge, 2011

[2] A.Maedche, B.Motik, L.Stanjanovic, Managing Multiple and Distributed Ontologies in the Semantic Web, *VLDB Journal* - Special Issue on Semantic Web, 12, 286-302, 2003.

[3] A.Maedche, B.Motik, L.Stojanovic, Managing Multiple and Distributed Ontologies in the Semantic Web, *VLDB Journal* - Special Issue on Semantic Web, 12, pp: 286-302, 2003.

[4] C. Jouis, *Contribution à la conceptualisation et à la modélisation des connaissances à partir d'une analyse linguistique de textes. Réalisation d'un prototype : le système SEEK*, Doctoral thesis in history (EHESS), University of Paris, 1993.

[5] D.E.Oliver, Y.Shahar, M.Musen, E.H Shortliffe, Representation of change in controlled medical terminologies. *Journal of Artificial Intelligence in Medicine (AI)*, Vol 15(1), pp. 53–76, 1999.

[6] D.McGuinness, "*Conceptual Modeling for Distributed Ontology Environments",* Proceedings of International Conference on Computational Science (ICCS), Germany, (2000).

[7] D.Rogozan, *Gestion de l'évolution des ontologies : méthodes et outils pour un référencement sémantique évolutif fondé sur une analyse des changements entre versions d'ontologie,* Doctoral thesis, Québec University, 2008.
[8] F.Rousselot, *Extracting concepts and relations from corpora*, Proceedings of corpus oriented semantic analysis workshop of ECAI, Budapest, Hungary, August, 1996.
[9] H.Assadi, D.Bourigault, Analyse syntaxique et statistique pour la construction d'ontologies à partir de textes, 2000, *Ingénierie des connaissances. Tendances actuelles et nouveaux défis*. Editions Eyrolles/France Telecom, Paris, pp. 243-255, 2000.
[10] J.Heflin, J.Hendler, *"Dynamic Ontology on the Web", Proceedings of the Seventeenth National Conference on Artificial Intelligence AAAI/MIT*, pp. 443-449. Canada, 2010.
[11] J.Heflin, J.Hendler, S.Luke, "Coping with Changing Ontologies in a Distributed Environment Ontology Management", Workshop of Association for the Advancement of Artificial Intelligence (AAAI), pp 74-79. Florida,1999
[12] L.Stanjanovic, *Methods and Tools for Ontology Evolution*, Doctoral Thesis, University of Karlsruhe, Germany, 2004.
[13] L.Stanjanovic, N.Stojanovic, S.Handschuh, "*Evolution of the Metadata in the Ontology based Knowledge Management Systems", Conference on Experience Management (EM'02)*, Germany, 2002.
[14] L.Stojanovic, B.Motik, *Ontology Evolution within Ontology Editors. Conf. Knowledge Acquisition, Modelling and Management (EKAW)*, Siguenza, Spain, 2012.
[15] M.Chagnoux, N.Hernandez, N. Aussenac, *From texts to ontologies : non taxonomical relation extraction*, Proceedings of « Journées Francophones sur les Ontologies », JFO, pp.126-134, ISBN :978-1-60558-373-0, Lyon, 2008.
[16] M.KLEIN, C*hange Management for Distributed Ontologies,* Doctoral Thesis of VRIJE University, Amsterdam, 2004.
[17] M.Klein, N.Noy, A component-based framework for the Ontology evolution, Workshop on Ontologies and Distributed Systems, IJCAI-03, Acapulco, Mexico, 2003.
[18] N.Aussenac-Gilles, S.Despres, S.Szulman, The terminae method and platform for Ontology engineering from texts, *IOS Press*, pp. 199-223, Janvier 2008.
[19] N.Aussenac-Gilles. A.Condamines, F. Sedes, Evolution et maintenance des ressources termino-ontologiques : une question à approfondir. *Information-Interaction-Intelligence,Cépaduès Editions*, Numéro spécial Ressources termino-ontologiques, Vol. Horssérie, p. 7-14, 2006.
[20] N.Noy, M.Klein, *Tracking Complex Changes during Ontology Evolution*. Proceedings of the Second International Semantic Web Conference (ISWC-2003) Poster Session, Florida, 2003.
[21] N.Taleb , Sellami M, Simonet M, *knowledge acquisition for the construction of an evolving Ontology: application to augmented surgery*, Proceedings of World Academy of Science , Engineering and Technology (WASET), Vol 40, pp.631-640, ISSN: 2070-3740 , Italy, April 2009.
[22] N.Taleb , Sellami M, Simonet M, *The Management of the Knowledge Evolution by Using Text Mining Techniques*, Proceedings of 5th International Conference on

Semantic Systems (ISEMANTICS), pp.663-669 Graz, Austria, September 2 – 4, 2009.

[23] N.Taleb, M. Sellami, *Acquisition des connaissances pour la construction d'un système à base de connaissances*, Thesis, University of Annaba, Algeria, 2003.

[24] P.Cimiano . (2006). Ontology Learning and Population from Text: Algorithms, Evaluation and Applications, Springer Verlag, 2006.

[25] P.Luong, *Gestion de l'évolution d'un web sémantique d'entreprise*, Doctoral thesis, University of Paris, 2007.

[26] P.Séguéla, *Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques*, Doctoral thesis, Paul Sabatier University, Toulouse, 2011.

Clinical Decision Support Systems to Prevent Domestic Accidents

Baya Naouel Barigou, Fatiha Barigou, Baghdad Atmani
Equipe de recherche « Simulation, Intégration et Fouille de données »
Laboratoire d'informatique d'Oran
Université d'Oran
barigounaouel@gmail.com
fatbarigou@gmail.com
atmani.baghdad@gmail.com

Abstract: the focus of this paper is talking about domestic accidents in Algeria, and solutions that decision-support systems will provide to reduce mortality and handicap. In the case of prevention of domestic accidents, decision support systems have evolved in Western countries. These systems provide physicians with high quality decisions that will help them to reduce errors in diagnosis and accelerate patient care. As Algeria is suffering from damage of such accidents, it is time to apply such systems for medical prevention. To motivate the interest of decision support systems, we give in this paper the state of the art of clinical decision support systems.

1 Introduction

Domestic accidents cause thousands of death and handicap in Algeria each year. The statistics are alarming; referring to the figures from the Ministry of Health, Population and Hospital Reform, about 340000 children were victims of domestic accidents in 2010. And according to a survey conducted by the civil protection services 99% of accidents occur due to thermal burns in 2011. In fact, it is considerably higher, especially in the poorest regions[1].

In the city of Oran and according to the Service Management Health and Population, at least 5832 domestic accidents affecting children under 15 years were reported during the first semester of the year 2012. These cases which are taken care of at different sanitary structures in Oran accounted for 3% of all children consulted or treated during the same period which is equivalent to 168797 cases, confirmed the same source and (SPDSP, 2012).

Common causes of childhood injuries include: falls, fires or burns, poisoning, and suffocation. These accidents are occurring in the home, including the kitchen, the balcony or the surroundings of the house during the holidays and the summer sun. Children less than 15 years are the most affected. Injuries in different forms come first with 2123 cases, followed by falls (1518 cases), skin burns (820 cases) and finally the accidents related to the ingestion of abrasives which are estimated to 99 cases (SPDSP, 2012).

Lack of safe storage for chemicals, complexity of the child's environment, and smallness of the housing and the lack of a culture of health in many families all expose children to

[1] http://www.nessnews.com/choc-buzz/algerie-340-000-enfants-victimes-d-accidents-domestiques-1109

higher levels of risk. Also, the lack of accessible, affordable emergency health services increases the number of deaths and long-term deficiencies.

We believe that measures must be made to face this epidemic of accidents, and such measures must not rely only in terms of health, but must use an interdisciplinary mobilization of different specialists: doctors, computer scientists, architects, engineers, parents, etc. indeed, we need qualified people to implement all the actions required to solve this emerging public health problem.

In industrialized countries, domestic accidents are a leading cause of mortality and morbidity. Generally, they represent the 5th cause of mortality among the children less than 15 years. For example, in France, domestic accidents are responsible for about 50% of accidental deaths of children under 14 years[2]. For this reason, these countries have invested in programs and systems for medical prevention (e.g. clinical decision support systems). One of the solutions that have been widely studied and approved is by using decision support systems. In fact, decision support systems have in recent years experienced a great evolution in medical domain, precisely in Europe and the United States, thanks to the quality provided by these decisional systems to clinicians who have reduce the number of deaths and handicap among injured children.

It is now time that; Algeria follows the same way to minimize the number of emerging mortality and handicap. We must highlight the urgent need for solutions. In fact, there is a need for more research that contributes to effective analysis of the situation, which is most likely to contribute to awareness and to practical prevention measures.

We believe that decision support systems can contribute to support managerial judgment and to improve the effectiveness and the efficiency of the decision process.

Benefits of such systems include an increased number of alternatives examined, a better understanding of the subject, and a faster response to unexpected situations, improved communication, improved control, cost savings, more objective decision making, more effective teamwork, time savings and better use of data resources Keen (1981).

The objective of this paper is to review the state of the art in clinical decision support systems and identify guidelines for designing and implementing such a system, with a focus on domestic accidents. Based on these findings, we present possible directions for future development of a CDSS to treat domestic accidents.

2 Motivation

An accident is an event independent of the human will. It's provoked by an outside force which acts quickly and is manifested by corporal or mortal damage. Domestic accidents which commonly occur in children represent a real public health problem. These accidents have serious consequences in terms of mortality and handicap, not forget the psychological impact on the child and his family. According to studies from the direction of prevention of the Algerian Ministry of Health[3], falls, burns and poisoning are the most common in young children and occurs most frequently at home.

In Algeria, the phenomenon of domestic accidents is not quantified because of the difficulty for data gathering. Accidents, however, as elsewhere in the world, are a major

[2] http://www.rmc.fr/editorial/316678/accidents-domestiques-cause-nationale-2013 (February 1st 2013)
[3] http://www.sante.dz/Dossiers/direction-prevention/accidents-domestiques.PDF (February 10th 2013)

source of morbidity and mortality. That their care is too late and insufficient aggravates the situation.

Domestic accidents throughout Algeria country have reached major epidemic proportions. They form one of the main causes of children death and have become more common in recent years (see figure 1). But the medical profession and the general public are still slow to recognize the facts and their effects. For example, the study of cause and prevention measures of domestic accidents is still insufficient and inadequate; there is a lack of data, statistics on mortality and handicap caused by this type of accident. And therefore, all this formulate the goal of this research.

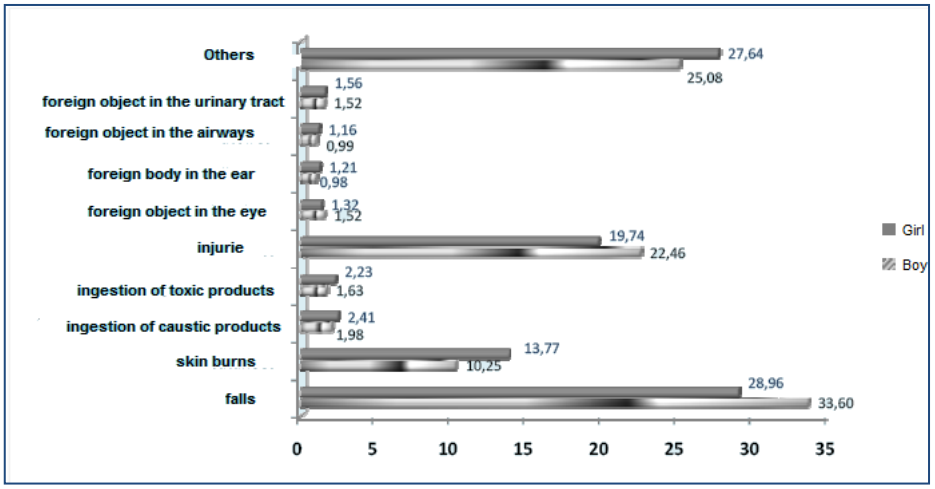


Fig.1- distribution of domestic accidents of children aged 0-15 years by accident type and gender (INSP, 2009)

3 Clinical decision support system

In this section we first, define the concept of clinical decision support systems then we discuss different types of these systems.

3.1 Definition

Clinical Decision Support Systems (CDSS) are computer systems dedicated to the decision making task, i.e. to support clinicians in practice (Shortliffe and Cimino, 2006). They use relevant knowledge, rules within a knowledge base and relevant patient and clinical data to improve clinical decision making on topics like preventive, acute and chronic care, diagnostics, specific test ordering, prescribing practices (Pearson et al., 2009).

Typically, CDSS are of two main types: Knowledge-Based and non-Knowledge-Based. The most frequently used type in Health Care settings today is the Knowledge-Based CDSS, also known as “Expert Systems” (Coiera 2003). Most knowledge-based systems are comprised of three parts: the knowledge base, the inference engine, and the user interface

Berner and Lande (2006). Systems that do not use a knowledge base use machine learning, recognizing patterns in data.

CDSS have been developed to assist with a variety of decisions. They are designed to support various medical functions such as:

- alerting (e.g. highlighting abnormal values),
- reminding (e.g. to schedule a surgery),
- critiquing (e.g. reviewing a prescription),
- interpreting (e.g. electrocardiogram interpretation),
- predicting (e.g. risk of mortality),
- diagnosing (e.g. producing a differential diagnosis),
- assisting (e.g. in selection of antibiotics), and
- suggesting (e.g. generating suggestions for adjusting medical equipment).

The table 1 provides examples of CDSS interventions by target area of care. We can see that CDDS interventions can assist health care providers at different stages in the care process, that is, from preventive care through diagnosis and treatment, all the way to monitoring and follow-up.

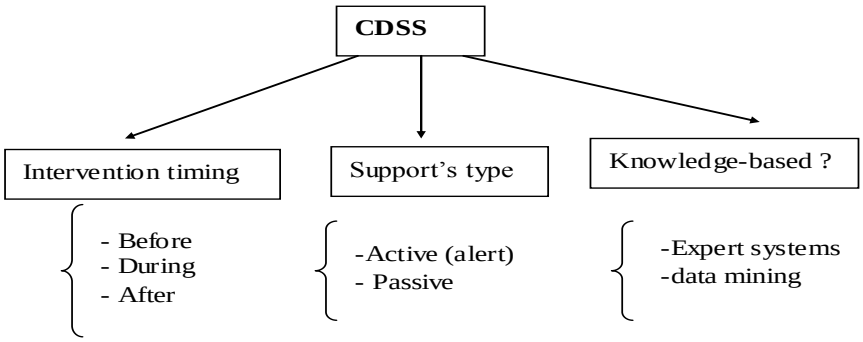
Target area of care	Example
Preventive care	Immunization, disease management, guidelines for secondary prevention
Diagnosis	Suggestions for possible diagnosis that match a patient's signs and symptoms
Planning or implementing treatment	Treatment guidelines for specific diagnosis, drug dosage recommendations, alerts for drug-drug interactions
Follow up management	Corollary orders, reminders of drug adverse event monitoring
Hospital provider efficiency	Care plans to minimize length of stay, order sets
Cost reduction and improved patient convenience	Duplicate testing alerts, drug formulary guidelines

TAB. 1 – Examples of CDS Interventions by Target Area of Care[4]

3.2 Types of clinical decision systems

There have been diverse descriptions of the types of CDSS systems and their characteristics. Many of the early CDSS are systems expert-based which are used by the clinicians for diagnosis and medication selection. As illustrated in table 2, there are a variety of CDSS. We can describe those using different dimensions. According to the timing at which they provide support, and how active or passive the support is. CDSS also differ in which they are knowledge-based systems or non knowledge-based systems.

[4] http://www.philblock.info/hitkb/c/clinical_decision_support_systems_part1.html



TAB.2- CDSS classification

3.2.1 Knowledge-based

Some of CDSS reasoning engines are Rule-based systems. A rule-based system uses different expert knowledge bases in form of expressions that can be evaluated as IF-THEN rules. They can have clinical knowledge about a specially defined task, or can even be able to work with case base reasoning. Sometimes, the knowledge based is used with variance management to execute patient care process and provide high quality health care services dynamically. This knowledge based management system is implemented using the object oriented analysis Ye and Tong (2009).

This approach was first used in the MYCIN (Shortliffe, 1976), with the goal of choosing appropriate antimicrobial therapy for a patient.

3.2.2 Non knowledge-based

These systems instead, used a form of artificial intelligence, called machine learning. They are then further divided into two main categories:

- Neural networks to derive the relationship between diagnosis and symptoms, neural networks use the nodes and weighted connections. But these systems fail to explain the reason for using the data in a particular way. Therefore, its reliability and accountability can be a reason. It has been observed that the self organizing process of training the neural network in which it isn't given any priory information about the categories it is required to identify, is capable of extracting relevant information from input data in order to generate clusters correspond to class Y. Kim, Y. Cho (1995). In identifying pain in infant child, neural networks extract the features from infant cry and are fed them into recognition module. The accuracy rate of this system under different parameters reported as 57% to 76,2% Y. Abdulaziz, S. Mumtazeh (2010). The neural networks are also very important especially in complex multi-variable systems to avoid costly medical treatment and for diagnosis of pain A. Soriano Paya and all (2006).
- Genetic algorithms: They are based on evolutionary process. Selection algorithm evaluates components of solutions to a problem. Solution that comes on top are recombined and the process runs again until a proper solution is observed. The

genetic system goes through an iterative procedure to produce the purpose the best solution of a problem.

3.2.3 Alerts and Reminders

The benefits to alerts and reminders include providing immediate notification of errors and risks related to new data or orders entered by clinical information system, or passage of the time interval during which a critical event should produce, help to impose standards of care. There are two subtypes[5]:

- Alerts to prevent potential omission, commission errors or risks, for example potential omission error detection, such as checking for a result from a follow-up test that is indicated after a medication is given.
- Alerts to promote best care, for example, disease management, alert for needed therapeutic intervention based on guidelines or evidence and patient-specific factors.

3.2.4 Clinical practice guidelines

Clinical practice guidelines are central to determining the care plan for a patient, and are considered to be the preferred process for care.

Clinical practice guidelines are a foundational part of the knowledge base. The Quality Assurance Project (QAP), funded by the U.S. Agency for International Development, includes a glossary of useful terms. Within the glossary is the following definition for clinical practice guidelines: "*A set of systematically developed statements, usually based on scientific evidence, to assist practitioners and patient decision making about appropriate healthcare for specific clinical circumstances"*. Synonyms for clinical practice guidelines include practice guidelines, guidelines, practice parameters[6]. A similar definition is that from the National Library of Medicine (NLM). NLM defines clinical or practice guidelines as works consisting of a set of directions or principles to assist the health care practitioner with patient care decisions about appropriate diagnostic, therapeutic, or other clinical procedures for specific clinical circumstances.

4 Our proposal

Since domestic accidents in Algeria is becoming increasingly growing, and as it has said in this paper, it is time to use CDSS for our children's preventive health.

As cited in AMIA's *A Roadmap for National Action on Clinical Decision Support*, "(CDSS) provides clinicians, staff, patients or other individuals with knowledge and person-specific information, intelligently filtered or presented at appropriate times, to enhance health and health care." Thereby the growing multiplicity of CDSS and their effectiveness certified in the decision making tasks at the time and the location of care, motivated us to study CDSS and their impact on the prevention and intervention in the case of domestic accidents.

[5] http://www.philblock.info/hitkb/c/clinical_decision_support_systems_part1.html
[6] http://www.philblock.info/hitkb/c/clinical_decision_support_systems_part1.html

The problem being addressed in this research can be resumed according to the following three questions: “how to design and develop a CDSS which considers domestic accidents? How do we integrate the CDSS in the medical system? How should we use the CDSS at different stages?

- Before: as a guidelines to prevent accidents.
- During: how to intervene and take actions.
- After accident: to look after the accident victim.

Our findings are the guidelines systems will be more practical for home accidents. They will be used before, during and after the accident. We are oriented towards this type of system because they are characterized by a number of features[7]:

- Guidelines system intent is an automatic intervention: they are reminder of actions a user intends to do but should not have to remember. As one would expect, timing is a key issue.
- Guidelines intent is an on demand intervention: they provide information when a user is unsure of what to do, or a request for consultation. In this instance, it is speed and ease of access that the user is looking for. If access is too difficult or time-consuming, potential users may choose not to use the CDS.
- Intervention’s intent, to correct user’s errors and/or recommend a user change plans, could be either an automatic or on demand intervention. For an automatic intervention, the key issues are timing, autonomy, and user control

Starting at home, the place of the accident, guidelines’ CDSS aimed at parents who are the first people reacting during the accident, can help them to face the accident efficiently by giving them advices and what to do in the emergency.

At the emergency department, guidelines’ CDSS are very important to guide the medical staff to the correct decision of treatment and gain time in their intervention. This will save the accident victim and consequently reduce mortality and handicap.

5 Conclusion

Accidents at home cause many injuries and a substantial number of deaths each year. Although dramatic, most of these accidents are predictable and preventable. The fight against this type of accident involves many actors, but so far no common set of objectives, strategies or actions exist to help guide a coordinated national effort.

As computer scientists, and to lower these figures, our objective is to discuss solutions supporting clinical decision support system. The findings of this study will be used in the development of the future clinical decision support system for treating domestic accidents. This study was an important first step towards designing the system. We are convinced that CDDS can add value to the existing medical care by providing advice at the point of decision making.

The preventive approach will be based on 3 axes:

- A good understanding of the epidemiological situation.
- An efficient strategy of information, education and communication that aims to change attitudes and behaviors of the population.

[7] http://www.philblock.info/hitkb/c/clinical_decision_support_systems_part2.html

- A revision concerning the regulation and safety of the products causing injury seen in children.

Understanding the circumstances and details of child accidents at home would help us identify key contributors to accident and promising preventive and care strategies.

References

Abdulaziz, Y. Mumtazeh, S. (2010) "Infant Cry Recognition System: A Comparison of System Performance based on Mel Frequency and Linear Prediction Campestral Coefficients.", *International Retrieval & Knowledge Management, IEEE, pages 260 - 263.*

Berner, E, and Lande, T.J. (2006) Overview of Clinical Decision Support Systems (updated version of Chapter 36 in Ball MJ, Weaver C, Kiel J (eds). Healthcare Information Management Systems, 3rd Edition, vol. 6, ch. updated ve, pp. 463-477

Coiera, E. (2003). *Guide to Health Informatics*, ARNOLD, Oxford University Press, New York

INSP. (2009). Analyse des données sur les accidents domestiques de l'enfant 0-15 ans. (2009), Institut national de la santé publique (INSP).

Keen, P. G. W. (1981). Value analysis: Justifying decision support systems. MIS Quarterly, 5, 1.

Kim, Y. Cho, Y. (1995) Correlation of Pain Severity with Thermography. *Engineering in medicine and biology Society, IEEE, pages 1699-1700.*

Pearson, S., Moxey, A. et al. (2009). "Do computerised clinical decision support systemsfor prescribing change practice? A systematic review of the literature (1990-2007)." BMC Health Services Research 9(1): 154.

Shortliffe, E. H. (1976). Computer-based medical consultations, MYCIN. New York, Elsevier.

Shortliffe, E., Cimino, J. (2006). Biomedical informatics: computer applications in health care and biomedicine. 3rd edition, New York, published by Springer.

Soriano Paya P., A. Fernandez, D. Gill Mendez, D. Hernandez, C. (2006). Development of an artificial neural network for helping to diagnose diseases in urology. *Bio-Inspired Model of Network, Information and Computing Systems, IEEE, pages 1-4*

SPDSP. (2012). Service de prévention de la direction de la santé et de la population, par Agence (24/07/2012) pages 15-48.

Ye, Y., Tong, S.J. (2009) "A Knowledge-Based Variance Management System for Supporting the Implementation of Clinical Pathways.", *Management and Service Science, IEEE, pages 1 -4*

Système d'information voyageurs global basé sur la décomposition de Voronoï pour l'aide à la mobilité

Zakaria Bendaoud*, Karim Bouamrane*

*Université d'Oran, Algérie
cherad@hotmail.com
kbouamranedz@yahoo.fr

Résumé. Le secteur des transports connait une expansion fulgurante, les entreprises de transports en commun ne cessent d'innover afin de venir aux besoins des voyageurs. Néanmoins le nombre important de ces opérateurs de transports poussent parfois les voyageurs à des confusions pour trouver et mémoriser toutes les informations et itinéraires possibles. Dans ce papier nous proposons un système d'informations voyageurs global en se basant sur les systèmes multi-agents afin d'éviter à l'utilisateur de consulter plusieurs sites web.

1 Introduction

Les récents développements du web ont apporté un atout considérable dans le quotidien des gens. Dans le domaine des transports, les entreprises de transports en commun (ETC) exploitent ces développements en mettant leur système d'informations à la disposition des voyageurs via des applications ou des interfaces web dans l'optique d'assurer leur confort.

Un système d'informations voyageurs (SIV) peut être de deux types : uni-modal ou multimodal. Le premier type permet de gérer un seul mode de transport alors que le deuxième inclue au minimum deux modes.

Avec la forte croissance qu'est entrain de connaitre le secteur des transports, le voyageur doit avoir une forte connaissance sur les itinéraires des lignes, les sites web et les pôles d'échanges proposés par chaque ETC. En effet, sur un territoire géographique étendu, il est très probable que le voyageur ait besoin de composer entre plusieurs entreprises pour avoir l'information recherchée.

Afin d'éviter aux voyageurs ces recherches manuelles et d'assurer leurs conforts nous proposons dans ce papier d'intégrer plusieurs systèmes d'informations voyageurs (monomodaux et multimodaux) dans un système d'information global. Nous proposons une décomposition fictive du réseau global afin d'alléger les processus de recherches en invoquant qu'une partie de ce super réseau. Il est important de noter que la majorité des SIV existants ne prennent pas en compte la personnalisation de l'information. Pourtant une information personnalisée est nettement plus pertinente qu'une information présentée telle qu'elle est stockée, elle encourage clairement le voyageur à utiliser les transports en commun aux dépens de la voiture personnelle (Petit-roze et all., 2004). Pour cela, nous proposons de personnaliser les réponses selon le profil du voyageur en utilisant des mesures de similarités.

L'objectif principal de ce travail est de fournir aux voyageurs l'information personnalisée de manière transparente. La structure du document est la suivante : la section 2 présente des connaissances et concepts de base, quelques recherches et projets réalisés sont présentés dans la section 3, la section 4 présente notre contribution, le modèle et les algorithmes que nous proposons sont présentés dans la section 5, enfin la section 6 concerne les conclusions et les perspectives.

2 Concepts généraux

Afin de modéliser un système d'information global, il est primordial de prendre en compte deux aspects très importants :

2.1 Le type de l'information multimodale

L'information multimodale peut être définie de plusieurs manières néanmoins la définition suivante semble être la plus complète de la littérature : « La fonction essentielle d'un système d'information multimodale est de fournir à l'usager des transports toute l'information nécessaire à la réalisation de son voyage. Cette information vise à réduire l'incertitude des usagers sur les itinéraires, les modes de déplacement envisageables, la durée et le coût de ces déplacements selon le mode utilisé, les ruptures de charge éventuelles, et si possible, à orienter le comportement des usagers au bénéfice d'une utilisation optimale des infrastructures et d'une priorité aux transports collectifs. » (Danflous, 2000). A partir de cette définition, nous pouvons distinguer trois types d'informations multimodales :

- Les informations sur les itinéraires : elles concernent les coûts et les horaires des trajets.
- Les informations en ligne : elles ont pour but de tenir le voyageur au courant et en temps réel des perturbations qui peuvent se produire pendant le trajet, cette information est diffusée via un media mobile tels que les cellulaires.
- Les informations touristiques (statiques) : elles concernent les informations sur les monuments ou sur les événements culturels que le voyageur peut croiser durant son trajet.

L'information multimodale dépasse donc largement le cadre des calculs des itinéraires. Bien que les voyageurs souhaitent plus d'informations sur leur trajet la majorité des SIM ne traitent que l'information multimodale sur les itinéraires.

2.2 L'intégration des données

Pour composer l'information multimodale demandée par le voyageur, il faut intégrer les données entre les différents systèmes d'informations, trois stratégies sont possibles :

- Une architecture centralisée : consiste à créer une base de données globale entre tous les systèmes d'informations disponibles.

- Une architecture client-serveur : consiste à créer un serveur global sur lequel tous les systèmes d'informations multimodaux existants se connectent comme étant des clients.
- Une architecture orientée agents mobiles : basée sur les systèmes multi-agents (SMA), elle consiste à initier un agent mobile à partir d'un nœud hôte. Cet agent mobile a la capacité de migrer d'un nœud vers un autre. Les nœuds représentent les systèmes d'informations qui appartiennent au système d'aide à la mobilité globale. L'agent mobile se déplace donc d'un système d'information vers un autre pour récupérer l'information recherchée. La figure 1 montre les différences entre l'architecture client serveur et l'architecture orientée agents mobiles.

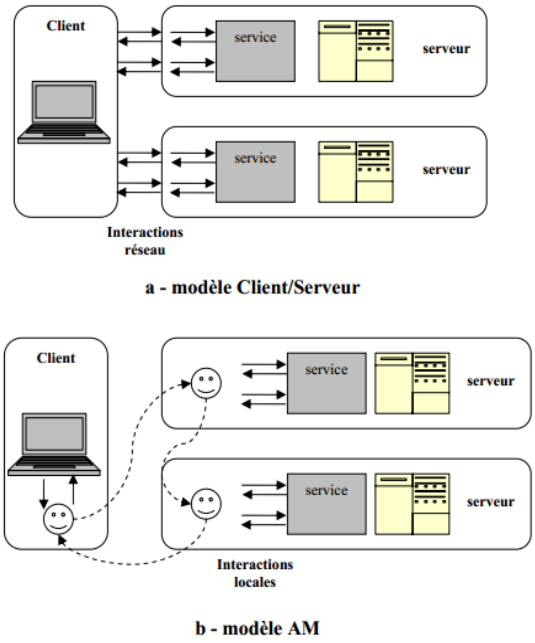


FIG. 1 - *Différences entre le modèle Client/serveur et le modèle Agent Mobile (Zgaya, 2007)*

L'avantage des deux dernières stratégies est que les données restent chez leur propriétaire. L'intégration de nouveaux systèmes d'informations est plus aisée car le système est ouvert, chaque opérateur gère son système d'information indépendamment des autres systèmes. De plus, ces architectures permettent une meilleure tolérance aux pannes dans le cas ou l'un des systèmes d'information devient indispensable.

3 Etat de l'art

Plusieurs travaux ont été effectués afin de mettre en place des systèmes d'informations voyageurs selon les besoins utilisateurs. Carlier, Fiorenzo, Catalano, Lindveld et Bovy ont

proposé une architecture qui inclue plusieurs composantes qui permettent d'optimiser le choix des routes selon le profil de chaque voyageur (Carlier, 2003). Petit-rose Anli, Strugeon, Abed, Uster et Kolski, ont proposé un service personnalisé pour l'aide à la mobilité multimodale, l'idée est d'aider les usagers des transports en commun à prendre une décision dans la planification de leur voyage tout en prenant en considération leurs besoins (Petit-roze et all., 2004). Beele a développé un assistant intelligent pour les voyages via les modes de transports urbains (Beele, 2004). Zidi a présenté un système interactif pour l'aide aux déplacements en mode normal et en mode dégradé, cela vise principalement à minimiser le temps d'attente des voyageurs (Zidi, 2004). Zgaya a proposé un système multi-agent pour la recherche et la composition des services liés au transport multimodal entre plusieurs opérateurs concurrents (Zgaya, 2007). Kamoun a proposé un système d'informations cooperatif pour la mobilité en automatisant la démarche des calculs d'itinéraires (kamoun, 2007). Grabener a présenté une approche pour le calcul d'itinéraire multi-objectifs dans un réseau de transport multimodal incluant les modes de transport publics et privés (Grabner, 2010).

Bien que la notion de multi-modalité ne date pas d'aujourd'hui, l'aide à la mobilité continue à être un axe de recherche très important. Le gouvernement allemand fut le premier à lancer son propre projet pour un système d'information multimodal globale à travers le projet DELPHI [1]en 1996, s'en est suivi le modèle anglais par le biais du projet NIT[2]. Tous les systèmes et modèles globaux cités traitent le réseau comme une seule entité, nous proposons donc une approche pour segmenter le réseau en plusieurs secteurs afin d'optimiser le traitement des requêtes et alléger les serveurs de calculs.

4 Notre Contribution

Lors de la réception d'une requête d'un voyageur, les systèmes d'informations globaux existants sélectionnent les SIV concernés par cette requête puis effectuent le traitement. Cette technique, bien qu'efficace peut s'avérer rapidement très couteuse, en effet l'invocation du système d'information d'un opérateur donné engendre la manipulation de tout son réseau même si le traitement ne concerne qu'une petite partie de ce réseau. Comme alternative, nous proposons de décomposer le réseau global en zones [3] selon la densité du trafic. Pour cela nous proposons d'utiliser la décomposition de Voronoï.

Soit S un ensemble de n sites de l'espace euclidien en dimension d. Pour chaque site p de S, la cellule Voronoï de p V(p) est l'ensemble des points de l'espace qui sont plus proches de p que de tous les autres sites de S. Le diagramme de Voronoï de V(S) est la décomposition de l'espace formée par les cellules de Voronoï des sites (Slimani, 2011). Le traitement des

[1] DELFI (Durchgängige Elektronische Fahrplanlnformation) : Information horaire électronique continue.

[2] NIT (National Integrated Transport) : plus connu sous le nom de "Public Transport Information".

[3] Sous domaines du réseau, l'union de ces zones forment le réseau global.

requêtes dépendra alors des sous-réseaux appartenant aux zones concernées et non pas de la totalité du réseau.

Le deuxième point que nous soulevons concerne le type de l'information multimodale à traiter. A travers la littérature, nous remarquons que tous les SIM existants ne traitent que les informations multimodales sur les transferts, seul (Zgaya, 2007) a traité l'information multimodale comme un service sans spécifier la nature de l'information. Dans notre travail, nous proposons de gérer les informations multimodales sur les itinéraires et les informations touristiques.

Enfin, les réponses aux requêtes des voyageurs sont personnalisées selon leurs profils afin de satisfaire au mieux leurs attentes.

La majorité des modèles étudiés et disponibles reposent sur une architecture SMA, ce qui nous permet de persévérer dans cette optique car elle répond aux besoins de l'information multimodale en termes de fiabilité et de distribution des données.

5 Vers un système d'informations voyageurs global pour l'aide à la mobilité (SIVGM)

5.1 SMA et applications dans les transports

Les systèmes multi-agents consistent à diviser un processus entre plusieurs agents. Au lieu d'avoir un seul programme qui gère la totalité du système. On divise le problème en taches, chaque agent aura pour mission de résoudre ce sous-problème. La solution globale est obtenue par l'interaction entre les différents agents. On parle alors d'intelligence artificielle distribuée (Ferber, 1995)

Plusieurs travaux dans le domaine des transports ont exploité les SMA à savoir ;

- La modélisation des réseaux de transports urbains.
- La régulation des perturbations dans les réseaux de transports.
- La personnalisation de l'information multimodale.
- L'optimisation multicritères dans les réseaux de transports.

5.2 Le système d'information voyageur globale pour l'aide à la mobilité

Afin de répondre aux requêtes des voyageurs et produire l'information multimodale, le système d'informations voyageurs global pour l'aide à la mobilité (SIVGM) doit accéder aux différents systèmes d'informations qui le composent afin d'élaborer la réponse finale. Le SIGM devient alors le médiateur entre les différents systèmes d'informations, il doit trouver les bonnes sources dans un environnement distribué, gérer l'hétérogénéité des informations et proposer les solutions aux voyageurs de manière transparente.

