Actes de la 6ème édition

Atelier des Systèmes Décisionnels

ASD 2012

1-3 avril 2012
Université Saad Dahlab, Blida, Algérie

Edités par :
Nadjia Benblidia
Saliha Oukid Khouas

6ème édition des

Atelier des Systèmes Décisionnels

ASD 2012

ASD 2012

Actes de la 6ème édition des

Atelier des Systèmes Décisionnels

Edités par

Nadjia Ben Blidia, Seliha Oukid-Khaous

1-3 avril 2012

Université Saad Dahlab, Blida, Algérie

Préface

Les technologies des entrepôts de données et analyses en lignes 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 nouvelles manifestations scientifiques venant enrichir le panorama des rencontres entre chercheurs et industriels.

Dans le prolongement des trois éditions précédentes (Agadir–Maroc 2006, Sousse, Tunisie 2007, Mohammedia, Maroc 2008, Jijel, Algérie 2009 et Sfax, Tunisie 2010), ASD 2012 (Atelier sur les Systèmes Décisionnels) ambitionne de consolider les expériences conduites par les chercheurs, industriels et utilisateurs issus de communautés travaillant avec les systèmes décisionnels. L'objectif de cette troisième édition de l'atelier, en particulier après le succès des deux premières 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, de renforcer les liens existants et de 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 sixième édition ASD. ASD 2012 a reçu 60 soumissions d'articles de nombreux pays (Algérie, Canada, France, Maroc, Suisse, Tunisie). Après évaluation par les membres du comité scientifique, composé par 54 experts internationaux du domaine, 20 articles longs et 4 articles courts ont été retenus. Ces papiers couvrent différents thèmes de recherche et d'application sur les systèmes décisionnels.

ASD 2012 a reçu le soutien de différentes institutions publiques d'enseignement et de recherche : le laboratoire ERIC de l'université Lumière Lyon2 (France), le laboratoire MIRACL de l'Université de Sfax (Tunisie), l'université HASSAN II Mohammedia, la Faculté des Sciences et Techniques de Mohammedia, Nous sommes reconnaissants de leur soutien.

Le succès de cette troisième édition de ASD n'aurait pas été réalisé sans la coopération étroite du comité scientifique et des membres du comité d'organisation, que nous tenons également à remercier très chaleureusement.

Les éditeurs
N. BEN BLIDIA, S. OUKID-AOKAS

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
- ABED Hafida, LRDSI Blida Algérie
- 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
- BADRI Abdelmajid, FST Université Hassan II Maroc
- BELDJILALI Bouziane, Université Oran Algérie
- BELLAFKIH Mostafa, INPT Rabat Maroc
- BELLATRECHE Ladjel, ENSMA Poitiers France
- BEN BLIDIA Nadjia, LRDSI Blida Algérie
- BEN MESSAOUD Riadh, UTC Tunis Tunisie
- BENHARKAT Nabila, LIRIS Lyon France
- BENMOHAMED Mohamed, LIRE Constantine Algérie
- BENSLIMANE Djamel, LIRIS Lyon France
- BIMONTE Sandro, Cemagref, Clermond-Ferrand, France
- BOUAZIZ Rafik, MIRACL, Université de Sfax Tunisie
- BOUFAIDA Mahmoud, LIRE Constantine Algérie
- BOUFARES Faouzi, LIPN Paris France
- BOUKERRAM Abdellah, Université Sétif Algérie
- BOUKHALFA Kamel, LSI, USTHB
- DARMONT Jérôme, ERIC Lyon France
- EL HEBIL Farid, INPT Rabat Maroc
- FAVRE Cécile, Université Lyon 2, France
- GUESSOUM Abderrezak, LATSI Blida Algérie
- HARBI Nouria, université Lyon 2, France
- KABACHI Nadia, Université Lyon1 , France
- KHOLLADI Med-khireddine, LIRE Constantine Algérie
- KHROUF Kais, MIRACL, Université de Sfax Tunisie

- LALAM Mustapha, Université Tizi-Ouzou Algérie
- LEMIRE Daniel, UQ Montréal Canada
- MAHDAOUI Latifa, LSI, USTHB
- MALKI Mimoune, USB Sidi Bel Abbes Algérie
- MARGHOUBI Rabia, INPT Rabat Maroc
- MELIT Ali, LAMEL Jijel Algérie
- MEZIANE Abdelkrim, CERIST - Alger
- 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

Comité d'organisation :

- ABED Hafida, Université Saad Dahlabdah - Blida
- BALA Mahfoud, Université Saad Dahlab - Blida
- BENBLIDIA Nadjia, Université Saad Dahlab - Blida
- BOUAYED Noureddine, Université Saad Dahlab - Blida
  BOUSTIA Narhimene, Université Saad Dahlab, Blida
- DOULKIFLI Boukraa, Université de Jijel - Algérie
- HAMMOUDA Mohamed, Université Saad Dahlab - Blida
- MADANI Amina, Université Benyoucef Benkhedda - Alger
- MASSIED Mohamed, Université Saad Dahlab - Blida
- MEZIANE Abdelkrim, CERIST - Alger
- MOALLA Mohamed Sahbi, ISET Sfax - Tunisie
- OUKID Saliha, Université Saad Dahlab - Blida
- REGUIEG F. Zohra, Université Saad Dahlab - Blida
- SIDDOUMOU Mohamed, Université Saad Dahlab - Blida

Sommaire

Personnalisation dans les entrepôts de documents

Yahia Hamdi, Kaïs Khrouf, Jamel Feki

MIR@CL, Université de Sfax, BP 1088, 3018 Sfax, Tunisie,
YahiaHamdi@hotmail.fr, Khrouf.Kais@isecs.rnu.tn, Jamel.Feki@fsegs.rnu.tn
http://www.miracl.rnu.tn

Résumé. Les entrepôts de documents doivent permettre aux dirigeants d'organisations d'extraire l'information utile à la prise de leurs décisions. Habituellement, l'accès aux informations (voire documentaires) pour des besoins décisionnels s'effectue en appliquant des techniques d'analyses en ligne OLAP (On-Line Analytical Processing). Cependant, les résultats de ces requêtes analytiques risquent d'être larges en quantité d'information. L'objet de cet article est de traiter la personnalisation dans les entrepôts de documents afin d'adapter les réponses en fonction des préférences des décideurs. Plus précisément, nous proposons une extension du méta-modèle d'entrepôt de documents, elle permet d'intégrer les profils utilisateur et la recommandation de documents.

1 Introduction

De nos jours, les décideurs souhaitent disposer d'outils leur permettant d'extraire l'information pertinente à leur processus de prise de décisions. En réalité, une grande partie de cette information est contenue dans des documents (Tseng et al., 2006) qui sont généralement collectés à partir des sources disséminées et hétérogènes (Internet, Workflow, bibliothèques numériques, etc.). Parmi les exigences de ces décideurs est la mise à leur disposition de nouveaux outils pour exploiter cette grande masse de documents souvent hétérogènes. Dans ce contexte, le concept d'entrepôt de documents a été proposé, il constitue pour toute organisation un outil de base à toute activité de veille économique et technique voire décisionnelle pour les documents (Sullivan, 2001). Dans le cadre de nos travaux antérieurs (Khrouf et al., 2012), nous avons proposé un méta-modèle permettant de regrouper et de classifier les documents selon des structures génériques (i.e., identiques ou similaires) et d'appliquer ainsi les techniques d'analyse en ligne OLAP (On-Line Analytical Processing) sur ces documents.

Dans le domaine décisionnel, les entrepôts de documents permettent de répondre à un ensemble de besoins d'analyse recensés auprès des décideurs à un moment donné. Cependant, ces décideurs expriment des besoins très variés auxquels l'entrepôt n'est pas forcément en mesure de les satisfaire convenablement. Le problème n'est pas dans la disponibilité de l'information mais dans sa pertinence relative aux besoins d'analyses. C'est pourquoi ces travaux s'orientent actuellement vers l'intégration du décideur comme composante intrinsèque d'un système (de recherche ou d'analyse) et ce dans le but de lui délivrer une information pertinente et adaptée à ses besoins et préférences.

Le travail présenté dans cet article s'inscrit dans le cadre de la personnalisation dans les entrepôts de documents en se focalisant sur les *besoins* des utilisateurs (i.e., décideurs) et les *contextes partagés* entre ces décideurs. Pour ce faire, nous procédons à une extension du méta-modèle d'entrepôts de documents qui consiste à adapter la partie *Ontologie* et à ajouter la composante *Utilisateur*.

Cet article est organisé comme suit. La section 2 présente un état de l'art des travaux traitant la personnalisation dans les entrepôts de données. La section 3 présente le contexte de nos travaux, à savoir : les entrepôts de documents. Dans la section 4, nous proposons une extension du méta-modèle d'entrepôts de documents. Enfin, nous décrivons dans la section 5 les interfaces réalisées et intégrées à notre outil *DocWare* (*Doc*ument *Ware*house) et permettant de valider les propositions étudiées.

2 Etat de l'art

La personnalisation a été abordée dans plusieurs domaines, à savoir : la recherche d'information, les bases de données, les interfaces Homme-Machine et récemment par les entrepôts de données. Nous nous intéressons dans cette section aux travaux de la personnalisation dans les entrepôts de données puisqu'à notre connaissance, il n'existe pas des travaux qui ont abordé cet aspect pour les entrepôts de documents.

(Favre et al. 2007) autorisent les décideurs d'enrichir un entrepôt de données avec l'ajout de l'information, en se basant sur les règles d'agrégation « si-alors » qui définissent les préférences des décideurs. Cependant, l'ajout de ces règles peut mettre en cause les analyses existantes.

(Jerbi et al., 2010) proposent de personnaliser les schémas multidimensionnels en fonction des besoins analytiques individuels en utilisant une approche qualitative. Plus précisément, ils utilisent des ordres de préférence (sous forme de prédicats), c'est-à-dire ordonner les préférences des décideurs, plutôt que de recourir à des poids. L'inconvénient de ces travaux réside dans le fait que ces ordres de préférence dépendent du contexte d'analyse.

(Khemiri et al. 2010) présentent une approche pour personnaliser le contenu des entrepôts de données, l'idée principale consiste à créer une vue matérialisée d'entrepôt pour chaque utilisateur en respectant son profil, décrit dans un document XML. Ainsi, le nombre de vues matérialisées construites dépend du nombre des décideurs. De plus, le coût de construction de ces vues matérialisées est élevé.

(Bellatreche et al. 2005) se sont intéressés à l'affichage du résultat multidimensionnel selon les préférences décideurs, décrites sous forme de contraintes de visualisation. Les travaux proposés se limitent à visualiser un seul attribut par dimension et à appliquer l'opérateur de sélection pour calculer le cube résultat.

Dans ces travaux abordant les entrepôts de données, les profils utilisateurs prennent plusieurs formes (règles d'agrégation, ordre de préférence qualitatifs, contraintes de visualisation) ; ces différentes formes permettent une définition des profils très individuels. Dans nos travaux, nous proposons de contextualiser cette définition de profils afin de dégager des intérêts communs entre les utilisateurs, ce qui permettra une collaboration et un échange de connaissances entre eux. Plus précisément, nous proposons de décrire les profils sous forme d'ontologies adaptées aux besoins des utilisateurs, en se basant sur les ontologies de domaines stockées dans l'entrepôt.

3 Contexte de nos travaux

Les entrepôts de documents doivent permettre d'appliquer les techniques d'analyses multidimensionnelles (habituellement appliquées aux données factuelles) aux documents. A cette fin, nous utilisons le méta-modèle de la figure 1.