5.3 Organisation multi-agents pour le SIVGM

Le SIGM est organisé autour de cinq types d'agents : les agents interfaces(AI), les Agents identificateurs(AId), les agents annuaires sélectionneurs(ASS), les agents zones (AZ) et les

agents de fusion (AF). Les agents AI jouent le rôle d'interface entre le système d'un coté et les voyageurs ou les ETC qui souhaitent s'inscrire d'un autre, les agents AId identifient les requêtes d'itinéraires des requêtes touristiques et demandent le domaine de recherche afin de questionner les agents zones responsables de cette requête, les agents AAS ont une vision globale sur le réseau, ils permettent d'inscrire les nouveaux opérateurs de transports, de décomposer le réseau en zones selon le diagramme de Voronoï et de fournir les domaines de recherche en sélectionnant les zones et les pôles d'échange concernés. Dans le cas des requêtes d'itinéraires la requête peut être fragmentée suivant le nombre de zones et les stations d'échanges. Les agents AZ gèrent les zones dont ils sont responsables, chaque agent est responsable d'une zone précise et a les capacités de questionner les systèmes d'informations composant la zone sur leurs sous-réseaux qui appartiennent à cette zone. Enfin les agents AF s'occupent de la composition des réponses selon le profil des utilisateurs. La figure 2 présente l'architecture globale du système proposé.

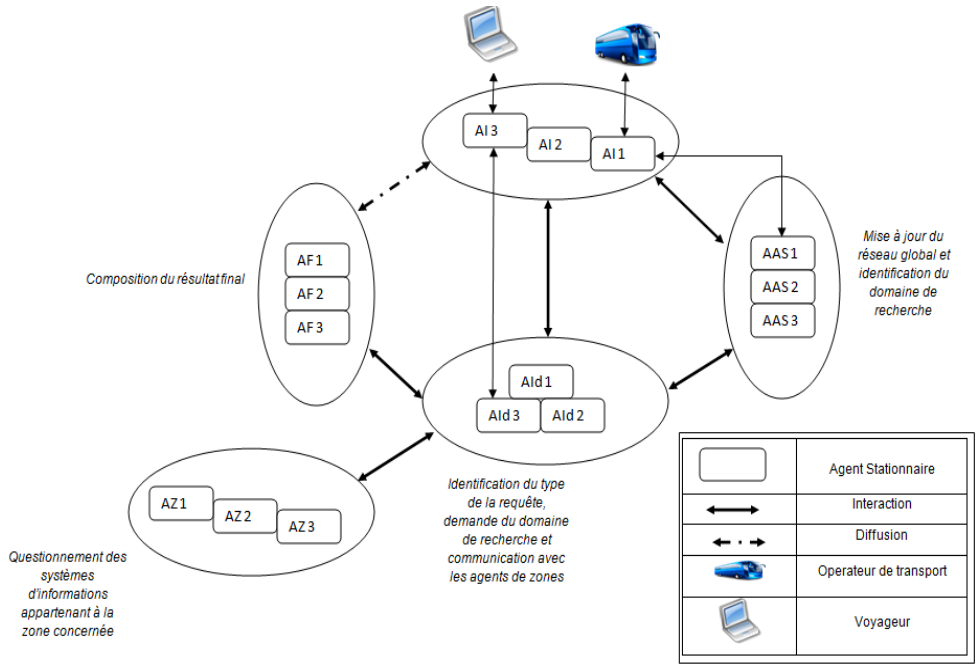


FIG. 2 - *Architecture globale du SIVGM.*

Agent Interface (AI)

Gere deux situations différentes :

- Soit il reçoit une requête d'un opérateur de transport qui veut intégrer le réseau globale et la transmet à l'ASS.
- Soit il reçoit une requête d'un utilisateur. Ce dernier peut s'inscrire afin d'adhérer au système. Dans le cas ou l'utilisateur qui émet la requête est identifiée, son profil est extrait pour que les réponses soient traitées et présentées selon ses besoins par l'agent

fusion. Dans le cas contraire, on le considère comme étant un profil standard, enfin la requête est transmise à l'agent identificateur.

La figure 3 présente le diagramme d'activité de l'agent interface.

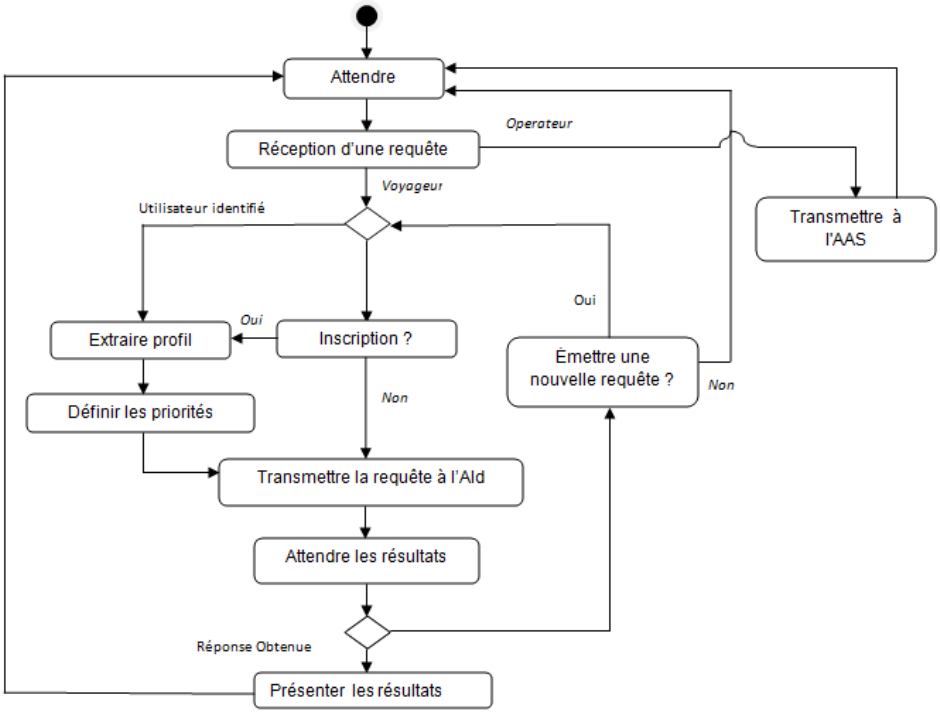


FIG. 3 - *Diagramme d'activité de l'agent interface (AI).*

Agent Annuaire Sélectionneur (ASS)

Il présente deux situations de fonctionnement :

- Il reçoit une requête d'inscription d'un nouvel opérateur à partir de l'AI, un ensemble de données sur les lignes, les stations d'échanges ainsi que son adresse réseau seront fournies. Afin de mettre le réseau global à jour, un nouveau découpage de zones est lancé par le biais de l'algorithme de décomposition de Voronoï. Comme cité plus haut, la décomposition de Voronoï consiste à créer des zones par rapport à des points fixés au préalable. Une zone est définie par l'ensemble des points les plus proches d'un des points présélectionnés au départ. Dans notre cas, les points sélectionnés au préalable représentent des stations dont le trafic est dense, pour des raisons de simplicité, nous faisons appel à des experts du domaine afin de localiser ces stations. Le nombre de zones peut donc varier selon le nombre des systèmes d'informations qui adhérent au système global. Une fois les nouvelles zones construites, une mise à jour des agents responsables des zones est lancée.

- Il reçoit une requête de la part de l'agent identificateur qui concerne le domaine de recherche pour une requête voyageur, cette requête peut être de deux types :
  - Une requête touristique : il localise le monument ou building en question pour déterminer la zone à laquelle il appartient, et par le même coup l'agent responsable.
  - Une requête d'itinéraire : il détermine à quelles zones appartient la station d'arrivé et la station de départ de cette requête. Si elles appartiennent à la même zone alors le domaine de recherche se limite à celle ci sinon un graphe d'adjacence est dressé afin de déterminer les zones qui peuvent intervenir pour satisfaire cette requête. Le graphe d'adjacence est représenté par une matrice d'adjacence des zones, les paires des stations voisines représentent l'intersection entre les différentes zones dans le tableau d'adjacence. l'algorithme djikstra du plus court chemin est lancé afin de sélectionner les zones concernées puis la requête est fragmentée en plusieurs sous requêtes selon les zones et les stations d'échange. La figure 4 présente le diagramme d'activité de l'ASS.

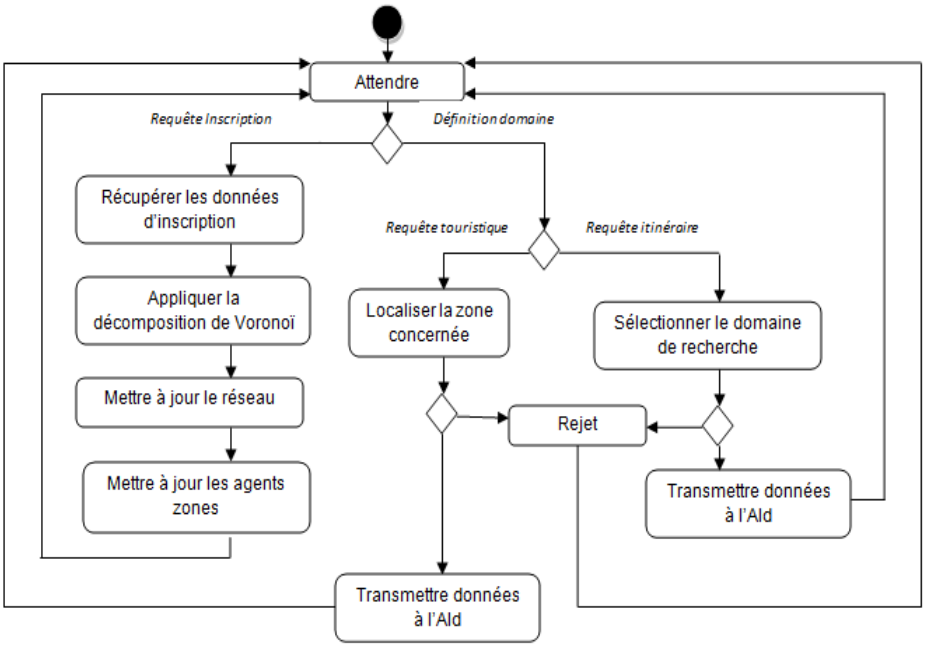


FIG. 4 - *Diagramme d'activité de l'agent annuaire sélectionneur.*

Agent identificateur (AId)

Il reçoit une requête de l'AI, cette requête peut être de deux types, soit une requête itinéraire ou une requête touristique. Si c'est une requête itinéraire, il détermine la station de départ et la station d'arrivé et demande à l'ASS de déterminer le domaine de recherche. Si

c'est une requête touristique, le nom du monument ou du building est extrait et une requête pour déterminer le domaine de recherche est envoyée à l'AAS. Si la demande aboutit, il contacte les AZ concernés pour leur demander de questionner les sous systèmes d'informations. Une fois les résultats obtenus, il les transmet à l'agent fusion. La figure 5 présente le diagramme d'activité de l'AId.

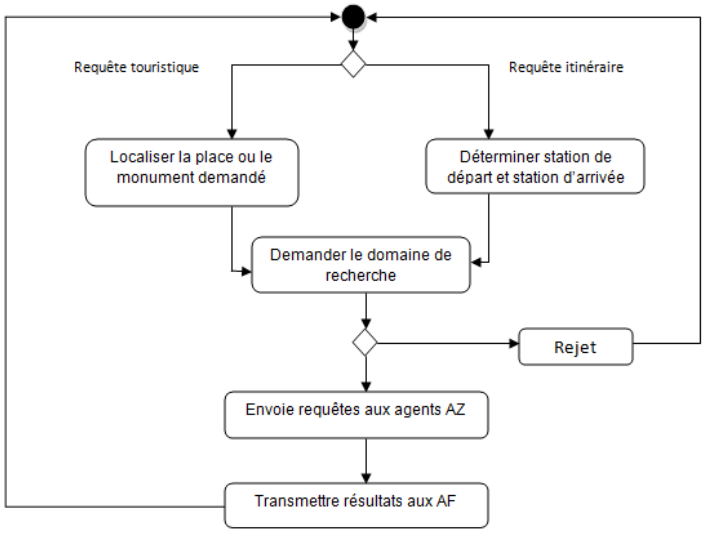


FIG. 5- D*iagramme d'activité de l'agent Identificateur.*

Agent zone

Il reçoit une requête de la part de l'agent AId pour récolter une information, la requête peut être de deux types :

- Requête touristique : l'AZ localise les stations les proches du point concerné en calculant les distances orthodromiques[4] par la formule suivante :

$$\mathrm{gc}(\delta, \lambda, \delta', \lambda') = 2R \arcsin \sqrt{\sin^2\left(\frac{\delta' - \delta}{2}\right) + \cos\delta \cdot \cos\delta' \cdot \sin^2\left(\frac{\lambda' - \lambda}{2}\right)}$$

R est le rayon de la sphère (Rayon de la Terre $\approx$ 6367000 mètres). δ est la latitude (en radians). λ est la longitude (en radians). Il sélectionne l'ensemble des opérateurs responsables des stations voisines puis les questionne afin d'avoir la réponse à la requête.

[4] Le chemin le plus court entre deux points d'une sphère, On considère la terre comme une sphère parfaite.

- Requête itinéraire : elle peut être la requête initiale lancée par le voyageur (dans le cas ou le départ et l'arrivée appartiennent a la même zone) ou une sous-requête (dans le cas ou la station de départ et la station d'arrivée n'appartiennent pas a la même zone) Dans les deux cas, elle a pour arguments la station de départ et la station d'arrivée. Si la station de départ et la station d'arrivée appartiennent à un même système d'informations alors il le contacte afin d'avoir l'itinéraire en question sinon un graphe d'adjacence des systèmes d'informations est dressé, le tableau d'adjacence possède comme intersection les pôles d'échanges. L'algorithme de yen (Yen, 1971) est lancé afin d'avoir les k plus courts chemins.

Enfin, l'ensemble des résultats est communiqué à l'AId. La figure présente le comportement de l'agent zone.

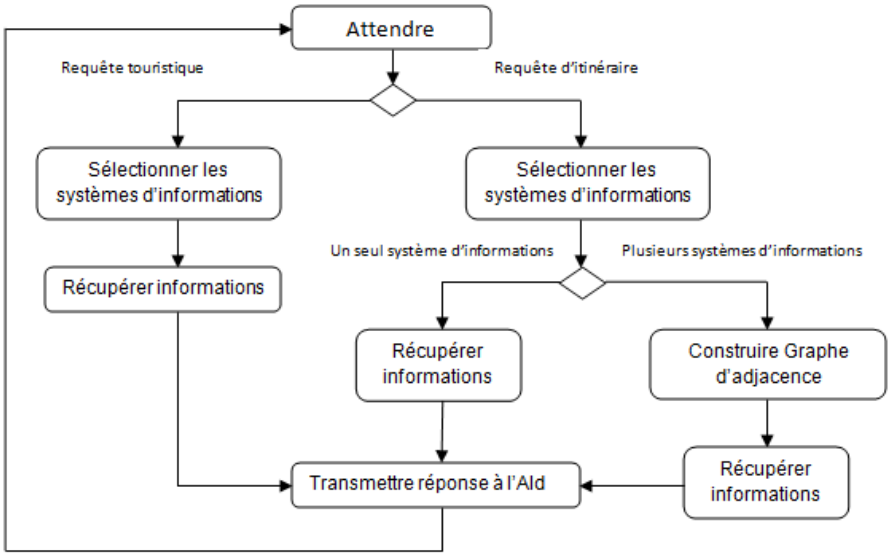


FIG. 6- *Diagramme d'activité de l'agent zone.*

Agent fusion (AF)

Cet agent a pour mission de fusionner les bouts de réponses par les agents AZ, si la requête est touristique, la réponse est de nature statique, la réponse est donc unique pour tous les profils. Si la requête concerne les itinéraires, la réponse peut varier selon le profil du voyageur. Afin de venir au mieux aux attentes des voyageurs. Nous avons considéré deux critères de choix que l'utilisateur spécifie lors de son inscription : l'âge et les préférences.

Lors de la fusion des itinéraires possibles, nous calculons la similarité entre le profil du voyageur et le profil estimé de la solution fusionné. Nous avons joint la notion de l'âge au nombre de changement lors d'un trajet, les préférences aux modes de transports utilisés. La figure 7 présente le diagramme d'activité de l'agent fusion.

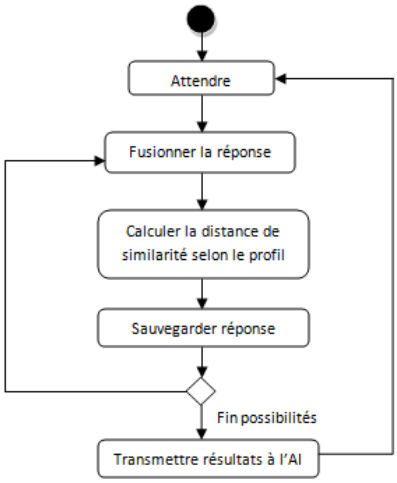


FIG. 6 -. *Diagramme d'activité de l'agent fusion.*

Afin de valider l'approche proposée, nous avons fait une simulation sur une amalgamation entre des données réelles (tableaux de marche de l'entreprise de transport de la ville d'Oran) et des données fictives. Nous avons travaillé sur une machine avec un processeur i3-380, 4giga de ram et 750giga de disque dur. Le langage utilisé pour la l'implémentation est le langage Java. Pour les systèmes multi-agents, nous avons utilisé la plateforme Jade. L'application est projetée sur un site web, la prise d'écran suivante permet de visualiser le résultat d'une requête concernant un itinéraire entre deux zones différentes.

Fig. 7. Simulation d'une requête d'itinéraire.

6 Conclusion et perspectives

Dans ce travail, nous avons présenté un système d'information global basé sur la décomposition de Voronoï afin d'aider les voyageurs dans leurs trajets, pour le moment nous considérons que la sélection des stations denses pour le lancement de l'algorithme de zonage se fait à l'aide des personnes du domaine. Nous comptons faire appel à l'analyse des données

pour automatiser cette tache. L'idée de la décomposition permet la manipulation de sous-réseaux et non du réseau complet, ce qui allège les processus de recherches. Enfin nous comptons intégrer un protocole de négociation entre les agents pour la gestion du réseau dans un état dégradé.

Références

Beele. M, (2004). *Personal intelligent travel assistant: a distributed approach*. Thèse de master, Université de technologie de Delft Hollande.

Carlier. K., Fiorenzo-Catalan. S, Lindveld. C et Bovy. P (2003). *A supernetwork approach towards multimodal travel modeling.* TRB 2003 Annual Meeting cd-rom.

Danflous. D, (2000). *Déploiement national des systèmes d'information multimodale*. Centre de documentation du CERTU. France

Ferber. J (1995). *Les systèmes multi-agents : vers une intelligence collective.* InterEditions

Grabener. T, (2010). *Calcul d'itinéraire multimodal et multiobjectif en milieu urbain.* Thèse de doctorat, université de Toulouse, France.

Kamoun. M.A, (2007). *Conception d'un système d'information pour l'aide au déplacement multimodal : Une approche multi-agents pour la recherche et la composition des itinéraires en ligne*. Thèse de doctorat, université de Lille, France

Petit-Roze. C, Anli. A, Grislin-Le Strugeon. E, Abed. M, Uster. G,(2004). *Système d'information transport personnalisée à base d'agents logiciels* Revue Génie Logiciel 70 pp. 29-38.

Slimani. H, Najjar. F, Slimani. Y,(2011). *Voronoi-Neighboring Regions Tree for Efficient Processing of Location Dependent Queries.* International Journal of Advanced Science and Technology Vol 33.

Yen. J. Y,(1971). *Finding the K shortest loopless paths in a network*. Management Science, Vol. 17, pp. 712–716

Zgaya. H, (2007). *Système Conception et optimisation distribuée d'un système d'information d'aide à la mobilité urbaine : Une approche multi-agent pour la recherche et la composition des services liés au transport*. Thèse de doctorat, université de Lille, France

Zidi. K, (2006). *Système Interactif d'aide au Déplacement Multimodal*. Thèse de doctorat, université de Lille, France

Summary

The transport sector knows a rapid expansion, and transportation companies are constantly innovating to come to the needs of travelers. However, the important number of these transport operators sometimes lead to confusion travelers to find and store all informations and possible routes, in this paper we propose a global passenger information based on multi agent systems, in order to prevent the user to view multiple websites

Enterprise Organization Assessment through Structural Analysis Framework

Azedine BOULMAKOUL, Zineb BESRI

LIM/IDS Lab. Computer Sciences Department, Mohammedia Faculty of Sciences and Technology, B.P. 146 Mohammedia, Morocco
azedine.boulmakoul@gmail.com, z.besri@gmail.com

Abstract. It is a continuous challenge for IT leaders and strategists to exploit the opportunities for optimally improving the enterprise's organization by establishing an agile and adaptable enterprise system architecture that not only facilitates new development. But also allows for leveraging the existing IT infrastructure assets through reengineering. This paper discusses first elements to implement for structural analysis framework using an enterprise organization example. Storage repository is through *NoSQL* system *Neo4J* graph database. Extraction using *Cypher Neo4J's* query language, export results in *XML* file. The use of different measures like eccentricity and complexity of a system are given. Structural analysis based on Q-analysis method proves how to ensure synchronization between formal organizational structure and the emergent one, due to perceived changes in business processes. The proposed solution architecture improves organizational structure of an enterprise in order to be more efficient and more aligned with current processes organization.

1 Introduction

All organizations have a management structure that determines relationships between functions and positions, subdivides and delegates roles, responsibilities, and authority to carry out defined tasks. Organization is a set of constraints on the activities performed by a set of collaborating agents (Weber 1987). Using the unified foundational ontology, organization is considered as a system including organizational activities structured in business process and services, information systems supporting organizational activities (Mark, Fox, Mihai, Gruninger. 1996) (Paulo Sergio et al) underlying information technology infrastructures and organizational structures. Intelligent organization is concerned about organization structure. In other hand, a learning organization is one skilled in acquiring, creating, transferring and retaining knowledge as well as transforming that knowledge into improved performance or innovative products and services. All these activities depend on human interaction that are members on it and are, on average, intelligent and capable of learning. i.e., organizational intelligence cannot simply be equated with human intelligence. (Schwaninger. 2009) Therefore, how can we conceive enterprise organization so as to be adapted to an intelligent and dynamic behavior? In our view, structural analysis with its foundation and holistic practices based on algebraic topology contributes to organizational intelligence paradigm. This work permits to establish a framework for design and development of intelligent organiza-

tions founded on advanced models of enterprise's architecture and complexity management. Next we give some existing enterprise modeling practices.

Among different reasons why strategic plans fail, we found failure to coordinate i.e., reporting and control relationships are not adequate so organizational structure is not enough flexible. Then failure to manage change because of lack of vision on the relationships between processes, technology and organization. Companies tend to improve their organization structures to be more effective and efficient. To carry out change toward a structure that aligns with projects and strategies of a company, we need to define the ontology's organization then make its structural analysis to implement a meta-model of a learning organization that aligns and meets strategic objectives of the company. Therefore, organizational issues are of high priority and should be of general interest. Also enterprises need to know how far their organization is stable. This paper present a banking study case which takes into consideration the proposed metamodel and solution architecture, to show our suggestion to audit or assess the effectiveness of the organizational strategy and how it could be reengineered so as to be more stable and aligned to the enterprise project.

The paper is organized as follow: Section 2 discusses existing enterprise modeling and our proposed meta-model for organizational matters. Section 3 outlines the proposed solution architecture. Section 4 focuses on q-analysis method for assessing organizational structure of an enterprise. Section 5 presents a Moroccan bank's organizational structure study case and its business process. This study case is projected onto the meta-model and the system architecture predefined. Using q-analysis to improve enterprise organization into a new organizational structure with less complexity. Finally, in section 6 we conclude the paper and we emphasize our future works.

2 Enterprise modeling

Enterprise models have a critical role in this study, enabling better designs for enterprises, analysis of their performances, and management of their operations. Modeling is at once organizational, informational and human. We study initially existing modeling techniques, to locate the standardization and normalization efforts (Kirikova 1982) (Oluwole, Olatunji 2010). There are many enterprise models such as, IDEF (Integrated DEFinition Methods) used for modeling activities necessary to support system analysis, design, improvement or integration (IDEF official site). Then GIM and GRAY models, here an enterprise consists of a physical system, a decision system and an information system. An enterprise can be described using four views: functional, physical, decisional and informational view. Also CIMOSA (Computer Integrated Manufacturing Open Systems Architecture) defines a model-based enterprise engineering method which categorizes manufacturing operations into generic and specific functions (CIMOSA. official site). The advanced models are COBIT, ISO 19440 extend in works (Boulmakoul et al 2012), ARIS framework (Architecture of Integrated Information Systems), etc. In our study we have been inspired by ARIS method, it provides unified organization foundation ontology. ARIS is an approach to enterprise modeling. It offers methods for analyzing processes and taking a holistic view of process design, management, work flow, and application processing. See works given in (ARIS toolset website). ARIS enterprise architecture framework defines organization as a system including:

organizational activities structured in business processes and services; Information system supporting organizational activities; Information technology infrastructures and Organizational structures. Organizational view in the requirement definition layers includes modeling concepts for enterprise's structure. Figures 1.a, 1.b, 1.c and 1.d, show fragments of our proposed organizational metamodel (Package classes). It defines the following packages. **Organizational package** (figure 1.a) which includes the generic classes: ***Organizational Unit***: entity responsible for achieving organizational goals; ***Position***: the smallest organizational unit; ***Performer:*** represents a person assigned to an organization; ***Location***: a geographical location of an organization unit, person, position or organization cell; Then **Objective package** (figure 1.b) that introduces ***Objective***: include explicit goals and targets set by the enterprise, while indicators are associated with assessing the enterprise's progress towards its objective. Finally **Process and Resource packages** (figures 1.c and 1.d) which define ***Activity:*** the fundamental business entities that represents actions taken by the enterprise. Activities can be composed of sub-activities thus can be combined with other business to represent business ***Process*****;** besides ***Resources*****:** business entities that can be used or consumed during the performance of an activity

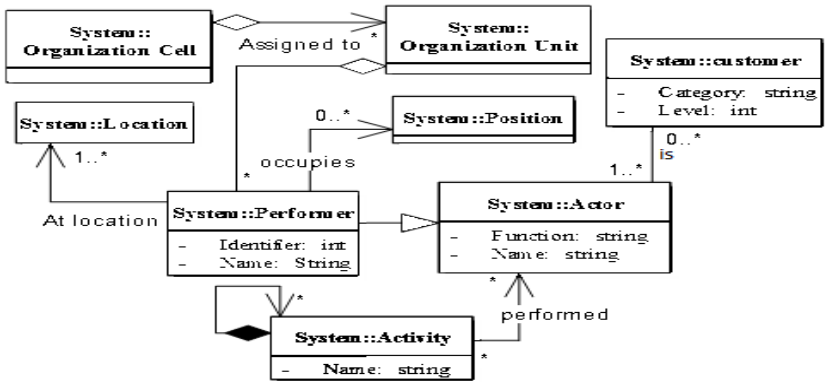


FIG. 1.a – *Organization view package*

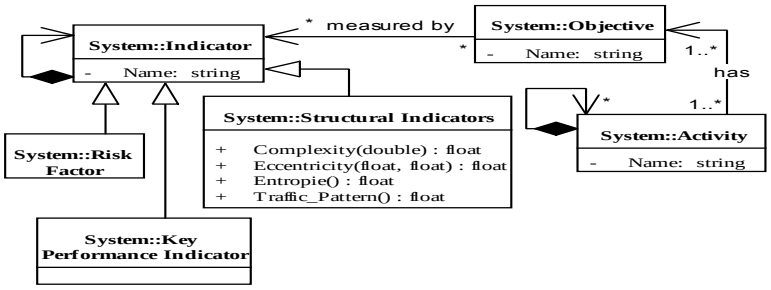


FIG. 1.b – *Objective view package*

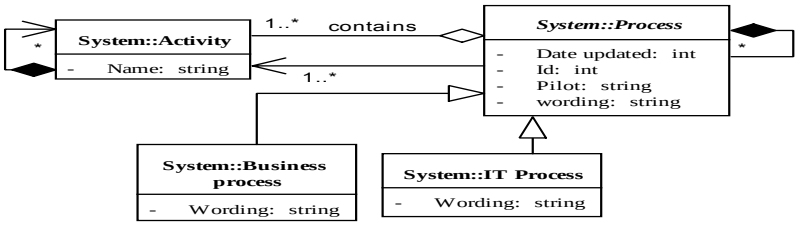


FIG. 1.C – *Process view package*

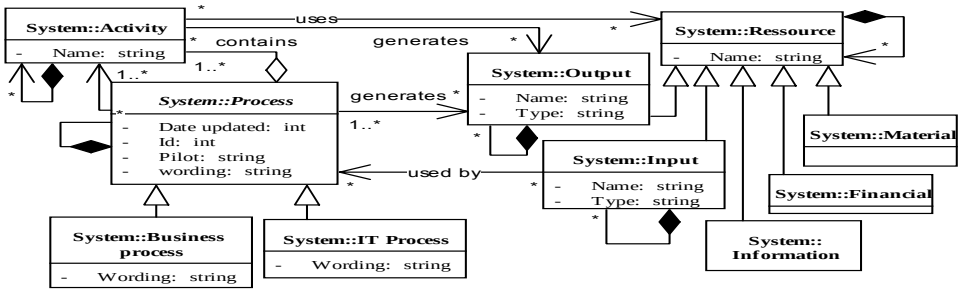


FIG. 1.d – *Resource view package*

3 System architecture

Our proposed system architecture for an intelligent organization based on structural analysis framework is organized in five layers: Repository, Extractor, Structural Analysis, Viewer and Organizational Structure Database. In the following we describe each layer of the proposed architecture shown in figure 2.

• **Repository** includes organization structure, processes, activities and different kinds of resources. We use big-data to store bank agency information's in a graph database *Neo4J* (Neo4J. official site). This layer will be the input of our system.

• **SETL**, Structural Extract Transform and Load, allow us to extract cleaned and useful information for a given analysis. It also provides the possibility to visualize the result of SETL processing using *Neoclipse* (Neoclipse website). We can query with *Cypher* language specialize for Neo4J graph database, visualize and export the result in XML JSON or CSV file.

• **Structural Analysis Framework** is the aim layer in our proposed system architecture. It takes as input the extract useful information from the repository so as to do structural analysis. It consists on measure of complexity, eccentricity and other organizational indicators in order to make diagnosis of current state of the enterprise organization and see if it's stable or requires improvement to make it more stable and aligned with the enterprise goal. In this stage we use java programming language to implement the framework. Using *Neo4J Java API* for data access, talk directly to Neo4j's graph engine directly in JVM based application. There is full feature parity with Neo4j Server, including HA clustering.

• **Viewer/Selector** to display and show results of the structural analysis framework, for the visualization

• **Organizational structure** database where we save data and future results of **new stable organizational structures**. Both are persisted in graph database of eventual re-engineering.

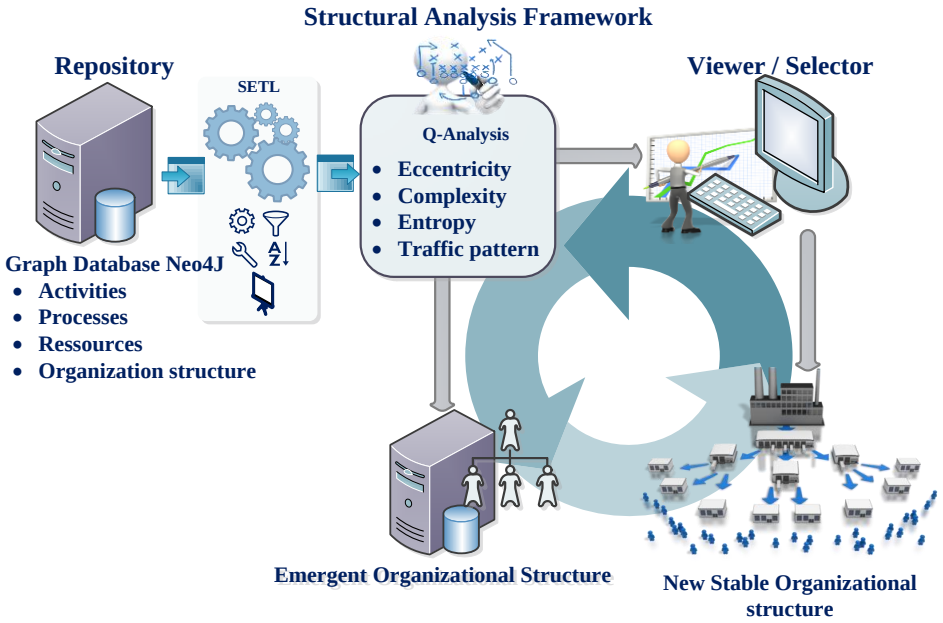


FIG. 2 – *Global solution architecture*

4 Structural Analysis : Case Study

Structural analysis provides an interactive, analytical environment for a user to scan an information system from multiple dimensions for analyzing the qualitative or structural aspects. The concept of structural analysis of enterprise is a simple notion of showing the user different views of an enterprise: who does what, where and how. It provides an interactive, analytical environment for a user to view the different entities in an enterprise in many ways (Ying et al (2009). We focus on Q-analysis method to improve the organizational structure of an enterprise. Let P be a set of processes and R a set of resources. D a database of business processes (**BP**), where each BP has a unique identifier (**rid**) and contains a set of processes. The set of all rids is denoted as R. The input database is a binary relation $\lambda \subseteq PxR$. The example given in Table.1 represents an illustration of database and its adjacency matrix of a BP.

Resource	Process	Adjacency matrix
r_1	P_3	$\begin{pmatrix} & P_1 & P_2 & P_3 & P_4 & P_5 \\ r_1 & 0 & 0 & 1 & 0 & 0 \\ r_2 & 1 & 0 & 0 & 1 & 0 \\ r_3 & 1 & 1 & 1 & 0 & 1 \\ r_4 & 1 & 0 & 1 & 1 & 1 \end{pmatrix}$
r_2	P_1P4	
r_3	$P_1P_2P_3P_5$	
r_4	$P_1P_3P_4P_5$	

TAB. 1 – *Structural presentation of Business process example*

Simplicial Complexes K: is a set of vertices, $X=\{x_1 \ldots x_n\}$ and a set of subsets of X. The subset σpi with p+1vertices is called a p-simplex. σ_{pi} is said to have dimension p (one less than the number of vertices). The superscript i is an index (more than one simplex has dimension p). A simplex σ_q is said to be a q-dimensional face of σ_p, if and only if every vertex

of σ_q is also a vertex of σ_p. K satisfies the condition that all the faces of its simplicies are also in K. The dimension of K is the largest value of p for which there exists σ_{pi}. The simplicies can be represented with a spatial structure usually shown as a polyhedral one. Gluing such polyhedra of mixed dimension forms the complex (Atkin 1974, 1977). A complex KY(X;λ) can be represented in Euclidean space E^H in the following way, for a suitable choice of H. Each p-simplex, typically $\sigma_p = <x_1,\ldots,x_{p+1}>$, is made to correspond to a convex polyhedron in E^H with (p+1) vertices which themselves correspond to $x_1,\ldots,x_{p+1}$. Thus, in an intuitive sense, in E^H the simplex σ_p is represented by the solid polyhedron with (p+1) vertices. The complex K is then represented by collection of polyhedra suitably connected to each other by sharing faces (or sub-polyhedra).

Chain of q-connection in K : Given two simplicies σ_p, σ_r in K we shall say they are joined by a chain of connection if there exists a finite sequence of simplicies $\sigma_{\alpha 1},\sigma_{\alpha 2},\ldots,\sigma_{\alpha h}$ such that : (i): $\sigma_{\alpha 1} \leq \sigma_p$; (ii): $\sigma_{\alpha h} \leq \sigma_r$; (iii): $\sigma_{\alpha i}$, $\sigma_{\alpha i+1}$ share a common face (say) $\sigma_{\beta i}$ (i=1,…,h-1).This sequence is a chain of q-connection (q-connectivity) if q is the least of the integers $\sigma_{\alpha 1},\sigma_{\alpha 2},\ldots,\sigma_{\alpha h}$. The length of the chain will be taken as (h-1) and, when needed the chain may be denoted by $[\sigma_p,, \sigma_r]_q$.