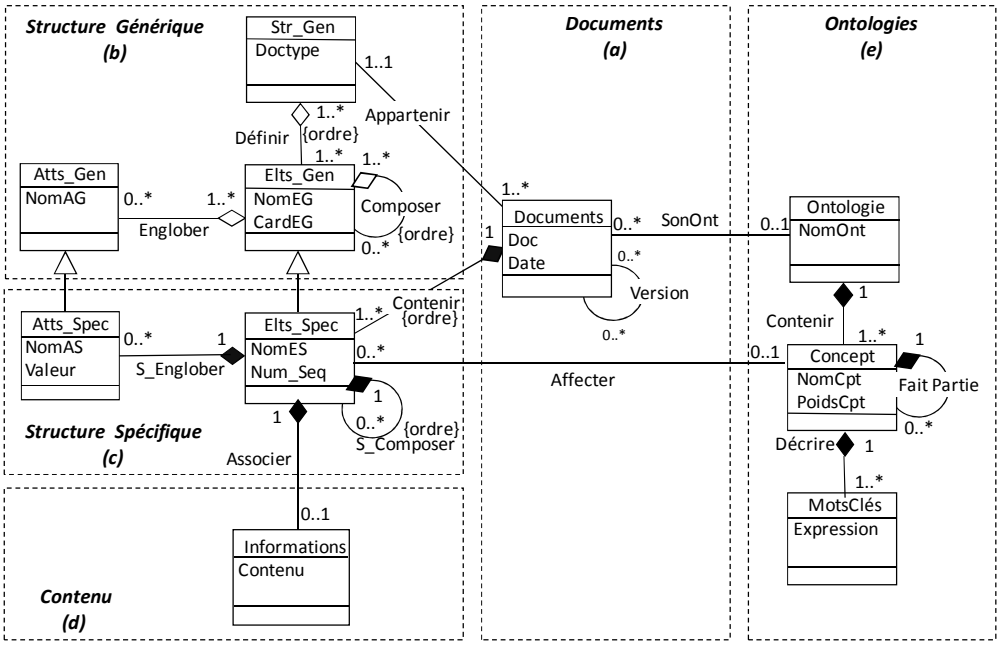


FIG. 1 – *Méta-modèle d'entrepôts de documents (Khrouf et al., 2012)*

Ce méta-modèle comporte les composants suivants :

1. Les documents intégrés dans l'entrepôt (cf. FIG 1.a).

2. La partie structurelle : Elle décrit la structure hiérarchique des documents. Nous distinguons deux types :
   - Une structure générique : c'est une structure commune pour un ensemble de documents (cf. FIG 1.b). Elle est définie par un ensemble d'éléments génériques pouvant être composés d'autres éléments génériques et/ou décrits par des attributs génériques.
   - Une structure spécifique : c'est une structure propre à un document et doit être conforme à la structure générique à laquelle est rattaché le document (cf. FIG 1.c). Elle est définie par un ensemble d'éléments spécifiques pouvant englober des attributs spécifiques.

3. La partie contenu : c'est la description du contenu textuel des éléments de la structure spécifique (cf. FIG 1.d).

4. La partie sémantique : Cette partie est définie par un ensemble d'ontologies. Une ontologie est composée d'un ensemble de concepts, décrits par des mots-clés (cf. FIG 1.e). L'affectation des concepts aux éléments spécifiques de l'entrepôt s'effectue en quatre étapes : (1) Extraction des mots-clés, (2) Choix de l'ontologie, (3) Affectation des concepts aux éléments feuilles et (4) Propagation des concepts aux éléments non-feuilles (Ben Meftah et al., 2012).

La figure 2 est un exemple d'instanciation du méta-modèle de la figure 1.

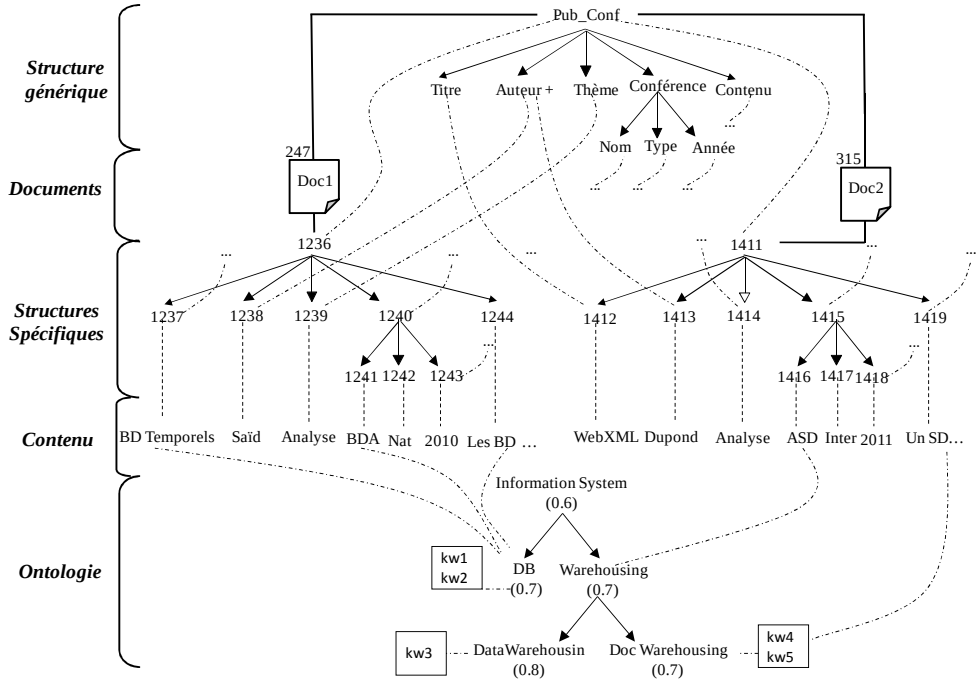


FIG. 2 – *Exemple d'instanciation du méta-modèle d'entrepôt de données de la figure 1*

La modélisation multidimensionnelle considère un sujet analysé « Fait » comme un point dans un espace à plusieurs axes « dimensions ». Pour spécifier une requête, l'utilisateur spécifie ces critères d'analyse puis, le système génère les vues correspondantes et affiche le résultat comme une table multidimensionnelle (Khrouf et al., 2012).

A titre d'exemple, supposons que l'entrepôt contienne des articles scientifiques, nous souhaitons analyser ces articles par *Type*, *Année* et *Thème*. L'utilisateur doit spécifier les éléments d'analyse, comme indique la FIG. 3.

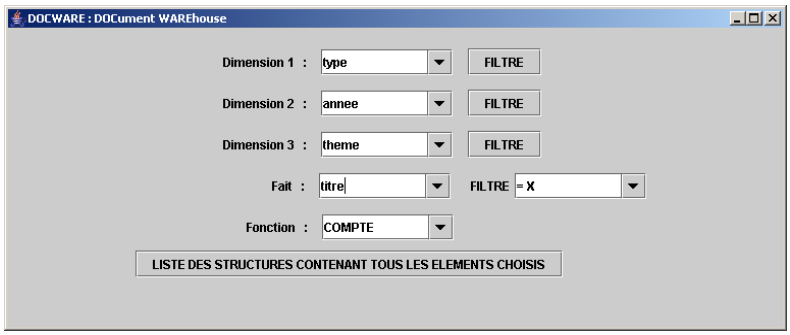
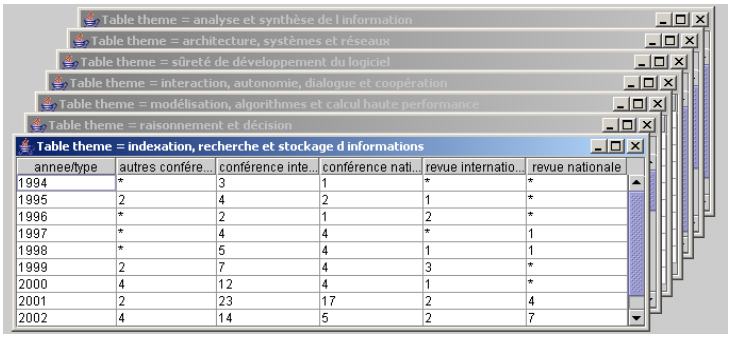


FIG. 3 – *Spécification des critères d'analyse*

Le système affiche toutes les structures génériques contenant ces éléments. Pour cet exemple, le système récupère *Pub-Conf* (Publications dans des conférences) et *Pub-Revue* (Publications dans des revues). L'utilisateur a le choix de faire ses analyses sur une seule structure générique ou les deux simultanément.

Si l'utilisateur a choisi de faire son analyse sur les deux structures génériques, le système génère automatiquement les vues correspondantes et affiche la table multidimensionnelle suivante.



annee/type	autres confére...	conférence inte...	conférence nati...	revue internatio...	revue nationale
1994	*	3	1	*	*
1995	2	4	2	1	*
1996	*	2	1	2	*
1997	*	4	4	*	1
1998	*	5	4	1	1
1999	2	7	4	3	*
2000	4	12	4	1	*
2001	2	23	17	2	4
2002	4	14	5	2	7

FIG. 4 – *Table multidimensionnelle*

Les résultats obtenus dans cette table multidimensionnelle concernent tous les documents de l'entrepôt conformes aux deux structures génériques *Pub-Conf* et *Pub-Revue*. Notons qu'à ce nouveau, les travaux effectués ne tiennent pas compte des préférences de l'utilisateur. En effet, deux utilisateurs qui posent la même requête récupèrent la même réponse. Il est fortement intéressant d'introduire des aspects de personnalisation. La suite de cet article abordera la personnalisation dans un entrepôt de documents.

4 Personnalisation

Afin de rendre ces entrepôts de documents personnalisés, nous proposons d'étendre le méta-modèle d'entrepôts de documents en adaptant la partie *Ontologie* et en ajoutant la composante *Utilisateur* qui décrira les besoins et les préférences des utilisateurs. Dans cette sec-

tion, nous commençons par la présentation du concept « Profil utilisateur » dans le contexte d'entrepôts de documents. Ensuite, nous présentons l'extension du méta-modèle que nous proposons. Enfin, nous présentons comment affecter les nouveaux documents, insérés dans l'entrepôt, aux profils utilisateurs.

4.1 Notion de profil utilisateur

L'entrepôt de documents, que nous proposons, permet de regrouper les documents selon des structures génériques (sous forme de classes). Afin d'apporter de la sémantique à certains éléments textuels (*Résumé*, *Paragraphe*, *Section*…), nous avons utilisé les ontologies de domaine. Une ontologie peut être définie comme étant un ensemble de concepts, ainsi que des relations entre ces concepts. Dans cet article, nous nous limitons aux relations d'agrégation car les ontologies utilisées décrivent des domaines partant du général vers le spécifique.

Une ontologie O_i est définie par $O_i = \{O_i^{Nom}, C\}$ où
- O^{Nom} : Nom de l'ontologie
- C : Un ensemble des concepts de O_i, noté $C = \{C_1, C_2, \ldots , C_m\}$

Un concept $C_i = \{C^{Nom}, C^{Poids}, C^{père}, T\}$
- C^{Nom} : Nom du concept
- C^{Poid} : Poids du concept (affecté par des experts)
- $C^{père}$: Concept père de C_i
- T : Un ensemble de termes (Mots-clés) associés à C_i, noté $T = \{T_1, T_2, \ldots , T_k\}$

La définition d'un profil utilisateur consiste à adapter une ontologie de domaine existante par rapport à ses besoins et ses préférences. Plus précisément, l'utilisateur peut :
- Sélectionner une partie de l'ontologie ou sa totalité.
- Ajouter d'autres concepts qu'il juge intéressants.
- Ajouter des termes aux concepts de l'ontologie.
- Supprimer des concepts de la sous-ontologie sélectionnée.

Un profil, noté $P_i = \{P^{Nom}, O_j, C^{Racine}, C^{Pi}\}$
- P^{Nom} : Nom du profil
- O_j : Ontologie sélectionnée sur laquelle se base le profil
- C^{Racine}: Nom du concept racine décrivant le profil sous forme d'ontologie, $C^{Racine} \in O_j.C$
- C_{Pi} : L'ensemble des concepts définissant le profil P_i, $C_{Pi} \subset O_j.C$

Un concept *du* profil P_i, *noté* $C_{Pi,j} = \{ C_{Pi,j}^{Nom}, C_{Pi,j}^{Type}, C_{Pi,j}^{Poids}, C_{Pi,j}^{Père}, T_{Pi,j}\}$
- $C_{Pi,j}^{Nom}$: Nom du concept j du profil P_i
- $C_{Pi,j}^{Type}$: Type du concept j du profil P_i = { O : Provenant de l'*Ontologie* $P_i.O_j$; U : Ajouté par *l'Utilisateur* }
- $C_{Pi,j}^{Poids}$: Poids du concept, affecté par l'utilisateur
- $C_{Pi,j}^{Père}$: Nom du concept père
- $T_{Pi,j}$: Un ensemble de termes (mots-clés) décrivant le concept j du profil P_i

Un terme $T_k \in T_{Pi,j} = \{ T_k^{Nom}, T_k^{Type} \}$
- T_k^{Nom} : Nom du terme
- T_k^{Type} : Type du terme = $\begin{cases} O : \text{Provenant de l'}\textit{Ontologie } P_i.O_j \\ U : \text{Ajouté par } \textit{l'Utilisateur} \end{cases}$

4.2 Méta-modèle étendu

Le méta-modèle initial (cf. FIG. 1) permet d'appliquer les techniques d'analyses multidimensionnelles sur les documents de l'entrepôt. Afin de personnaliser ces analyses, nous étendons ce méta-modèle. Plus précisément, nous *modifions la partie Ontologie* et nous *ajoutons la composante Utilisateur* (cf. FIG. 6).

Concernant la partie Ontologie, nous avons distingué deux types de concepts : Ceux appartenant à l'ontologie de base, et ceux ajoutés par l'utilisateur et qui lui sont spécifiques. Ces concepts utilisateurs ne sont accessibles que par leur créateur. Egalement, nous avons autorisé l'utilisateur d'affecter des poids à ses concepts spécifiques (cf. FIG. 6-e).

Quant aux extensions de la composante *Utilisateur*, nous avons défini un ensemble de profils dont chacun est décrit par le concept racine C_{racine} (sélectionné par l'utilisateur). Le choix de ce concept racine entraîne l'affectation automatique de ses concepts feuilles au profil considéré comme une sous-ontologie. Pour ce profil, l'utilisateur peut l'adapter à ses préférences en *i)* supprimant les concepts qu'il juge inutiles, et/ou *ii)* ajoutant de nouveaux concepts spécifiques en les pondérant et en leur associant des termes. Ces extensions sont décrites dans la (cf. FIG. 6-f).

La FIG. 5 présente un exemple d'instanciation de la composante *Utilisateur* du méta-modèle (Cas d'ajout d'un nouveau profil P_2). Dans cet exemple, l'utilisateur a choisi le concept *Warehousing* comme racine, il n'est pas intéressé par le concept *DataWarehouse* (absence de lien entre le profil P_2 et ce concept). Pour compléter la définition de son profil, il a ajouté les deux concepts *Textual Warehouse* (pondéré 0.7) et *XML Warehouse* (pondéré 0.8) au concept *DocWarehouse*. Il a associé une liste de mots-clés (L1 composée de Kw6, Kw7 et Kw8) à son nouveau concept *XML Warehouse*.

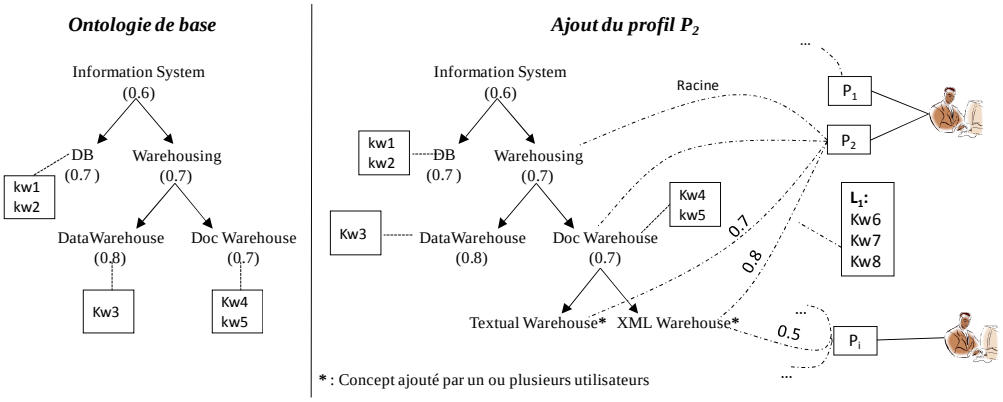


FIG. 5 – *Exemple d'instanciation de la composante Utilisateur*

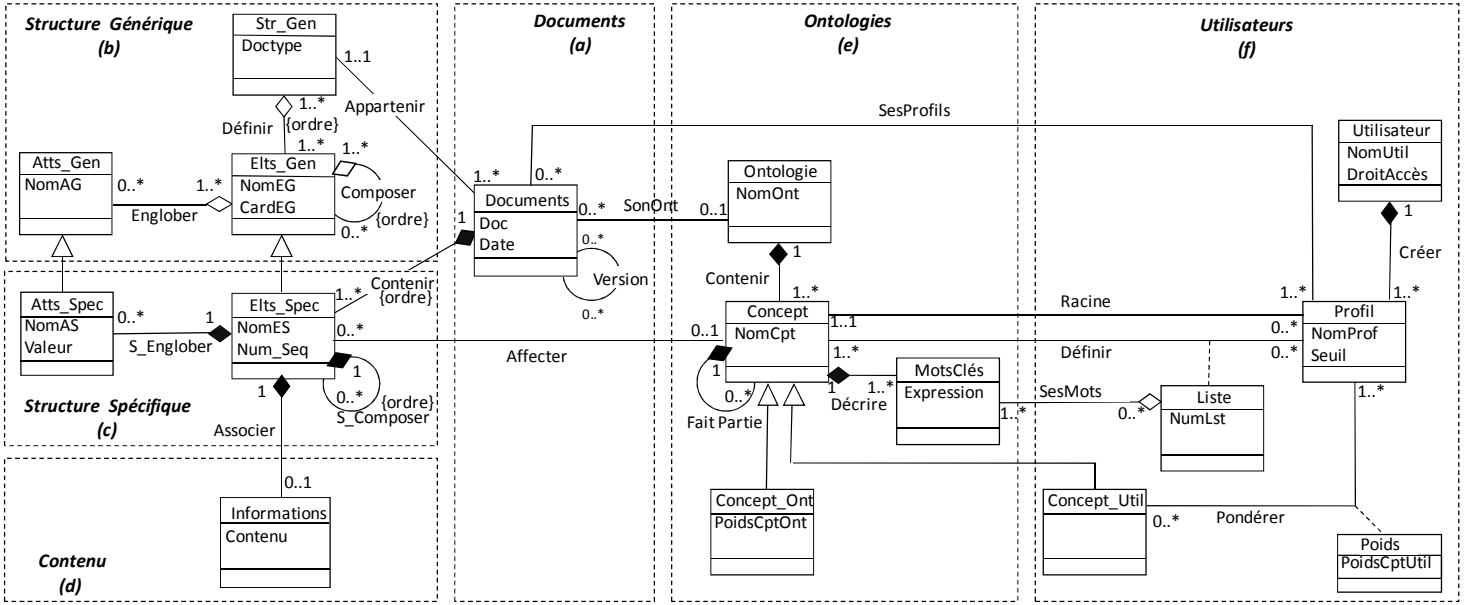


FIG. 6 – *Méta-modèle étendu d'entrepôts de documents*

4.3 Affectation des documents

Une fois que les profils utilisateurs ont été définis, il s'agit maintenant d'affecter les nouveaux documents aux profils. Plus précisément, si un utilisateur ajoute un nouveau document *d* dans l'entrepôt, le système prépare automatiquement un degré de similarité *Sim* (cf. Formule 1) entre le document *d* et chacun des profils de tous les utilisateurs. Ultérieurement, un système de recommandation des documents ajoutés informera tout utilisateur de la présence du nouveau document *d* et des degrés de similarité de *d* par rapport à ses profils. Notons qu'un utilisateur ne sera informé du document *d* que si le degré de similarité *Sim* soit supérieur à un seuil, fixé par lui-même (cf. FIG. 6.f, attribut *Seuil* de la classe *Profil*).

$$Sim(d, P_i) = \frac{\sum_{j=1}^{|Pi|}\left(\sum_{k=1}^{|Cpi,j|} freq(d, T_k) * C_{Pi,j}^{Poids}\right)}{\sum_{j=1}^{|Pi|}\left(\sum_{k=1}^{|Cpi,j|} freq(d, T_k)\right)} \quad [1]$$

Avec :

- $Sim(d, P_i)$: Degré de similarité du document *d* par rapport au profil P_i,
- $|C_{Pi,j}|$: Nombre de termes du concept *j* appartenant au profil P_i,
- $|P_i|$: Nombre de concepts du profil P_i,
- $Freq(d, T_k)$: fréquence d'apparition du terme T_k du concept $C_{Pi,j}$ dans *d*,
- $C_{Pi,j}^{Poids}$: Poids du concept $C_{Pi,j}$.

5 Implantation et validation

Nous avons intégré dans notre outil *DocWare* (*Doc*ument *Ware*house) la personnalisation présentée dans cet article.

La définition des profils se fait à travers les interfaces graphiques du prototype. Tout utilisateur peut créer ses propres profils. Pour cela, il commence par choisir une ontologie, parmi celles stockées dans l'entrepôt. Ensuite, le système la visualise graphiquement. En utilisant un menu contextuel, l'utilisateur choisit le concept racine pour son profil et procède aux opérations souhaitées :

- Ajout et suppression des concepts.
- Ajout et suppression des termes.
- Affectation et modification des poids aux concepts ajoutés.

La FIG. 7 présente un exemple de visualisation d'une ontologie simplifiée de l'entrepôt, à partir de laquelle l'utilisateur définira son profil.

Une fois l'ontologie de la FIG. 7 affichée, l'utilisateur a choisi le concept racine *OLAP*, supprimé le concept *Storage* du profil (ses concepts feuilles ont été automatiquement supprimés) et ajouté le concept *Operators* au dessous de *OLAP* et les trois concepts fils : *Slice*, *Dice* et *Rotate* (cf. FIG. 8).