Q-analysis: is based on the q-nearness and q-connectivity relations between the simplicies of a given complex (or simplicial complex) (Duckstein et al (1988), (Jiang et al 2004, 2006). A Q-analysis of a complex K determines the number of distinct equivalence classes, or q-connected components, for each level of dimension q ranging from 0 to q-1. The equivalence classes are decided by a rule as follows. If two simplicies are q-connected (either q-near or q-connected), then they are in a same class. To see this we introduce, for a fixed q, a relation γ_q on the simplicies of K, defined by: $(\sigma_p, \sigma_r) \in \gamma_q$ if and only if σp is q-connected to σ_r. This γ_q is reflexive, symmetric and transitive and therefore an equivalence relation. The equivalence classes, under γ_q are the members of the quotient set K/γ_q, and constitutes a partition of all simplicies of K which are of order ≥ q. We denote the cardinality of $K/\gamma q$ by Qq. This equals the number of distinct q-connected components in K. When we analyze K by finding all the values of $Q_0,Q_1,Q_2,\ldots,Q_N$ where N = dimK , we say that we have performed a Q-analysis on K. To find the shared face q-value between all pairs of the Y's in KY(X;λ)., the following steps could be performed: (i) form ΛxΛT, (ii) evaluate $\Lambda \times \Lambda T - \Omega = (\omega_{ij})$ and $\omega_{ij} = 1$. For example, the Q-analysis of the complex in example 1 leads to the following equivalence classes at the different dimensional levels of q=0, q=1, q=2 and q=3. Each equivalence class is enclosed in the curly brackets. The sign "-" in the matrix stands for -1, and shows that r_1 and r_4 are disconnected.

At **q = 3** we have $Q_3 = 2$; $\{r_3\}$, $\{r_4\}$
At **q = 2** we have $Q_2 = 1$; $\{ r_3, r_4\}$
At **q = 1** we have $Q_1 = 2$; $\{ r_2, r_4\}$, $\{ r_3, r_4\}$
At **q = 0** we have $Q_0 = 1$; all $\{r_1, r_2, r_3, r_4\}$

$$\Lambda \times \Lambda^T - \Omega = \begin{pmatrix} & r_1 & r_2 & r_3 & r_4 \\ r_1 & 0 & - & 0 & - \\ r_2 & & 1 & 0 & 1 \\ r_3 & & & 3 & 2 \\ r_4 & & & & 3 \end{pmatrix}$$

TAB. 2 – *Q-analysis of KY(X,λ) given in example **1***

4.1 Case study

Our case study is conducted from an entity-oriented bank that centralizes the processing of all its operations. In fact, our analysis takes as its target the structural organization of the

branch banking business linked to a regional management companies. The latter is attached to the commercial pole. In this section, we list all organizational units. Processes undertaken by this structure, as well as the activities that arise and resources used to achieve the desired objectives. In our example, we apply a structural analysis of relationships existing among the major organizational units in order to assess the degree of strategic alignment within this structure. Figure 3 Organization chart typifies points of sale as undertaken Agency. It shows the organization of the agency representing the banking business. There are processes represented by activities, resources, objectives, indicators and organizational units.

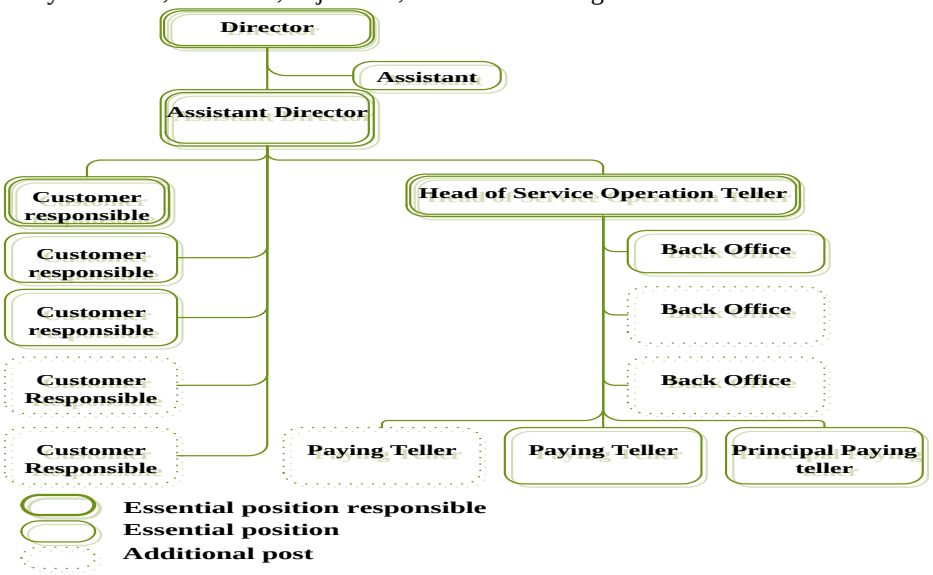


FIG. 3 – *Organization chart typifies points of sale (type: undertaken Agency)*

The proposed meta-model has been instantiate with this example using new concept of persistency database, Big data.

Big Data: is the frontier of a firm's ability to store, process, and access all the data it needs to operate effectively, make decisions, reduce risks, and serve customers (Gualtieri 2012), (Bixo labs 2012). Then we store repository of this case study in No SQL System. The importance of Big Data is when we consider enterprise organization as social network. its business processes model requires projections on real experience, which can only be delivered by capturing and using all the available data about a performer or organization unit.

Why use No-SQL system? No-SQL, is an alternative to using traditional Database Management System (DBMS). It provides flexible schema than the rigid relational model which can be useful when it not easy to get a data into structure table format. It tends to be quicker/ cheaper to set up. They are design for massive scalability, both on amount data also with the efficiency of the operation on that data. They don't necessarily have transactional guarantees, in general what they do is relaxed consistency offered by the system and in turns in higher performance and higher availability of the system. Several incarnations of No-SQL systems are divided into four categories *Map Reducer framework like OLAP*, *key-value stores like OLTP*, *Documents store* and *Graph database systems* (Garcia et al2009)

Graph database system for enterprise organization matters? Graphs are everywhere. Organizations of all sizes, from large enterprise to new startups, are embracing graph databases as the fastest way to query and store graph data. Figure 3 is represented by graph structure with nodes, edges and properties. As our case study takes a professional network so it is evident to choose a graph database system. With a graph database, the focus is on the connections between data. Telling the database in advance that things are connected and how, and representing those relationships physically, as opposed to storing them in tables and relating them through indexes. Several graph databases used to store big data (*Neo4J*, *InfiniteGraph* (InfiniteGraph website 2012), *OrientDB* (*OrientDB website 2012*) …etc). To represent and store information with large scale, we choose Neo4J graph database with the following advantages.

Neo4J graph database: It is an open source solution to manage our information's study case. As a robust, scalable and high-performance database, Neo4j is suitable for full enterprise deployment or a subset of the full server can be used in lightweight projects. It features true ACID (Atomicity, Consistency, Isolation and Durability) transactions; high availability; scales to billions of nodes and relationships and high speed querying through traversals

It runs as server, as embedded java, scales to 34bn nodes, licensed like *MySQL*. Besides provides ACID transactions and integrates *Lucene* (Lucene website 2012) index and it traversals 1 M/s. There is a special query language for No4J NoSQL graph database called *Cypher*. They goals are: declarative, pattern matching, ASCII art pattern, closures, SQL familiarly and external DSL.

Figure 4 shows the stored repository in Neo4J graph database using graphic interface neoclipse. The graph representation allows us to instantiate the metamodel using nodes and relationships with index and an id of each element of the graph. For example Organizational unit is a node with specific Id and properties with values from the case study (Org-Unit-label, type). It has a relationship with performer node as *ASSIGNED_TO* relationship_type. Relationship_type represents associations between classes of the instantiated meta-model. The different relationship_types are: *ASSIGNED_TO, PERFORMED_BY, OCCUPIES, AT_LOCATION, UPPER_HIERARCHICAL, HAS, GENERATE*.

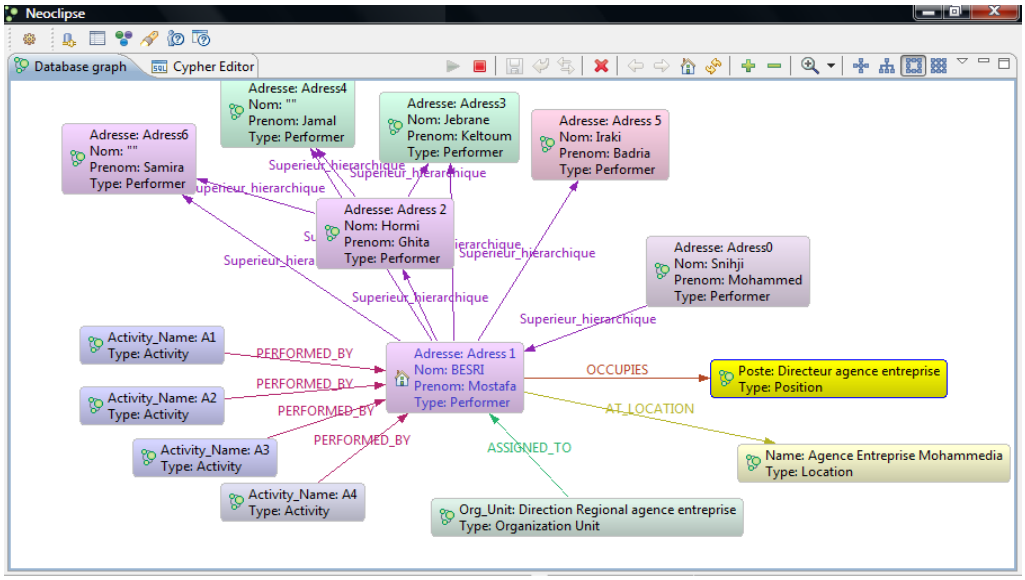


FIG. 4 – *Case study graph database using Neo4J with neoclipse*

4.2 Structural analysis application

Using Cypher graph database query language (table 3), we extract records to get adjacency matrix of relationship $\boldsymbol{\lambda}$ between performers $\boldsymbol{P}$ and activities $\boldsymbol{A}$. The query has as inputs all performers' nodes, organizational unit's nodes, activity's node and objectives of activity too. The output of this query is for each organization units have its performers and attached activities. The query has Match clause to indicate relationships between nodes .Query result can be extract as XML, CSV or JSON file. We use q-analysis algorithm with shared face matrix for *KP(A;λ) in* input (table 4, second column) from the extracted XML file, , we get q-connectivities (also called q-chains). Q-connectivities are revealed formally by a Q-analysis of the complex and this is given in table 5. It also generates the structured vector Q (last line in table 4 first column) of the simplicial complex.

```
START p1=node(*),p3=node(*),p2=node(*),p4=node(*)
MATCH  p3-[*]->p1 <-[*]-p2<-[*]-p4
WHERE    p1.Type="Performer"    AND    p3.Type="Organization    Unit"    AND
p2.Type="Activity" AND p4.Type="Objective"
RETURN    p1.Prenom    as    Performer,    p3.Org_Unit    as    Organiza-
tion_Unit,p2.Activity_Name  as  Activity,  p4.Objective_label  as  Objective,
count(p3) as Occurrences
```

TAB. 3 – Cypher query to extract relationships between performers of each organizational unit and activities

At **q = 7** we have Q_7 = 2; {P_2} {P_9}
At **q = 6** we have Q_6 = 2; {P_2} {P_9}
At **q = 5** we have Q_5 = 2; {P_2} {P_9}
At **q = 4** we have Q_4 = 2; {P_2} {P_9}
At **q = 3** we have Q_3 = 6; {P_1} {P_2, P_6 ,P_7,P_8} {P_3} {P_4}{P_5}{P_9, P_{10} ,P_{11}}
At **q = 2** we have Q_2 = 6; {P_1} {P_2, P_6 ,P_7,P_8} {P_3} {P_4}{P_5}{P_9, P_{10} ,P_{11}}
At **q = 1** we have Q_1 = 6; {P_1} {P_2, P_6 ,P_7,P_8} {P_3} {P_4}{P_5}{P_9, P_{10} ,P_{11}}
At **q = 0** we have Q_0 = 1;{ P_1,P_2, P_3, P_4, P_5,P_6 ,P_7,P_8,P_9, P_{10} ,P_{11}}

$$KP(A;\lambda)\begin{pmatrix} & P_1 & P_2 & P_3 & P_4 & P_5 & P_6 & P_7 & P_8 & P_9 & P_{10} & P_{11} \\ P_1 & 3 & - & - & - & - & - & - & - & - & - & - \\ P_2 & & 7 & - & - & - & 3 & 3 & 3 & - & - & - \\ P_3 & & & 3 & - & - & - & - & - & - & - & - \\ P_4 & & & & 3 & - & - & - & - & - & - & - \\ P_5 & & & & & 3 & - & - & - & - & - & - \\ P_6 & & & & & & 3 & 3 & 3 & - & - & - \\ P_7 & & & & & & & 3 & 3 & - & - & - \\ P_8 & & & & & & & & 3 & - & - & - \\ P_9 & & & & & & & & & 7 & 3 & 3 \\ P_{10} & & & & & & & & & & 3 & 3 \\ P_{11} & & & & & & & & & & & 3 \end{pmatrix}$$

TAB. 4 – Q-Analysis of KP(A;λ) first column. Shared face matrix for KP(A;λ) second column

Performance diagnosis:

Eccentricity: It describes status of an individual simplex within the entire complex. It indicates the degree of integration of a specific simplex into the whole complex K. Atkins suggest a measure of eccentricity (Atkins 1974); denoted as ecc (equation 1):

$$ecc(\sigma) = \frac{\hat{q} - \check{q}}{\check{q} + 1} \qquad (1)$$

Where top-q "$\hat{q}$" the dimensional level at which a simplex first appears in the simplicial complex. Bottom-q "$\check{q}$"is the level at which simplex first becomes connected in a component with another simplex. A simplex is eccentric when it is badly embedded within the complex. (Beaumont,2007), (Kasiphan et al 2007), (Linton et al 1980) and (Duckstein et al 1988,1997)suggest another measure of eccentricity called ecc' (equation 2)

$$ecc'(\sigma) = \frac{2\sum_i q_i / \sigma_i}{q_{max}(q_{max} + 1)} \qquad (2)$$

Where q_i each q-level where σ appears, σi is the number of elements in σ_i's equivalence class at level q_i and q_{max} the maximum level of the complex. In the proposed business process example and using eccentricity measures we found the following results. The difference between ecc and ecc' is that ecc depends on the other simplicies and takes values in interval of [0, ∞]. In this case study, we obtain following values given in table 5 and graphic result (figure 6):

Simplex	Ecc	Ecc'
$\sigma_3(P_1)$	3	0.333
$\sigma_7(P_2)$	1	0.428
$\sigma_3(P_3)$	3	0.333
$\sigma_3(P_4)$	3	0.333
$\sigma_3(P_5)$	3	0.333
$\sigma_3(P_6)$	0	0. 166
$\sigma_3(P_7)$	0	0.166
$\sigma_3(P_8)$	0	0.166
$\sigma_7(P_9)$	1	0.428
$\sigma_3(P_{10})$	0	0.166
$\sigma_3(P_{11})$	0	0.166

TAB. 5 – *Eccentricity of each performer*

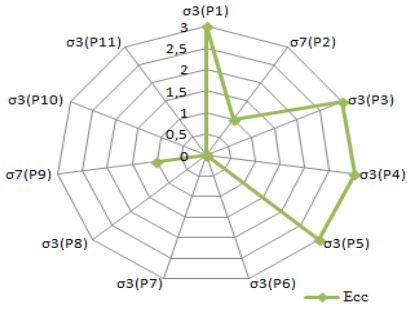


FIG. 6 – *Eccentricity of each performer*

In our example the simplicial complex is one piece at q=0 and simplexes with maximum eccentricity are P_1, P3, P4, P5, the most distinctive or least well connected in the simplicial complex (table 5, figure 6). It is due to three sources (q top = 3), of which at most are shared with other activities.

Complexity: The complexity of the system structure of this example can be described by the complexity measure Ψ(K) suggested by (Jiang et al 2006) in the equation 3 :

$$\Psi(K) = 2\left[\sum_{k=0}^{dimk} (k+1)Q_k \,/\, (dimK+1)(dimK+2)\right] \qquad (3)$$

Where Q_k is the k^{th} component of Q. In the proposed example the complexity of system's structure is: $\Psi(K) = 10.928$

This case shows that several simplicies are poorly embedded within the complex we might say they are eccentric. Instead of treating the performers as simplicies and activities as vertices, we might look at the question of organization structure from the viewpoint of the facilities and treat them as simplicies defined by the organizational units they impact. This means that we make a reverse engineering of the organization to get less eccentric components with less complexity.

In this paper, we have addressed the using of Q-analysis to assess enterprise organization complexity. By using an organizational mining approach based on business process organization, we can discover the organizational structure adequate to the actual activities of the company. In order to study how the simplicies conform to the complex (enterprise organization) and to determine whether there are any simplicies that are totally disconnected, the following indicators have been proposed: q-connectivity, structure and obstruction vectors, eccentricity, and complexity. Q-Connectivity describes the global relationship among equivalence classes. The structure and obstruction vector indicate the potential for simplifying the representation of the relationships. Q-analysis provides a canonical view of the structural relationship between performers and organization units.

5 Conclusions &Future work

It is of great importance for IT organizations to bring about technological improvements to the enterprise systems. The process organization obtained by the analysis of business processes observed in practice will be used to generate a new organization (Boulmakoul, Besri 2012). This emerging organization must be confronted with the formal organization to measure conformance. Our Framework tends to achieve a re-engineering of enterprise organization by the use of a canonical method of Q-Analysis and advantages provided by NoSQL system especially graph database for enterprise organization structure. The paper shows the use of NoSQL system to instantiate the metamodel for structural analysis of the banking case study

Future works will complete implementation of the software solution for mining enterprise organization and will also allow reengineering processes to ensure conformance between organization structure and process organization.

References

ARIS toolset. Available at http://www.ids-scheer.com accessed (Mai 15, 2012).

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

Atkin, R. (1977). Combinatorial Connectivities in Social Systems. Basel, Birkhäuser Verlag.

Beaumont J.R., A.C. Gatrell. An introduction To Q-Analysis. ISSN 0306-6142.

Bixo labs. Elastic web mining November 01, 2009. Available at http://fr.slideshare.net/kkrugler/elastic-web-mining-2407818 Accessed (December 05, 2012.)

Boulmakoul, A., N. Falih and R. Marghoubi. (2012). Deploying holistic meta-modelling for strategic information system alignment. Asian network for scientific information. Information Technology Journal. ISSN 1812-5638 / DOI: 10.3923/ITJ.2012. Asian network for scientific information.

Boulmakoul. A, Z. Besri. (2012). Intelligent organization, structural analysis framework for enterprise organizational reengineering. Innovation and new trends in information systems 2nd edition ISBN 2168/2008 pp 157.

CIMOSA. official site Available at http://cimosa.cnt.pl/ accessed (Mai 15,2012)

Duckstein Lucien, Steven A. Nobe. (1997). Q-analysis for modelling and decision making. European journal of operational Research 103 411-425.

Duckstein, L., Bartels, P. H. and Weber, J. E. (1988). Organization of a knowledge base by Q-analysis. Applied Mathematics and Computation vol. 26(4), 289-301.

Hector Garcia Molina, Jeffrey D.Ullman, Jennifer Widom. (2009). Database Systems – The Complete Book, 2nd Edition. ISBN: 0131873253, 136067018

IDEF official site. Available at http://www.idef.com/ accessed Mai 15, (2012).

InfiniteGraph tool Available at http://objectivity.com/ accessed (December 10, 2012).

Jiang B. and Claramunt C. (2004). Topological Analysis of Urban Street Networks, Environment andPlanning B: Planning and Design, Pion Ltd, Vol. 31, pp. 151-162.

Jiang Bin, Itzhak Omer. (2006). Spatial Topology and its Structural Analysis based on the concept of Simplical Complex. 9th AGILE Conference on Geographic Information Science, Visegrad, Hungary 204-212.

Kasiphan Masakul, Suchai Thanawastien, Pakorn Sermsuk. (2007) Ontological Automation of Strategic information System Planing. 24th South east Asia Regional computer conference, November. Bangkok, Thailand.

Linton C. Freeman. (1980). Q-Analysis and the structure of friendship networks. International journal Man-Machine studies 12, 367-378.

Lucene core. Available at http://lucene.apache.org/core/ accessed (December 10, 2012)

Marite Kirikova. (1982). Flexibility of Organizational Structures for flexible Business Processes.

Mark S. Fox, Mihai Baruceanu, Micheal Gruninger. (1996). An organization ontology for enterprise modeling: Preliminary concepts for linking structure and behavior. Computers in Industry 29 123-134.

Mike Gualtieri. The Pragmatic Definition Of Big Data . Available at http://blogs.forrester.com/mike_gualtieri/12-12-05-the_pragmatic_definition_of_big_data (accessed December 8, 2012)

Neo4J. official site Available at http://www.neo4j.org/ accessed (December 10, 2012)

Neoclipse. Available at https://github.com/neo4j/neoclipse accessed (December 10, 2012)

Oluwole Alfred, Olatunji. (2010). Modelling organizations' structural adjustment to BIM adoption: Apilot study on estimating organizations. Journal of information technology in construction ISSN 1874-4753.

Orientdb Available at http://www.orientdb.org/ accessed (December 9, 2012)

Paulo Sergio Santos JR., Joao Paulo A.Almeida, Giancarlo Guizzardi, An ontology-Base Semantic Foundation For Organizational Structure Modeling in ARIS Method. Ontology & conceptual modeling research group (NEMO).

Schwaninger Markus. (2009). Intelligent organization, Powerful models for systemic Management. ISBN 978-3-540-85161-5

Weber, M. (1987). Economy and Society, University of California Press, Berkeley, Calif.

Ying Tat Leung, Jesse Bockstedt. (2009) Structural Analysis of a business Enterprise. Service Science 1(3), pp. 169-188

Framework Structural pour un alignement stratégique des systèmes d'information multipoints de vue

Noureddine Falih*

* FST Mohammedia, Département informatique, B.P. 146 Mohammedia Maroc
nourfald@yahoo.fr , http://www.fstm.ac.ma/

Résumé. Depuis les années 1990, les chercheurs s'intéressant au management des organisations ont reconnu que les stratégies d'alignement des métiers avec les Technologies d'Information (IT) exigeraient également l'alignement structural entre les systèmes d'information (SI) et l'organisation. L'alignement structural met l'accent sur l'importance de la cohérence et l'harmonie entre les métiers de l'entreprise et les technologies de l'information qui les supportent, en particulier dans le domaine du décisionnel, et gouvernance des SI. Dans cet article, et faisant suite à nos derniers travaux en la matière, nous proposons un Framework structural doté d'une architecture informatique concrétisant cette pensée et susceptible de fournir une bibliothèque de méta-connaissances enrichissant l'ensemble des tableaux de bord communs à l'entreprise pour un alignement stratégique multipoints de vues.

1 Introduction

L'orientation stratégique de l'entreprise est cruciale à sa performance et compétitivité dans le marché (Atkinson 1990). En effet, la structure de l'entreprise est considérée en tant que fondement de ses choix stratégiques et technologiques (Ettlie et Bridges 1984). Par ailleurs, les SI constituent un axe de développement de nouvelles activités génératrices de rentabilité, et donc d'avantages concurrentiels, par son rôle primordial de répartition des processus/activités et propagation de l'information, en temps réel, à tous les niveaux hiérarchiques. L'amélioration de la performance globale de l'entreprise doit être le but ultime de la stratégie d'entreprise et des SI (Delone et McLean E.R 1992). De ce fait, la question de synchronisation des SI avec la stratégie globale de l'entreprise est un prés-requis indispensable pour cette finalité. En effet, l'alignement stratégique des SI est un processus continu et dynamique, qui fournit des solutions et des infrastructures technologiques à l'entreprise lui permettant de rencontrer les objectifs de performance fixés par sa stratégie. Toutefois, cette notion d'alignement stratégique est considérée comme l'un des plus grands défis auxquels sont confrontés les directeurs des SI dans l'entreprise, vu son caractère nébuleux, difficile à comprendre et à mesurer. En effet, les recherches actuelles ayant trait à l'alignement stratégique proposent des démarches purement managériales, ne s'intéressent pas assez aux aspects techniques et opérationnels. En revanche et afin d'accompagner l'entreprise dans un projet d'alignement stratégique multipoints de vues, nous proposons une approche intitulée S2A (Structural Strategic Alignement) susceptible d'apporter une information pertinente permettant d'évaluer le niveau de cohérence entre les axes stratégiques majeurs de l'entreprise et son SI. Cette approche est centrée sur le Standard de l'entreprise ISO/DIS 19440 (ISO 19440, 2007), intégrant des structures spécifiques au référentiel Cobit dans son cadre de gouvernance des SI et permettant de porter des outils systémiques issus du paradigme struc-

tural en vue d'une meilleure visibilité de l'alignement stratégique. Une multitude de matrices structurales découlant des interactions combinatoires entre les différentes composantes du Méta-modèle, pourra être étudiée en vue d'un alignement stratégique multipoints de vue. Cet article est subdivisé en quatre parties : Les deux premières parties sont dédiées à l'introduction et l'état de l'art concernant cette notion d'alignement stratégique, une troisième partie mettant en relief notre Framework structural basé sur une architecture de mise en œuvre pour la contribution à la résolution de la problématique soulevée, puis une dernière partie réservé à la conclusion tirée de cette technique et perspectives futures à développer.

2 Etat de l'art

2.1 Définition de l'alignement stratégique

L'alignement stratégique a de nombreux pseudonymes. Il a été dénommé "Coordination" (Lederer et Mendelow 1986), "Harmony" (Luftman 1996), "fit" (Porter 1996), "Linkage" (Reich et Benbasat 1996), "Bridge" (Ciborra 1997), "fusion" (Smaczny 2001). Cependant, dans tous les cas, il s'agit de l'intégration des stratégies relatives aux Métiers et Technologies de l'Information. De nombreux chercheurs ont essayé de fixer une définition pour l'alignement stratégique (CIGREF 2009), (Henderson et Venkatraman 1993), (Lederer et Sethi 1992), (Benbasat et Reich 1996), (Ward et Peppard 2002). Toutefois, l'ensemble de ces définitions ont convergé vers le fait que c'est une démarche qui vise à faire coïncider la stratégie SI sur la stratégie métiers de l'entreprise. Cette démarche a pour finalité de renforcer la valeur d'usage du SI et de faire de celui-ci un actif principal de l'entreprise.

2.2 Démarches existantes de l'alignement stratégique

Les premières approches top-down en matière d'alignement stratégique donnent naissance à des modèles de planification stratégique basés sur l'hypothèse que la stratégie IT peut être planifiée et souvent étroitement associée à la stratégie métiers. L'ensemble des travaux portant sur cette question de cohérence entre les stratégies métier et SI ont été catégorisés en deux grandes visions :

a. SISP: Strategic IS planning

La démarche SISP ou planification stratégique des SI consiste au développement de différentes méthodologies intégrant les objectifs stratégiques de l'entreprise dans les feuilles de routes établies pour les SI, tout en essayant de créer des applications dédiées à la gestion des SI (MIS) susceptibles d'améliorer la position concurrentielle de la société (Ang et al. 1995), (Newkirk et Lederer 2006), (Wilson 1989), (Vargas, et al. 2008).

b. **IS/IT Alignement**

L'alignement métiers et IT est le degré auquel la stratégie IT supporte et est supportée par la stratégie métiers (Luftman et al. 1993), (Silvius 2007). L'alignement IT/Métiers est défini comme étant le degré auquel les missions, objectifs et plans liés aux technologies de l'information (IT) supportent et sont supportés par les missions, objectifs et plans liés aux métiers (Reich et Benbasat 1996), (Issa-Salwe et al. 2010). L'ensemble des approches ayant trait à l'alignement stratégique proposent des modèles de diagnostic, de prescription, d'action et d'intégration tels que, par exemple:

- **Le modèle MIT de Scot Morton** (Scott 1991) : comprend 5 «forces» qui interagissent au sein d'une organisation face à un environnement technologique et socio-économique externe de plus en plus exigeant. Ces forces sont axées sur la stratégie, La structure organisationnelle, le management des processus, les IT et les acteurs et leurs rôles.
- **Le modèle SAM** (Henderson et Venkatraman 1993) : définit à la fois les concepts clés et les processus de mise en cohérence des quatre domaines suivants : La Stratégie d'affaire, La conception de l'organisation, la Stratégie IT et les SI.
- **Le modèle de Luftman** (Luftman 2003) : propose 6 critères pour l'évaluation du niveau d'alignement une organisation : degré de maturité de communication, capacité à mesurer, gouvernance, partenariat métier/IT, architecture et maturité des connaissances.
- **La démarche SEAM «Systemic Enterprise Architecture Methodology»** (Wegmann et al. 2005) : intègre, selon une approche unifiée basée sur la pensée systémique, les méthodes les plus populaires utilisées en management et en engineering pour résoudre des problèmes spécifiques. Cette méthode permet de réduire les risques et optimiser les ressources de développement dans l'entreprise.
- **La démarche ACEM « Alignment Correction and Evolution Method »** (Etien 2006) : vise à faire évoluer conjointement un système d'information et les processus de l'entreprise. Cette démarche vise à rétablir l'alignement entre le modèle des processus et le modèle du SI par une adaptabilité de l'un ou l'autre, voire l'ensemble de ces deux composantes. Ensuite, adopter un aspect évolutif de cet alignement en tenant en compte des exigences d'évolution.
- **La démarche INSTAL « Intentional Strategic Alignment »** (Thevenet 2009) : a pour objectif d'assurer la cohérence entre un ensemble d'éléments appartenant aux niveaux stratégique et opérationnel dans l'entreprise: la stratégie, le SI et les processus métiers. La méthode INSTAL utilise à son tour un modèle pivot intentionnel susceptible de faciliter la mise en correspondance entre les éléments stratégiques et les éléments opérationnels.

Malgré l'acceptation généralisée que la stratégie métiers devrait être alignée à la stratégie SI, la recherche dans le domaine est principalement conceptuelle et manque de considérations pratiques (Campbell et al. 2005). La nature de l'alignement est insuffisamment clarifiée et explicitée dans la littérature. En effet, les approches actuelles d'évaluation, bien que principalement focalisées au niveau stratégique, offrent un peu de finesse aux niveaux tactiques et opérationnels qui sont identifiés comme des plans importants pour la réalisation de l'alignement. Dans la section suivante, nous détaillons notre approche de méta-modélisation holistique intégrant une analyse structurale considérée comme étant une démarche pratique en vue d'une lecture formelle et multipoints de vues de l'alignement stratégique.

3 Approche S2A « Structural Strategic Alignement »

3.1 Méta-modélisation holistique

Dans cette section, nous rappelons notre méta-modélisation holistique basée sur une extension formelle du méta-modèle ISO/DSI 19440, en vue d'un alignement stratégique multipoints de vue. Cette méta-modélisation étendue supporte un agencement de construits permettant d'entamer une analyse structurale pour une meilleure synchronisation Stratégie/SI. Notre proposition d'extension formelle du méta-modèle a été évoquée et argumentée dans nos derniers travaux en la matière (Boulmakoul et al. 2009), (Falih et al. 2012). La figure 1

présente cette méta-modélisation holistique axée conjointement sur une analyse systémique et une vision holistique dans des domaines particuliers de l'entreprise. L'explicitation de ces notions fondamentales se renvoient l'une à l'autre et s'établissent simultanément dans une pensée paradigmatique pour une meilleure représentativité de l'entreprise. Dans cet article, notre analyse va au-delà des définitions explicites des entités fonctionnelles, organisationnelles, informationnelles et de ressource pour se mêler aussi des relations et associations liant ces entités afin de mieux situer cette notion d'alignement multipoints de vue.

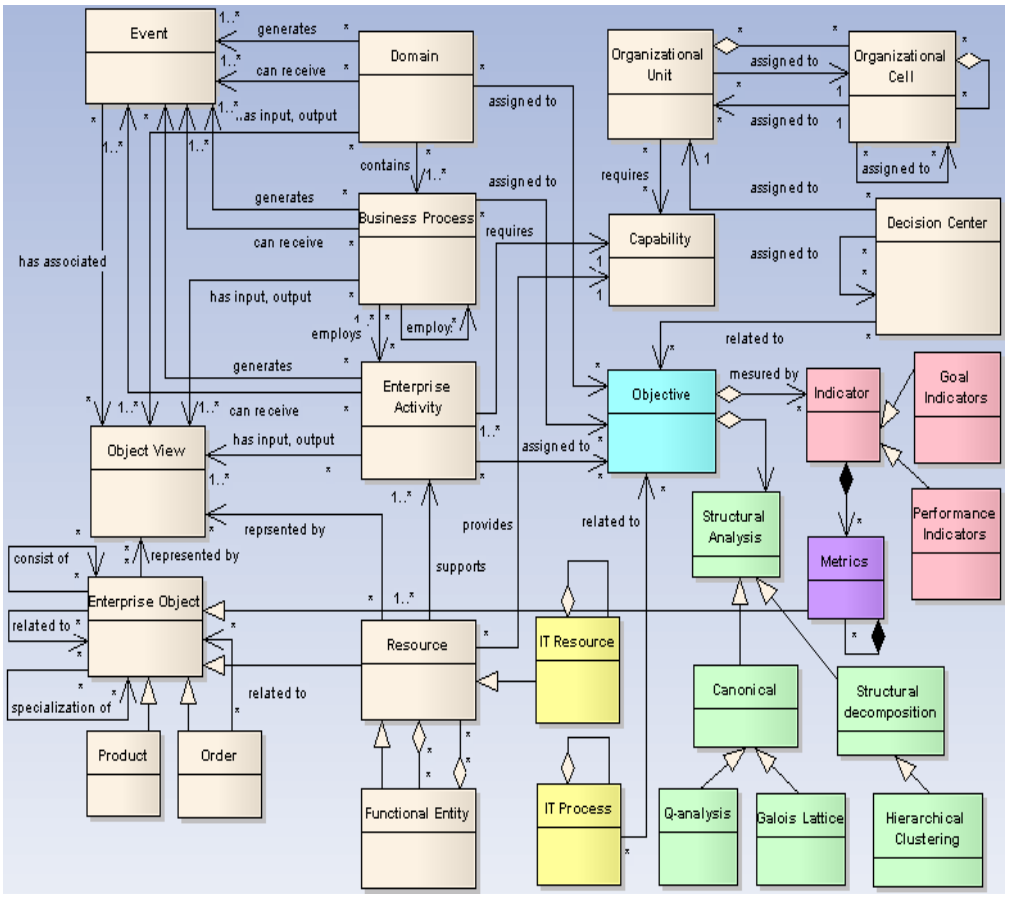


FIG. 1 – *Méta-modélisation Holistique pour l'Alignement Stratégique des SI*

3.2 Analyse structurale

Le méta-modèle ISO 19440 fournit un cadre pour une description formelle de l'organisation interne d'une entreprise. Il offre la plus large vue d'une entreprise parmi toutes les techniques communes, mais se concentre sur une représentation quelque peu statique pour soutenir la conception et l'intégration de l'aspect relationnel entre ses différents constituants. La méta-modélisation holistique proposée dans ce travail s'articule autour du concept de l'ana-

lyse structurale. Cette technique permet d'identifier et analyser les relations entre l'ensemble des entités constituant l'entreprise. Notre démarche d'analyse structurale établit les liens et les relations entre ces entités pour effectuer des analyses qui sont utiles pour un certain nombre de besoins, bien au-delà de l'alignement stratégique. En effet, l'analyse structurale fournit un environnement interactif d'analyse pour les décideurs afin d'examiner l'entreprise à partir de plusieurs perspectives, c'est une approche qui sort de l'ordinaire à travers la technique classique OLAP « On-Line Analytical Processing », elle permet une analyse des aspects qualitatif ou structural de l'entreprise afin de généraliser l'alignement sur tous les points de vue (Leung and Bockstedt 2009). L'analyse structurale permet un drill-down sur une dimension qualitative spécifique d'une entreprise d'où le basculement rapide entre les différents points de vue, où chaque vue peut contenir des combinaisons de différentes dimensions et relations. Ces capacités dans OLAP se sont avérées très utile dans la gestion d'une entreprise et nous nous attendons à la même chose avec l'analyse structurale. Typiquement, l'instanciation du méta-modèle étendu proposé permet d'appliquer notre approche S2A (Structural Strategic Alignement) sur des matrices diversifiées issues de différentes classes du méta-modèle de l'entreprise. Dans ce contexte, nous obtenons une collection complète d'entités ainsi que leurs relations constituant une méta-ontologie pertinente à partir de laquelle nous pouvons définir les éléments de l'analyse structurale pouvant refléter le niveau de synchronisation des composants appropriés à chaque domaine.