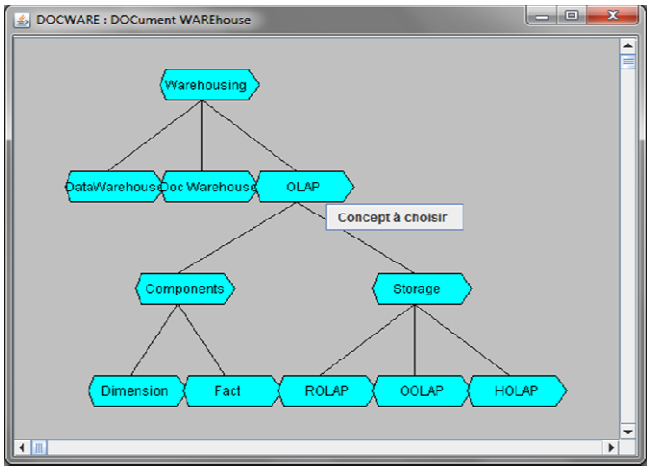


FIG. 7 – *Visualisation graphique d'une ontologie simplifiée de l'entrepôt*

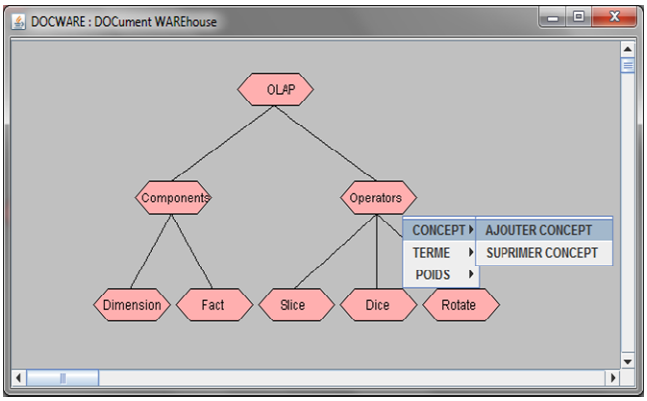


FIG. 8 – *Profil utilisateur construit à partir de l'ontologie de la FIG 7*

Nous avons aussi réalisé la partie affectation des documents aux profils. L'utilisateur choisit et ajoute un document à l'entrepôt. Le système calcule le degré de similarité entre ce document et les profils de cet utilisateur, puis il affiche le ou les profils correspondants avec les degrés calculés. Il est à noter que chaque utilisateur dispose automatiquement d'un profil *Other* contenant les documents ajoutés et qui ne correspondent à aucun de ses profils. Ces documents peuvent être repris ultérieurement par le système en cas d'ajout d'un nouveau profil par l'utilisateur.

Dès qu'un autre utilisateur accède à son compte, il sera informé de l'ajout du document et s'il correspond à un de ses profils. Il peut accepter ce document dans son profil ou l'affecter à *Other* ou même le rejeter.

La FIG. 9 présente un exemple de notification qu'affiche le système pour recommander à l'utilisateur *Yahia Hamdi* les deux documents insérés par *Kaïs Khrouf* et *Jamel Feki*.

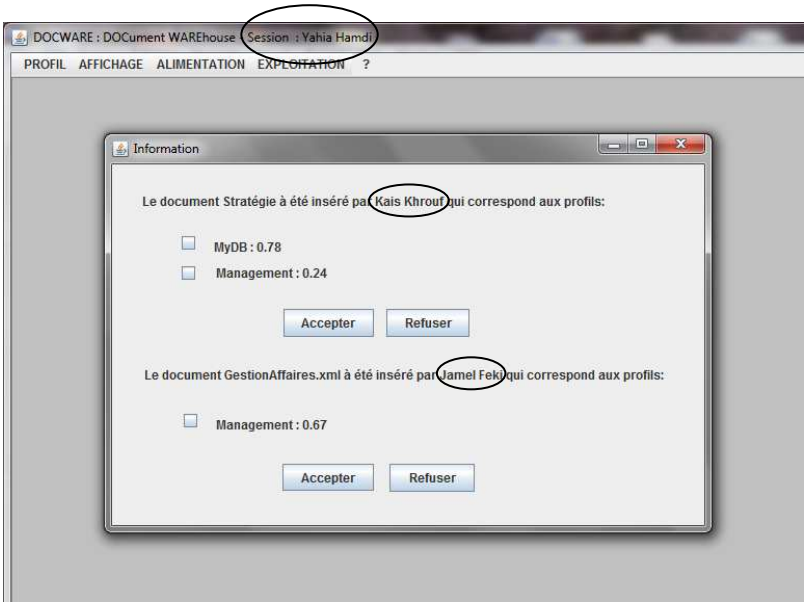


FIG. 9 – *Notification d'ajouts de documents*

6 Conclusion

Ce travail présente l'intégration de la « personnalisation » dans les entrepôts de documents. Dans nos travaux, le profil n'est pas une suite de mots-clés mais plutôt une ontologie réduite et adapté aux besoins et préférences de l'utilisateur. Pour ce faire, nous avons étendu le méta-modèle proposé par l'ajout de la composante *Utilisateur* et la modification de la composante *Ontologie*. Enfin, nous avons proposé une formule pour le calcul du degré de similarité entre tout nouveau document inséré et les profils utilisateurs existants. Ces travaux ont été réalisés et intégrés dans notre outil *DocWare* (*Doc*ument *Ware*house).

Plusieurs perspectives sont envisageables. Dans l'immédiat, nous comptons effectuer des tests sur une collection de documents concernant la formule proposée du calcul de degré de similarité pour l'affectation de documents aux profils et ce afin de la vérifier et de la valider. A court terme, nous envisageons intégrer la personnalisation dans la phase d'interrogation (au niveau des schémas multidimensionnels). Plus précisément, nous comptons adapter le fonctionnement des opérateurs OLAP à la personnalisation.

Références

Bellatreche L., Giacommetti A., Marcel P., Mouloudi H., Laurent D. (2005). *A Personalization Framework for OLAP Queries.* International Workshop on Data Warehousing and OLAP (DOLAP'05), pp. 9-18, Bremen, Germany.

Ben Meftah S., Khrouf K., Feki J., Ben Kraiem M., Soulé-Dupuy C. (2012). *Document Warehouse: Integration of Semantic Structures*, International Conference on Information Systems and Economic Intelligence (SIIE'12), Djerba, Tunisia, To appear.

Favre C., Bentayeb F., Boussaïd O. (2007). *Evolution et personnalisation des analyses dans les entrepôts de données : une approche orientée utilisateur*. Congrès Informatique des organisations et systèmes d'information et de décision (Inforsid'07), pp. 308 – 323, Perros-Guirec, France.

Jerbi. H., Pujolle G., Ravat F., Teste O. (2010) *Personnalisation de Systèmes OLAP Annotés*, Congrès Informatique des organisations et systèmes d'information et de décision (Inforsid'10), Marseille, France.

Khrouf K., Feki J., Soulé-Dupuy C. (2012). Document Warehouse: From Text to Knowledge, *Eletronic Journal of Digital Enterpise* (EJDE), To appear.

Khemyri R., Bentayeb F. (2010). *Utilisation des vues matérialisées pour la personnalisation des entrepôts de données*, Atelier des Systèmes Décisionnels (ASD'10), pp. 121-129, Sfax, Tunisie.

Sullivan D. (2001). *Document Warehousing and Text Mining*, Wiley Edition.

Tseng F.S.C., Chou A.Y.H. (2006). The concept of Document Warehousing for Multi-Dimensional Modeling of Textual based Business Intelligence. *Decision Support Systems*, Vol.42, PP. 727-744, Elsevier.

Summary

The document warehouses allow to decision-makers the extraction of useful information from the managed documents. For that, we have proposed a meta-model of document warehouses for applying the OLAP (On-Line Analytical Processing) techniques to documentary information. In this paper, we have focused on the personalization of document warehouses in order to adapt their content according to users' needs and preferences. Specifically, we proposed an extended meta-model for integrating the *User* component that describes the personalization. We have developed a software prototype that implements the users' profiles definition and document recommendation.

Entrepôts de Données Spatiales : Vers une Classification et Évaluation des Modèles Conceptuels

Boubaker Boulekrouche*, Zaia Alimazighi*,
Nafaa Jabeur**
* Faculté Electronique et Informatique, Laboratoire des Systèmes Informatiques (LSI)
USTHB, BP 32 EL Alia, Bab Ezzouar Alger, Algerie, 16111
Boubakernour@gmail.com
alimazighi@wissal.dz
** Computer Science Department, College of Arts and Applied Sciences, Dhofar University, Salalah, Postal Code: 211, Sultanate of Oman
nafaa_jabeur@du.edu.om

Résumé. On assiste ces dernières années à l'émergence de nouveaux systèmes d'aide à la décision basés sur les Entrepôts de Données Spatiales (EDSs). Plusieurs modèles conceptuels ont été proposés pour la mise en œuvre de ces EDSs dans différents domaines d'applications. Les quelques tentatives qui ont été faites pour la classification de ces modèles se sont particulièrement basées sur des propriétés liées au contexte d'étude présenté. Étant donné le manque de recherches sur l'état actuel de ces modèles ainsi que leurs projections futures, nous proposons dans cet article une revue ainsi qu'une classification de ces modèles conceptuels avec l'objectif ultime de faciliter leur évaluation et favoriser l'identification de nouveaux axes de recherche. Notre étude se distingue particulièrement par la classification, l'évaluation et la comparaison de ces modèles selon les caractéristiques normatives de la modélisation conceptuelles des EDSs ainsi que les propriétés multidimensionnelles spatiales.

1 Introduction

Les Systèmes d'Aide à la Décision (SAD) sont des systèmes d'information qui permettent aux décideurs d'effectuer des analyses complexes. Lorsqu'il s'agit de grand volume de données, les entrepôts de données (ED) sont probablement les systèmes d'aide à la décision les plus utilisés étant donné leurs outils d'analyse et de découverte de connaissances. Un entrepôt de données est une collection de données portant sur des sujets touchant une organisation, intégrée, variant dans le temps, et non-volatile pour supporter le processus de prise de décision, Inmon (1996). Afin de supporter le traitement analytique enligne, les entrepôts de données utilisent des techniques de type OLAP (On-Line Analytical Processing). Ces techniques ne présentent pas d'outils pour la gestion des données de type spatiales et ce malgré leur importance. En effet, il est reconnu qu'environ 80% des données présentes dans les différentebases de données mondiales ont une composante spatiale qui est souvent inexploitée, Franklin (1992). De nombreux SIG (Systèmes d'Information Géographique) offrent des fonctionnalités d'analyse permettant d'exploiter ces composantes spatiales. Cependant, ils ne sont pas adaptés pour l'agrégation de données, Keenan (1996). Afin de pouvoir bénéficier du rôle capital des données spatiales dans le processus décisionnel, et notamment son pouvoir d'expression et d'analyse, les systèmes d'aide à la décision spatiale (SADS), basés sur des entrepôts de données spatiales, ont été émergés.

Un entrepôt de données spatiales (EDS) est une collection de données spatiales et thématiques, intégrées, historiées et non volatiles pour la prise de décisions spatiales, Stefanovic et al.(2000). Cet entrepôt adopte le paradigme multidimensionnel pour l'analyse de grands volumes de données spatiales et alphanumériques et reposent sur de nouveaux concepts de dimensions spatiales, mesures spatiales, hypercube et SOLAP (Spatial OLAP). Les architectures des systèmes d'entrepôts de données spatiales sont des architectures à trois niveaux, comme illustré en Figure 1. Ces architectures se constituent par un entrepôt de données spatiales, un serveur SOLAP et un client SOLAP.

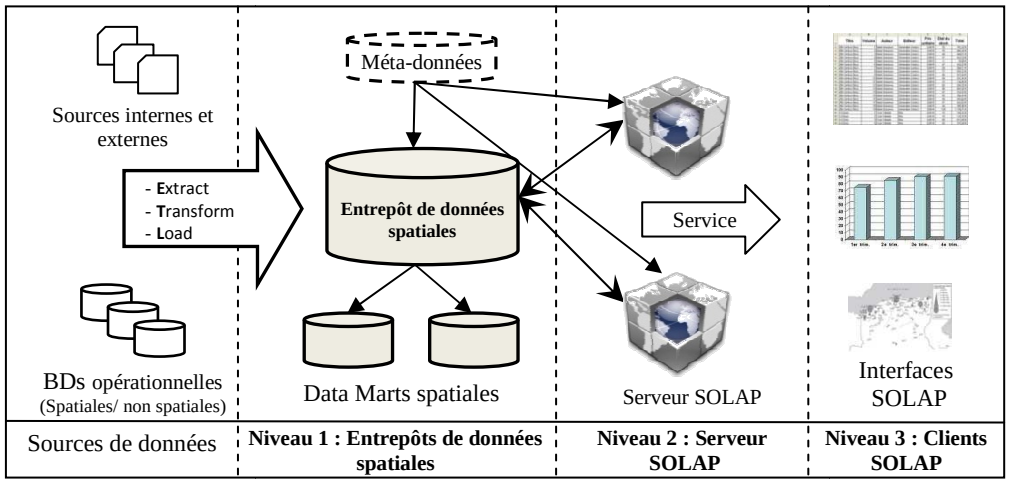


FIG. 1 – Architecture d'un système d'entrepôt de données spatiales (adaptée de Bimonté (2007)).

Plusieurs modèles conceptuels ont été proposés pour la mise en œuvre des EDSs. Néanmoins, les quelques tentatives pour la classification de ces modèles se sont particulièrement basées sur des propriétés liées au contexte d'étude présenté. Afin de montrer une meilleure idée sur les modèles conceptuels existants, nous proposons dans cet article une revue, une classification, une évaluation et comparaison de ces modèles selon les caractéristiques de la modélisation conceptuelles des EDSs ainsi que les propriétés multidimensionnelles spatiales. Dans la Section 2 de notre travail, nous présentons les travaux liés à la classification des modèles des EDSs. Dans la Section 3, nous décrivons les propriétés de la modélisation conceptuelles des EDSs. Dans la Section 4, nous proposons une classification des travaux liés à la modélisation conceptuelle des EDSs. Dans la Section 5, nous abordons une comparaison et une évaluation de ces travaux. Nous concluons l'article dans la Section 6.

2 Travaux connexes

Plusieurs travaux liés à la mise en œuvre et la modélisation des systèmes basés sur les EDSs ont été proposés dans la littérature. Néanmoins, il y a très peu de travaux qui se sont intéressés à une revue des travaux liés à la modélisation conceptuelle des EDSs ainsi que leur classification et comparaison.

Malinowski et Zimányi (2008) ont proposé une revue des différents travaux liés à la modélisation et la mise en œuvre des systèmes basés sur les entrepôts de données spatiales. Dans ce

cadre, ils ont présenté des concepts et des définitions des propriétés multidimensionnelles spatiales selon ces travaux sans aucune classification.

Da Silva et al. (2009) ont fait un survol de quelques travaux sur les métas modèles des entrepôts de données spatiales en faisant une comparaison par rapport aux propriétés du méta modèle proposé. Cette comparaison a été faite en fonction: (i) des définitions formelles des concepts; (ii) des langages et standard de modélisation; (iii) des standards liés aux SIG et entrepôts de données et enfin (iv) de la définition de pictogrammes. Néanmoins, ce travail n'a pas proposé une classification de ces modèles. Gómez et al. (2010) ont fait un survol des travaux liés à la modélisation et mise en œuvre des entrepôts de données spatio-temporelles et spatiales. Les auteurs ont présenté les concepts multidimensionnelles spatiales (dimensions, mesures, faits, agrégation, hiérarchie, …) selon les points de vue de ces différents travaux. Néanmoins, ils n'ont pas proposé une classification de ces travaux ni des critères de comparaison et d'évaluation de ces modèles.

Del Aguila et al. (2011) ont présentés un méta modèle pour la conception et la validation des schémas des entrepôts de données spatiales. Dans ce travail, les auteurs ont évalués quelques travaux liés à la modélisation des EDSs par rapport aux caractéristiques spécifiques du méta modèle proposé. Cette évaluation s'est basée sur: (i) la séparation entre la modélisation des entrepôts de données et OLAP; (ii) le développement des outils de type CASE (Computer-Aided Software Engineering); (iii) le support de dimensions de type : dégénérées; jeux de rôle et les relations plusieurs à plusieurs; (iv) le support des attributs spatiaux ; (v) le support de mesure spatiale; (vi) l'adoption de pictogrammes et enfin (vii) la définition des attributs spatiaux normalisés et /ou partagés. Néanmoins, ce travail n'a pas proposé une classification de ces modèles.

Viswanathan et Schneider (2011) ont décrit un méta-Framework pour la modélisation des entrepôts de données spatiales. Les auteurs ont proposé une classification des modèles conceptuels des EDS qui se résume comme suit :(i) extension du modèle Entité/Association (E/A), (ii) basé sur les diagrammes UML (Unified Modeling Language) et (iii) les approches de modélisation Ad-hoc. Quelques travaux relatifs à cette classification ont été énumérés sans aucune comparaison.

Les différentes tentatives de classification et de comparaison des travaux liés aux EDSs se sont basées, particulièrement, sur des propriétés intrinsèques des modèles proposés. Ceci rend ces classifications spécifiques au contexte d'étude présenté. Nous pensons qu`une revue plus exhaustive des travaux existants pourrait faciliter la classification et la comparaison des EDSs et par la suite identifier de nouveaux axes de recherche particulièrement ceux visant une meilleure prise en compte des données spatiales dans les systèmes d`aides à la décision.

3 Caractéristiques de la modélisation des EDSs

La modélisation des EDs est une étape cruciale car elle assure leur développement selon les exigences des utilisateurs. La plupart de méthodologies de modélisation des EDs intègrent trois phases indispensables : (i) la modélisation conceptuelle ; (ii) la modélisation logique et (iii) la modélisation physique. Les caractéristiques de chacune de ces phases subissent des adaptations et des extensions pour tenir compte des spécificités des données spatiales.

La modélisation conceptuelle est une étape clé dans le développement des EDSs. Elle hérite des propriétés de la modélisation conceptuelle des entrepôts de données conventionnelles. Trujillo et al.(2001) définissent deux types de propriétés de la modélisation conceptuelle

multidimensionnelle pour les entrepôts de données conventionnelles: (i) propriétés niveau structurel qui concernent la définition de dimensions, faits, mesures, et (ii) propriétés niveau opérationnel qui concernent les propriétés relatives à l'analyse de données, telles que les opérateurs OLAP. Afin d'évaluer les performances des approches adoptées pour la modélisation conceptuelles des EDSs nous pouvons classifier leurs propriétés de modélisation en deux catégories: (i) propriétés liées aux concepts multidimensionnels spatiaux et (ii) propriétés liées à la modélisation conceptuelle.

1. **Les propriétés liées aux concepts multidimensionnels spatiaux:** une approche de modélisation des EDSs est évaluée en fonction de sa capacité à supporter les concepts suivants:
   (a) Les dimensions spatiales: plusieurs définitions de dimension spatiale existent dans la littérature. Ces définitions représentent plusieurs points de vues liées à l'intégration de l'information spatiale en tant qu'axe d'analyse dans un système décisionnel.
   (b) Les hiérarchies spatiales: une hiérarchie spatiale est une hiérarchie qui contient au moins un niveau spatial. Un niveau est dit spatial s'il contient au moins un attribut géométrique. Les niveaux spatiaux sont liés entre eux par des relations topologiques (intersection, inclusion, imbrication, égalité, etc.). Malinowski et Zimányi (2008) ont défini plusieurs types des hiérarchies spatiales que nous pouvons catégoriser de simples et complexes. Dans les hiérarchies spatiales simples, les relations entre les membres des niveaux peuvent être représentées par un arbre. Ce type d'hiérarchie est associé à un seul critère d'analyse, et peuvent être symétrique, asymétrique ou généralisé. Les hiérarchies spatiales complexes sont caractérisées par des relations complexes entre les membres de niveaux. Ces relations peuvent être de type: alternatives multiples, parallèles, non-strictes.
   (c) Les faits spatiaux : Un fait est considéré comme le sujet analysé dans un entrepôt de données. Il est décrit par un ensemble de mesures. Un fait est dit spatial s'il intègre plus d'une dimension spatiale, Malinowski et Zimányi (2004).
   (d) Les mesures spatiales: le concept de mesure spatiale peut être vue de multiples façons: (i) une collection de pointeurs vers des objets spatiaux Stefanovic et al.(2000), Rivest et al. (2005), Malinowski et Zimányi (2004); (ii) le résultat d'opérateurs spatiaux tels que les d'opérateurs topologiques spatiaux (union, intersection, …) Rivest et al. (2005), et les opérateurs métriques (distance entre deux régions) et enfin (iii) comme un membre spatial d'une dimension, Marchand et al. (2003). Une mesure spatiale dérivée peut être calculée à partir d'autres mesures spatiales.
   (e) Additivité des mesures spatiales: toute modélisation d'un EDS doit explicitement supporter l'agrégation de mesures spatiales selon les membres de dimensions, Viswanathan et Schneider (2011).
   (f) Les relations topologiques entre les niveaux spatiaux: si les niveaux liés par une relation de type père-fils sont spatiaux, alors il y a plusieurs relations topologiques entre ces deux niveaux (intersection, inclusion, imbrication, égalité,…), Pedersen et Tryfona (2001). Par conséquent, le calcul des agrégations des mesures spatiales doit tenir compte de ces relations afin d'éviter de prendre en compte plusieurs fois la même mesure spatiale.
   (g) Les relations plusieurs à plusieurs entre fait et dimensions (N-N) : généralement, la relation entre un fait et les dimensions est de type (N-1). Cependant, il y a des

cas qui nécessitent la prise en charge des relations de type (N-N) entre un fait est une dimension particulière, Trujillo et al. (2001).

2. **Les propriétés liées à la modélisation conceptuelle:** ces propriétés facilitent aux concepteurs et aux utilisateurs la création, la compréhension et la gestion des schémas conceptuels. Ces propriétés peuvent être résumées comme suit:
   (a) Définition formelle des concepts multidimensionnels: ces définitions facilitent la spécification des propriétés multidimensionnelles spatiales, ce qui facilite ainsi la modélisation et l'implémentation des EDSs.
   (b) Représentation graphique des structures et concepts: il s'agit de l'adoption ou la définition de notation graphiques pour la représentation des propriétés multidimensionnelles spatiales et les relations y afférentes.
   (c) Définition des icones spatiales (pictogrammes) pour la représentation des faits, des mesures, et les relations topologiques entre les niveaux de dimensions.
   (d) Définition de Méta-modèle pour faciliter la description des schémas et structures du modèle conceptuel.
   La figure 2 présente les deux catégories de propriétés relatives à la modélisation conceptuelle des entrepôts de données spatiales.

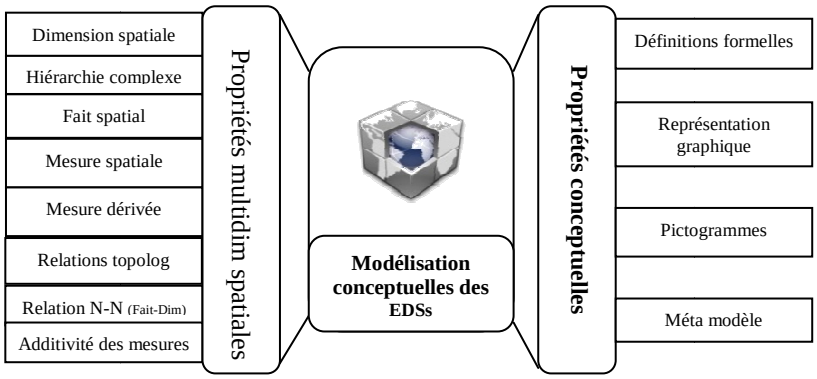


FIG. 2 – Propriétés de la modélisation conceptuelle des EDSs.

4 Revue et classification des modèles conceptuelles

Dans cette section nous présentons les différents travaux liés à la conception des systèmes basés sur les Entrepôts de Données Spatiales (EDSs), proposés dans la littérature. Il n'existe à ce jour aucun consensus sur la modélisation conceptuelle des entrepôts de données spatiales comme cela peut être le cas avec la méthode MERISE pour la conception des bases de données relationnelles. Ainsi, les différents travaux proposés dans la littérature peuvent être classifiés selon le modèle conceptuel adopté :(i) basés sur l'extension du modèle Entité/Association (E/A), (ii) basés sur le modèle orienté objets et (iii) les modèles spécifiques.