3.3 Exemples

Dans cette section, nous présentons quelques exemples de connexions matricielles constituant les éléments de base de différentes analyses structurales menées de part et d'autres de chaque domaine de l'entreprise. Des figures appropriées sont proposées, à titre d'illustration, pour situer les démarches S2A spécifiques pouvant être déclenchées afin de fournir l'information utile à toute fin d'alignement.

3.3.1 Analyse du marché

Dans les entreprises où les clients sont identifiés selon certains critères (sexe, groupe d'âge, code postal, etc.), savoir qui achète quels produits sont des informations de base du marché permettant d'identifier les opportunités commerciales, augmenter l'efficacité et aider à la Gestion des portefeuilles produits. Sachant qu'un client ou fournisseur peuvent constituer des inputs ou événements du méta-modèle, un point de vue Produit-Client, ou à un niveau plus agrégé un point de vue par segment Produit-Marché, donne un aperçu sur la destination des produits. Les lacunes évidentes relevées fournissent des conseils sur les opportunités commerciales potentielles. Un point de vue Produit-Prix permet de représenter le positionnement du portefeuille de produits en termes de prix de vente. Un point de vue Client-Prix donne un aperçu sur les niveaux de prix privilégiés des différents clients. De même, les points de vue Produits-Promotion et Client-Promotion offre une vision sur la couverture du marché par les promotions actuelles (ou passée). Ça permet d'identifier les lacunes et chevauchements des promotions courantes ou antérieures. Les points de vue servant à l'analyse du marché mentionnés sont une pratique assez courante dans le marketing. Pour soutenir cette vision de rationalisation et optimisation des offres de produits conformément aux attentes clients, nous pouvons mener des analyses S2A axées sur des combinaisons différentes mais qui convergent vers le même souci client. On y trouve, par exemple, des relations intra-vues comme le couplage (Domain, Business Process) qui situe le mode de partitionnement des processus par

rapport à chaque domaine de l'entreprise et le couplage (Business Process, Enterprise Activity) qui fournit des métadonnées propres à la catégorisation des activités par rapport à chaque contexte métier. Ces interactions évoluent par rapport aux objectifs opérationnels ou stratégiques influencés à leurs tours par les événements socio-économiques pouvant influencer l'entreprise. D'autres combinaisons inter-vues peuvent être remarquées pour considérer le mode de synchronisation entre les domaines, les processus métiers et les activités d'une part et les objets de l'entreprise d'autre part. L'analyse structurale entre ces différentes composantes est susceptible de mesurer l'alignement fonctionnel/informationnel au sein de l'entreprise. La figure 2 récapitule le positionnement des différentes analyses S2A pouvant fournir des informations pertinentes pour cette fin d'alignement.

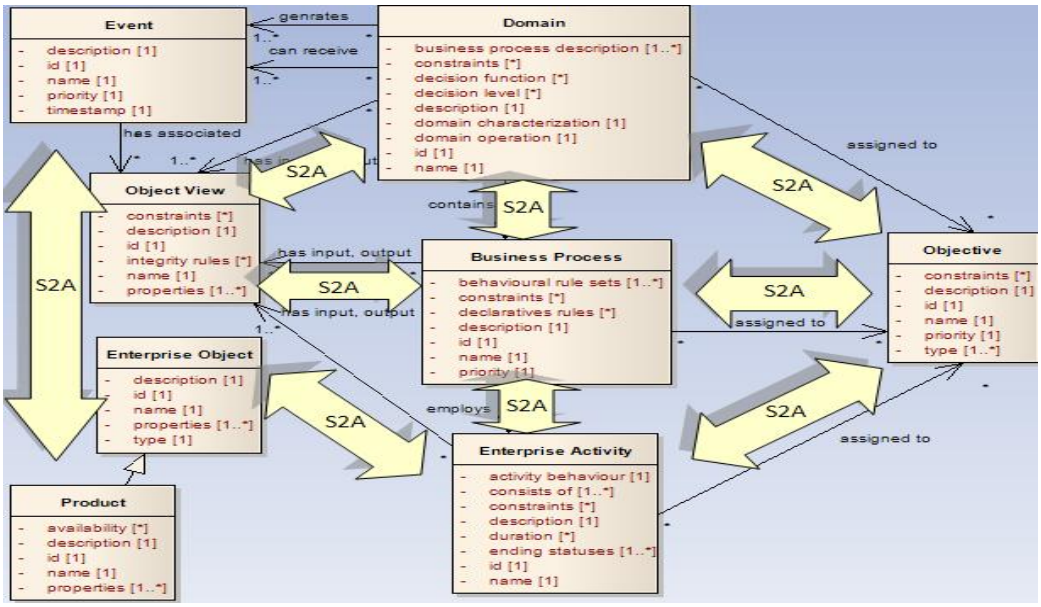


FIG. 2 – E*xemples de S2A appliquées pour la gestion du marché*

3.3.2 Gestion des ressources

L'analyse structurale des produits et ressources permet de les classer en termes de contribution au chiffre d'affaires. C'est une pratique courante basée sur la correspondance entre les produits et les ressources fournie par l'analyse de la matrice structurale (Resource, EO/produit). Nous pouvons étendre l'analyse pour trouver les ressources ou activités liées aux produits du plus haut rang dans les revenus et gains. Ces ressources sont essentielles pour l'entreprise et représentent la plus haute priorité pour l'amélioration des considérations d'investissement. De manière alternative, nous pouvons classer les ressources en fonction de leur utilisation par le nombre de produits/clients. Même si elles ne peuvent pas contribuer à la plus haute fraction de revenus ou de gains, les ressources qui contribuent au plus grand nombre de produits ou de clients peuvent être encore très importantes pour l'entreprise. Une fois que les produits les mieux classés sont identifiés, un point de vue produit/ressource peut être généré pour identifier les ressources utilisées pour produire tels ou tels produits. Cette démarche fournit une information pertinente sur l'alignement « Informationnel/Ressource »

basé, pour notre exemple, sur la correspondance des construits « product » et « ressource » de l'ISO 19440 étendu. Par ailleurs, une analyse structurale peut s'appliquer sur la matrice (Enterprise Activity, Resource) pour dégager les activités les plus consommatrices en termes de ressources et celles à fort impact sur le déroulement des processus métiers mais ne déployant pas énormément de ressources. Une agrégation des données peut servir ici à une répartition des tâches équitable entre collaborateurs, une utilisation optimale du matériel bureautique et une rationalisation des budgets alloués à l'externalisation de certaines activités. D'autres analyses peuvent surgir pour évaluer les compétences des ressources pour la réalisation des activités sensibles et à for impact sur les objectifs assignés. Ces résultats sont issus par exemple des couplages (*Resource/Capability*), (*Resource/Objective*), (Capability, Objective) etc. La figure 3 reprend ces exemples d'analyse structurale pouvant extraire des informations pertinentes en vue d'un alignement Fonctionnel/Ressource, Informationnel/Ressource, Organisationnel/ressource etc.

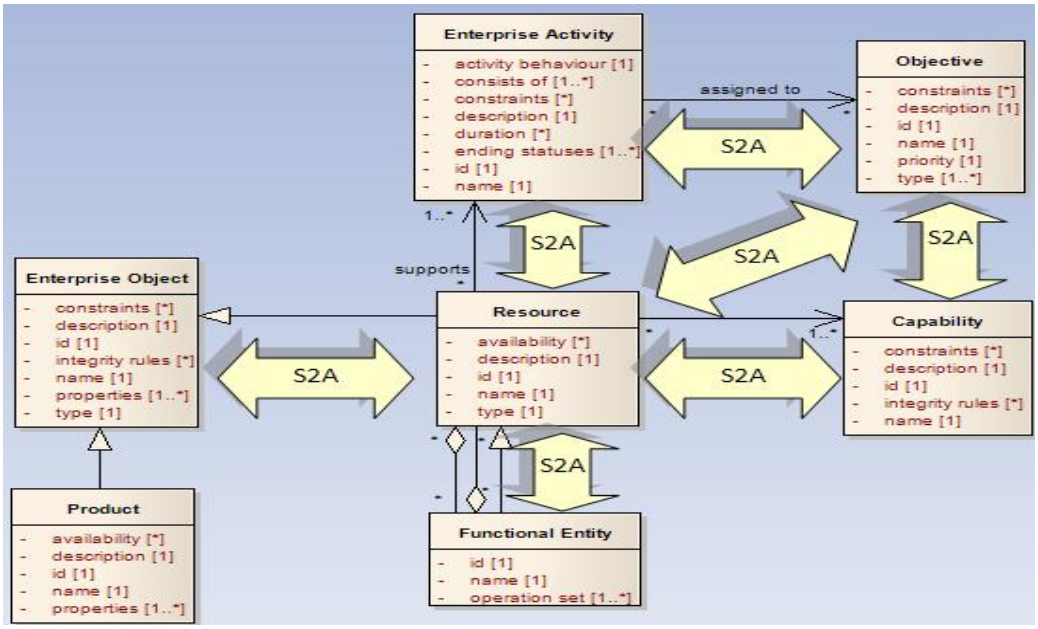


FIG. 3 – E*xemples de S2A appliquées pour la gestion des ressources*

3.3.3 Diagnostic de la performance

L'entreprise est mesurée par un certain nombre d'indicateurs de performance clés (KPI) qu'il faut surveiller minutieusement pour s'assurer de l'exécution efficace des processus/activités. Lorsque l'un ou plusieurs KPI sont jugés insatisfaisants, des mesures correctives sont déclenchées. L'analyse structurale fournit une vue d'ensemble des processus, ressources, activités ou objectifs directement liés aux KPI en question. Dans le cas de notre exemple, la figure 4 montre une vue (Enterprise Activity, Indicator) qui indique quelles sont les activités qui sont mesurés par un KPI pour réaliser un ensemble d'objectifs. Si, par exemple, le temps pour servir un client est trop long, une ou plusieurs activités pourraient en être responsables. Dans ce cadre, nous pouvons généraliser pour relever toutes les compo-

santes responsables de la dégradation des KPI, ensuite prendre les mesures nécessaires. On peut citer, par exemple, les produits en vertu d'un processus sélectionné d'une combinaison Processus-KPI, ou une combinaison Ressource-KPI. Cette analyse multipoints de vue donne un aperçu sur l'ampleur d'une combinaison particulière Processus-KPI, en termes de nombre de produits touchés. Un KPI insatisfaisant associé à de nombreux processus (ou ressource) impactant de nombreux produits est naturellement au cœur d'une réflexion de redressement pour améliorer le niveau d'alignement (Leung et Bockstedt 2009). En outre, les Tableaux de bord de performance communs de l'entreprise montrent des mesures quantitatives de haut niveau relatives aux produits (par exemple, nombre d'unités vendues, le montant des revenus générés, le montant de la marge brute, etc) et aux clients (par exemple, le nombre de clients, nombre de commandes ou montant du chiffre d'affaires par client/région, etc). La sortie d'un tableau de bord de performance métier de l'entreprise peut être combinée avec celui de l'analyse structurale pour montrer les mesures des ressources, acteurs et activités. Ces informations seront utiles dans les reporting périodiques ou surveillance en temps réel. En effet, les « Decision Center » composantes clés de prise de décision dans la structure organisationnelle de l'entreprise, détiennent toute l'information utile cumulée dans des tableaux de bord de performance. Ces connaissances permettent d'évaluer si l'ensemble des constituants de l'entreprise coïncide, ou en harmonie, avec les objectifs stratégiques et opérationnels, en vue d'un alignement multi-vues au sein de l'entreprise.

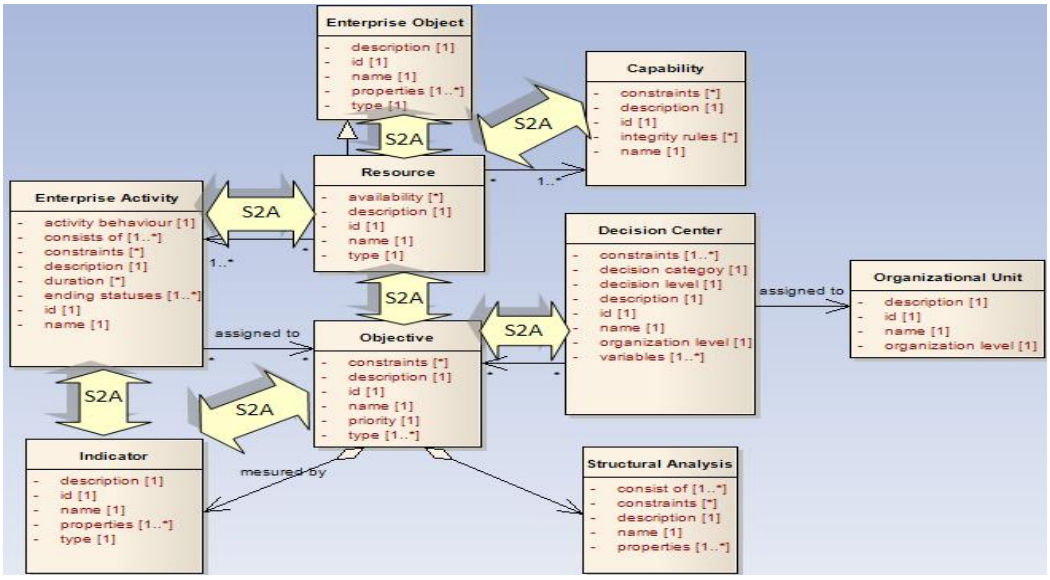


FIG. 4 – E*xemples de S2A appliquées pour la mesure de la performance*

3.4 Architecture

Dans cette section, nous proposons une architecture de mise en œuvre de l'approche S2A qui fournit un outil générique susceptible d'être modélisé par un système informatique en vue d'entamer une évaluation pratique et réaliste de l'alignement de différents composants de l'entreprise. Cette architecture se déroule en trois phases essentielles : une phase de confi-

guration, une phase d'analyse et une phase d'exploitation des résultats (Figure 5). Nous apportons dans ce qui suit une description de chacune de ses phases.

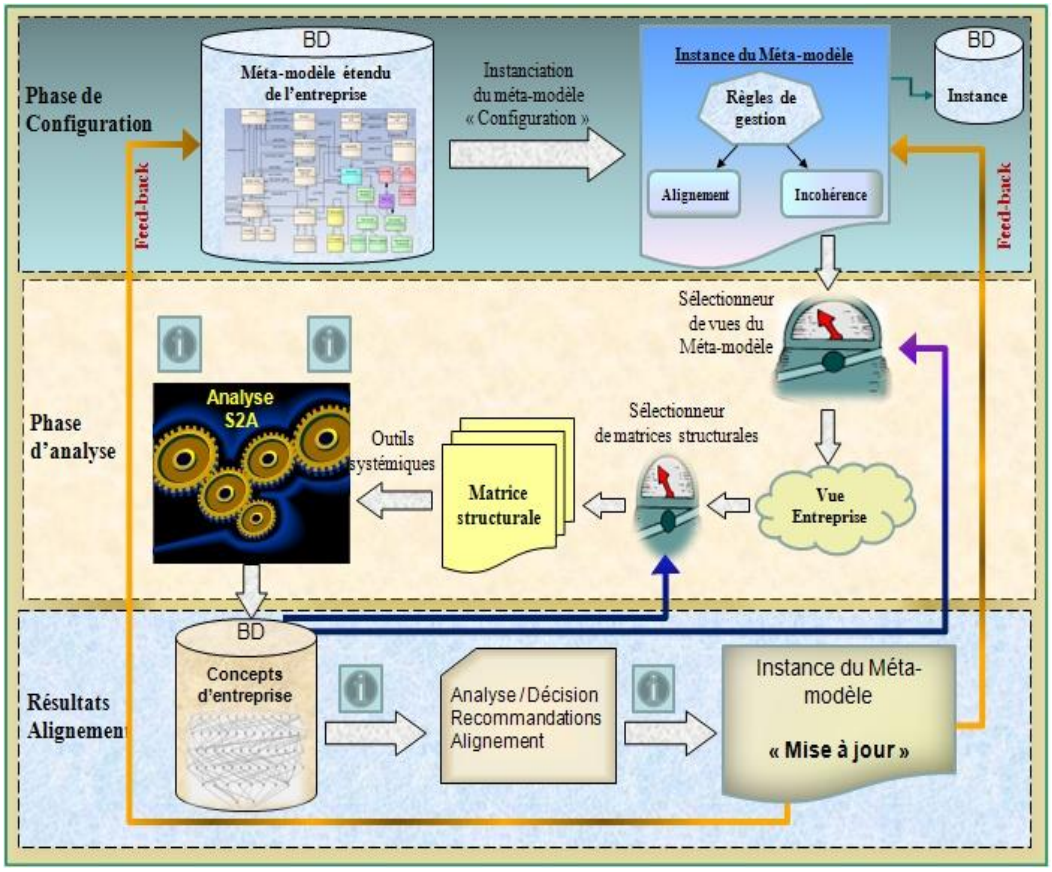


FIG. 5 – *Architecture de mise en œuvre de l'approche S2A*

3.4.1 Phase de configuration

Cette première étape propose la mise en place d'une base de données regroupant toutes les tables issues du méta-modèle. Selon le contexte choisi, l'instanciation du méta-modèle fournit un modèle spécifique qui reflète l'entreprise dans un domaine particulier. Cette instance est ensuite candidat à une configuration du modèle visant à définir explicitement l'ensemble des règles de gestion régissant la notion d'alignement ou incohérence au sein de l'entreprise. En effet, l'identification des paramètres de congruence ou incohérence entre les constituants clés de l'entreprise est incontournable pour toute fin d'alignement. Cela permet d'apporter une réponse ferme à la question fondamentale « Les composantes X et Y sont-elles alignées, oui ou non ? ». Ci-après, un exemple de règles de gestion caractérisant les relations bilatérales liant deux composantes, voire plus, du modèle étudié:

- La réalisation de l'activité A nécessite au plus 2 ressource humaines ;
- L'atteinte de l'objectif O ne doit pas utiliser plus que 3 processus ;

- Le budget des ressources pour la réalisation du processus Pi est limité à 1M DH ;
- La réduction des ressources ne doit pas impacter les indicateurs de satisfaction client ;
- Un processus métier peut déployer au maximum 3 ressources ;

Le résultat de l'analyse structurale des différentes matrices pouvant être générées se renvoie à la définition explicite des règles de gestion prédéfinies pour relever les éventuelles incohérences ou désalignements. Ainsi, nous arrivons à établir une configuration du modèle de l'entreprise à étudier, en vue d'entamer la phase d'analyse et dégager toute incohérence nécessitant des plans d'actions urgents.

3.4.2 Phase d'analyse

Cette étape est à vocation technique permettant d'évaluer l'alignement entre les différents constituants du modèle. En effet, nous procédons à une projection de l'entreprise sur un ensemble de vues afin de réduire la complexité. Nous faisons, ainsi, appel à un outil informatique qui joue le rôle d'un sélectionneur de vues capable de dégager des composantes spécifiques à un contexte donné. Cette alternative permet de générer des vues entreprises selon le domaine visé. Une fois qu'on aura choisi une vue particulière, nous utilisons également un autre sélectionneur qui génère, cette fois-ci, une matrice structurale issue des combinaisons liant deux composantes de la vue sélectionnée. Chaque matrice structurale est soumise à une analyse basée sur des outils systémiques d'ordre treillis susceptible de fournir une information pertinente tirée de l'étude des fermés ainsi générés. L'ensemble de ces treillis est stocké dans une base de données spécifique pour d'autres analyses. Cette phase repose essentiellement sur une analyse structurale susceptible de relever toutes les connaissances utiles aux décideurs pour une meilleure visibilité fonctionnelle et une aide à la prise de décision.

3.4.3 Exploitation des résultats

La base de données comportant l'ensemble des treillis générés à l'issue de l'analyse structurale constitue une ontologie solide permettant d'évaluer l'alignement entre les différentes composantes de l'entreprise. Sur la base des règles de gestion définies dans le modèle initial, nous pouvons, ainsi, détecter d'éventuelles incohérences de part et d'autre de chaque point de vue. Les résultats obtenus offrent une bonne opportunité pour appréhender dans les différentes facettes de l'alignement. Cette technique aboutit finalement à la constitution d'un ensemble de notes et recommandations pouvant aider les hautes instances de l'entreprise à reconsidérer le statut de leur structure moyennant des plans d'action multidimensionnels.

4 Conclusion

Dans cet article, nous proposons une autre vision de l'alignement stratégique basée sur une analyse structurale intégrée formellement dans le Méta-modèle de l'entreprise ISO 19440. Cette démarche constitue une causalité instrumentale pour l'alignement stratégique qui est une partie intégrante de l'ingénierie des SI et se veut une validation empirique de plus en plus convaincante pour en opérationnaliser le concept. L'analyse structurale des couplages liant les différents composants de l'entreprise est susceptible de subvenir aux besoins des stratèges en termes de requêtes et demande d'informations pertinentes permettant une évaluation multipoints de vue de l'alignement stratégique. Cette technique pourrait être utilisée conjointement avec les outils analytiques pour l'organisation, la conception ou la réingénierie des processus.

Références

1. Ettlie J. E., Bridges W.P. et O'Keefe R.D., Organization Strategy and structural Differences for Radical versus Incremental Innovation, (Management Science), Volume 30, pp.682-695. (1984)
2. Ang J., Shaw N. and Pavri F., Identifying Strategic Management IS planning parameters using case studies, International Journal of Information Management, pp. 463-474. (1995)
3. Boulmakoul A., Falih N. et Marghoubi R., Meta-Modelling and Structural Paradigm for Strategic Alignment of Information Systems, Education, Policy and Practice in the Mediterranean Region, University of Economics and Business, Athens-Greece. ISBN: 978-9609-8566-7-6. (2009)
4. Campbell B., Kay R., David Avison, Strategic alignment: a practitioner's perspective, (Journal of Enterprise Information Management) Volume 18(6), pp.653-664. (2005)
5. Ciborra C.U., De profundis? Deconstructing the concept of strategic alignment, (Scandinavian Journal of Information Systems) Volume 9(1), pp. 67-82. (1997)
6. CIGREF, Système d'information éco-responsables: l'usage des TIC au service de l'entreprise durable, (2009)
7. Delone W.H. & McLean E.R., Information system success: The quest for the dependent variable, Information System Research, Vol 3, n°1, pp. 60-95. (1992)
8. Falih N., Boulmakoul A. and Marghoubi R., Déploiement d'une Méta-modélisation Holistique pour l'aide à la prise de décision, ASD'12, Université Saad Dahlab, Blida, Algérie, 1-3 avril, (2012)
9. Henderson J. and Venkatraman N., Strategic Alignment: Leveraging IT for Transforming Organizations (IBM Systems Journal) Volume 32 No. 1, pp. 4-16. (1993)
10. ISO 19440, Enterprise integration, Constructs for enterprise modelling, Edition 1, (2007)
11. Issa-Salwe A., Munir A., Aloufi K. and Kabir M., Strategic IS Alignment: Alignment of IS/IT with Business Strategy, (Journal of Information Processing Systems) Volume 6, No.1, March (2010)
12. Lederer A.L. and Sethi V., Root Causes of Strategic Information Systems Planning Implementation Problems, (Journal of Management Information Systems) Volume 9. N°1, pp. 25-45. (1992)
13. Lederer A.L., Mendelow A.L., Issues in information systems planning, (Information & Management - Inform management) Volume 10(5), pp. 245-254. (1986)
14. Leung and Bockstedt, Structural Analysis of a Business Enterprise, (Service Science 1(3)), pp. 169-188. (2009)
15. Luftman J., Competing in the Information Age: Strategic Alignment in Practice, New York: Oxford University Press, (1996)
16. Luftman J., Lewis P.R. and Oldach S.H., Transforming the enterprise, The alignment of business and information technology strategies, IBM Systems Journal 32(1), 198, (1993)
17. Luftman J., Assessing IT/business alignment (Information systems management) Volume 20(4), pp. 9-15. (2003)
18. Porter M. E., What is strategy? (Harvard Business Review), Volume 74(6), pp.61-78. (1996)
19. Reich B. and Benbasat I., Measuring the linkage between Business and Information Technology Objectives MIS Quarterly March, pp. 55-81. (1996)
20. Silvius G.A.J., Business and IT alignment in theory and practice, Proceedings of the 40th Hawaii International Conference on Systems Sciences, Hawaii, (2007)
21. Smaczny T, Is an alignment between business and IT the appropriate paradigm to manage IT in today's organisations?, (Management Decision) Volume 39(10), pp. 797-802. (2001)
22. Vargas N. Plazaola L., Ekstedt M, A consolidated strategic business and IT alignment representation: A framework aggregated from literature. Proceedings of the 41st Hawaii International Conference on System Sciences, (2008)
23. Ward J. and Peppard J., Strategic planning for IS, 3rd ed. New York:Wiley, (2002)
24. Wilson T, Towards an information management curriculum, (Journal of Information Management) Volume 15, pp. 203-209. (1989)

Optimisation des outils d'aide à la décision par SBML

Dalila Hamami*,Baghdad Atmani**
Equipe de recherche Simulation, Intégration et Fouille de données
Laboratoire d'Informatique d'Oran
*Université de Mostaganem
dhamami8@gmail.com
**Université d'Oran
atmani.baghdad@gmail.com

Résumé. De nombreux travaux théoriques et outils présents sur le marché témoignent l'importance accordée à la fois par la santé publique ainsi que la communauté scientifique au domaine épidémiologique et aux outils décisionnels, qui continue de s'accroitre.
En effet, dans le domaine épidémiologique, les outils de modélisation s'avèrent un moyen très important à l'aide à la décision. Cependant, la variété et le volume important des données et la nature même des épidémies nous conduisent à chercher des solutions pour alléger la lourde tâche imposée à la fois aux experts et aux développeurs.
Dans ce papier, nous présentons une nouvelle approche pour le passage d'un modèle épidémique réalisé en Bio-PEPA en un langage narratif utilisant les bases du langage SBML. Notre but est de permettre d'une part, aux épidémiologues de vérifier et valider le modèle et d'autres part aux développeurs d'optimiser le modèle en question afin d'aboutir à un meilleur modèle de prise de décision. Nous présentons également quelques résultats préliminaires et d'éventuelles suggestions pour améliorer le modèle.

1 Introduction

La biotechnologie a permis, au cours des dernières années, d'améliorer les connaissances sur les agents pathogènes épidémiologiques, et de développer des moyens de lutte efficaces contre ces épidémies. Actuellement, plusieurs épidémies sont en vogue, et les facteurs les développant ont permis de constituer des banques de données énormes (Mansoul, Atmani, 2008). De ce fait, les quantités de données brutes disponibles sont déjà trop importantes pour pouvoir être analysées manuellement par les experts d'une part et les informaticiens qui doivent comprendre le domaine d'une autre part.

Du fait de l'inefficacité des méthodes adoptées due à la variété des données biologiques, et à la nature même des épidémies (Ciocchetta, 2008), une nouvelle approche est utilisée : c'est le développement d'une interface entre expert et informaticien qui n'est plus obligé de démarrer du « tout » pour arriver au modèle « parfait ».

Cette interface permet de convertir le modèle réalisé par l'informaticien à la demande de l'expert en un langage plus compréhensible par l'expert, afin qu'il puisse vérifier la validité

du modèle de par les: paramètres, règles, contraintes…etc. et ainsi de permettre à l'informaticien de revoir et d'optimiser le modèle, pour qu'enfin, toute mise en œuvre de mesures de prévention et de lutte, soient effectuées pour des traitements appropriés.

Le reste de l'article est structuré comme suit. Section 2 présente une brève revue de la modélisation épidémiologique et pourquoi avons-nous besoin de passer aux langages narratifs à partir des modèles informatiques ? Une description de notre modèle en SBML (Bio-PEPA) et comment effectuer sa traduction en langage narratif dans la section 3. Section 4 décrit les détails des informations sur les tests et l'évaluation. La section 5 résume le travail effectué et présente également d'éventuelles suggestions pour améliorer le modèle.

2 Du langage narratif au modèle

Développer et utiliser un bon modèle épidémiologique, reste jusqu'à ce jour une idée très attractive et pour y parvenir de nombreux chercheurs se débattent entre le fait de choisir les meilleurs outils et méthodes ou bien d'effectuer une formation approfondie dans le domaine en question, et bien souvent ils se retrouvent vacillés entre les deux. D'autres au contraire ne donnent guère d'importance ni à l'un ni à l'autre ils préfèrent plutôt économiser leur énergie et adopter une technique tout à fait originale qui est de transformer le contexte exprimé par un expert directement en un modèle simulable. Tel qu'il a été présenté par Georgoulas et Guerriero (2012) pour la traduction du langage narratif en un modèle formel « Bio-PEPA », Guerriero et all (2007, 2009) qui ont étudié la traduction du langage narratif en un modèle « Beta-binders », et « a bio-inspired process calculus », les auteurs sont partis du principe qu'il serait plus judicieux de simplifier la communication entre experts et développeurs en leur offrant une interface simple et conviviale qui permettrait à la fois à l'expert de saisir ses informations et au développeur de manipuler uniquement son code sans trop se soucier de tout comprendre. Cette approche a été baptisée passage du langage narratif au modèle. Bien que ce travail soit considéré comme une grande ouverture dans le domaine de la modélisation, toutefois, en se pose la question, que deviennent les modèles déjà existants?.

3 Du modèle au langage narratif

Afin de pouvoir répondre à la problématique posée dans la section précédente et en s'inspirant du principe définis ci-dessus, nous proposons une approche dont le but est d'une part préserver les modèles existants et d'autre part de les optimiser au mieux et ainsi de permettre une implémentation incrémentale du modèle.

Après une large recherche bibliographique, qui s'est focalisée sur les méthodes de modélisation offrant à la fois, des outils d'analyse et d'aide à la décision, ainsi que la traduction du modèle en d'autres formats spécifiques permettant de se rapprocher du langage narratif, nous avons pu faire ressortir Bio-PEPA (Ciocchetta et Ellavarason 2008), (Hamami et Atmani 2012), qui est un langage formel basé sur les algèbres des processus préconisé pour les systèmes biochimiques et qui a parfaitement été adapté aux domaines épidémiologiques. Au-delà de cette définition, Bio-PEPA est muni d'une extension qui permet de traduire tout modèle en Bio-PEPA en un format XML mieux connu sous l'appellation SBML.

3.1 Bio-PEPA

Bio-PEPA est un outil, méthode et langage basé sur les algèbres des processus. Ces derniers sont décrits par des formalismes mathématiques utilisés dans l'analyse des systèmes concurrents (Ciocchetta et Ellavarason, 2008), (Milner, 1999), (Baeten, 2005) qui sont composés d'un ensemble de processus s'exécutant en parallèles, pouvant aussi être indépendant ou encore partager des tâches communes.

Tel qu'il a été défini dans (Hamami et Atmani, 2012), le langage Bio-PEPA est un 7-uplet (V, N, K, FR, Comp, P, Event), Où :

- V est un ensemble de locations,
- N est un ensemble d'informations auxiliaires,
- K est un ensemble de paramètres,
- FR est un ensemble de taux fonctionnels
- Comp est l'ensemble des espèces,
- P est le composent modèle.
- Event est l'ensemble des événements.

3.1.1 Caractéristiques de Bio-PEPA

Les principales caractéristiques dont est muni Bio-PEPA sont:

- offre une abstraction formelle des systèmes biochimiques et en outre les systèmes épidémiologiques.
- Permet d'exprimer tout type de loi d'interaction exprimée à l'aide de taux fonctionnels.
- Permet d'exprimer l'évolution des espèces et leur interaction.
- Défini par une syntaxe et une sémantique structurelle basée sur une représentation formelle.
- Offre la possibilité d'effectuer différents types d'analyse à partir du modèle (chaîne markovienne à temps continu, les algorithmes à simulation stochastique, les équations différentielles).

3.1.2 Syntaxe de Bio-PEPA

La syntaxe de Bio-PEPA est définie comme suit (Ciocchetta et Guerriero, 2009) :

S:= (α,k) op S:=S ; S:=S+S; S:=C avec

op = ↓ I ⊖ I ⊕ I ↑ I ⊙ And

S::= S S I S(x)

Où, S: décrit les espèces (différents types d'individus); P: le modèle décrivant le système et l'interaction entre les espèces. Le terme (α, k) op S, l'action α est décrite par un taux k et est exécutée par l'espèce, "op" représente le rôle de S. Op={ ↓: acteur, ↑: producteur, ⊕ : activateur, ⊖ : inhibiteur, ⊙ : modificateur}.

3.2 Systems Biology Markup Language : SBML

SBML (The Systems Biology Markup Language) est un langage au format dédié (Ciocchetta et Ellavarason, 2008), (Hucka, 2004), décrivant qualitativement et quantitativement un modèle à la base, biochimique. SBML est basé sur du XML (the eXtensible Markup Language (XML), pour exprimer les données structurées de façon générique. Il permet aux biologistes du système de partager, d'évaluer et de développer des modèles de coopération entre différents formalismes.

Tel qu'il a été défini dans la section 3.1, un modèle épidémiologique est défini dans Bio-PEPA par un ensemble de compartiments, d'espèces ainsi que des réactions décrites par des taux et des paramètres. SBML décrit ces composants par l'utilisation des tags et attributs (Hucka et all, 2007). Tel qu'il a été présenté par Ciocchetta et Ellavarason (2008), la structure d'un code SBML est illustrée dans la figure 1 et détaillée comme suit (Beurton-aimar, 2007):

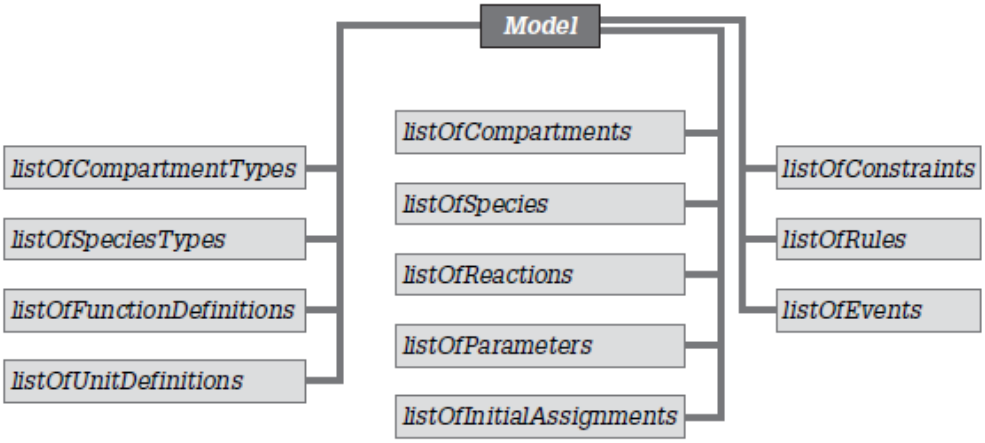


FIG. 1 – ***Organisation générale du langage SBML.***

- Model : il couvre tout les autres niveaux.