4.1 Travaux basés sur l'extension du modèle Entité/Association (E/A)

Le modèle (E/A) ne permet pas de représenter les concepts multidimensionnels car il n'offre pas de moyens pour la description de la modélisation des entrepôts de données et no-

tamment ceux relatifs à l'analyse dans un contexte décisionnel, Torlone, (2003). Par conséquent, les travaux basés sur l'extension du modèle (E/A) doivent étendre ce modèle pour tenir compte des concepts : fait spatial, dimension spatial et mesure spatiale. Dans ce qui suit nous présentons deux travaux qui adoptent ce type de modèle pour la modélisation des EDSs.

Dans ce contexte, le modèle MultiDimER (Malinowski et Zimányi, 2004 ; Malinowski and et Zimányi (2005), a étendu le Entité/Relation, avec les concepts de dimension, hiérarchie et mesure spatiales. Ce modèle présente une représentation graphique pour les propriétés multidimensionnelles, ainsi que des icones spatiales (pictogrammes) pour la représentation des faits, des mesures spatiaux, des types de données spatiales et les relations topologique entre les niveaux de dimensions spatiales. En outre. Il offre des symboles pour représenter les opérateurs d'agrégation de mesures spatiales. Il défini également un modèle formel correspondant aux concepts multidimensionnels ainsi qu'aux opérateurs SOLAP Finalement. MlutiDimER, présente une classification des différents types des hiérarchies spatiales complexes, souvent ignorées dans les autres modèles.

Le travail proposé dans Pedersen et Tryfona (2001) vise l'amélioration des techniques de pré agrégation dans les EDSs par l'exploitation et l'analyse des propriétés des relations topologiques entre les objets spatiaux. Ce travail propose une représentation formelle et un modèle de données adaptés à l'agrégation de mesures spatiales. Il donne également une représentation graphique basé sur le modèle E/A et propose quelques définitions formelles des faits, dimensions, relations faits-dimensions, hiérarchies, et leurs propriétés relatives à l'agrégation. Cependant, ce travail n'a pas proposé des définitions formelles des concepts ni de méta modèle.

4.2 Travaux basés sur le modèle orienté-objets (OO)

Dans le cadre de la modélisation conceptuelle des EDs, le paradigme orienté-objets (OO) supporte plusieurs types de dimensions (généralisation/spécialisation, classification/instanciation, agrégation/composition, …) (Abelló et al., 2000). Il adopte le standard UML qui est naturellement extensible pour la représentation et la modélisation des concepts multidimensionnels. Par conséquent, les travaux de conception des EDSs basés sur le modèle orienté-objets profitent de la popularité de la modélisation UML et sa puissance d'expression des aspects statiques et dynamiques afin de tenir compte de la composante spatiale. Dans ce qui suit, nous présentons une panoplie de travaux, qui adoptent le modèle OO pour la modélisation des EDSs.

Da Silva et al. (2009) définissent un méta modèle permettant la spécification des schémas des EDSs par la définition des concepts (dimensions conventionnelles et géographiques, mesures géographiques) de ces schémas. Ce méta modèle étend les définitions du GeoDWFrame défini dans Fidalgo et al. (2004) par la prise en compte des mesures spatiales. Il se base sur le langage UML ainsi que les standards CWM et OGC. Aussi, ce modèle utilise des pictogrammes pour améliorer son expressivité, cependant, il n'a pas proposé de définitions formelles et ne tient pas compte des hiérarchies complexes.

Del Aguila et al. (2011) proposent un méta modèle pour la conception et la validation de schémas des entrepôts de données spatiales. Contrairement aux métas modèles proposés dans la littérature, ce dernier sépare entre la modélisation des entrepôts de données (dimensions, faits) et la modélisation de cubes OLAP (hiérarchies et niveaux). Il permet aussi de spécifier si la géométrie d'un attribut spatial peut être normalisée et /ou partagée et il propose un ensemble de pictogrammes pour la représentation concise des attributs spatiaux et non spatiaux des propriétés multidimensionnelles. Cependant, il n'a pas proposé des définitions formelles

pour les propriétés multidimensionnelles et ne tient pas compte des hiérarchies complexes et des relations topologiques entre niveaux spatiaux.

Le modèle proposé par Fidalgo et al. (2004) se base sur une architecture appelé GOLAPA (Geographical On-Line Analytical Processing Architecture) visant l'intégration des technologies liées aux entrepôts de données, OLAP et les SIGs. Cette architecture offre un cadre conceptuel dénommé (GeoDWFrame) pour faciliter la conception de schémas de dimensions géographiques. Ce modèle propose deux types de dimensions: dimension géographique et dimension hybride. Ce modèle supporte uniquement les mesures numériques. Ce travail adopte le modèle orienté objets (UML) pour la représentation graphique des dimensions spatiales seulement. En outre, ce modèle n'a pas proposé de modèles formels, ni de définitions pour les faits, hiérarchies, mesures et hiérarchies spatiaux. Il n'a pas aussi proposé de pictogrammes et ne tient pas compte des relations topologiques entre les niveaux spatiaux. Cependant ce travail a proposé des métas modèles basés sur UML et CWM (Ommon Warehouse Metamodel) et les standards OGC (Open Geospatial Consortium).

Glorio et Trujillo (2009) définit une méthode basée sur une architecture dirigée par les modèles (MDA : Model Driven Architecture) pour le développement et la conception des entrepôts de données spatiales. Cette approche a été définie par extension du méta modèle du langage UML à travers des profiles UML afin de l'adapter pour la conception des SDW. Ces profiles contienent des pictogrammes permettant la représentation concise des propriétés multidimensionnelles (dimensions, mesures, faits, …) au niveau conceptuel. cependant, ce travail n'a pas défini de modèles formels pour les concepts multidimensionnels spatiaux.

4.3 Modèles spécifiques

Les modèles spécifiques adoptent généralement des approches de modélisation multidimensionnelle propriétaires et qui sont généralement non connu par les concepteurs. Ainsi, les travaux qui adoptent les modèles spécifiques pour la modélisation conceptuelle des EDSs peuvent être résumés comme suit.

Bimonte (2007) propose un modèle formel appelé GeoCube et une algèbre associée, qui vise la reformulation des principaux concepts de SOLAP, pour définir un nouveau concept dénommé OLAP Géographique. Ce concept introduit la composante spatiale et la composante sémantique de l'information géographique dans l'analyse multidimensionnelle. Selon ce modèle, une mesure géographique est similaire à un niveau d'une dimension géographique. Il s'agit d'une vision complètement symétrique entre mesures et dimensions géographiques. Ce modèle adopte la notation graphique du modèle conceptuel pour les EDSs du Malinowski et Zimányi (2004) avec l'ajout d'une représentation graphique pour une mesure géographique et des attributs dérivés. En outre, ce modèle n'a pas donné une définition explicite pour les faits spatiaux, et n'a pas proposé de méta modèle pour ces concepts.

Damiani et Spaccapietra (2006) présentement le modèle dénommé MuSD (Multigranular Spatial Data warehouse) et définit une algèbre pour modéliser et interroger les EDSs. Ce modèle définit les concepts de dimension, fait, mesure, et cube spatiaux. Il a défini les entités spatiales selon les standards de la modélisation géographique (OGC). Ce modèle présente une innovation relative à la représentation de la mesure spatiale à plusieurs granularités géométriques. Cependant, ce modèle n'a pas présenté une représentation graphique, ni de pictogrammes ni de méta modèle. En outre, il ne prend pas en compte les hiérarchies spatiales complexes et les relations topologiques.

Le travail proposé par Jensen et al (2004) défini un modèle et une algèbre pour les applications multidimensionnelles pour les services basés sur la localisation, permettant de capturer les relations topologiques d'inclusion partielle ou totale entre les membres d'une dimension spatiale. Ce modèle utilise le concept de schéma de fait pour la définition de la structure de l'entrepôt de données. Ce travail a défini les concepts de dimension, fait, hiérarchies. Cependant, il n'a pas proposé ni de pictogrammes ni de méta modèle. En outre, il ne permet pas de modéliser les mesures spatiales car il utilise les niveaux de dimensions comme mesures.

Contrairement aux travaux cités précédemment, Ahmed et Miquel (2005) ont introduit une vision continue de l'espace géographique dans les entrepôts de données spatiales. Le modèle proposé est basé sur la notion de « cube de base discret » qui est représenté par une liste de dimensions comprenant une mesure M, une liste des niveaux les plus détaillés des dimensions et un ensemble de cellules. Des valeurs estimées sont dérivées du cube de base et sont calculées grâce à des fonctions d'interpolation. Ce travail a partiellement présenté quelques propriétés relatives à la modélisation conceptuelle car il introduit une vision continue de l'espace qui est différente par rapport à la vision discontinue adoptée par les autres approches.

Ferri et al.(2000) proposent une approche conceptuelle permettant de supporter des requêtes nécessitant l'accès simultané à une base de données géographique (BDG) et un entrepôt de données, à travers l'extension de la BDG avec des données multidimensionnelles. Cette extension se base sur l'introduction des attributs appelés " attributs fonctionnels dans la BDG. A chaque attribut fonctionnel de la BDG correspond un et un seul cube. Cependant, ce modèle n'a pas présenté une représentation graphique, ni de pictogrammes ni de méta modèle. Néanmoins, nous n'avons pas assez d'information pour évaluer l'approche proposée par rapport aux propriétés: hiérarchies spatiales complexes, relations topologiques.

Salehi et al. (2010) propose un modèle formel au niveau conceptuel pour les cubes de données spatiales tout en mettant l'accent sur la définition précise et formelle des différents composants spatiaux (dimensions spatiales, mesures spatiales, fait spatial, cube spatial) au niveau schéma et au niveau instance. Ce travail n'a pas défini explicitement une représentation graphique et des pictogrammes pour la modélisation conceptuelle des cubes spatiaux. En outre, il n'a pas proposé de méta modèle pour le schéma conceptuel et ne prend pas en compte les hiérarchies spatiales complexes.

5 Comparaison et évaluation des modèles

Les travaux présentés ont été évalués selon les propriétés de la modélisation conceptuelle des EDSs, présentées dans la Section 3. Le tableau 1 présente une matrice d'évaluation et de comparaison de ces travaux. Nous avons utilisé quatre symboles dans cette matrice pour l'évaluation de ces modèles. Le symbole ☑ signifie que le modèle supporte cette propriété, le symbole ± indique que le modèle supporte partiellement cette propriété et le symbole – indique que le modèle ne supporte pas cette propriété. Cependant, Le symbole **?** indique que nous n'avons pas assez d'information pour évaluer le modèle par rapport à cette propriété.

		Propriétés conceptuelles				Propriétés multidimension-nelles spatiales							
Classe	**Modèles**	Définitions formelles	Représentation graphique	Pictogrammes	Méta-modèle	Dimension spatiale	Hiérarchie complexe	Fait spatial	Mesure spatiale	Mesure spatiale dérivée	Relations topologiques	Relation N-N (fait- dim)	Additivité des mesures
Basés sur le modèle E/A	Malinowski et Zimányi (2004)	☑	☑	☑	☑	☑	☑	☑	☑	–	±	–	☑
	Pedersen et Tryfona (2001)	☑	±	±	–	☑	☑	☑	☑	?	☑	–	☑
Basés sur Le modèle OO	Da Silva et al. (2009)	☑	☑	☑	☑	☑	–	☑	☑	–	±	–	☑
	Del Aguila et al. (2011)	–	☑	☑	☑	☑	–	☑	☑	?	–	☑	☑
	Fidalgo et al (2004)	–	±	–	☑	☑	–	–	–	–	–	–	–
	Glorio et Trujillo (2009)	–	☑	☑	±	☑	☑	☑	☑	?	–	–	☑
Modèles spécifiques	Ahmed et Miquel (2005)	☑	±	±	–	±	–	±	±	☑	–	–	☑
	Bimonté (2007)	☑	☑	☑	–	☑	☑	±	☑	☑	±	–	☑
	Damiani et Spaccapietra (2006)	☑	–	–	–	☑	–	±	±	–	–	?	☑
	Ferri et al. (2000)	±	±	–	–	☑	?	–	–	–	–	?	–
	Jensen et al (2004)	☑	±	–	–	☑	☑	☑	–	?	☑	☑	☑
	Salehi et al. (2010)	☑	±	±	–	☑	–	☑	☑	–	±	–	±

TAB. 1 –Evaluation et comparaison des modèles

L'évaluation et la comparaison des modèles proposés et leurs classes peuvent se résumer comme suit :

1. Evaluation et comparaison des classes des modèles:

Nous pouvons dire initialement que la plupart des travaux se basent sur des modèles spécifiques. Ces modèles offrent des définitions formelles des concepts mieux que les modèles des deux autres classes. Malgré que les modèles de cette classe supportent partiellement les propriétés conceptuelles et multidimensionnelles spatiales, nous pouvons dire que les modèles spécifiques supportent beaucoup plus les propriétés multidimensionnelles spatiales que les propriétés conceptuelles.

Cependant, nous constatons qu'il y a très peu de travaux qui adoptent le modèle E/A. Ces modèles supportent beaucoup plus les propriétés multidimensionnelles spatiales que les propriétés conceptuelles à l'exception du modèle proposé par Malinowski et Zimányi (2004) qui supporte presque toutes les propriétés. Selon le tableau 1, les travaux basés sur le modèle orienté-objets supporte mieux les propriétés conceptuelles que les autres modèles. On peut constater aussi, que les travaux appartenant à cette catégorie définissent des métas modèles, qui représentent une propriété souvent ignorées par les modèles des

autres classes. Enfin, nous pouvons déduire que les travaux basés sur le modèle orienté objets supporte les propriétés conceptuelles et multidimensionnelles spatiale mieux que les deux autres classes de modèles. Ceci est dû principalement aux techniques de modélisation offertes par le paradigme orienté-objets ainsi que l'extensibilité du langage UML qui permet la création de diagrammes personnalisés.

2. Evaluation et comparaison des modèles :

Selon le tableau 1, il y a que les modèles proposés par Del Aguila et al. (2011), Jensen et al. (2004) qui supportent les relations de type plusieurs à plusieurs (N-N) entre un fait et une dimension. Les relations topologiques entre les niveaux spatiaux des dimensions sont supportées partiellement par quelques modèles, néanmoins, seulement les deux modèles proposés par Pedersen et Tryfona (2001) et Jensen et al (2004) supportent cette propriété. Enfin, et selon la matrice de comparaison et d'évaluation des modèles proposés, nous pouvons dire que les quatre meilleurs travaux sont ceux proposés par Bimonté (2007), Da Silva et al. (2009), Del Aguila et al. (2011) et Malinowski et Zimányi (2004). Néanmoins, le modèle conceptuel proposé par Malinowski et Zimányi (2004) semble légèrement meilleur que les trois autres modèles.

6 Conclusion

Ces dernières années ont été marqués par l'émergence de nouveaux systèmes d'aide à la décision basés sur les Entrepôts de Données Spatiales (EDSs). Cependant et malgré l'importance de la conception de ces systèmes, il n'existe à ce jour aucun consensus sur la modélisation conceptuelles des EDSs. Plusieurs modèles conceptuels ont été proposés pour le développement des EDSs. Néanmoins, les quelques tentatives pour la classification de ces modèles se sont particulièrement basées sur des propriétés liées au contexte d'étude présenté. Vue le manque de recherches sur l'état actuel de ces modèles ainsi que leur évaluation, nous proposons dans cet article une revue ainsi qu'une classification des ces travaux selon le modèle conceptuel adopté: (i) basés sur l'extension du modèle Entité/Association (E/A), (ii) basés sur le modèle orienté objets et (iii) les modèles spécifiques. En outre, nous avons évalué et comparé les classes de ces modèles ainsi que les modèles selon les propriétés de la modélisation conceptuelle des EDSs ainsi que les propriétés multidimensionnelles spatiales.

En effet, l'évaluation de ces modèles, nous a permis de définir les critères normatifs pour la mise en œuvre d'un modèle conceptuel pour les EDSs et favorise l'identification de nouveaux axes de recherche et défis en fonction des propriétés non supportées par les modèles proposés dans la littérature. En perspective, nous somme en train de définir une étude de cas pour la représentation des différents modèles étudiés afin d'améliorer leur évaluation et comparaison. Nous envisagerons aussi l'extension de notre méthode d'évaluation de ces modèles afin de tenir compte des propriétés liés à la modélisation logique et physique des Entrepôts de Données Spatiales.

Références

Abelló, A., J. Samos, F. Saltor (2000) Benefits of an Object-Oriented Multidimensional Data Model", In Proceedings of the 14th European Conference on Object-Oriented Programming (ECOOP'00), 141-152.

Ahmed T., M. MIQUEL (2005). Multidimensional Structures Dedicated to Continuous Spatiotemporal Phenomena. In : JACKSON Mike, NELSON David et STIRK Sue. 22th British National Conference on Databases, 5-7 Juillet 2005, Sunderland, UK. Berlin Heidelberg : Springer,, 29-40 p. (Lecture Notes in Computer Science 3567)

Bimonte, S. (2007) Intégration de l'information géographique dans les entrepôts de données et l'analyse en ligne : de la modélisation à la visualisation. Thèse de doctorat, INSA de Lyon.

Damiani, M.L. et S. Spaccapietra. (2006). Spatial Data Warehouse Modeling. Processing and Managing Complex Data for Decision Support, J. Darmont and O. Boussaid (Eds.), Idea Group Inc.,1-27.

Da Silva J, A.G. de Oliveira, R. N. Fidalgo, A.C.Salgado, V.C. Times (2009): Modelling and querying geographical data warehouses, in Information Systems journal

Del Aguila, P.S., R. N. Fidalgo, A. Mota. (2011): Towards a More Straightforward and More Expressive Metamodel for SDW Modeling, in Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP DOLAP 11

Fidalgo, R. N., V. C.Times, J.da Silva, F. F.de Souza, A. C. Salgado (2004). Providing Multidimensional and Geographical Integration Based on a GDW and Metamodels. Journal of information and Data Management, Vol. 1, No. 1, 93–106.

Franklin, C. (1992). An Introduction to Geographic Information Systems: Linking Maps to databases. Database , Vol. 15, n° 2, 13-21.

Glorio O., J. Trujillo. (2009) Designing data warehouses for geographic OLAP querying by using MDA. In ICCSA(1),505–519.

Gómez, L., B.,Kuijpers, B.,Moelans, A.,Vaisman (2010). "A Survey of Spatio-Temporal Data Warehousing." Business Information Systems: Concepts, Methodologies, Tools and Applications 4 vol. IGI Global, 949-977.

Inmon, W. H. (1996). The Data Warehouse and Data Mining , communication of the ACM, , Vol. 39, N° 11 .

Jensen, C.S., A.Kligys , T.B. Pedersen et I. Timko (2004). Multidimensional data modeling for location-based services. International Journal on Very Large Data Bases, Vol.13, n° 1, 1-21.

Keenan, P. (1996). Using a GIS as a DSS Generator. In : DARZENTAS John, DARZENTAS Jenny et SPYROU Thomas. Perspectives on Decision Support System. Grèce : University of the Aegean, 33-40.

Malinowski, E et E. Zimányi (2004) . Representing spatiality in a conceptual multidimensional model. In : PFOSER Dieter, CRUZ Isabel F. et RONTHALER Marc. 12th ACM International Workshop on Geographic Information Systems, 12-13 Novembre, 2004, Washington, DC, USA. New York, USA : ACM Press, 12-22 .

Malinowski, E et E. Zimányi (2005). Spatial hierarchies and topological relationships in the Spatial MultiDimER model. In Proc. of the 22nd British Nat. Conf. on Databases, 17–28.

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

Marchand, P., A. Brisebois, Y. Bédard et G. Edwards (2003). Implementation and evaluation of a hypercube-based method for spatio-temporal exploration and analysis. Elsevier : Journal of the International Society of Photogrammetry and Remote Sensing, Vol. 59, n° 1, 6-20.

Pedersen, TB, Tryfona N (2001) Pre-aggregation in spatial data warehouses. In: Proc. 7th international symposium on advances in spatial and temporal databases, Redondo Beach, CA, 12–15 July 2001, 460-478

Peralta V., A. Illarze, R. Ruggia (2003). On the Applicability of Rules to Automate Data Warehouse Logical Design, In Proceedings of the 15th Conference on Advanced Information Systems Engineering Klagenfurt, Velden, Austria, 329-340.

Rivest, S., Y. Bédard, M.J. Proulx, M. Nadeau, F. Hubert and J. Pastor.(2005) SOLAP: Merging Business Intelligence with Geospatial Technology for Interactive Spatio-temporal Exploration and Analysis of Data. Journal of International Society for Photogrammetry and Remote Sensing (ISPRS), 60(1),17-33.

Salehi, M., Y. Bédard, S. Rivest, (2010). A Formal Conceptual Model and Definition Framework for Spatial Datacubes, Geomatica, Vol. 64, No. 3, pp.

Torlone, R. (2003), "Conceptual Multidimensional Models", Multidimensional Databases, .69-90.

Trujillo, J., M. Palomar, J. Gomez, and I., Song. (2001), Designing Data Warehouses With OO Conceptual Models", IEEE Computer, Vol. 34, No. 12. pp. 66-75.

Stefanovic N, J. Han et K. Koperski. (2000). Object-Based Selective Materialization for Efficient Implementation of Spatial Data Cubes. IEEE Transactions on Knowledge and Data Engineering, Vol. 12, n° 6, 938-958.

Viswanathan, G., Schneider, M (2011). On the Requirements for User-Centric Spatial Data Warehousing and SOLAP. Database Systems for Advanced Applications 144–155

Summary

In recent years there has been a growing interest to new category of Decision Support System based on the Spatial Data Warehouse (SDWs). Several conceptual models have been proposed for the SDWs in different fields of applications. The few attempts that were made for the classification of these models are based on particular properties related to the context presented study. Given the lack of research on the current state of these models and their future projections, we propose in this paper a review and a classification of these models with the ultimate goal of facilitating the evaluation and facilitate the identification of new research areas. Our study differs particularly by the classification, evaluation and comparison of these models according to the characteristics of conceptual modeling of SDWs and the multidimensional spatial properties.