<model id="My_Model" > ….</model>

- listOfFunctionDefinitions: cette structure est définie par un identificateur associés une définition de fonction. L'identificateur peut ensuite être utilisé dans tous les éléments subséquents.
- listOfUnitDefinitions : ce type d'unités permettent de spécifier explicitement: les constantes, les conditions initiales, les symboles dans les formules et les résultats des formules.
- listOfCompartments : Représente un espace clos dans lequel les espèces (species) sont localisées.
- listOfSpecies : Permet de spécifier les différentes entités du modèle quelle que soit leur nature, où un type d'espèce « listOfSpeciesTypes » peut être spécifié.
- listOfReactions : tout processus permettant le transfert d'une espèce d'un compartiment à un autre.

La représentation et sémantique des expressions mathématiques sont définies dans le SBML en utilisant le MatHML.

4 Implémentation

Pour implémenter notre approche, nous avons repris le travail que nous avions déjà entamé dans (Hamami et Atmani, 2012) qui consistait à reproduire le principe de la propagation et le protocole de vaccination de la varicelle, tel qu'il est illustré sur la figure 2.

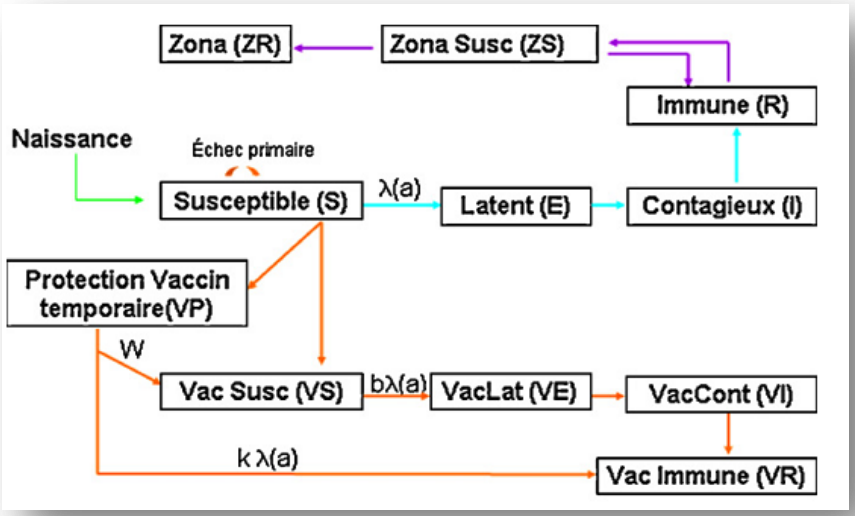


FIG. 2 – ***Structure du modèle (tiré de** Bonmarin, 2008)*.

Le schéma global de notre approche est défini par trois principales étapes :

- Formulation du modèle épidémique en Bio-PEPA : définition des espèces et des réactions.
- Exportation du fichier SBML.
- Représentation en langage narratif : analyse du fichier SBML, affichage d'un rapport détaillé, validation par l'expert.

4.1 Description de la structure du modèle

Notre approche telle qu'elle a été structurée, nous permet de partager notre travail en deux principales étapes, la première est de développer un modèle avec Bio-PEPA (Formulation Du Modèle épidémique En Bio-PEPA), un travail qui a été déjà réalisé (Hamami et Atmani, 2012) et a démontré l'importance de l'utilisation d'un tel outil.

La deuxième partie (Exportation Du Fichier SBML à partir de Bio-PEPA, Représentation du texte SBML en Langage Narratif) est le développement d'un module qui permettrait de traduire le code Bio-PEPA en un langage compréhensible par l'expert, qui pourra facilement vérifier si le contenu du modèle est adéquat à l'exemple et de ce fait le valider.

4.1.1 Formulation Du Modèle épidémique En Bio-PEPA

Pour mieux comprendre le processus de traduction, nous avons repris et explicité dans cette section les plus importantes partie du code Bio-PEPA du modèle de la varicelle (Bonmarin, 2008). (pour la clarté du document, nous n'avons repris ici que quelques exemples).

1. **Location:** dans notre modèle nous avons eu besoin de représenter sept tranches d'âges pour ceci nous les avons représenté sous forme de compartiments.

```
location Age1 in world : size = sizeLeeds, type = compartment;
.........
location Age7 in world : size = sizeLeeds, type = compartment;
```

2. **Les taux fonctionnels**: décrivent les lois d'interactions entre les différents compartiments.

Exposition = λ . S . I; décrit le contact entre un susceptible(S) et un infecté(I) à un taux d'infection λ.

…….

LostVaccin = W . VP; définie le taux de perte de l'immunité(W) d'un protégé par la vaccination(VP).

3. **les espèces:** sont les entités du système décrites par les opérations de leur évolution.

S = [(Exposition,1)↓ S + (Vaccination_1,1) ↓ S + (Vaccination_2,1)↓S;

La figure 3 illustre une partie du code Bio-PEPA, et qui nous permet de constater que même si le langage est aujourd'hui de ce qu'il y a de plus facile à implémenter pour un développeur, reste toutefois, une partie ambigüe face à laquelle est mis l'épidémiologue.

Nous pouvons extraire à partir de cette figure, deux points importants, d'une part la représentation du modèle de la varicelle en Bio-PEPA, et d'une autre part la visualisation des résultats de simulation par un graphe résumant l'état des différentes espèces.

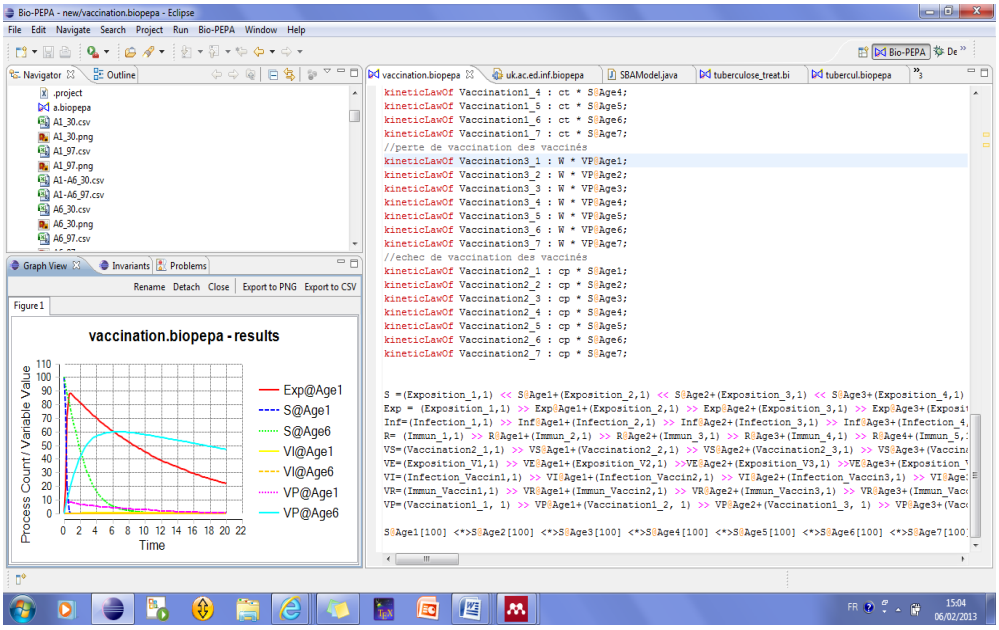


Fig. 3 – ***Vue globale d'un modèle en Bio-PEPA***

4.1.2 Exportation Du Fichier SBML à partir de Bio-PEPA

Bio-PEPA offre la possibilité d'exporter le modèle sous un fichier SBML, par un simple parcours dans son menu.

Tel qu'il est illustré dans la figure 4: (pour une meilleure vus nous n'avons représenté qu'une partie du fichier), le texte ainsi obtenu décrit l'ensemble des tags et attributs tels qu'ils ont été présentés dans la section 3.2, correspondant à notre modèle de la varicelle.

Il est à rappeler que pour étudier une épidémie, nous devons prendre en considération: l'environnement «espace», le temps, et différentes autres fonctions.

SBML permet d'exprimer parfaitement chaque partie décrivant ces éléments définis en Bio-PEPA.

4.1.3 Représentation en langage narratif

Afin de travailler avec du SBML, nous avions besoin d'effectuer une recherche dans la littérature des outils analysant et interprétant ce type de descripteur, cette dernière nous a révélé l'outil **JDOM (**Hunter, 2002**).**

Les principales caractéristiques du modèle DOM sont les suivantes:

- Le modèle DOM (contrairement sur ce point à une autre API fameuse: SAX), représente une spécification qui puise ses origines dans le consortium w3C.
- Le modèle DOM est non seulement une spécification multi-plateformes, mais aussi multi-langages: il existe des liaisons avec Java, Javascript, et d'autres langages encore.

- DOM présente les documents sous forme d'une hiérarchie d'objets, à partir desquels d'autres interfaces plus spécialisées sont elles-mêmes implémentées: Document, Element, Attribut, Text,... Grâce à ce modèle, on peut traiter tous les composants DOM soit par leur type générique, Node, soit par leur type spécifique (Elément, Attribut): de nombreuses méthodes de navigation, permettent ainsi une navigation dans l'arborescence sans avoir à s'inquiéter du type spécifique de composant traité.

```
<?xml version="1.0" encoding="UTF-8"?>
<sbml version="3" level="2" xmlns="http://www.sbml.org/sbml/level2/version3">
<model id= "vaccination_biopepa">
<listOfCompartmentTypes> <compartmentType id="Compartment"/>
<compartmentTypeid="Membrane"/>
</listOfCompartmentTypes>
<listOfCompartments>
<compartmentid="Age7" outside="world" size="100000.0" compartmentType="Compartment"/>
……..
<compartment id="Age5" outside="world" size="100000.0" compartmentType="Compartment"/>
</listOfCompartments>
<listOfSpecies> <species id="Exp_Age1" hasOnlySubstanceUnits="true" substanceUnits="item" compartment="Age1" name="Exp"/>
……
<species id="VS_Age7" hasOnlySubstanceUnits="true" substanceUnits="item" compartment="Age7" name="VS"/>
</listOfSpecies>
<listOfParameters> <parameter id="landa1" value="0.17241"/>
……..
<parameter id="W" value="0.021"/>
</listOfParameters>
…….
```

*Fig.4 – **Codage du modèle sous SBML.***

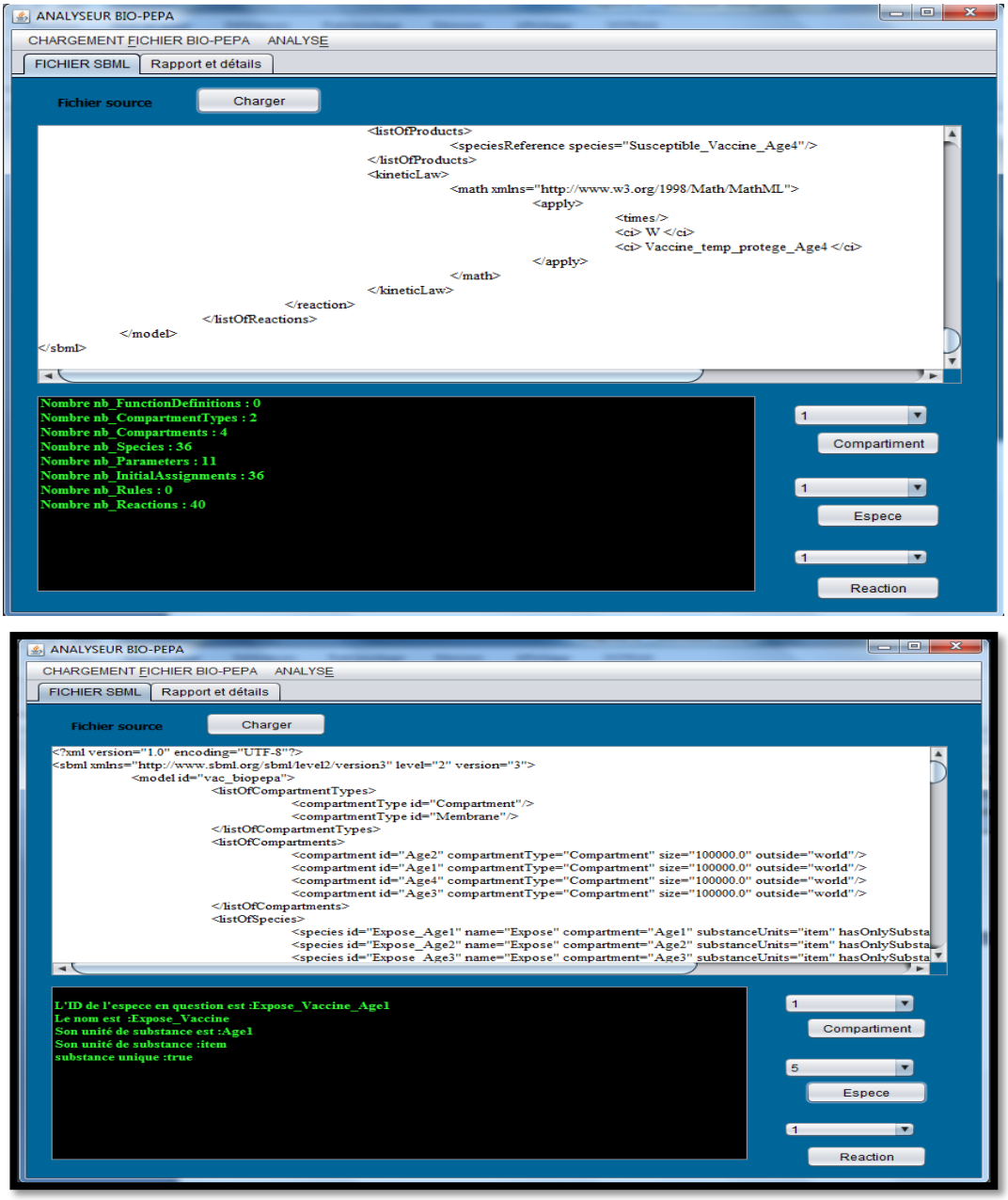


Fig.5 (a, b) – ***Traduction du code*** **SBML en langage narratif**

La figure 5(a,b), illustre l'interface de notre application en se basant sur le modèle JDOM et regroupant ainsi les étapes précédemment définies. L'espace blanc visionné dans la figure correspond au chargement du fichier SBML, quand à l'espace noir, correspond à la traduction et l'analyse du SBML en langage narratif compréhensible par l'expert, de cette façon l'expert n'a aucune difficulté à vérifier la validité du modèle. La convivialité de

l'interface lui permet de surfer sur les différents éléments composant le modèle (espèces, fonction d'interaction, locations…etc)

Afin de valider notre application, nous avons effectué une modification dans le code initiale (Bio-PEPA) où nous avons volontairement causé une erreur de modélisation, la génération de ce dernier, tel qu'illustré dans la figure 6, démontre clairement à l'expert que des espèces et des actions sont manquantes, et de ce fait il pourra facilement les détecter et nous les signaler. (La ligne encadrée en rouge permet de détecter l'erreur, par le nombre d'espèces qui a diminué)

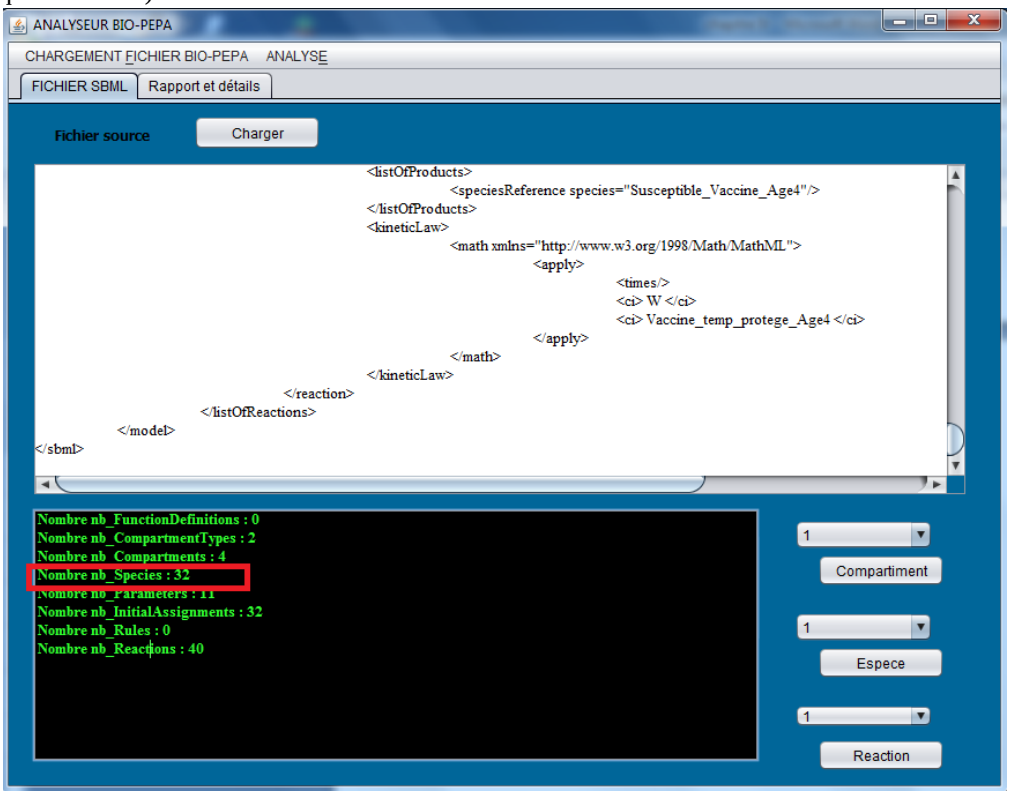


Fig.6 – ***Détection des anomalies après traduction***

5 Conclusion

La modélisation et la simulation sont très utiles pour comprendre et prédire la dynamique des différents phénomènes biologiques. L'approche Bio-PEPA semble être une approche intéressante et un outil puissant pour traiter ce type de problèmes. Grâce à ses différentes caractéristiques elle permet une élaboration facile du modèle informatique et un passage transparent pour les biologistes entre le système réel est celui construit ce qui aide à une représentation fidèle du phénomène étudié. Toutes fois, dans le cas de l'apparition d'un nouvel événement, qui a été mal assimilé par le développeur et donc omis, la correction du modèle est considérée comme une tâche fastidieuse pour les deux. C'est la raison pour la-

quelle, nous avons introduit une interface, où l'expert pourra facilement détecter cette omission et de ce fait revenir vers le développeur, ce dernier pourra discerner l'erreur et la positionner rapidement sur son modèle en Bio-PEPA.
Comme perspective pour le renforcement de ce travail, pourquoi ne pas le rattaché à celui qui a été cité dans la section 2, et ainsi dériver vers un modèle cyclique, qui nécessiterai même pas la présence du développeur, toutefois, après réflexion, que deviendra l'expert devant ses multitudes d'informations, et par quoi va-t-il démarrer pour y arriver ? Après une brève recherche bibliographique l'idée d'intégrer tous ça avec le monde du Data Mining serais une bien meilleure idée à concrétiser.

Références

Baeten J.C.M.(2005). A Brief History of Process Algebras. Theoretical Com- puter Science, Volume 335, Issue 2-3, Pages 131-146.

Beurton-aimar.M (2007). Langage de modélisation des réseaux biochimiques, 1–16, ECRIN-Biologie syst, Chap. 07, Page 7·.

Bonmarin.I, Santa-Olalla.P, Lévy-Bruhl.D(2008), « Modélisation de l'impact de la vaccination sur l'épidémiologie de la varicelle et du zona », Revue d'Epidémiologie et de Santé Publique 56 323–331.

Ciocchetta, F. and M. Guerriero (2009), Modelling Biological Compartments in Bio-PEPA, ENTCS 227, pp. 77–95.

Ciocchetta, F., & Ellavarason, K. (2008). An Automatic Mapping from the Systems Biology Markup Language to the Bio-PEPA Process Algebra.

Georgoulas, A., & Guerriero, M. L. (2012). A software interface between the Narrative Language and Bio-PEPA, 1–9.

Guerriero, M. L., A. Dudka, N. Underhill-Day, J. K. Heath and C. Priami (2009), Narrative-based computational modelling of the Gp130/JAK/STAT signalling pathway, BMC Systems Biology 3, p. 40.

Guerriero, M. L., J. K. Heath and C. Priami, (2007),An Automated Translation from a Narrative Language for Biological Modelling into Process Algebra, in: Proceedings of Computational Methods in Systems Biology (CMSB'07), LNCS 4695, pp. 136–151.

Hamami.D, Atmani.B. (2012). Modeling the effect of vaccination on varicella using Bio-PEPA. Proceeding *Iasted* , 783-077, doi:978-0-88986-926-4

Hucka.M, Finney.A, Bornstein.B.J, Keating.S.M, B.E. Shapiro, J. Matthews, B.L. Kovitz, M.J. Schilstra, A. Funahashi, J.C. Doyle and H. Kitano. Evolving a Lingua Franca and Associated Software Infrastructure for Computational Systems Biology (2004): The Systems Biology Markup Language (SBML) Project, Systems Biology, Volume 1, Pages 41-53.

Hucka.M, Finney.A, S. Hoops, S. Keating and N. L. Novere (2007). Systems Biology Markup Language (SBML) Level 2: Structures and Facilities for Model Definitions. Systems Biology Markup Language, Release 2.

Hunter, J. (2002). JDOM Makes XML Easy.Sun's 2002 Worlwide Java Developer Conference.

Mansoul, A., & Atmani, B. (2009). Fouille de données biologiques : vers une représentation booléenne des règles d'association, In Proceedings of CIIA.

Milner.R. (1999). Communicating and Mobile Systems: the π-calculus. Cam- bridge University Press.

Summary

The importance given by the scientific community and the industrial to the business intelligence continues to grow, as evidenced by the number of theoretical works and tools on the market. Indeed, in the field of epidemiological modeling tools are found to be a very important way to aid in the decision. However, the variety of data and the nature of epidemics lead us to seek solutions to alleviate the difficult task imposed on both the expert and the developer.

We present a new tool for the translation of biological model using SBML from Bio-PEPA into the narrative language. Our goal is to allow in one hand, biologists to verify and validate the model and in second hand the computer scientist to optimize his model, to make a good decision model

We also present some preliminary results. Finally, we suggest some improvements to this work.

Les problèmes de sécurité liés aux architectures de l'entrepôt de données dans le Cloud

Hana Gara Kort*, Jalel Akaichi**

Département d'informatique,Institut supérieur de Gestion Tunis, Le Bardo, Tunisie,2000.
*hannougr@yahoo.fr
**j.akaichi@gmail.com

Résumé. Avec le développement de l'informatique dans les nuages les offres de Cloud BI se sont multipliées. Et comme chaque avancée technologique, il commence à révolutionner le modèle décisionnel. A cet égard, le Cloud Computing apporte également son lot de risque qu'il convient de prendre en compte pour protéger les entrepôts de données de tous les risques. Par conséquent dans le contexte spécifique des architectures décisionnelles dans les nuages, l'objectif de ce papier est d'aborder les problèmes de la sécurité en discutant les différents problèmes de sécurité liés aux architectures de l'entrepôt de données dans le nuage et en analysant les différents scénarios possibles de migrations des fonctionnalités de l'entrepôt de données vers le Cloud.

Mots clés : Entrepôt de données, Cloud Computing, Sécurité.

1 Introduction

La pertinence de l'entrepôt de données pour un système d'aide à la décision a augmenté au cours de ces dernières années, en plus avec le développement de l'informatique dans les nuages ou Cloud Computing, les offres de Cloud BI se sont récemment multipliées et comme chaque avancée technologique, il commence également à s'affirmer dans le domaine de la recherche. Toutefois, les architectures décisionnelles dans les nuages apportent un lot de risques qu'il convient de prendre en compte avant de pouvoir bénéficier de tous les avantages de la solution.

Le Cloud Computing se présente donc aujourd'hui comme une réponse satisfaisante aux problématiques rencontrée par les entreprises. Il résout aussi bien les problématiques liées aux données par sa haute disponibilité, espace de stockage virtuellement infini (Rosenthal and Sciore, 2000). Il propose d'assurer le traitement et l'hébergement de leurs informations numériques via une infrastructure entièrement externalisée. Le Cloud Computing apparait donc comme une opportunité formidable pour les entreprises qui peuvent réduire rapidement le TCO (Total cost of ownership) de leurs matériels et services informatiques (Priebe and Permul, 2000) mais certains commencent pourtant à pointer du doigt la sécurité du système (Harmonium 2010). C'est pourquoi il nous paraît important de proposer une solution qui prend en charge l'aspect sécurité.

Dans le contexte spécifique des architectures décisionnelles dans les nuages, l'objectif de ce papier est (1) : d'aborder les problèmes de la sécurité liés aux architectures de l'entrepôt de données dans le nuage, et(2) : analyser les différents scénarios possibles de migrations des fonctionnalités de l'entrepôt de données vers le Cloud.

Ce document décrit les différents problèmes de sécurité de l'entrepôt de données dans le cloud. Il est organisé comme suit : La section 2 décrit les problèmes de sécurité liés aux architectures des entrepôts de données dans le cloud. La section 3 décrit les différents scénarios de migration des fonctionnalités de l'entrepôt de données vers le cloud. La section 4 énumère quelques unes des solutions actuelles des deux domaines sécurité de l'entrepôt et sécurité cloud computing. La section 6 comporte une conclusion.

2 Les problèmes de sécurité liés aux architectures des entrepôts de données dans le Cloud

2.1 La localité des données

L'abstraction des infrastructures physiques rend la localisation de données spécifiques significativement plus compliquée. Toutefois la dématérialisation touche à ses limites lorsqu'on s'intéresse au lieu où se trouve implanté un site de stockage. Cette incertitude induit une sécurité pas forcément amoindrie mais considérablement complexifiée (Benkemoun, 2010). Le client des infrastructures de Cloud accorde une confiance totale aux prestataires en lui livrant toutes ses données. Le client devient alors spectateur de la sécurisation de ces données et il n'a aucune information : sur quel serveur, dans quel centre et, surtout, depuis quel pays?

2.2 La virtualisation

La virtualisation est l'ensemble des outils et méthodes qui permettent de rendre indépendants l'infrastructure et les services. C'est subdiviser les vastes ressources d'un ordinateur sans garanties d'isolement ou de performance (Barham and al, 2003), d'ou son utilisation pose de multiples problèmes de sécurité avant même qu'elle soit envisagée d'être utilisée pour le Cloud Computing. En ajoutant chaque nouvelle machine virtuelle, il y aura un ajout d'un système d'exploitation supplémentaire, ce qui implique des risques de sécurité supplémentaires d'où chaque système d'exploitation doit être sécurisé, maintenu et contrôlé comme il convient pour son utilisation prévue. La virtualisation doit s'assurer que les différentes instances d'entrepôt de données fonctionnant sur la même machine physique sont isolées les unes des autres ce qui est une tâche très complexe.

2.3 Le choix du modèle de déploiement

L'IaaS est sujette à des problèmes de sécurité posés par les modèles de déploiements grâce auxquels elle est livrée. Disposant de matériel dédié, les Clouds privés peuvent être exploités sur site ou via un fournisseur de services. Ils permettent un contrôle, une personnalisation et une sécurité plus élevés que les Clouds publics qui se composent de ressources informatiques mises à disposition par un fournisseur de services via une infrastructure multi-

locataire ou partagée stimulant les utilisateurs malveillants. Mais c'est le cloud hybride qui propose le plus de flexibilité et d'avantages, en fin de compte (Gartner Group, 2009).

2.4 La multi-location

La colocation consiste en l'hébergement sur le Cloud des applications et données de multiples clients au sein d'une seule et unique infrastructure physique. Le partage de l'infrastructure entre plusieurs clients engendrent des risques accrus et nécessitent un renforcement de la politique de sécurité. En effet, les sociétés-clientes du Cloud veulent être rassurées sur le fait que leurs données et traitements seront bien isolés et protégés des autres environnements hébergés sur l'infrastructure partagée. L'isolation des données sous leurs différentes formes est réalisée au moyen de différents services de sécurité ou techniques de sécurisation, tel que le contrôle d'accès et le chiffrement (Syntec numérique, 2010).

Le nuage est construit sur l'internet et toutes les préoccupations de sécurité en internet sont également posées par le Cloud (Subashini S. and Kavitha, 2011).

3 Fonctionnalités et migration vers le Cloud

Un entrepôt de données inclut des fonctionnalités diverses (Valentin, 2008). Toutefois, la migration vers le Cloud peut poser plusieurs interrogations.

- Quels sont les services de l'entrepôt de données à faire immigrer sur le Cloud ?
- Quels sont les scénarios possibles ?
- Quels sont les risques de l'externalisation de point de vue sécurité ?

3.1 Scénario 1 : One out two in

Dans ce scénario, le service d'intégration réside en dehors des limites du Cloud Computing, dans les frontières de l'entreprise et seulement les deux services de modélisation et d'analyse sont externalisés chez un fournisseur de services en Cloud. La sécurité de l'intégration sera attribuée à l'entreprise et celle des deux autres sera assurée par le fournisseur Cloud. L'entreprise pour avoir une meilleure sécurité de ses données pendant le processus ETL, doit garantir la sécurité de l'environnement et la sécurité du stockage. Donc l'outil ETL doit être examiné et maintenu, puisque certain ETL utilisent par défaut la sécurité du système d'exploitation du serveur ETL pour l'identification des utilisateurs. La figure 1 illustre le premier scénario.

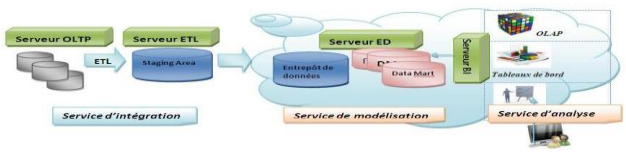


FIG. 1 – *Senario1: One out two in.*

3.2 Scénario 2 : Two out one in

Dans cette architecture le service de modélisation et le service de l'intégration ne sont pas externalisés vers le Cloud, ils sont gérés par l'entreprise elle-même qui aura un contrôle totale et une visibilité absolue sur la manière dont ses données sont stockées. En effet, toutes les opérations d'administration de l'entrepôt de données, création ou suppression de plusieurs entrepôts, lecture et écriture sur entrepôt et accès seront traitées par l'entreprise. D'une autre part, elle doit s'assurer qu'elle obtient les aspects de sécurité, car c'est elle qui va assumer la responsabilité si les choses tournent mal. comme illustré par la figure 2.

Fig. 2 – *Senario2: Two out one in.*

3.3 Scénario 3 : One out one in one out

Pour ce modèle l'entreprise a choisi de confier le stockage de ses données et le service de modélisation, en général, à un prestataire de service Cloud. Ainsi elle s'engage de la phase d'alimentation de l'entrepôt et de la phase de pilotage et d'analyse. Toutefois, les données stockées dans le Cloud ne constituent pas pour autant une connaissance sur l'activité de l'entreprise cliente qui préfère parfois garder son activité confidentielle. Les outils du data mining et de l'analyse multidimensionnelle permettent d'extraire des informations à forte valeur ajoutée à partir des différentes données résidentes dans le nuage afin de faire apparaitre des connaissances utiles pour le développement de l'entreprise.comme illustré par la figure 3.

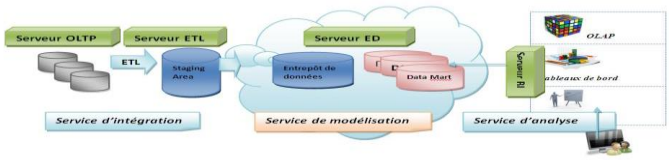


Fig. 3 – *Senario3: One out one in one out.*

3.4 Scénario 4 : All in

Ce dernier scénario illustré par la figure 4 envisage la migration de tout l'environnement décisionnel vers le Cloud Computing, l'entreprise cliente doit dépendre du fournisseur pour les mesures de sécurité appropriées. Le prestataire de service doit faire le travail pour tenir la confidentialité des données et faire face à tout les problèmes de sécurité liés au Cloud surtout la localité de données. En effet, les données de l'entreprise sont stockées dans les Data Center du fournisseur, elles peuvent être répliquées dans plusieurs pays, ainsi que les données des autres entreprises colocataires, par ailleurs, si le fournisseur de services est exploitant d'un Cloud public les risques de violations de données s'aggravent en posant toutes les préoccupations liées à la sécurité sur internet. Cependant, les risques sont extrêmement élevés dans le nuage, la multi location implique que plusieurs instances d'entrepôt de données sont exécutées ensemble.

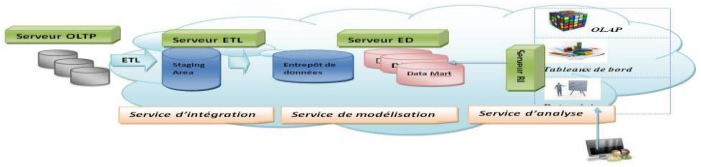


Fig. 4– *Senario4: All in.*

Le fournisseur de service Cloud quelque soit la formule choisie : SaaS, PaaS ou IaaS, et quel que soit le scénario adapté, doit pouvoir offrir une garantie d'une quasi-absolue impossibilité de contourner ou de perdre les données d'où il doit s'assurer de la pleine sécurité dans le nuage, de même pour une entreprise cliente qui a choisi de garder la totalité ou une partie de ses données dans ses frontières ; elle doit mettre en places des techniques et des pratiques de sécurité fiables pour répondre aux exigences de sécurité et garantir sa confidentialité.

4 Etat de l'art

La sécurité dans le Cloud Computing était le sujet de plusieurs travaux de recherche, ce domaine a connu un essor incomparable pendant les dernières années et s'est intéressé au développement des solutions de sécurité pour le Cloud. Il en est de même pour l'entrepôt de données. Mais peux de travaux pour la sécurité de l'entrepôt de données dans le cloud !

Dans (Danwei and H. Yanjun, 2010). les auteurs proposent une stratégie de stockage sécurisée de données applicable au système distribué dans le Cloud Computing capable de combler les lacunes des méthodes traditionnelles de sécurité. Leurs solution divise les données d en k sections, en utilisant l'algorithme de fractionnement de données(data splitting algorithm) ce qui garantit une haute sécurité des données en simplifiant les solutions par des équations k, et dans le même temps, assure la fiabilité en utilisant les coefficients générés par l'algorithme de division. Cependant, cette solution a aussi des défauts tels que la redondance des données.

Des auteurs (Cong and al., 2009). se concentrent aussi sur la sécurité de stockage de données en Cloud. Pour assurer l'exactitude des données des utilisateurs dans le nuage, ils proposent un régime efficace et flexible distribué en utilisant le jeton homomorphique pour la vérification d'effacement des données distribuée codées ; la solution réalise l'intégrité des données stockées ,la localisation des erreurs de données, à savoir, l'identification du mauvais fonctionnement des serveurs et l' efficacité des opérations dynamiques sur les blocs de données.

Smith and Brightwell (1997), présentent une méthode de protection de l'information basée sur un schéma de chiffrement qui conserve le type de données de la source en clair. Ils pensent que cette méthode est particulièrement bien adaptée pour les environnements complexes de données d'entrepôt. Leurs approche de chiffrement consiste en quatre étapes incluant l'attribution d'une valeur à l'indice (index), l'ajout d'une position sensitive au décalage (offset) ensuite mélange de la chaine « string » des valeurs d'index et enfin la conversion d'un retour au type de données souhaité (Sreedhar et al., 2011).

5 Conclusion

L'apparition du Cloud Computing a apporté plusieurs avantages en termes de puissance de calcul, de rapidité d'exécution et de réduction des couts. Avec un tel modèle, les entreprises sont assez satisfaites pour l'entreposage de ces données. Nous avons décrit les défis auxquels font face les organisations qui veulent profiter du nuage pour l'entreposage de leurs données sensibles.

La sécurité dans le Cloud Computing a beaucoup de bouts ce qui fait fuir beaucoup d'utilisateurs potentiels. Sans une solution de sécurité adéquate, les utilisateurs potentiels ne seront pas en mesure de tirer parti des avantages de cette technologie.

Les premières réflexions se sont orientées à qualifier les différentes architectures de l'entrepôt de données en précisant les problèmes de sécurité qui y sont liés. Ce qui nous a amené à discuter les différents scénarios possibles et constater que l'empêchement majeur des entreprises pour l'externalisation de leurs systèmes d'information vers le Cloud est le manque de confiance accordé aux fournisseurs de services, la sécurité des données, la sécurité logique et la sécurité physique.

Le Cloud Computing a été conçu pour traiter de grandes quantités de données donc a un potentiel de régler des problèmes de BI, mais il faut bien convenir qu'il est particulièrement difficile de mettre en place des moyens de parer toutes les attaques.

Références

A. Rosenthal and E. Sciore, "View Security as the Basis for Data Warehouse Security," in Proceedings of the International Workshop on Design and management of Data Warehouse (DMDW'2000), Juin 2000.

A. Benkemoun. (2010, Février) Dangers du cloud computing : la sécurité des données. [Online]. http://www.antoinebenkemoun.fr/2010/02/dangers-du-cloud-computing-la-securite-des-donnees/

C.Danwei and H. Yanjun, "A Study on Secure Data Storage Strategy in Cloud Computing," Journal of Convergence Information Technology, vol. 5, no. 7, September 2010.

H. Smith and M. Brightwell, "Using Datatype-Preserving Encryption to Enhance Data Warehouse Security," in 20th NISSC Proceedings, Octobre 1997, pp. 141–149.

T. Priebe and G. Permul."Towards OLAP Security Design – Survey and Research Issues," in Proceedings of the 3rd ACM international workshop on Data warehousing and OLAP(DOLAP 2000), New York, 2000.

Subashini S. and Kavitha V., "A survey on security issues in service delivery models of Cloud Computing," Journal of Network and Computer Applications, vol. In Press, no. Corrected Proof , 2011.

P. Barham and al, "Xen and the art of virtualization," in SOSP, 2003.

Gartner Group. (2009, mai) France EMC. , http://france.emc.com/collateral/emc-perspective/h8558-cloud-trust-ep.pdf

Syntec numérique. (2010) Livre Blanc. [Online]. http://www.lesechos-conferences.fr/co/catalogue/conferences/medias-ntic/securite-information-numerique-201.html?partenaire

P.Valentin. (2008) Association DotNet France. [Online]. http://www.dotnet-france.com/Documents/SQLServer/BI/Introduction%20%C3%A0%20la%20BI%20avec%20SQL%20Server%202008.pdf

J.Harmonium. (2010, Mars) Master creation. [Online]. http://www.euresis.com/download/brainsfeed_coex11_78_pages.pdf

R. Sreedhar et al., "A Schematic Technique Using Data type Preserving Encryption to Boost Data Warehouse Security," IJCSI, vol. 8, no. 1, Janvier 2011.

W. Cong, W. Qian, and R. Kui, "Ensuring Data Storage Security in Cloud Computing," in 17th International Workshop on Quality of Service, IWQoS 2009, Charleston, South Carolina, USA, Juillet 2009, pp. 13-15.

Summary

With the development of could computing, Cloud BI offerings have multiplied. And like every technological advance, there starting to revolutionize the decision model. In this regard, cloud computing also brings a lot of risk to be taken into account to protect all data warehouse risks. Therefore in the specific context of BI architectures in the clouds, the objective of this paper is to address the security issues in discussing the various security issues related to the architecture of the data warehouse in the cloud and analyzing different scenarios of migration features from the warehouse data to the cloud.

A virtual data integration approach in datawarehouses

Fatima Lahmar Boulçane
Faculty of New Technologies of Information and Communication
University Constantine2
BP 325, Route Ain El Bey 25017 Constantine Algeria
Boulsane_f@hotmail.com
http://www.umc.edu.dz

Abstract. One feature for classification of the approaches to information integration is whether the data are materialized in a Data Warehouse, or else the data are kept in the sources, in which case the approach is called virtual. This paper presents a virtual approach of integration of heterogeneous data which can be used as well for the construction of the global schema of mediation systems as for the datawarehouses which take advantage of the virtual approach of data integration to materialize their data. The main purpose of the approach proposed in this paper is to support and to improve the conception of the mediation schema on which will be materialize the data of the Datawarehouse. This approach is baptized HAV (Hybrid As View). HAV rises from combining the best of Global As View (GAV) and Local As View (LAV) approaches, reducing thus the problem of query rewriting while being easily extensible.

1 Introduction

The notion of datawarehouse is widened today to support the distributed and heterogeneous nature of the distributed data. To be exploitable, all the data resulting from distributed systems must be organized, coordinated, integrated and finally stored to give a global view of the information. The researches in this domain followed closely the offer of the market, by trying to define architecture of reference, to give a good understanding of the concepts, and to propose a formal framework to resolve numerous technical problems such as the integration of data.

An integration system can be characterized by its architecture and its model of integration Calvanese et al (2001), Leonidas et al (2003).

We can distinguish between two fundamental types of architecture for the integration of data. The first is said the mediator approach. It aims to offer a uniform view on a set of heterogeneous sources for end-users or applications. It is based on the definition of transformation rules allowing the translation of requests corresponding to the user request.

The second is the datawarehouse approach. It applies the principle of materialized views and integrates the data in agreement with the global schema. The result is a datawarehouse which can be directly questioned through an adapted language.

The constraints that are typical of datawarehouse applications restrict the large spectrum of approaches that have been proposed for integration for more details see, Calvanese et al (2001, 2011) Inmon (1996), Jarke et al (1999), Bogdan (2008), Xu and Embley (2004). As in mediation systems, the data in the sources and in the datawarehouse can be defined by using the Global As View (GAV) or the Local-As-View (LAV) approaches. The GAV approach requires, for each information need, to specify the corresponding query in terms of the data at the sources. On the contrary in LAV, the source schemas are defined as views over a fixed global schema of the datawarehouse. This makes it easy to add a new source, but query transformation has exponential time complexity. In contrast, the global schema is defined as a view over local schemas in the GAV approach. Here, the query transformation can be reduced to rule unfolding, but the global view must be modified whenever a new source is added.

In this paper, we present a novel approach named HAV (Hybrid AS View) to data integration in a datawarehouse which can be used as well for the construction of the global schema of mediation systems as for the datawarehouses which take advantage of the virtual approach of data integration to materialize their data.

We organize the contributions in this paper as follows:

Section 2 presents the general architecture we use in our approach to data integration in datawarehouses and summarizes the basic notions and concepts with regards to data integration referring to the proposed architecture. Section 3 defines formally the HAV approach and describes our mediation formalism. We conclude in section 4 by recapitulating contributions and discussing possible future developments.

2 Presentation of the HAV approach

The approach suggested in this paper is baptized HAV (Hybrid As View). It is the result of the combination of best of GAV and LAV described above, thus reducing the problem of rewriting of requests while being easily extensible.

In HAV, the schema of the datawarehouse is not directly related to the data sources. Indeed, a set of partial schemas play the role of sources with respect to the datawarehouse schema. A GAV mapping defines the relationship between the datawarehouse schema and the partial schemas. Every partial schema is defined over a single data source type that store materialized data, by means of a LAV mapping. The HAV approach overcomes the drawbacks of both GAV and LAV; this because HAV data integration systems, having a small number of partial schemas, allow for a simple and stable design of the GAV mapping between the global schema and the partial schemas. On the other hand, the LAV mappings are defined on a small number of sources (which are expressed in the same data model). So the query processing does not need to scale and will be more easy. Then, HAV combines the best of the two basic approaches: GAV's simple query reformulation and LAV's scalability.

Figure 1 illustrates the system architecture that complies with the HAV approach where the specialized mediators place themselves in the heart of this architecture. They are considered as virtual sources to be questioned by the global mediator.

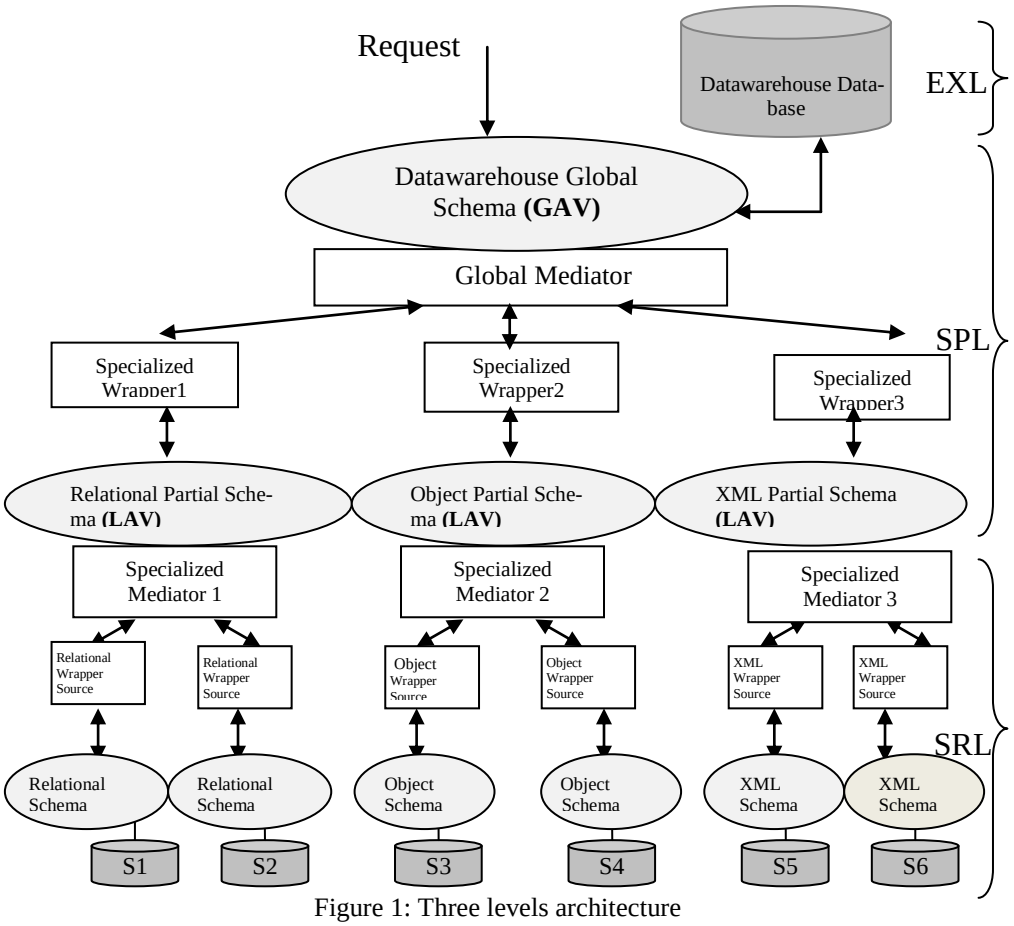


Figure 1: Three levels architecture

As Figure 1 shows, the HAV architecture is made up of three levels: the source level (SRL), the specialized level (SPL) and the external level (EXL).

The SRL: is a set of sources, source wrappers and specialized mediators. The interaction between each specialized mediator and the sources is made easy by software modules, called source wrappers. Every specialized mediator will be concerned by a set of autonomous, independent and homogeneous data sources (in the sense that the source schemas are expressed in the same data model). According to the HAV approach, every partial schema concerning a specialized mediator is built according to the LAV approach

The SPL: This middle-layer is the heart of the HAV mediation architecture. There are two basic components: the global mediator and one specialized wrapper per partial schema. Partial schemas are considered as virtual sources to be requested by the global mediator via the specialized wrappers. The global mediator is built according to the GAV approach.

The EXL: this level permits the interaction of the datawarehouse administrator with the system by querying a datawarehouse schema. The system will carry out the task of dealing with the sources via the partial schemas to retrieve the data satisfying the administrator request. We note that these data serve to materialize the datawarehouse.

3 HAV formal definition

In this paragraph, we present a logical formalism for defining a data integration system based on the HAV approach. For this, we recursively use the definition suggested by Lenzerini (2002) for a data integration system based on a global schema.

3.1 Formal definition

Definition 1: A data integration system I in HAV is a triplet (G, Ip, Mg,p) where:

- G is the global schema expressed in a language L_G, over alphabet A_G. The language L_G determines the set of constraints that can be defined over it.
- Ip, is a set of data integration systems, each one is like a triplet (Sp, S, MSp,S) where:
  - Sp is the partial schema of a specialized mediator expressed in a language L_p on analphabet A_p. The language determines the set of constraints that can be defined over it.
  - S is the schema of the source with the same model as Sp on an alphabet As.
  - MSp,S represents the mapping rules between S and Sp and reciprocally. It is constituted by a set of assertions of the form: $q_{Sp} \rightarrow q_{;} q_s \rightarrow q_{Sp}$

  Where q_{Sp} and q_s are two queries of the same arity, respectively over the partial schema Sp and over the source schema S.
- Mg,p represents the mapping rules between G and S_p and reciprocally. It is constituted by a set of assertions of the form: $q_{Sp} \rightarrow q_{G;} q_G \rightarrow q_{Sp}$

  Where q_{Sp} and q_G are two queries of the same arity, respectively over the partial schema S_p and over the global schema G.

In order to specify the semantics of data integration system according to HAV, we have to specify which data satisfies the global schema. Firstly, we start with a set of data at the sources and specify which data satisfies the corresponding partial schema. We consider a source database for Ip, the database D for the source schema S. Based on D, we have to specify which information content of the corresponding partial schema is. Secondly, we consider a set of data at the partial database (obtained by the execution of LAV on the sources) and specify which data satisfies the global schema. Let a partial database P for the partial schema Sp. Based on P, we have to specify which information content of the global schema is.

Any database for G is called a global database for I. A global database B for I is said to be legal with respect to D via P if:

- B is coherent with G, i.e., every constraint in global schema G is satisfied by B.
- P is coherent with partial schema Sp, i.e. every constraint in partial schema Sp is satisfied by P.
- B satisfies the mapping with respect to P.
- P satisfies the mapping with respect to D.

Definition 2: The semantics of I with respect to D, denoted Sem (I,D), defined on a source database D for I is defined as:

Sem (I,D) = {B | B is a legal global database for I w.r.t P and P is a legal global database for Ip w.r.t D}

3.2 Mapping in HAV

As discussed earlier, the HAV approach combines global and local approaches. In this approach, the specialized schemas are defined independently of the local sources schemas. Each source is described as materialized view of the specialized schema with the same data model. The global schema is defined in terms of the specialized schemas. That is, the global schema is defined as a view over the specialized schemas.

Definition 3: Mappings $M_{Sp,S}$ and Mg,p in HAV approach are in the form:

$M_{Sp,S}$: $S_{Si}(X) \longleftarrow S_1(X_1), S_2(X_2) \ldots S_k(X_J, Z_J)$

Where: $X = \cup_i X_i$, S_i *are relations of partials schemas,* S_{Si} are local relations

$M_{g,p}$: $Gi(X) \longleftarrow S_1(X_1), S_2(X_2), \ldots S_k(X_k)$

Where: $X = \cup_i X_i$, *Gi are global relation,* S_i *are relations of partials schemas*

The mapping $M_{Sp,S}$ is constituted by the LAV correspondences, associating (in the case of relational model) to each source relation a conjunctive query over the corresponding partial schema.

The mapping Mg,p is constituted by the GAV correspondences, associating to each global relation a conjunctive query over the partial schema.

4 Conclusion

Based on the fact that a datawarehouse integrates informations resulting from sources of data which are often heterogeneous and distributed, we proposed an integration approach of heterogeneous and distributed data sources by combining both GAV and LAV. This task includes two problems as the one or other one of the approaches is used: i) the update of the global schema during the addition or during the deletion of a source. ii) the complexity of rewriting of the requests in terms of views

To try to solve in a satisfactorier way these problems relative to the integration of data, we showed in this paper the interest and the feasibility of the combination of GAV and LAV used as well in a virtual integration for the definition of the global schema of a system of mediation as in one datawarehouse based on a virtual integration.

The contribution of the approach HAV is on one side its efficiency and its practicability because the number of partial schemas concerned by GAV is small and stable. On the other hand, HAV is very flexible with the addition (or of the deletion) of data sources to be integrated. Indeed, when a new source is added (or deleted), it will be taken care by the specialized mediator which is concerned by the model of data of this source. The stability of the higher layer and the flexibility of the lower level of the architecture make that the approach HAV is evolutionary when the number of sources increases. Besides, it will be less complex for the specialized mediators to make the sub-requests (because each one has a reduced number of sources to be integrated and these sources are homogeneous), that if the global schema was built itself with the LAV approach; thus the complexity to reformulate the requests is reduced. Indeed, a potentially important characteristic of our approach is that the underlying sources to a specialized mediator are in the same model.

The use of specialized wrappers in our approach allows to benefit from every typology of model of data and guarantees us an optimization of the processing cost relative to the homogeneous nature of the underlying sources. Beyond these first realizations, several aspects

remain to land before the construction of a prototype integrating all the new proposals relative to an integration system built on HAV.

References

Bogdan, A., Chiticariu, Miller, L.R., Tan, W.c (2008), *Muse: Mapping Understanding and design by Example*, International Conference on Data Engineering (ICDE).

Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini and Moshe Y. Vardi (2011), Simplifying schema mappings , International Conference on Database Theory.

Diego Calvanese, De Giacomo, Giuseppe, Maurizio Lenzerini, Daniele Nardi and Riccardo Rosati" (2001), Data Integration in Data Warehousing. Int. J. of Cooperative Information Systems, 10:237-271.

W. H. Inmon (1996), Building the Data Warehouse, 2nd end. (John Wiley & Sons).

M. Jarke, M. Lenzerini, Y. Vassiliou and P. Vassiliadis (1999), Fundamentals of Data Warehouses (Springer-Verlag).

Lenzerini, M., Data Integration (2002), *A Theoretical perspective. Università di Roma "La Sapienza".* Proceeding of the twenty-first ACM SIGMOD-SIGACT-SIGART, Symposium on Principles of Database Systems, pages 233–246.

Leonidas R. J., Galanis, Y.W and DeWitt, D.J (2003), *Locating data sources in large distributed systems*, in the 29th VLDB Conference, Berlin, Germany.

Résumé

Une caractéristique de classification des approches d'intégration des informations est de savoir si les données sont matérialisées dans un entrepôt de données, ou bien elles sont conservées dans les sources, dans ce cas, l'approche est appelée virtuelle. Cet article présente une approche virtuelle de l'intégration de données hétérogènes qui peut être utilisée aussi bien pour la construction du schéma global des systèmes de médiation que pour les entrepôts de données qui tirent avantage de l'approche virtuelle pour la matérialisation de leurs données. L'objectif principal de l'approche proposée dans cet article est de soutenir et d'améliorer la conception du schéma de médiation sur lequel seront matérialisées les données de l'entrepôt. Cette approche est baptisée HAV (Hybrid As View). HAV permet de combiner le meilleur des approches Global As View (GAV) et de Local As View (LAV), réduisant ainsi le problème de la réécriture de requêtes tout en étant facilement extensible.

Service Web pour la fragmentation horizontale des entrepôts de données

Abdelaziz Ettaoufik*, Ladjel Bellatreche**, Mohamed Ouzzif*, Elhoussaine Ziyati*, Hicham Belhadaoui*
*Labo RITM, CED Ensem, Université Hassan II. Casablanca, Maroc
aziztaoufik@hotmail.com, ouzzif@uh2c.ac.ma, ziyati@uh2c.ac.ma, belhadaoui@gmail.com
**LISI/ENSMA – Université de Poitiers. France
bellatre@ensma.fr

Résumé. Un entrepôt de données(EDD) est une collection de données orientées sujet, intégrées, non volatiles et historisées. L'EDD est modélisé suivant différents schémas(en étoile, en flocon de neige, en constellation). Il est alimenté à partir de différentes sources de données via des requêtes transactionnelles et il propose des données analytiques à travers des requêtes décisionnelles. L'EDD doit être capable de gérer la diversité des données, la complexité des requêtes et la charge du travail afin de répondre aux requêtes décisionnelles complexes dont le temps de traitement peut prendre plusieurs heures. Pour améliorer la performance des entrepôts de données, l'administrateur utilise différentes techniques d'optimisation : l'indexation, les vues matérialisées, la fragmentation et le parallélisme. Dans ce papier, nous traitons d'abord les différentes techniques d'optimisation des requêtes dans les EDDs. Deuxièmement, nous présentons des travaux qui traitent l'utilisation des services web dans le domaine des EDDs. Finalement nous proposons une approche de fragmentation horizontale des EDDs en se basant sur la technologie des services web.

1 Introduction

Avec la grande révolution de l'internet, beaucoup de services traditionnels sont décalés vers l'internet. Les services web sont souvent utilisés dans le monde d'internet et d'intranet, ils permettent de partager des informations et de proposer ou d'obtenir des services.

Les auteurs ont déjà parlés de la combinaison des entrepôts de données et des services web, ZHONG et al. (2005) ont proposé un modèle d'entrepôt de données répartie basé sur un service Web, cette renaissance de l'architecture d'entreposage de données est appelée le Data WebHouse, Etienne et al. (2008) ont proposé une architecture orientée service Web pour la constitution de mini-cubes SOLAP pour clients mobiles. Mehedintu et al. (2008) ont traité les étapes de construction d'un entrepôt de données compatible Web.

Dans cet article, nous proposons une nouvelle approche qui propose, à l'administrateur, différents schémas de fragmentation horizontale d'un entrepôt de données en se basant sur la technologie des services web.

Dans la section 2 nous présentons un état de l'art sur les différentes techniques d'optimisation et les différents algorithmes utilisés pour sélectionner un schéma d'un entrepôt de données. La section 3 présente la technologie des services web. Dans la section 4 nous présentons notre stratégie de sélection d'un schéma d'entrepôt de données en utilisant la

technologie des services web. Nous évoquons nos perspectives dans la section 5. Nous concluons finalement ce travail dans la section 6.

2 Techniques d'optimisation des requêtes dans les entrepôts de données

Afin d'améliorer le coût d'exécution des requêtes et le temps de traitement dans les EDD, les administrateurs utilisent différentes techniques d'optimisation à savoir : les index, les vues matérialisées et la fragmentation (Aouiche et al. 2005). La sélection d'une technique d'optimisation constitue l'axe de recherche de plusieurs travaux. Deux types de sélection existent (Bouchakri et al.2009) : (i) sélection isolée qui consiste à implémenter une technique d'optimisation à la fois. (ii) sélection combinée consiste à implémenter conjointement ou séquentiellement deux ou plusieurs techniques en exploitant les dépendances entre eux.

La sélection d'un schéma des EDDs est un problème NP-complet (Bellatreche, 2003), il n'existe pas d'algorithme qui propose une solution optimale en un temps fini. Plusieurs techniques d'optimisation existantes parmi les quelles nous citons: l'optimisation combinatoire, l'optimisation métaheuristique, l'optimisation par apprentissage et l'optimisation évolutionnaire guidée par les connaissances (Pitiot, 2009). Dans les travaux, il ya plusieurs algorithmes et méthodes, nous citons : Colonies de fourmis, Recuit simulé, Recherche tabou, méthode des essaims particulières (Pitiot, 2009). Algorithme génétique, K-Means (Debbat et Bendimerad, 2005), Algorithme glouton (Maiz et al.2007), hill-climbing (Boukhalfa et al. 2008).

L'Indexation. Un Index est une structure de données permettant l'accès direct et rapide aux n-uplets d'une relation volumineuse. L'indexation constitue une option importante dans la phase de conception physique des EDDs (Benameur et Youcef, 2006). Plusieurs types d'index sont utilisés dans le domaine des bases de données. Les index de jointure binaire (IJB) sont les plus efficaces dans les EDDs (Bouchakri et al.2009), ils constituent une combinaison entre l'index de jointure et l'index binaire, Il a été proposé pour précalculer les jointures entre une ou plusieurs tables de dimension et la table des faits dans les entrepôts de données modélisés par un schéma en étoile.

Les vues matérialisées. Une Vue matérialisée est une relation contient le résultat de l'exécution d'une requête, elle est stockée physiquement, elle améliore l'exécution des requêtes en précalculant et en stockant les opérations complexes et les plus couteuses, mais elle introduit le problème de leur maintenance (Bellatreche, 2003)

La fragmentation. Consiste à répartir un ensemble de données d'un EDD en plusieurs partitions disjoints, la combinaison de ces partitions produit l'intégralité des données source, sans perte ou ajout d'information, trois types de fragmentation possible : fragmentation horizontale(FH) consiste à partitionner une table suivant un prédicat de sélection. Fragmentation verticale(FV) permet de partitionner une table suivant une requête de projection. Fragmentation mixte (FM) permet le partitionnement d'une table en combinant la FH et la FV. La FH ne duplique pas les données, elle est devenue un candidat sérieux pour améliorer la performance des EDDs (Boukhalfa et al. 2008), puisqu'elle réduit le temps de stockage et de maintenance. Dans les entrepôts de données en parle de la fragmentation horizontale primaire et la

fragmentation horizontale dérivée (Boukhalfa et al. 2008), la première consiste à fragmenter les tables de dimensions et la deuxième à fragmenter la table des fais suivant les fragments des tables des dimensions.

Similitudes. Il ya des similitudes entre les index et les vues matérialisées : Les deux appartiennent à la structure redondante, partagent la même ressource de stockage, nécessitent des mises à jour régulières. La présence d'un index sur une vue matérialisée peut rendre celle-ci plus « attractive » et vice versa. (Maiz et al.2007)

Les IJB et la FH permettent de répartir les données d'une table de fait suivant la répartition d'une dimension, afin de réduire le coût d'exécution des jointures en étoile (Bellatreche et al.2007). De plus les deux partagent les mêmes attributs de sélection. Ils sont donc sélectionnés de manière combinée. (Bouchakri et al.2009)

3 Technologie des services web

Un service est décrit comme un ensemble de fonctionnalités accomplissant des tâches spécifiques, accessible par un réseau informatique. Les composants d'une architecture orientée service se classent en trois catégories : les *fournisseurs* de services, les *consommateurs* de services et les *médiateurs* (Dubé et al. 2008). Par leurs caractéristiques, les architectures orientées services favorisent la réutilisation de composants, et ainsi la réduction des efforts et du coût de développement de systèmes logiciels complexes. (Dubé et al. 2008)

Un Web Service est un composant logiciel identifié par une URL. C'est une technologie permettant à des applications de dialoguer à distance via Internet ou Intranet indépendamment des plates-formes et des langages sur lesquelles elles reposent, et ce utilisant une norme d'échange de messages basée sur XML à savoir les normes SOAP et WSDL (Dubé et al. 2008). Il s'agit donc d'un ensemble de fonctionnalités exposées sur internet ou sur un intranet, par et pour des applications ou machines, sans intervention humaine, et en temps réel.

4 Problème de sélection d'un schéma de fragmentation Horizontale

La sélection d'un schéma de fragmentation horizontale consiste à (1) déterminer un ensemble de tables de dimension à fragmenter, et (2) répartir les tuples de données d'une table de dimension suivant ces attributs, (3)fragmenter une table des fais suivant les schémas de fragmentation des tables de dimension (Boukhalfa et al. 2008).

Notons que le nombre de schémas de fragmentation possibles d'une table des faits peut être très grand, $N=\prod_{i=1}^{g} mi$ ou mi représente le nombre de fragments de la table de dimension et g représente le nombre de tables de dimension qui ont participées dans le processus de fragmentation. (Boukhalfa et al. 2008). Pour éviter l'explosion de ce nombre, Boukhalfa et al. (2008) ont traité le problème de sélection d'un schéma de fragmentation comme un problème d'optimisation sous contrainte, ils ont utilisé des heuristiques comme le Hill Climbing et le Recuit simulé.

Pour sélectionner un schéma de fragmentation horizontale nous proposons une architecture orientée web service, ce dernier utilise un algorithme heuristique(AH). Les AHs ont été largement utilisés pour la conception physique des bases de données. On peut citer, le pro-

blème d'optimisation des requêtes de jointure, le problème de sélection des vues matérialisées, l'automatisation de la conception physique des bases de données parallèles et la sélection d'un schéma de fragmentation mixte dans les entrepôts de données (Ziyati et al. 2006).

5 Approche proposée

Nous abordons dans ce travail le concept d'amélioration de la performance des EDD à travers un réseau informatique, plus précisément l'utilisation des services web pour sélectionner un schéma optimale d'un EDD.

L'administrateur de l'EDD se connecte au serveur web et invoque le service web dédié. Après l'analyse de la charge de requêtes envoyée par l'administrateur, le service web lui propose un schéma de l'EDD (figure1).

Le problème de la fragmentation horizontale peut être formaté comme suit :

Entrées :

- Un fichier descriptif de l'EDD : EDDXml ;
- Q= {Q1, Q2,..,Qm) ;
- Contraints : NF ;

Intermédiaires :

P= {P1, P2,.., Ps} ;

Sortie :

Schéma de fragmentation horizontale.

Le service web envoie le schéma de fragmentation à l'administrateur.

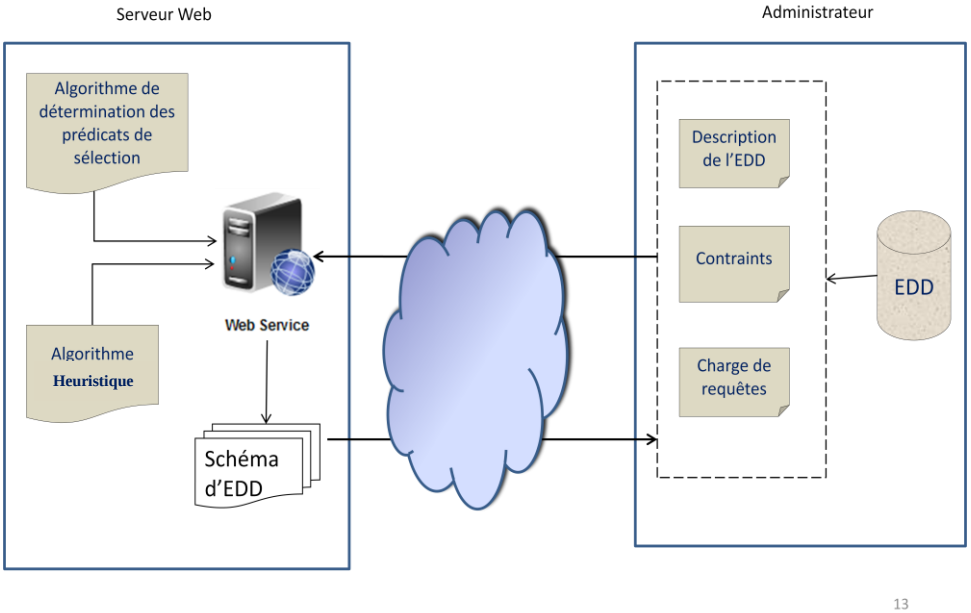


Figure 1 : processus d'invocation d'un Web Service de création d'un schéma d'EDD.

Le fichier EDDXml permet de décrire l'entrepôt de données en XML. Les auteurs ont déjà traité le lien entre l'EDD et l'XML. Mahboubi et Darmont (2008) ont considéré l'entrepôt de données comme un document XML. Ce document contient des métadonnées qui définissent des informations sur le stockage des données, la provenance (URL) des sources des données et les spécifications des vues. Boussaid et al. (2006) ont proposé une approche d'entreposage de données complexes contenues dans des documents XML, appelée X-Warehousing. Celle-ci définit une méthodologie pour concevoir des entrepôts de données complexes à l'aide du formalisme XML. Boukraâ et al. (2006) ont proposé un schéma physique d'un entrepôt XML en proposant une solution pour l'optimisation des performances de ce dernier.

Dans l'approche proposée, afin de proposer un schéma de fragmentation horizontale de l'EDD, le service web doit s'informer sur le schéma de ce dernier, notamment la liste de tables de faits, des tables de dimension, l'ensemble des jointures entre les tables de faits et les tables de dimension, la liste des attributs de chaque table et le nombre des tuplets de chaque table.

6 Perspective

Après une étude détaillée des approches et de différentes techniques d'optimisation existantes, nous proposerons une approche de sélection isolée des index de jointure binaire, des vues matérialisées ou des fragments horizontaux dans les entrepôts de données et ce en utilisant un service web. Nous projetons proposer un service web qui permet une sélection séquentielle et une sélection conjointe de différentes techniques d'optimisation. Notre approche consiste aussi à rendre l'amélioration de la performance des EDD automatique pour faciliter la tâche aux administrateurs.

Références

Aouiche,K, J. Darmont et O. Boussaid (2005). Sélection automatique d'index dans les entrepôts de données. Laboratoire ERIC Université Lumière Lyon 2 Ecole Doctorale de Sciences Cognitives.

Barr.M et L.Bellatreche(2012). Approche dirigée par les fourmis pour la fragmentation horizontale dans les entrepôts de données relationnels. Revue « Nature & Technologie ». janvier 2012.

Bellatreche,L (2003). Techniques d'optimisation des requêtes dans les data warehouse.

Bellatreche L., Boukhalfa K., Mohania M. K.,« Pruning Search Space of Physical Database Design », in 18 International Conference on Database and Expert Systems Applications (DEXA'07), 2007.

Benameur Z et O. Youcef (2010). Vers l'auto-sélection des index dans les entrepôts de données. LIM Université Amar Télidji Laghouat Algérie.

Bouchakri, R, L.Bellatreche et K.Boukhafa (2009). Administration et Tuning des Entrepôts de Données. Ecole Doctorale STIC ESI Alger Algérie.

Boukhalfa K, L. Bellatreche et P. Richard (2008). Fragmentation primaire et dérivée : étude de complexité, algorithme de sélection et de validation. Université de Poitiers.

Boukraâ.O, R.Ben Messaoud, O.Boussaid (2006), Proposition d'un modèle physique pour les entrepôts XML .

Boussaid.O, R.Ben Messaoud, R.Choquet et S.Anthoard (2006). Conception et construction d'entrepôts en XML. Laboratoire ERIC, Université Lyon 2

Debbat.F et Bendimerad.F.T (2005). Les Algorithmes d'Optimisation Globale. 3rd Internatio-nal Conference : Sciences of Electronic. Tunisia.

Etienne.D, T.Badard et Y Bédard (2008). Une architecture orientée service Web pour la cons-titution de mini-cubes SOLAP. Université Laval Québec (Québec) G1V 0A6 Canada.

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

Maiz, N, K. Aouiche et J.Darmont 2007). Sélection simultanée d'index et de vues matérialisées. Université Lumière Lyon 2.

Mehedintu,A, I.Buligi et C. Pîrvu.(2008). Web-enabled Data Warehouse and Data Webhouse. Faculty of Economics and Business Administration,University of Craiova.

Pitiot, P (2009). Thèse Amélioration des techniques d'optimisation combinatoire par retour d'expérience dans le cadre de la sélection de scénarios de Produit/Projet. Université de Toulouse.

Zhong, L, K.Zhang, H.Xia et K.Zhang(2005). The Data Warehouse Model Based on Web Service Technology. Wuhan University of Technology.

Ziyati, E, Ladjel Bellatreche et Driss Aboutajdine (2006). Un Algorithme génétique pour la sélection d'un schéma de fragmentation mixte dans les entrepôts de données. GSCM_LRIT Université Mohammed V-Agdal - Rabat Maroc.

Summary

A data warehouse (DWH) is a collection of information subject oriented, integrated, non-volatile and logged. DWH is modeled following different patterns (star, snowflake, and constellation). It is powered from different data sources through transactional queries and provides analytical data through BI queries. DWH should be able to handle the diversity of data, query complexity and workload to meet the complex BI queries whose processing time can take several hours. To improve the performance of DWH, the administrator uses different techniques of optimization: indexing, materialized views, fragmentation and parallelism. In this paper, we first discuss the different techniques for optimizing queries in the DWH. Second, we present the operation of web services. Finally we propose an approach for improving the performance of DWH based on the combination of optimization techniques and web services technology.

Performances de requêtes OLAP dans les bases de données en colonnes

Khaled DEHDOUH*, Fadila BENTAYEB*
Nadia KABACHI**

*Laboratoire ERIC, Université de Lyon 2
5 avenue Pierre Mendes-France, 69676 Bron Cedex, France
Khaled.Dehdouh@univ-lyon2.fr, Fadila.Bentayeb@univ-lyon2.fr
**Université Claude Bernard-Lyon 1
43, Boulevard du 11 Novembre 1918 - 69622 Villeurbanne Cedex, France
Nadia.Kabachi@univ-lyon1.fr

Résumé. Les bases de données orientées colonnes stockent les données d'une table colonne par colonne contrairement aux bases de données orientées lignes qui stockent les données ligne par ligne. Le stockage de données en colonnes permet ainsi de stocker dans un même bloc disque les valeurs appartenant à une même colonne. Dans le cadre des bases de données multidimensionnelles, ce type de stockage permet de réduire considérablement les accès au disque car le nombre de colonnes lues pour répondre aux requêtes décisionnelles est faible. Dans cet article, en se basant sur des expérimentations effectuées sur un entrepôt de données en étoile, nous montrons que la construction d'un cube OLAP (*On-Line Analys Process*) est plus performante lorsque l'entrepôt de données est stocké en colonnes que lorsqu'il est stocké en lignes.

1 Introduction

Un entrepôt de données présente une modélisation dite dimensionnelle qui se compose classiquement d'une table de faits centrale et d'un ensemble de tables de dimension. Cette modélisation conceptuelle a pour objectif d'observer les faits à travers des mesures (appelées aussi indicateurs), en fonction des dimensions qui représentent les axes d'analyse. Ce modèle est qualifié de modèle en étoile. Les systèmes de gestion de bases de données relationnelles (SGBDR) ont généralement été utilisés pour l'implémentation des entrepôts de données (E.F.Codd, 1970). Cependant, les SGBDR sont confrontés à leurs limites en matière de stockage et d'analyse en ligne de données à grande échelle. C'est dans ce contexte que sont nées les bases de données orientées colonnes. Néanmoins, le respect des propriétés ACID (Atomicité, Cohérence, Isolation et Durabilité) rend difficile et coûteux le passage à l'échelle (X.Borderiel, 2006). Cette situation a favorisé l'apparition du modèle non-relationnel orienté colonnes qui s'inscrit dans la technologie des bases de données NoSQL (Not Only SQL). L'utilisateur peut extraire des cubes de données correspondants à des contextes d'analyse grâce aux opérateurs OLAP. Dans un SGBDR, la construction d'un cube OLAP via une requête décisionnelle nécessite le calcul d'agrégats issus de l'extraction des données à partir des n-upelts

stockés en lignes. Pour réduire le temps d'exécution des requêtes OLAP, des techniques d'optimisation ont été proposées dans la littérature. En opposition, un SGBD orienté colonnes se présente comme une solution alternative au stockage de données en lignes (G.Matei, 2010). Cette technique offre au processus analytique l'opportunité de réduire les accès au disque en accédant uniquement aux blocs contenant les valeurs des colonnes sollicitées par les requêtes décisionnelles (D.J.Abadi, 2008). L'objectif de cet article est de montrer l'avantage d'utiliser les bases de données en colonnes pour le stockage des entrepôts de données relationnels comparées à un stockage en lignes. Nous présentons dans un premier temps le principe de stockage en colonnes. Dans un second temps, nous montrons grâce aux expérimentations que nous avons menées sur un entrepôt de données en étoile appelé SSBM (Star Schema Benchmark) (P.O'Neil et al., 2009) implémenté en colonnes, que la construction d'un cube OLAP est plus performante en termes de temps d'exécution comparé à un entrepôt de données implémenté en lignes. La suite de cet article est organisée de la manière suivante : la section 2 présente un état de l'art des travaux relatifs aux bases de données en colonnes. La section 3 décrit le principe de stockage et de traitement en particulier l'opération de la jointure invisible dans les bases de données en colonnes. La section 4 est consacrée aux différentes expérimentations que nous avons menées ainsi que les résultats que nous avons obtenus. Enfin, nous concluons cet article et présentons quelques perspectives de notre travail dans la section 5.

2 État de l'art

Dans les années 2000 le stockage des données en colonnes est apparu comme une alternative au stockage en lignes et adopté par les SGBDR notamment pour le stockage des bases de données multidimensionnelles. Selon le mode de stockage et le type de moteur d'exécution utilisés, trois approches de bases de données en colonnes sont proposées dans la littérature. La première approche s'appuie sur un SGBDR tout en simulant le stockage en colonnes au-dessus de la couche de stockage orienté lignes. Cette approche utilise l'une des trois techniques suivantes, à savoir, le partitionnement vertical, le plan d'index et les vues matérialisées (D.J.Abadi, 2008). Néanmoins, cette approche connaît des limites et des difficultés à maîtriser les performances du système. La deuxième approche propose un stockage physique des données réellement en colonnes. Toutefois, elle utilise le moteur d'exécution du SGBDR orienté lignes, ce qui nécessite la reconstruction des n-upelts pour pouvoir effectuer des traitements (D.J.Abadi, 2008). Cette approche n'a pas connu de succès mais elle a servi comme étape de transition à la troisième approche. Cette dernière adopte le même mode de stockage que l'approche précédente et applique un réel traitement des données en colonnes (M.Stonebraker et al., 2005). Par ailleurs, cette troisième approche offre de meilleures performances relatives à l'opération de jointure en se basant sur la technique de la matérialisation tardive (D.J.Abadi et al., 2007). En outre, grâce à la technique de compression des données qui agit efficacement sur des valeurs de même type, elle diminue considérablement l'espace de stockage des données sur le disque (D.J.Abadi et al., 2006). C'est pourquoi, cette troisième approche a été adoptée par les éditeurs des SGBD en colonnes tels que *C-Store* [1] *MonetDb* [2] et *Vertica* [3].

1. http ://db.csail.mit.edu/projects/cstore/
2. www.monetdb.org
3. http ://www.vertica.org/

3 Base de données en colonnes

Afin de mieux comprendre le principe du stockage en colonnes, nous présentons tout d'abord la méthode de stockage en lignes. Les n-uplets d'une table relationnelle sont stockés dans un même espace de stockage sous forme de blocs de données. Dans le cadre d'un stockage en lignes, la recherche d'une valeur d'un attribut donné passe nécessairement par la lecture de tout le n-uplet et donc des valeurs des autres attributs décrivant ce même n-uplet. De plus, la lecture se fait par bloc de données qui peut contenir plusieurs n-upelts d'une même table. Ce mode de stockage sollicite donc des ressources mémoires importantes et un coût d'accès au disque élevé. En opposition, une table relationnelle implémentée en colonnes est stockée dans des blocs de données qui contiennent les valeurs des colonnes de la table, colonne par colonne. Lors de la lecture des valeurs d'un attribut, seuls les blocs contenant les valeurs de cet attribut seront chargés en mémoire. Par conséquent, nous avons une meilleure utilisation de ressources, notamment la mémoire. En outre, le fait de regrouper ensemble des valeurs de même type, offre une meilleure compression de ces données.

Dans les bases de données multidimensionnelles stockées selon l'approche relationnelle, la construction d'un cube OLAP se fait grâce à l'opérateur <*CUBE*> qui optimise le temps d'exécution de la requête de construction. Cependant, Il n'existe pas d'opérateur OLAP approprié à la construction de cubes OLAP à partir d'entrepôts de données stockés en colonnes. La construction d'un cube peut alors être généré en combinant des requêtes de regroupement en utilisant "*UNION ALL*", ce qui sollicite d'avantage l'opération de jointure. Cette dernière est la plus coûteuse en matière d'entrées/sorties. Dans les bases de données en colonnes, l'opération de jointure diffère totalement de celle utilisée dans les bases de données orientées lignes. Elle utilise des vecteurs de bits de position pour sélectionner les données. Cette technique est appelée la jointure invisible. Cette opération s'appuie sur la stratégie de la matérialisation tardive (D.J.Abadi et al., 2007). Cette dernière est appelée tardive car elle permet de retarder la construction de n-uplets jusqu'à la fin des traitements et se base sur la manipulation des vecteurs de bits de position qui représentent les positions des valeurs d'attributs dans la table. Cette opération s'effectue selon trois phases. La première phase consiste à appliquer les prédicats sur les tables de dimension. Le résultats est un ensemble de tables de hachage qui contiennent les clés des n-uplets vérifiant les prédicats. Cette opération est effectuée pour chaque condition de la requête. La deuxième phase consiste à identifier, au niveau de la table de faits, les positions des clés de n-uplets des tables de hachage pour générer un vecteur de bits de position globale qui représente les positions des n-uplets qui vérifient l'ensemble des prédicats au niveau de la table de faits. Enfin , la troisième phase consiste à sélectionner les valeurs à agréger et les valeurs d'attributs qui interviennent dans la requête en fonction des nouvelles listes de position.

4 Expérimentations

Pour comparer le temps d'exécution d'un cube OLAP selon les différentes approches que nous avons présentées dans cet article, nous avons procédé à des tests de performance. Nous avons alors implanté l'entrepôt de données en étoile issu du schéma SSBM (Star Schema Benchmark) (P.O'Neil et al., 2009) selon l'approche relationnelle orientée lignes et l'approche relationnelle orientée colonnes. L'entrepôt de données SSBM gère les lignes de commandes en

fonction du produit, du fournisseur, du client et de la date. Il est constitué d'une table de faits appelée "*LINEORDER*" composée de dix sept attributs pour renseigner une commande dont la clé primaire est composée de "*ORDERKEY*" et de "*LINENUMBER*" et des clés étrangères provenant des tables de dimension "*CUSTOMER*", "*SUPPLIER*" et "*DATE*". La première approche est orientée lignes et stocke l'entrepôt SSBM sous forme de base de données relationnelle au sein du SGBD Oracle 11g. Nous avons baptisé cet entrepôt LSSBM. La deuxième approche est orientée colonnes et stocke l'entrepôt SSBM en colonnes sous le SGBD MonetDB. Nous avons baptisé cet entrepôt CSSBM. Les deux implantations LSSBM et CSSBM nous permettent de comparer le temps d'exécution des requêtes OLAP appliquées sur les deux entrepôts respectivement. Pour cela, nous avons mené des expérimentations pour évaluer le coût d'une requête OLAP sur les deux types d'implantation. Cependant, en l'absence d'opérateurs OLAP au sein des SGBD orientés colonnes, et en particulier dans MonetDB, nous avons simulé la construction d'un cube OLAP en utilisant l'union des requêtes de regroupement. A l'inverse, les SGBDR orientés lignes, quant à eux, disposent de l'opérateur <CUBE> pour construire un cube de données. L'environnement de tests que nous avons utilisé consiste en une machine intel-Core TMi5-3550 CPU@3.30 GHZ avec une mémoire "RAM" de 16 Go et un disque dur d'une capacité de stockage de 1.5 To. Cette machine fonctionne avec le système d'exploitation Microsoft Windows 7 de 64bits. Afin de montrer l'avantage d'utiliser les bases de données en colonnes dans le cadre de requêtes décisionnelles, nous avons mené deux expérimentations.

Expérimentation 1 : Dans la première expérimentation, nous fixons la taille des entrepôts de données LSSBM et CSSBM. Nous exécutons ensuite cinq requêtes décisionnelles avec une complexité qui augmente progressivement. Ces requêtes construisent des cubes OLAP en impliquant la table de faits *LINEORDER* et les tables de dimension *CUSTOMER*, *PART*, *SUPPLIER* et *DATE*. Dans cette expérimentation, nous avons mené trois types de tests pour exécuter les cinq requêtes décisionnelles : (1) sur l'entrepôt LSSBM en utilisant l'opérateur CUBE d'Oracle, (2) sur l'entrepôt LSSBM en utilisant l'union de requêtes de regroupement (*UNION ALL*), et (3) sur l'entrepôt CSSBM en utilisant également l'union de requêtes de regroupement. Les résultats que nous avons obtenus sont présentés dans la figure 1.

	Requête 1	Requête 2	Requête 3	Requête 4	Requête 5
ED- Relationnel avec Union ALL	15,01	15,3	15,44	26,88	39,5
ED- Relationnel avec Group by cube	4,08	4,09	4,10	4,14	4,50
ED en colonnes	0,10	0,11	0,15	0,37	1,08

FIG. 1 – *Comparaison des temps d'exécution des requêtes décisionnelles pour la construction d'un cube OLAP, en fonction de l'opérateur de CUBE et de UNION ALL sous Oracle et de UNION ALL sous MonetDB*

Concernant l'entrepôt de données relationnel LSSBM, les premiers résultats que nous avons obtenus sont prévisibles. En utilisant UNION ALL, nous constatons que plus la requête décisionnelle est complexes, plus le temps d'exécution de la requête est élevé. Par contre, avec l'utilisation de l'opérateur CUBE d'Oracle, le temps d'exécution des requêtes décisionnelles est quasi-constant quelque soit la complexité de la requête. En comparant le temps d'exécution des requêtes sur les deux entrepôts LSSBM et CSSBM, nous avons constaté que les bases de données en colonnes présentent des temps d'exécution des requêtes nettement meilleurs que ceux obtenus par les bases de données en ligne alors que ces dernières utilisent l'opérateur CUBE d'Oracle. En effet, L'architecture orientée colonnes de l'entrepôt de données présente une meilleure performance quant à la construction de cube OLAP et ce grâce au mécanisme de la jointure invisible qui se base sur la matérialisation tardive et l'utilisation des vecteurs de bits de position.

Expérimentation 2 : Dans la première expérimentation, nous avons réalisé des tests sur un entrepôt de données de petite taille. Afin de valider les résultats que nous avons obtenus lors de cette expérimentation, nous avons procédé à une deuxième expérimentation pour le passage à l'échelle. Ainsi, nous avons augmenté progressivement la taille des entrepôts de données LSSBM et CSSBM allant de 100 Go jusqu'à 1 To. Ensuite, nous avons exécuté la même requête décisionnelle pour construire un cube OLAP sur les deux entrepôts de données. Les résultats obtenus sont présentés dans la figure 2.

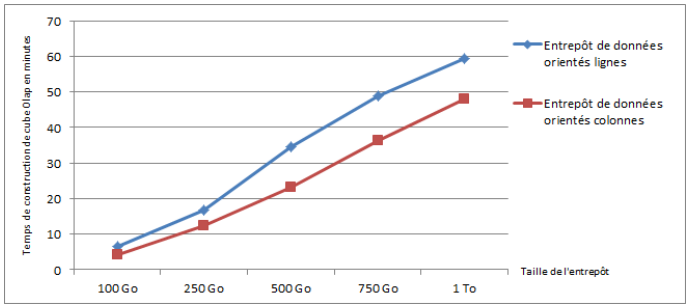


FIG. 2 – *Passage à l'échelle dans la construction de cube OlAP*

Nous constatons que la courbe qui représente le temps d'exécution de la requête de construction de cube OLAP à partir de l'entrepôt de données en colonnes (CSSBM) présente de meilleures performances comparée à celle obtenue par l'entrepôt de données en lignes (LSSBM). Aussi, nous pouvons dire que les bases de données en colonnes conservent leurs performances face au passage à l'échelle. Ceci est expliqué par le fait que dans une opération d'extraction de données pour la construction d'un cube OLAP, les SGBDR orientés lignes sont obligés de charger toutes les colonnes d'une table même celles qui ne sont pas concernées par les jointures. Par conséquent, il en découle une saturation de la mémoire. Cette dernière sollicite le disque pour gérer les résultats intermédiaires ; ce qui augmente considérablement le coût d'entrées/sorties qui se traduit par un temps d'exécution plus élevé et une dégradation des performances du système.

5 Conclusion

Dans cet article nous avons montré clairement, via des expérimentations que nous avons menées, la pertinence d'implanter l'entrepôt de données relationnel selon l'approche orientée colonnes puisqu'elle permet de réduire considérablement le temps d'exécution des requêtes OLAP. Enfin, ce travail ouvre plusieurs perspectives de recherche intéressantes. L'une des pistes de recherche consiste à étudier la possibilité d'intégrer les opérateurs OLAP dans les SGBDR orientés colonnes, en particulier MonetDB, afin d'améliorer le temps nécessaire à la construction de cubes OLAP notamment pour le passage à l'échelle.

Références

D.J.Abadi (2008). Query execution in column-oriented database systems, Massachusetts institute of technology, PhD.thesis.

D.J.Abadi, D.S.Myers, D.J.Dewitt, et S.Madden (2007). Materialization strategies in a column-oriented dbms, ICDE. pp. 466–475.

D.J.Abadi, S.Madden, et M.Ferreira (2006). Integrating compression and execution in column-oriented database systems, SIGMOD conference. pp. 671–682.

E.F.Codd (1970). A relational model of data for large shared data banks, Communications of the ACM 13. pp. 377–387.

G.Matei (2010). Column-oriented databases, an alternative for analytical environment, Database Systems Journal. pp. 3–16.

M.Stonebraker, D.J.Abadi, A. Batkin, X. Chen, M. Cherniack, M. Ferreira, E.Lau, A.Lin, S.Madden, E.O'Neil, P.O'Neil, A.Rasin, N.Tran, et S.Zdonik (2005). C-store: a column-oriented dbms, VLDB 05: Proceedings of the 31st international conference on very large data bases.

P.O'Neil, B.O'Neil, et X.Chen (June 5, 2009). The star schema benchmark (SSBM). http://www.cs.umb.edu/p̃oneil/StarSchemaB.PDF.

X.Borderiel (2006). Les propriétés ACID d'une base de données, Journal du Net.

Summary

Columnar databases store data with columns instead of rows as in relational databases. The major difference between traditional row-oriented databases and column-oriented databases is in regard with the performance and storage requirements. This can be efficiently exploited in the case of data warehouses and OLAP analysis. Hence, we present in this paper our practical research for evaluating the performance of data warehouses with respect to their physical implementations: relational storage and columnar storage. The results we obtained show that building OLAP cubes is more efficient in terms of time execution with column-oriented data warehouse than row-oriented data warehouses.

Vers des entrepôts de connaissances : Définition et architecture

Rim Ayadi*, Yasser Hachaichi**, Jamel Feki*

* Université de Sfax, Laboratoire MIRACL
Route de l'Aéroport km 4, B.P. 1088, 3018 Sfax, Tunisie
Jamel.Feki@fsegs.rnu.tn, rim.ayadi@yahoo.fr
** Université de Sfax, Laboratoire MIRACL
Institut Supérieur d'Administration des Affaires
Yasser.hachaichi@fsegs.rnu.tn

Résumé. Les connaissances constituent une source de pouvoir qui assistent les entreprises à mieux réussir leur processus de prise de décision. Cependant, la majorité des connaissances sont tacites, i.e., enracinées chez les individus. D'autres connaissances, extraites à partir des sources de données sont implémentées dans des systèmes informatiques. En outre, pour un secteur d'activité, ces connaissances concernent une ou plusieurs entreprises. Pour des raisons de veille économique, nous proposons de réunir ces connaissances dans un entrepôt de connaissances. Cet article se veut un tour d'horizon du concept d'entrepôt de connaissances, propose une définition et une architecture et discute quelques perspectives liées à ce concept.

1 Introduction

Les sources de données (SD) telles que les entrepôts de données (ED), les bases de données (BD) relationnelles constituent des gisements pour l'extraction de connaissances. Cependant, leur structure ne permettent pas directement la mémorisation de ces connaissances. Ces connaissances peuvent être explicitées et mémorisées dans des systèmes informatiques dispersés dans des entreprises appartenant à plusieurs secteurs d'activité. D'autre part, il existe des connaissances dites tacites : enracinées chez les individus et enrichies par les expériences partagées et la discussion. Ainsi, l'évolution de l'entreprise dans son environnement et ses nouveaux enjeux externes la pousse à mutualiser et à capitaliser ses connaissances. D'où est née le besoin de la gestion des connaissances (GC) qui est un processus organisationnel destiné à faciliter la création, la mémorisation et l'échange des connaissances (Nemati et al., 2002).

L'entrepôt de connaissances (EC) est une solution pour implémenter et améliorer les phases du processus de GC. Son architecture évoluée va fournir l'infrastructure nécessaire pour capturer et mémoriser des connaissances explicites ainsi qu'améliorer le partage et l'exploitation de ces connaissances pour des activités décisionnelles intelligentes (Rateni et Djebbar, 2009).

Le reste de cet article est organisé comme suit : La section 2 introduit l'état de l'art et l'objectif du processus de GC ; également critique quelques architectures d'un EC. La section

3 propose notre définition d'un EC. La section 4 présente l'architecture de base proposée d'un EC. Enfin, La section 5 conclut cet article et donne les perspectives attendues d'un EC.

2 État de l'art

Le processus de GC peut être défini comme un ensemble de procédures et outils organisationnels, appliqué pour faciliter la création, la mémorisation et l'échange de connaissances entre les individus. Il est destiné à gérer les connaissances existantes et à créer de nouvelles.

Généralement, les personnes sont les principaux dépôts de connaissances, mais ces travailleurs partent et prennent leurs connaissances avec eux (Dymond, 2002). De plus, même si la connaissance est extraite, il y a un problème de savoir qu'elle existe, où elle se trouve, comment la partager avec les autres et sous quelle forme elle est représentée ? D'autre part, les connaissances acquises lors de la collaboration entre les individus ou les connaissances extraites en utilisant des techniques de data mining, permettent d'obtenir des connaissances de divers formats. Ces nouvelles connaissances sont généralement utilisées pour une situation donnée afin de prendre une décision mais elles ne seront pas réutilisées dans des situations semblables. De plus, elles sont dispersées dans plusieurs emplacements sans qu'elles soient liées entre elles et sans l'existence d'un modèle prédéfini de représentation et de mémorisation.

Face à ces insuffisances et visant à satisfaire les exigences mentionnées dans le processus de GC, nous considérons que la collecte des connaissances en vu de leur utilisation rationnelle est une idée prometteuse pour assister les entreprises à profiter au mieux de ces connaissances. L'idée d'un EC est ainsi née. Plusieurs travaux dans la littérature (Kerschberg, 2001), (Dymond, 2002), (Nemati et al., 2002), (Qing-lan et Zhi-jun, 2009), (Irfan et uddin Shaikh, 2010) ont essayé de proposer des architectures de principe pour un EC. Cependant, tous ces auteurs n'ont pas proposé une définition précise ou une architecture détaillée.

Anthony Dymond (Dymond, 2002) considère qu'un EC est semblable à un ED décrit avec un processus à trois niveaux : la capture, le stockage et l'accès au contenu. De plus, il considère que la structure d'un EC est un arbre avec des objets aux noeuds. La plupart des architectures d'EC sont présentées d'une manière abstraite (Kerschberg, 2001) (Dymond, 2002) (Nemati et al., 2002) et (Irfan et uddin Shaikh, 2010). Elles montrent seulement les importantes phases du processus de GC. Ces architectures ne présentent pas l'interaction entre les décideurs et l'EC menant à mettre à jour ce dernier (ce module de mise à jour ne doit pas se limiter à l'insertion ou la mémorisation d'une nouvelle connaissance générée) et à exploiter ces connaissances. De plus, elles ne précisent pas les sources de ces connaissances : est-ce qu'elles existent déjà dans les entreprises, sont-elles extraites à partir des données existantes dans ces entreprises ou bien sont-elles acquises et capturées à travers la reformulation des connaissances tacites des individus travaillant au sein de ces entreprises. Principalement, l'architecture proposée dans la littérature est composée de trois couches (Kerschberg, 2001) (Irfan et uddin Shaikh, 2010) :

- Couche des sources de données : elle renferme des données internes à l'entreprise telles que les documents et les BD et des données externes telles que les services Web ;
- Couche de gestion de connaissances : elle permet la création du « knowledge repository » approprié à l'entreprise, en utilisant divers services tels que les services de data mining, les services de sécurité, les services d'ontologies, etc ;

– Couche de présentation de connaissances : elle permet aux « knowledge workers » d'obtenir des informations personnalisées via un portail. Ce dernier doit supporter la collaboration des utilisateurs afin de combiner leurs connaissances tacites avec celles explicites.

Tenant compte de ces travaux étudiés, nous proposons dans la suite de ce papier notre vision de ce que devrait être un EC. Ainsi, nous proposons une définition et une architecture.

3 Définition d'un EC

En l'absence de définition complète d'un EC et en se basant sur les travaux de (Kerschberg, 2001) (Dymond, 2002) (Nemati et al., 2002) (Rateni et Djebbar, 2009) (Suciu et al., 2012), nous donnons notre définition d'un EC : « Un Entrepôt de Connaissances rassemble des **connaissances explicites** pouvant provenir de **sources multiples**, de **formats hétérogènes** et se rapportant à un ou plusieurs secteurs d'activité. Ces connaissances sont intégrées et organisées pour **supporter un processus de prise de décisions intelligent** ».

Nous commentons dans ce qui suit les termes de cette définition.

Connaissances explicites : Ces connaissances peuvent être codifiées et transférées à travers des méthodes formelles et systématiques (Nonaka et Takeuchi, 1995). Ces connaissances sont découvertes à partir des données ou obtenues par la conversion des connaissances tacites.

Sources multiples : Les connaissances peuvent provenir d'entreprises appartenant à un ou plusieurs secteurs d'activité, prendre leur origine à partir des connaissances tacites des experts travaillant au sein de ces entreprises ou bien être extraites à partir des données moyennant les techniques de data mining. Ces données sont généralement hétérogènes ; elles peuvent être stockées dans diverses entités de stockage telles que : les BD, les ED, les référentiels de messageries électroniques, les référentiels des sites Web, les référentiels média, les tweets, etc.

Formats hétérogènes : Les connaissances explicites peuvent être exprimées formellement sous plusieurs formes tels que : des règles d'association, des arbres de décision, des réseaux bayésiens, des réseaux de neurones, des groupes de données homogènes avec leur caractéristiques, des règles de classification, etc. Un EC doit tenir compte de cette hétérogénéité soit en standardisant la représentation des connaissances, soit en les représentant de manières différentes et prévoir des règles de transformation permettant des passages inter-représentations.

Supporter un processus de prise de décisions intelligent : Les connaissances de l'EC seront exploitées par les décideurs pour trouver des solutions menant à résoudre les problèmes auxquels ils sont confrontés (Nemati et al., 2002) et prendre les décisions convenables dans des situations données. Ces connaissances seront réutilisables dans d'autres situations semblables. De plus, elles permettent d'augmenter les capacités cognitives des décideurs afin de convertir leurs connaissances tacites en de nouvelles connaissances explicites.

S'appuyant sur cette définition, nous présentons, dans la section suivante, une architecture de base pour un EC.

4 Architecture de principe d'un EC

La figure 1 propose l'architecture de base pour un EC. Dans cette architecture, on trouve :

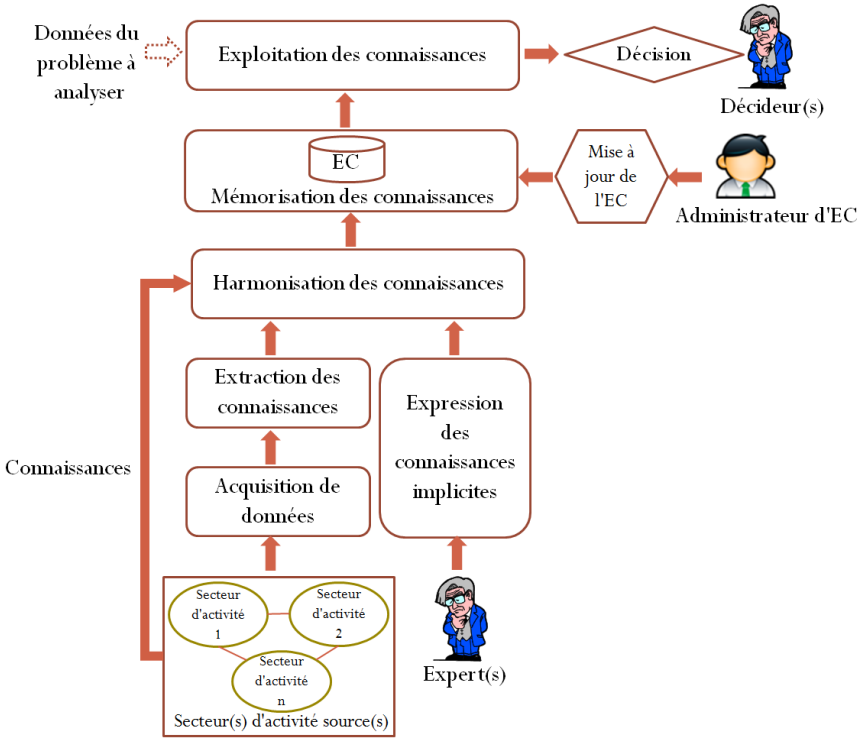


FIG. 1 – *Architecture de principe d'un EC.*

Secteur(s) d'activité : Selon la définition de l'Institut national de la statistique et des études économiques (Insee)[1] : « Un secteur d'activité regroupe des entreprises de fabrication, de commerce ou de service qui ont la même activité principale. L'activité d'un secteur n'est donc pas tout à fait homogène et comprend des productions ou services secondaires qui relèveraient d'autres items de la nomenclature que celui du secteur considéré ». De ce fait, nous nous intéressons à l'entreprise ainsi qu'à toutes les instances qui interagissent avec (i.e., concurrents).

Secteur(s) d'activité sources : Les entreprises d'un secteur d'activité constituent principalement les SD ou les sources de connaissances que l'on vise à utiliser afin de construire l'EC. Naturellement, comme il peut y avoir des relations entre les entreprises de différents secteurs ayant des activités proches alors l'élargissement de l'EC à des connaissances issues de ces secteurs ne peut qu'enrichir l'EC et, par conséquent, améliorer la prise de décisions.

Acquisition de données : C'est la couche qui s'intéresse à collecter l'ensemble des données initiales issues des entreprises. Ces données peuvent être stockées dans plusieurs entités de natures différentes (Kerschberg, 2001) telles que : les BD, les ED, les fichiers, etc.

Extraction des connaissances : Cette couche convertit les connaissances cachées dans les données initiales en des connaissances explicites exprimées formellement et transcrites dans des modèles appropriés. Ces dernières sont obtenues en utilisant plusieurs techniques d'extrac-

[1] http ://www.insee.fr/fr/

tion de connaissances (exemples : les techniques data mining telles que : les règles d'association, les arbres de décision, la segmentation, etc (Silwattananusarn et Tuamsuk, 2012)).

Connaissances implicites (tacites) : Les experts doivent pouvoir être guidés par des modèles de connaissances afin d'exprimer leurs connaissances implicites en connaissances explicites dans des formats exploitables par les décideurs et par un processus informatique.

Harmonisation des connaissances : Les connaissances explicites extraites à partir des données, les connaissances implicites des experts formalisées ou les connaissances existantes déjà dans le(s) secteur(s) d'activité sources ont des formats hétérogènes (e.g., règles de production, arbres de décision, modèles) (Nemati et al., 2002) (Suciu et al., 2012). Ces connaissances ont besoin d'être codifiées, cataloguées et homogénéisées avant d'être mémorisées.

Mémorisation des connaissances : Les connaissances harmonisées seront rassembler et organiser dans une entité de stockage (EC) visible et accessible à la fois aux êtres humains et aux machines (Dymond, 2002). Ainsi, les connaissances seront diffusées au sein d'une même entreprise ou entre plusieurs entreprises. Ces connaissances peuvent être mises à jour par l'administrateur de l'EC suite à des évolutions survenues au cours du temps sur les SD initiales et sur les connaissances tacites des experts ou suite à l'acquisition de nouvelles connaissances.

Exploitation des connaissances : Cette couche joue le rôle d'une interface utilisateur qui facilite l'accès à l'EC via des services appropriés. Ces services sont offerts pour faciliter le processus de prise de décisions. Les travailleurs du savoir peuvent collaborer afin de combiner leurs connaissances implicites avec les connaissances explicites de l'EC et ceci afin de résoudre des problèmes. Ils peuvent créer de nouvelles connaissances, effectuer des recherches pour obtenir des informations (Kerschberg, 2001), etc. À ce stade, l'expert peut utiliser les données du problème à analyser pour sélectionner la ou les connaissance(s) de l'EC à appliquer pour une situation donnée afin de prendre une ou plusieurs décision(s) qui peuvent être cohérentes ou incohérentes. Ainsi, il faut utiliser un autre indicateur de sélection de décisions convenables aux situations confrontées, autre que celui utilisé pour la sélection de connaissances.

Décision : C'est l'acte par lequel le décideur choisit les actions à entreprendre pour une situation donnée. Après avoir pris sa décision et après l'avoir évaluée, si le décideur est convaincu par cette décision alors il la valide, sinon il peut informer l'administrateur de l'EC pour mettre à jour les connaissances qui ont mené à des décisions invalides.

5 Conclusion et perspectives

Les connaissances acquises des individus ou extraites en utilisant des techniques automatisées sont utilisées pour trancher sur une situation donnée. Dans cet article nous avons proposé la définition et l'architecture d'un EC. Un EC rassemble des connaissances provenant de sources multiples, de formats hétérogènes et se rapportant à un ou plusieurs secteurs d'activité. Ces connaissances sont organisées pour supporter un processus de prise de décisions intelligent. En s'appuyant sur cette définition, nous avons proposé une architecture pour les EC. Cependant l'implémentation d'un EC soulève plusieurs défis à résoudre : Comment convertir les connaissances implicites en des connaissances explicites et vers quel modèle seront-elles converties ? Quelles sont les techniques d'extraction à appliquer sur les SD à explorer (i.e., indicateur de sélection de techniques) ? Avec la diversité des formats des connaissances extraites à partir de données ou exprimées par les experts, comment harmoniser et regrouper les connaissances qui peuvent se présenter d'une manière divergente ? Quel est le meilleur modèle

permettant la mémorisation et l'exploitation future des connaissances ? Quels sont les indicateurs permettant de sélectionner les connaissances convenables permettant la résolution d'un problème donné ? Quels sont les indicateurs permettant de sélectionner la meilleure décision dans le cas où une multitude de choix se présente ? Quels sont les opérateurs nécessaires pour maintenir la structure d'un EC ? Comment gérer l'évolution des connaissances de l'EC ?

Références

Dymond, A. (2002). The knowledge warehouse : The next step beyond the data warehouse. In *Data Warehousing and Enterprise Solutions*. SAS Users Group International 27.

Irfan, R. et M. uddin Shaikh (2010). Enhance knowledge management process for group decision making. In *Proceedings of the 2010 Second International Conference on Computer Engineering and Applications - Volume 01*, ICCEA '10, pp. 66–70. IEEE Computer Society.

Kerschberg, L. (2001). Knowledge management in heterogeneous data warehouse environments. In *International Conference on Data Warehousing and Knowledge Discovery*, pp. 1–10. Springer.

Nemati, H. R., D. M. Steiger, L. S. Iyer, et R. T. Herschel (2002). Knowledge warehouse : an architectural integration of knowledge management, decision support, artificial intelligence and data warehousing. *Decision Support Systems 33*(2), 143 – 161.

Nonaka, I. et H. Takeuchi (1995). *The Knowledge-Creating Company : How Japanese Companies Create the Dynamics of Innovation*. New York : Oxford University Press.

Qing-lan, H. et H. Zhi-jun (2009). Research on cost control dss based on knowledge warehouse. In *Proceedings of the 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery - Volume 07*, pp. 357–361. IEEE Computer Society.

Rateni, I. et B. Djebbar (2009). De l'ingénierie de connaissances à l'aide à la décision. In *Proceedings of the 2nd CIIA'09*, Volume 547 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Silwattananusarn, T. et K. Tuamsuk (2012). Data mining and its applications for knowledge management : A literature review from 2007 to 2012. *International Journal of Data Mining & Knowledge Management Process (IJDKP) 2*(5).

Suciu, I., C. Fernandez, et A. Ndiaye (2012). How to acquire scientific knowledge for university to industry knowledge transfer. In *Proceedings of the Fourth IEEE International Conference on Information Process and Knowledge Management*, eKNOW '12, pp. 24–27.

Summary

Knowledge is a source of power which helps companies to improve their decision making process and then deal with more competitive situations. However, most of knowledge are tacit, i.e., rooted in individuals. Other knowledge are extracted from data sources and implemented in computer systems. In addition, for a sector of activity, this knowledge interests several companies. For business intelligence purposes, we suggest to gather all sector knowledge into a knowledge warehouse. This paper suggests a definition and an architecture for the knowledge warehouse and it discusses also some relevant perspectives related to this concept.

Etude méthodologique de l'intégration de l'analyse multicritère aux systèmes OLAP: Modèle multidimensionnel

Omar Boutkhoum*, Mohamed Hanine*, Abdessadek Tikniouine*, Tarik Agouti**

*Laboratoire d'Ingénierie des Systèmes d'Information, Faculté des sciences Semlalia, Université Cadi Ayyad, Marrakech -Maroc
o.boutkhoum@uca.ma, m.hanine@uca.ma, tikninouine@uca.ma
**Equipe de Télécommunication et Réseau Informatique, Faculté des sciences Semlalia, Université Cadi Ayyad, Marrakech - Maroc
agoutitarik@gmail.com

Résumé. Les systèmes d'aide à la décision sont basés sur des technologies décisionnelles permettant aux organisations de faire un meilleur usage de leurs flots de données en simplifiant l'analyse de celles-ci. En effet, les outils OLAP (On Line Analytical Processing), comme technologies décisionnelles, offrent la possibilité d'archivage, de gestion, d'analyse et de modélisation multidimensionnelle. Cependant, ils sont limités au niveau de la prise en considération de l'aspect multicritère et qualitatif du problème décisionnel.

Pour surmonter ses limites, nous avons proposés dans cette recherche, une approche méthodologique d'intégration de l'analyse multicritère aux systèmes OLAP. Notre but visé de cette approche est de créer un modèle hybride de données (OLAP/AMC), ce qui va permettre d'élargir les capacités d'OLAP à travers l'apport de l'AMC comme outil d'analyse multicritère.

1 Introduction

OLAP est basé sur l'approche des bases de données multidimensionnelles, qui introduit des concepts qui diffèrent des concepts reliés aux bases de données communément utilisées. En effet, OLAP dispose d'un ensemble d'opérations élémentaires par rapport au processus transactionnel OLTP (On Ligne Transaction Processing), permettant une réorientation de points de vue (rotate/pivot, switch, split …) selon différentes dimensions de la vue multidimensionnelle d'un cube OLAP (Gray, et al., 1996) , et assurant aussi l'hiérarchisation de l'information en différents niveaux de détail (Granularité : roll up, drill down). L'agrégation des données en tant que notion fondamentale d'OLAP, contrôle et optimise les temps de calcul via ses opérations de somme et de moyenne. Ce qui minimise le temps d'accès à l'information et assure une exploitation facile à des données immenses de l'entrepôt de données. Cependant, jusqu'ici nous avons constatés d'une part, que cet outil ne traite que des valeurs mesurables et manque de critères qualitatifs, d'autre part, il ne permet pas la prise en considération de plusieurs critères simultanément.

Afin de remédier à ce problème, les méthodes multicritères apparaissent comme étant l'outil le plus adéquat à travers leurs capacités d'analyse multicritère qui prend en considération l'aspect qualitatif des données.

2 Apport de l'analyse multicritère d'aide à la décision

L'analyse multicritère est une méthode d'analyse visant la résolution progressive du problème de décision selon plusieurs critères, souvent contradictoires (Roy, 93). Un critère est une expression qualitative ou quantitative permettant d'apprécier des options ou des scénarios (Roy, 93) (Vinck, 89). Elle est utilisée pour porter un jugement comparatif entre des projets ou des mesures hétérogènes. Les données de la problématique auquel fait face l'analyse multicritère sont caractérisées par des indicateurs mesurables et qualitatifs par rapport à l'analyse OLAP. Elle est plus particulièrement utilisée dans l'élaboration des choix stratégiques d'intervention. La contribution de chaque critère dans le résultat final se fait par intrusion de différentes pondérations dans le processus d'évaluation, ce qui permet d'attribuer un poids d'importance selon l'effet de chaque critère sur l'action à choisir. Mais le manque que nous avons remarqué au niveau de cette analyse réside dans le temps, souvent lent, de réalisation d'un tel projet, et aussi le problème de disponibilité des données pendant tout le processus d'analyse.

3 Problématique

Comme déjà expliqué dans les sections I et II, l'implémentation unique d'une des deux analyses (OLAP et AMC); surtout avec l'impact de leurs limites, implique une faiblesse directe au niveau du poids de la décision à prendre.

Le choix d'une analyse multidimensionnelle traduit ici par l'utilisation de l'outil OLAP, assure la gestion des données détaillées d'un entrepôt de données, et y facilite l'accès en utilisant les axes d'analyse dans la structure des cubes.

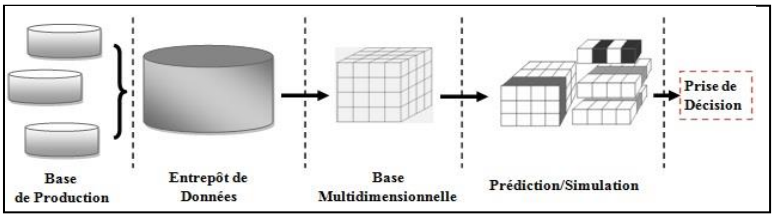


FIG. 1 – Architecture d'un système décisionnel

Chaque cube OLAP (Gray, et al., 1996) est constitué d'un ensemble de mesures organisé sous forme de dimensions. Chaque dimension représente un axe d'analyse sur lequel les données seront analysées. Mais, comme déjà déduit, le problème persiste toujours au niveau du model de donnée OLAP qui néglige l'aspect qualitatif des données, et ne permet pas la prise en considération de plusieurs critères simultanément lors de la résolution d'un problème décisionnel.

Nous pensons donc que la création d'un nouveau modèle de donnée intégrant l'analyse multidimensionnelle et l'analyse multicritère, peut aider à regrouper les caractéristiques quantitatives et qualitatives des données dans le même processus d'analyse.

4 Intégration des systèmes OLAP et méthodes AMC

L'accouplement des systèmes OLAP et méthodes AMC n'a pas suscité sa part légitime d'intérêt par rapport aux autres intégrations courantes sur le marché décisionnel. C'est ainsi que nous allons focaliser nos efforts sur les données multidimensionnelles, et baser nos recherches sur l'apport de la combinaison OLAP/AMCD (TAB.1) comme voie privilégiée permettant d'étendre les capacités analytiques et multidimensionnelles du système OLAP, et d'enrichir ensuite l'habilité de la méthode AMC.

	Avantages	Inconvénients
OLAP	- Analyse dynamique des données multidimensionnelles. - Accessibilité des données. - Habilité à présenter la hiérarchie des données et les détails des calculs. - Rapidité des traitements.	- Nombre exhaustif de dimensions lors de la création de la base de données OLAP. - Absence de l'aspect qualitatif lors de l'analyse et la gestion des données. - Manque de structuration des problèmes complexes.
AMC	- Capacités à simplifier les situations complexes. - Présence de l'aspect qualitatif lors de l'analyse des données. - La méthode constitue un outil de négociation utile aux débats.	- Manque de données fiables, sur une durée suffisante pour mettre en place et valider les méthodes. - Les analyses multicritères sont souvent basées sur des processus lents et itératifs nécessitant une longue durée.

TAB. 1 – Complémentarité OLAP et AMC

4.1 Approche Proposé

L'approche méthodologique qui consiste à simplifier l'intégration OLAP/AMC est tributaire d'une recherche préliminaire des actions réalisées lors de l'utilisation du système OLAP d'une part, et les critères de l'AMCD d'une autre part.

Les étapes opérationnelles d'intégration à suivre :

- Identification des actions à mettre en place lors de l'évaluation
- Construction des critères objectifs (quantitatifs) et subjectifs (qualitatifs) en respectant l'aspect d'hétérogénéités lors du choix des critères.
- Evaluation de l'ensemble des actions en se basant sur nos critères de choix et sur les capacités flexible du processus d'analyse OLAP.
- Visualisation ou restitution des résultats obtenus via les moyens analytiques d'OLAP et par le comité de pilotage (décideurs).

Nous utilisons deux processus pour cette intégration :

- Processus d'analyse OLAP : Ce processus intervient lorsque les décideurs spécifient les critères qui vont déterminer leur choix favorable, ces critères sont pris en compte par le processus d'analyse OLAP et vont nous permettre de préciser les actions candidates en prenant en compte les besoins existants des décideurs.
- Processus d'analyse multicritère : Comme il est déjà expliqué dans la section II, le processus d'analyse multicritère utilise dans notre cas des pondérations selon les

poids attribués aux différents critères. Les traitements effectués au niveau de cette étape se basent et s'enrichissent des informations résultant du processus OLAP et des préférences spécifiées par le comité de pilotage. L'objectif est d'évaluer les différentes actions considérées lors du processus décisionnel et offrir ainsi une solution comme proposition finale.

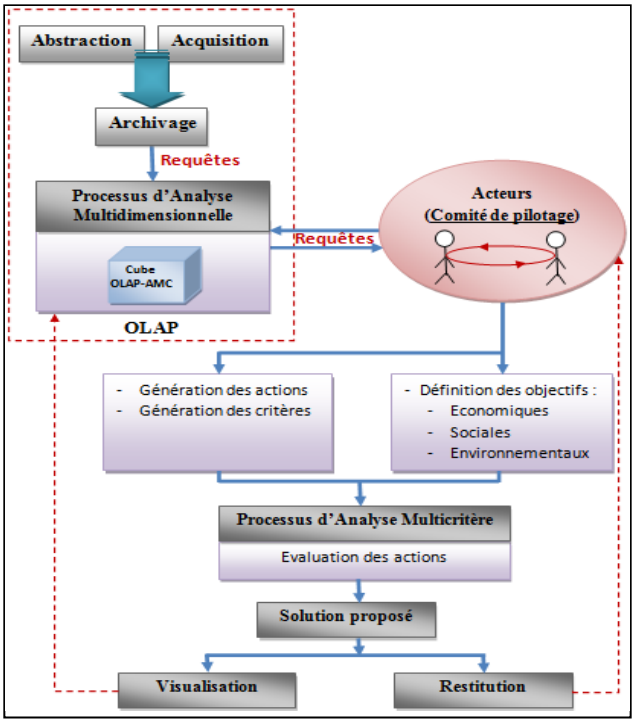


FIG.2 – Approche générale d'intégration OLAP-AMC

4.2 *Modèle* conceptuel d'intégration OLAP-AMCD

Notre modèle conceptuel sera basée sur une structure dimensionnelle en étoile (star schema) (Kimball, 1996), qui propose une table de faits (cube OLAP) comme table d'évaluation contenant des données observables, numériques et quantifiables (Kimball, 2003) entourées par un seul cercle de dimensions qui regroupent les besoins spécifiques des décideurs (FIG.3).

Ce cercle contient trois dimensions :

Dimension Critères : réunit l'ensemble des critères sélectionnés par le comité de pilotage lors de la définition des problèmes.

Dimension Action : représente l'ensemble des options ou des solutions à évaluer.

Dimension Temps : contrôle l'état et l'importance de chaque critère pendant une durée bien déterminée lors de l'évaluation des actions.

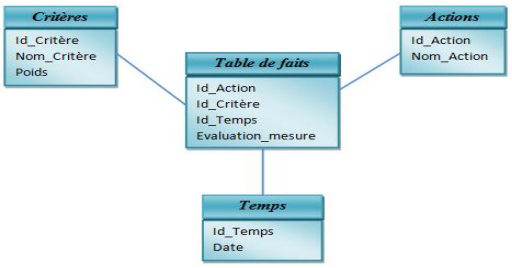


FIG.3 – Schéma multidimensionnel en étoile

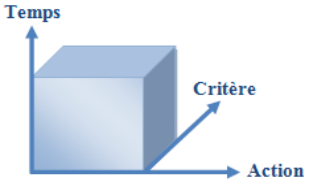


FIG.4–Présentation abstraite du cube

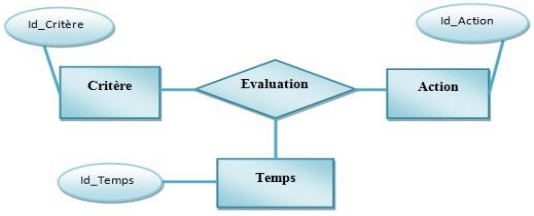


FIG.5 – Diagramme d'entité relation

Le cube OLAP sera construit via le modèle de la figure 3. Les données du cube seront exploitées lors de l'intégration de la fonction de la « somme pondérée » comme méthode multicritère au niveau de notre cube OLAP/AMC.

L'agrégation des valeurs de la dimension 'Critères' se réalisera en introduisant différentes pondérations dans le processus d'évaluation. Le principe de pondération (FIG.6) consiste à attribuer un poids à chaque critère selon son importance.

Les critères choisis concernent les trois volets du développement durable qui sont: l'économique, le social et l'environnemental. La pondération est effectuée en deux niveaux :

- Pondération 1 : les trois domaines déjà cités sont pondérés entre eux, c'est-à-dire qu'on attribue un poids à chaque domaine.
- Pondération 2 : C'est le niveau utilisé dans notre cas d'étude, elle consiste à pondérer les critères du même domaine entre eux.

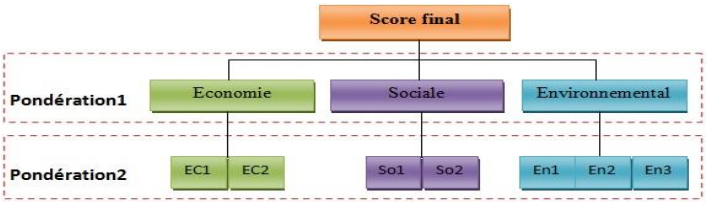


FIG.6– Type de pondération

Ou: $$u(a_i) = \sum_{j=1}^{k} v_j \cdot r_{ij}$$

$u(a_i)$ = utilité évaluée de i alternative (action)
v_i = poids de j critère
r_{ij} = utilité évaluée de i alternative pour j critère

5 Conclusion

Ce présent travail permet de fournir aux décideurs un nouveau modèle multidimensionnel et multicritère d'aide à la décision. Ce modèle intègre à la fois les fonctionnalités des systèmes OLAP et les caractéristiques analytiques des méthodes AMC. Cette synergie réunissant les systèmes OLAP et méthodes AMC est une solution conceptuelle et méthodologique permettant aux décideurs de comprendre et simplifier la complexité des problèmes décisionnels. Comme perspective, et afin de valider cette démarche analytique, nous allons l'appliquer sur une étude de cas traitant le choix de la meilleure solution pour l'implantation d'une grande surface de distribution.

Références

B. Roy, D. Bouysson (1993) Aide multicritère à la décision, Economica, Paris.

GRAY et al (1997). Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. J. Data Mining and Knowledge Discovery 1(1), 29–53.

Inmon, W. H. (1994). Building the datawarehouse. John Wiley and Sons.

Jerbi H. (2012). Personnalisation d'analyses décisionnelles sur des données Multidimensionnelles. Thèse de doctorat, Université Toulouse 1 Capitole (UT1 Capitole).

Kimball, R. (1996). The data warehouse toolkit. John Wiley and Sons.

Marcel, P. (1999). Modeling and querying multidimensional databases. Networking and information systems journal ISSN 1290-2926 2, 515–548.

Martel Jean-Marc. (1999), L'aide multicritère à la décision: Méthodes et Applications. CORS -SCRO. ANNUAL CONFERENCE ,WINDSOR, ONTARIO.

Vanderpooten, D. (2009). Introduction à l'aide multicritère à la décision. Précis de Recherche Opérationnelle.

Summary

The decisions support Systems are based on decision technologies enabling organizations to make better use of their data flow analysis by simplifying them. Indeed, OLAP (On Line Analytical Processing) tools such as decisional technology, offer the possibility of archiving, management, analysis and multidimensional modeling. However, they are limited in the consideration of the multicriteria and quality aspect of the decision problem.

To overcome these limitations, we proposed in this research, a methodological approach for integrating multicriteria analysis in OLAP systems. Our aim of this approach is to create a hybrid model of data (OLAP / MCA), which will expand the capabilities of OLAP through the contribution of MCA as a tool for multi-criteria analysis.

DWEv : Un prototype pour l'évolution partielle du schéma multidimensionnel

Noura Azaiez, Saïd Taktak, Jamel Feki

Laboratoire MIRACL
Université de Sfax, Faculté des Sciences Economiques et de Gestion de Sfax,
route de l'Aérodrome km 4.5, B.P. 1088 – 3018 Sfax, Tunisie.

Noura.azaiez@gmail.com, {Said.taktak, Jamel.feki}@fsegs.rnu.tn

Résumé. L'architecture décisionnelle met en jeu deux éléments essentiels : l'entrepôt de données (ED) et ses magasin(s) de données (MD). L'émergence de nouveaux besoins d'analyse fait apparaître la nécessité de faire évoluer le schéma de l'entrepôt et par conséquent celui de ses magasins de données. Dans ce contexte évolutif, nous étudions la propagation de l'évolution du schéma de l'ED vers ses magasins. Pour cette propagation nous avons défini un ensemble de règles «Si-Alors». Et Afin de valider la démarche d'évolution proposée, nous avons développé l'outil *DWEv* (*D*ata *W*arehouse *Ev*olution).

1 Introduction

De nos jours, la prise de décision est devenue cruciale pour les dirigeants d'entreprises. Les systèmes d'information opérationnels s'avèrent inadaptés pour le décisionnel. Face à cette inadéquation, la technologie des entrepôts de données (ED) a vu le jour afin d'aider les décideurs à la prise de décisions. Le processus d'entreposage consiste à extraire les données à partir de sources opérationnelles, puis de les transformer et nettoyer pour les charger dans les tables de faits et de dimensions qui forment les magasins de données (MD). Cette forte dépendance entre l'ED et ses MDs rend le problème de l'évolution très important dans l'environnement des systèmes d'information décisionnels (SID).

Plusieurs travaux de recherche ont abordé le problème d'évolution dans le contexte d'ED. On peut classer ces travaux en trois courants : *Mise à jour du schéma*, *Versionnement*, et *Maintenance des vues matérialisées*. Malgré l'intérêt que portent les chercheurs aux évolutions des schémas de l'ED, les systèmes commerciaux actuels n'intègrent pas la propagation de l'évolution des schémas d'un ED vers ses MDs.

Dans ce travail, notre objectif est de proposer une démarche permettant d'assurer l'évolution du schéma du MD suite à l'évolution du modèle de l'ED. Nous nous intéressons à étudier l'impact des opérations de changement de l'ED sur ses MDs.

Ce papier est organisé comme suit. Dans la section 2, nous faisons un tour d'horizon des travaux relatifs au problème d'évolution aux EDs. La section 3 décrit notre approche de propagation de l'évolution du schéma d'ED vers ses MDs. La section 4 présente le prototype développé *DWEv* supportant notre démarche. Enfin, la section 5 conclut le papier.

2 État de l'art

Le problème d'évolution des EDs a été sujet de plusieurs travaux de recherche. Hurtado et al. (1999) ont traité le problème d'évolution dans le contexte multidimensionnel. En effet, ils ont proposé es opérateurs d'évolution spécifiques à ce type de modèle et ont étudié leurs effets sur les structures dimensionnelles ainsi que sur leurs instances. Ils ont également abordé l'effet de ces mises à jour sur les vues matérialisées et ont proposé un algorithme pour réaliser leur maintenance.
Blaschka et al. (1999) se sont intéressés à l'évolution touchant non seulement les dimensions mais aussi les tables de faits. Ils ont également enrichi les opérations d'évolution proposées par Hurtado et al. (1999) en offrant la possibilité d'insérer un nouveau niveau de hiérarchie à n'importe quelle position. Les auteurs ont proposé l'outil FIESTA pour valider leur approche.
Guerrero et al. (2004) ont proposé un prototype WHES (WareHouse Evolution System) qui permet la création et l'évolution des EDs en se basant sur un modèle d'évolution et le langage MDL « Multidimensional Data Language ».
Favre et al. (2007) ont relié l'évolution de l'ED à une évolution des besoins d'analyse des décideurs. Les auteurs ont proposé ainsi une approche qui permet aux utilisateurs d'intégrer leurs propres connaissances afin d'enrichir les possibilités d'analyse de l'ED.
Papastefanatos et al. (2009) sont intéressés par l'étude du problème d'évolution sur le processus ETL (Extract-Transform-Load). En effet, ils ont étudié le problème des incohérences qui apparaissent sur les processus ETL suite à une évolution au niveau source de données.

Les travaux présentés traitent différentes opérations d'évolution touchant différents composants dans les MDs à savoir les tables de dimension et de fait, les niveaux d'hiérarchie, et le processus de chargement, etc. Nous soulignons qu'en général, ces opérations sont liées à une évolution au niveau des besoins d'analyse des décideurs ou bien à des changements structuraux au niveau des MDs. Or le schéma de MD peut évoluer aussi suite à une évolution au niveau du schéma de l'ED dont il dépend. En effet, les processus métiers des organisations peuvent évoluer au cours du temps ; cette évolution peut toucher les procédures de travail et plus fondamentalement les modèles de données de leur système opérationnel. Le processus d'entreposage associé ne peut échapper à cette évolution qui mérite d'être propagée vers les MDs. Nous avons constaté qu'à l'état actuel, ces impacts n'ont pas été étudiés.

3 Démarche proposée pour l'évolution des MDs

Nous nous intéressons à proposer une démarche qui vise à étudier les modifications qui affectent le schéma de l'ED ainsi que leurs impacts sur les MDs de l'entrepôt. L'approche proposée se base sur un ensemble de règles de type «Si-Alors » permettant de déterminer les mises à jour à appliquer sur les MDs. Ces règles prennent en entrée le type d'opération d'évolution appliquée au niveau de l'ED (ajout, suppression, modification) et les éléments affectés (table, colonne) ainsi que les différentes correspondances ED/MDs. Elles renvoient les opérations d'évolution à appliquer au niveau des MDs.
La figure 1 représente l'architecture de la démarche proposée.

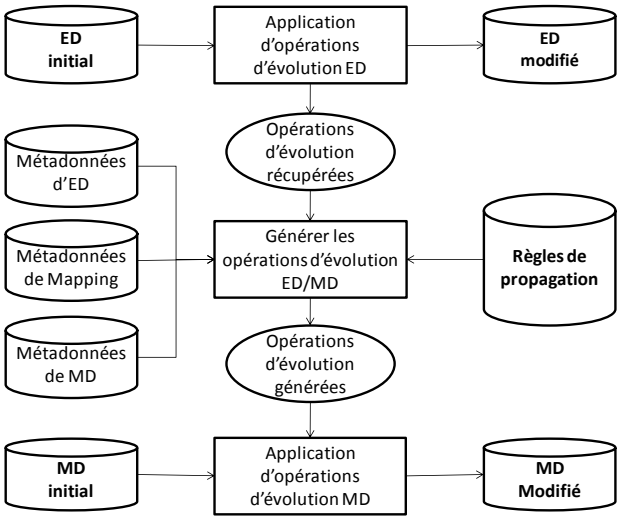


FIG. 1 – *Architecture proposée pour l'approche d'évolution*

Pour illustrer les différents cas des évolutions du schéma de l'ED et leurs impacts sur ses MDs, nous nous basons sur un modèle d'ED relationnel. La figure 2 montre un schéma d'ED construit à partir des tables de la base de données échantillon du SGBD Oracle. La figure 3 est un schéma en étoile du MD des commandes (Orders) que nous avons construit sur l'ED de la figure 2.

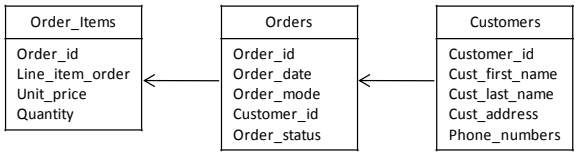


FIG. 2 – *Un exemple d'ED*

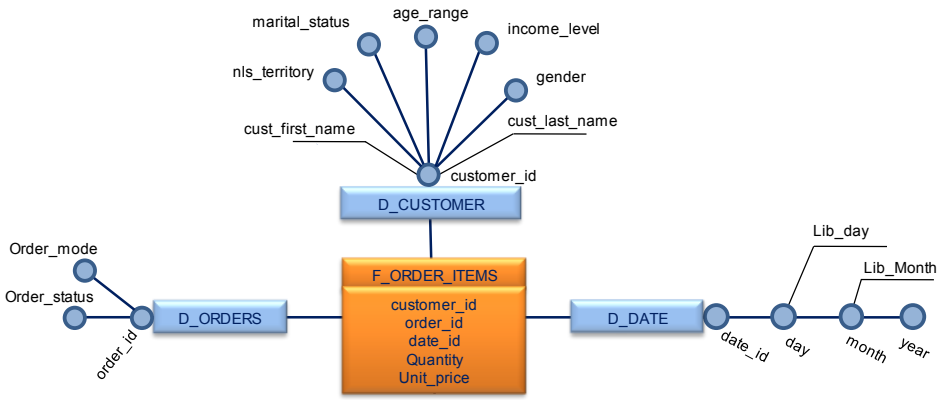


FIG. 3 – *MD pour l'Analyse des commandes construit sur l'ED de la FIG. 2*

Afin de définir les règles de propagation des évolutions sur les MDs, nous utilisons les notations suivantes :

- T_D : Une table *T* de l'ED qui alimente une dimension *D* ;
- T_{id} : L'identifiant d'une table *T* ;
- T_F : Une table *T* de l'ED qui alimente un fait *F*.

Ajout d'une table T. L'ajout d'une table *T* à l'ED peut alimenter un nouveau fait *F*, une nouvelle dimension *D* ou une nouvelle hiérarchie *H*. Nous prenons en considération le rôle que joue chaque table reliée à *T* dans le(s) MD(s), les types des colonnes de T et leur additivité.

- Règle *RT1* : Si *T* fait référence à n (n≥1) table(s) T_{D1}, T_{D2}…,T_{Dn} alimentant *n* dimensions et si *T* contient une/des colonne(s) additive(s), alors *T* peut alimenter un nouveau fait *F*. Les dimensions de *F* sont alors T_{D1}, T_{D2}…, T_{Dn}.
  Par exemple on se propose d'ajouter la table *BILL* (*bill_id, total_bil*) à l'ED qui contient une colonne *additive total_bil* et faisant référence à la table *CUSTOMERS* qui alimente la dimension *D_CUSTOMERS*. Selon la règle *RT1*, la table *BILL* peut alimenter un nouveau fait conventionnellement appelé *F_BILL ; D_CUSTOMERS* sera une dimension pour *F_BILL*.
- Règle *RT2* : Si *T* est référencée par une table T_F et si T_{id} est atomique, et si *T* contient des colonnes pouvant être des attributs dimensionnels (forts ou faibles), alors *T* alimente une dimension *D* pour *F*. Supposons que la table *ORDER_ITEMS* ne fait pas référence à la table *PRODUCT (product_id, product_name, product_desc*) et que nous ajoutons cette dernière en la reliant à la table *ORDER_ITEMS* qui alimente le fait *F_ORDER_ITEMS* . Alors puisque *PRODUCT* possède un identifiant non composé *product_id* et des colonnes susceptibles de devenir des paramètres (e.g., *product_name, product_desc*), elle satisfait la règle *RT2*. En conséquence elle se transforme en une nouvelle dimension *D_PRODUCT* pour le fait *F_ORDER_ITEMS*.
- Règle *RT3* : Si *T* est référencée par $T_{D,}$ et si T_{id} est atomique, et si *T* ne possède pas des colonnes additives, alors *T* complète la dimension *D* par une hiérarchie *H* reliant D_{id}, *à l'attribut* T_{id}. Les attributs faibles potentiels du paramètre T_{id} sont les attributs de type chaîne de caractères appartenant à *T*. Par exemple, l'ajout de la table *EMPLOYEES* (*employee_id, first_name, last_name*) référencée par *ORDERS* complète la dimension *D_ORDERS* par une nouvelle hiérarchie *H_EMPLOYEES (order_id, employee_id).* Les attributs faibles de *employee_id* sont les attributs *first_name, last_name* puisqu'ils sont du type chaîne de caractères.

Ajout d'une colonne *C.* L'ajout d'une colonne *C* à l'ED peut enrichir un fait existant avec une nouvelle mesure, une dimension existante avec un nouvel attribut ou un fait avec une nouvelle dimension. Nous prenons en considération le type de la colonne, son additivité et le rôle de chaque table contenant cette colonne dans les magasins de données. Nous définissons une règle pour chacune de ces cas.

- Règle *RC1* : Si la colonne *C* est additive et est ajoutée à une table de l'ED qui alimente un fait *F*, alors *C* devient une mesure pour *F* dans le MD.
  Par exemple on propose d'ajouter la colonne *TOTAL_PRICE* numérique à la table *ORDER_ITEMS* qui alimente le fait *F_ORDER_ITEMS*. La colonne *TOTAL_PRICE* satisfait la règle *RC1*, donc elle alimentera une nouvelle mesure *(appelée TOTAL_PRICE)* pour le fait *F_ORDER_ITEMS.*

- Règle *RC2* : Si on ajoute *C* à une table T_D qui alimente une dimension *D*, et si *C* est non additive, alors *C* sera un attribut pour la dimension *D* dans le MD. Le choix du rôle de l'attribut (faible ou fort) sera décidé par l'utilisateur. Par exemple on ajoute la colonne *ORDER_TYPE* de type chaîne de caractères à *ORDERS* qui alimente la dimension *D_ORDERS*. La colonne *ORDER_TYPE* satisfait la règle *RC2*, donc elle alimente un attribut faible ou fort pour la dimension *D_ORDERS*.
- Règle *RC3* : Si on ajoute *C* de type date à T_F, et si cette dernière ne comporte pas de colonnes de type Date, alors *C* peut alimenter une dimension temporelle dans le MD du fait *F*. Par exemple on ajoute la colonne *DAT* à la table *BILL* qui alimente *F_BILL*. La colonne *DAT* satisfait la règle *RC3*, donc elle enrichit le MD par une dimension *D_DATE*.

4 Prototype

Pour valider nos règles, nous avons développé le prototype *DWEv* (*D*ata *W*arehouse *Ev*olution) qui présente une implémentation authentique de la démarche proposée. La figure 4 représente l'architecture globale de *DWEv*.

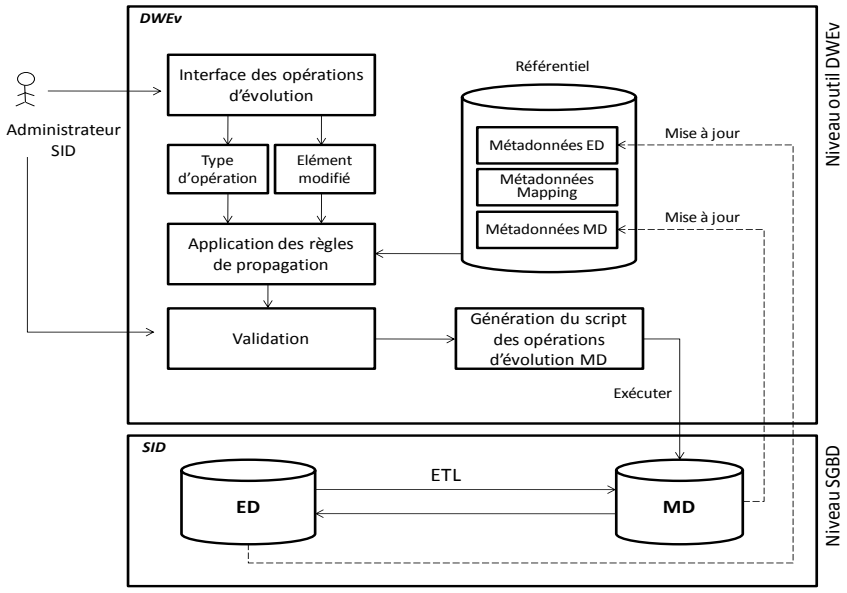


FIG. 4 – *Architecture de DWEv*

A travers l'interface des opérations d'évolution, l'administrateur d'ED à la possibilité de modifier le schéma de l'ED en identifiant le type d'opération d'évolution (Ajout, modification, suppression …) et l'élément affecté (Table, colonne…) par cette modification. Ces opérations d'évolution constituent le point de départ du processus d'altération du SID.
DWEv se base sur un référentiel qui effectue le stockage et la récupération des informations utilisées par les règles de propagation. Ces informations comprennent le schéma relationnel de l'ED, les schémas des MDs et les correspondances. L'application des règles de propagation présente le cœur du processus d'altération. En se basant sur les informations collectées à partir

du référentiel et l'opération d'évolution exécutée sur l'ED, *DWEv* applique la règle convenable pour identifier les éléments affectés et les opérations de mise à jour à appliquer sur les MDs. Le passage ensuite par une étape de validation est nécessaire notamment lorsqu'il s'agit d'une opération d'évolution destructive (i.e. suppression). Une fois l'administrateur valide les opérations d'évolution proposées, *DWEv* génère le script de mise à jour correspondant et l'exécute. Après chaque évolution du SID, le référentiel est mis à jour.

5 Conclusion

Dans cet article, nous avons abordé la problématique de l'évolution du schéma de l'ED et ses impacts sur les MDs en proposant une approche à base de règles « Si-Alors » pour la propagation de l'évolution de l'ED sur ses MDs. Par ailleurs, nous avons validé notre approche par un prototype logiciel nommé *DWEv* qui vise à appliquer des changements évolutives sur un ED et d'identifier leurs impacts sur le/les MD(s).

Références

C. A. Hurtado, A. O. Mendelzon, et A. A. Vaisman (1999). Maintaining Data Cubes under Dimension Updates. *In XVth International Conference on Data Engineering (ICDE 99), Sydney, Australia, pages 346–355. IEEE Computer Society.*

C. Favre, F. Bentayeb, O. Boussaid (2007). A Survey of Data Warehouse Model Evolution, *Encyclopedia of Database Technologies and Applications, Idea Group Publishing*.

E. Benitez-Guerrero, et C. Collet, M. Adiba (2004). The WHES approach to data warehouse evolution. *E-Gnosis [online], Vol.2Art.*

G. Papastefanatos, P. Vassiliadis, A. Simitsis, T. Sellis, et Y. Vassiliou (2009). Rulebased Management of Schema Changes at ETL Sources. *In: The International Workshop on Managing Evolution of Data Warehouses (MEDWa), Riga, Latvia.*

M. Blaschka, C. Sapia, et G. Hofling (1999). On Schema Evolution in Multi-dimensional Databases. *In Ist International Conference on Data Ware-housing and Knowledge Discovery (DaWaK), Florence, Italy, volume1676 of LNCS, pages 153–164.*

Summary

The architecture of decision involves two essential components: the data warehouse (DW) and data mart (DM). The emergence of new analytical needs highlighted the need to change the schema of data warehouse and therefore its data marts. In this evolving context, we study the propagation of DW schema evolution on its DMs. For this propagation we have defined a set of rules "If-Then". To validate the proposed evolution approach, we have developed the tool *DWEv* (*D*ata *W*arehouse *Ev*olution).

Keywords: Data warehouse, evolution operations, data mart, propagation rules.