Langage textuel pour interroger une base de données XML-OLAP

Ben Ltaief Soufien

70,Avenu Abdel Hamid El Ghadi Houmet Souk Djerba 4180
Soufien.beltaif@hotmail.fr

Résumé. La technologie OLAP *(On-Line Analytical Processing)* permet un accès facile aux données décisionnelles détaillées et agrégées et l'exploration interactive en vue d'aide à la décision. Ces systèmes permettent l'analyse des données de l'entreprise. Comme les systèmes OLAP et les entrepôts de données évolues, les données complexes sont de plus en plus utiliser. XML *(eXtensible Markup Language)* est un format de texte souple permettant l'échange et la représentation des données. En effet, les documents décisionnels sont de plus en plus représentés sous l'environnement XML. Afin de pouvoir effectuer des requêtes sur des données semi structurées, de nombreux langages de requêtes ont été proposés. XQuery *(XML Query Language)*, permet d'élaborer des requêtes complexes. L'objectif de notre travail consiste à étendre le langage du XQuery par les différents opérateurs OLAP (slice, dice, rollup et drill down) afin d'interroger les entrepôts de données XML-OLAP.
Mots-clé : XQuery, OLAP, entrepôts de données XML, slice, dice, rollup, drilldown

1 Introduction

De nos jours, les données actuellement utilisées et échangées par les applications décisionnelles OLAP sont de plus en plus diverses et hétérogènes. D'autant plus que ces systèmes se présentent peu flexibles et permettent difficilement de prendre en compte les métadonnées.
En outre, avec l'émergence de formats de données semi-structurés, le stockage de documents dans un entrepôt centralisé est apparu de façon naturelle comme une adaptation des entrepôts de données et une réponse possible à la problématique d'entreposage de données complexes qui s'appuie sur le langage XML. Ce langage est en effet de plus en plus utilisé pour représenter des données décisionnelles (Beyer et al., 2005) et se montre particulièrement adapté pour modéliser des données dites complexes (Darmont et al., 2005) issues de sources hétérogènes. Ce langage définit un standard d'échange de données complexes provenant de différentes sources et soutenues par des formats hétérogènes (Boussaid et al., 2008). Leurs analyse est possible grâce au langage XQuery qui est le langage principale pour accéder à ces informations (Beyer et al., 2005). Il représente le standard le plus important produit par le W3C. XQuery a emprunté les caractéristiques et les avantages de plusieurs autres langages, y compris XPath, SQL… malheureusement XQuery tel qu'il est normaliser par le W3C est incomplet et ne possède pas des opérateurs OLAP prédéfinis afin d'interroger un entrepôt de données XML-OLAP. Il fallut écrire des requêtes complexe pour remédier ce problème ce qui n'est pas facile pour tous le monde à le faire.

Dans ce papier, nous proposons une extension de la grammaire du langage XQuery par les différents opérateurs OLAP afin d'obtenir une expression XQOLAP (XQuery OLAP) prête à être exécutée et par la suite à produire des résultats ce qui facile la tache des informaticiens. Ce travail est organisé comme suit nous présenterons dans la première section les différentes tentatives d'extension du langage XQuery. En suite, nous proposerons, dans la deuxième section, notre propre syntaxe des opérateurs OLAP dans le cadre de la grammaire XQuery. Pour valider cette extension, nous présenterons, dans la troisième section l'implémentation et l'évaluation de notre prototype proposé. Enfin, nous concluons notre travail et nous présenterons nos futures pistes de recherches.

2 Travaux connexe

De nombreux travaux se sont focalisés sur l'optimisation des documents XML en vue d'une analyse en ligne des données. Dans cette étude nous allons classer ces approches en deux catégories ; ceux qui ont étudié le langage de requête XQuery, et ceux qui ont travaillés sur d'autres langages XML (XML-MDX, SQL_{XM}).

2.1 Langage de XQuery

Comme nous l'avons évoqué précédemment dans cet article, XQuery présente à ce jour quelques limites. Généralement les études s'adressant au besoin d'analyse en ligne en XQuery peuvent être classées en deux larges catégories : (i) Produire le support d'analyse en ligne au niveau physique ou logique. (ii) Etendre l'expression FLWOR de XQuery avec des constructions explicites similaires aux clauses *Group-By, Order-By* et *Having* de SQL.

2.1.1 Production de support d'analyse en ligne au niveau physique ou logique

(Jagadish et al.*,2001)* proposent une algèbre logique d'arbres nommée TAX (*Tree Algebra for XML*) comme une extension de l'algèbre relationnelle. Dans un arbre TAX, chaque nœud est modélisé par une paire attribut-valeur. Pour chaque opérateur, un modèle d'arbre (*pattern tree*) est défini, qui doit être suivi par l'arbre témoin (*witness tree*) résultat. TAX offre également des opérateurs de mise à jour telle que l'insertion d'un nœud, la suppression d'un nœud,…

(Hachicha et Darmont , 2010) ont proposés un opérateur rollup basé sur un modèle d'arbre et associé à un algorithme, permettant d'agréger les données XML multidimensionnelles représentes dans des hiérarchies complexes, dans le but de concevoir un algèbre XML-OLAP permettant d'exécuter des requêtes OLAP sur des données XML native et d'étendre les opérateurs OLAP dans l'expression sur des arbres de données XML.

(Chang et Hang, 2011) proposent un traitement modelé pour mettre en œuvre des documents XML codé efficacement en utilisant XQuery. Ce traitement exige le codage d'une partie de documents XML en éliminant les éléments non désirés. Donc, Le premier but est éliminer le décryptions inutile. Le Deuxièmes but est d'éviter d'exécuter la décryptions inutile pour cela ils éliminent la décryptions redondant parce que les encryptions peuvent casser la structure du document XML. Ce modèle exige des certains documents pour mettre en doute, y compris un DSL qui spécifie comment coder le document XML et le Schéma XML

des documents XML originaux. Les résultats expérimentaux ont présenté une bonne performance en résolvant requêtes XML.

2.1.2 Extension de l'éxpression FLWOR de XQuery

Certaines études ont essayé d'étendre l'expression FLWOR de XQuery par des constructions explicites semblables aux clauses *Group By, Order By* et *Having* de SQL.

Dans le but de généraliser la syntaxe de l'expression courante FLWOR de XQuery, (Borkar et Carey, 2004) propose d'ajouter trois opérateurs ;

(i) un groupement *Group by*. La clause *group-by* ajoutée a rendu la requête XQuery moins complexe et plus similaire à une requête SQL. (ii) un opérateur d'élimination de duplication *Distinct by* qui explore les occurrences de la variable concurrente, (iii) un opérateur de jointure « *Ofor* » qui établit une jointure entre les variables moyennant leurs identifiants.

(Deutsch et al., 2004) ont également proposé un nouveau opérateur de groupement pour XQuery ainsi qu'un ensemble de règles permettant de traduire les requêtes utilisant la fonction *distinct-values ()* en une forme minimisée et optimisée. Cette minimisation n'est possible que si un tel opérateur est implémenté dans XQuery.

En outre, (Witt et al., 2007) ont étendu l'expression FLWOR de XQuery pour qu'elle inclut un opérateur de cube X^3. Pour ce faire, ils ont proposé un entrepôt de treillis de cubes XML, et un mécanisme de spécification généralisé. Ils discutent également le mécanisme de construction de cube et comparent plusieurs algorithmes alternatifs. X^3 a été implémenté en C++ au sein du SGBD natif XML TIMBER[1]. Ce travail est une extension de l'algèbre TAX.

(Wang et al., 2005) présentent des concepts pour XOLAP (OLAP sur données XML). Ils définissent un opérateur général d'agrégation pour XML, *GXaggregation*, qui forme la base de *XCube*, une extension de l'opérateur traditionnel cube pour les données XML. Cet opérateur est mis en œuvre avec une extension de XQuery, *GXaggregation* permet l'extraction de propriétés à partir des dimensions et des mesures suivant leurs expressions de chemin XPath.

(Ben Messaoud et al,2006) proposent un opérateur d'agrégation OLAP basé sur une méthode automatique de classification (*clustering*) : *OpAC*. Cet opérateur permet des analyses précises et fournit les agrégats sémantiques sur des données complexes représentées dans des documents XML.

(Beyer et al., 2005) ont également proposés un opérateur de groupement pour XQuery. Cette clause a été implémentée au sein du SGBD natif XML eXist[2], qu'on décrira en détaille plus tard dans ce mémoire. L'opérateur Group by a réduit considérablement le temps de traitement des requêtes, et son utilisation a amélioré la lisibilité de la requête XQuery.

Outre l'opérateur Group by, certaines études ont cherchés à remédier un autre manque de XQuery, à savoir, la recherche documentaire, connue par IR *(Information Retrieval)*. En effet, (Bremer et Gertz, 2002) ont introduit un nouvel opérateur qui étend les requêtes XML par les capacités de la recherche documentaire. L'opérateur ajouté est le

[1] http://www.eecs.umich.edu/db/timber/
[2] http://exist.sourceforge.net/index.html

« *Rank* ». Il est employé dans une expression *Rank by* très similaire à l'expression *Sort by* de XQuery. L'originalité de cet opérateur est de permettre de trier les variables selon leurs poids, par exemple, le nombre d'apparitions d'un mot clés dans un document. L'opérateur *Rank* fournit des capacités très importantes pour interroger les documents riches en textes.

(Andrei et al., 2010) Proposent une extension de XQuery appelé XQuery Data Définition Facility (XQDDF). Elle étend XQuery avec trois artefacts : des collections, des index et des contraintes d'intégrité.

2.2 Autres langages

Certains travaux se sont focalisés sur l'optimisation d'autres langages XML que XQuery, ou sur la mise en ouvre d'autres langages adaptés à l'analyse en ligne des données.

(Park et al.,2005) proposent un langage d'expression multidimensionnel pour l'interrogation des entrepôts XML : XML-MDX. Ils ont étendues le langage relationnel MDX de Microsoft avec deux opérateurs : CREATE XQ-CUBE pour la création de cubes XML et SELECT pour leur interrogation.

De plus, les auteurs définissent sept opérateurs d'agrégation: ADD, LIST, COUNT, SUMMARY, TOPIC, TOP KEYWORD et CLUSTER. Certains de ces opérateurs sont inspirés du relationnel, d'autres utilisent des techniques de fouille de données textuelles pour agréger des valeurs non- additives.

Dans le cadre d'une approche de fédération des sources des données XML avec des cubes OLAP existants, (Pedersen et al.,2004) proposent un nouveau langage, nommée SQL_{XM}. Ce langage a été créé à l'aide d'une extension de SQL_M qui permet d'associer des requêtes XPath aux requêtes SQL_M. SQL_M est lui-même une extension de SQL pour le traitement des données multidimensionnelles. Parmi les opérateurs mis en œuvre dans ce langage, on trouve la *décoration* qui attache une nouvelle dimension à un cube en se basant sur les valeurs des éléments XML liés, la *sélection* qui sert à filtrer les mesures et l'opérateur de *projection généralisée* (*generalized projection*) qui a pour rôle l'agrégation des mesures ensemble d'opérateurs sur les données OLAP- XML.

(Tan et al., 2010) ont étudiés le problème de définir et calculer les réponses des requêtes logiques quand les requêtes sont posées au système de données XML virtuellement intégrés. Le système global et chaque source du données utilise XML comme leurs schémas, et la projection topographique entre le schéma global et les sources du données sont définies dans une approche comme vue locale. Ils ont proposés une approche à définir les vues XML. Ils ont fourni aussi un modèle de contrainte pour XML en définissant les contraintes globales qui peuvent exprimer les contraintes les plus communément discutées. Ils ont donné une définition cérémonieuse de réponses à la requête logique de données XML virtuellement intégré, en redéfinissant le concept de réparation et réponses de la requête logiques pour XML. Ils ont fourni aussi une méthode de calcul logique des réponses des requêtes.

3 Opérateurs OLAP pour XQuery

A la lumière de ce qui précède, nous constatons la nécessité de proposer un langage d'analyse multidimensionnelle des données XML. En effet, pour interroger les bases de données XML, le langage de requêtes XQuery (le SQL de XML), est tout à fait bien adap-

té. Par contre, XQuery n'est pas suffisant pour interroger les bases de données XML multi-dimensionnelles. Nous pouvons ainsi constater qu'il lui manque un mécanisme approprié destiné à effectuer des opérations d'analyse en ligne. Dans cette partie, nous proposons la syntaxe des opérateurs OLAP proposés comme une extension de la grammaire de XQuery.

3.1 Opérateurs OLAP

OLAP *(On line Analytical Processing*) est un terme qui désigne les bases de données multidimensionnelles (aussi appelées cubes ou hyper-cubes) destinées à l'analyse (Codd et al., 1993). Niguel pendes a redéfinie ce terme en 1995 par un système d'analyse rapide d'information multidimensionnelles partagées; FASMI (Fast Analysis Shared Multidimensional Information). Cet acronyme permet de résumer la définition des produits OLAP. Ainsi ils devraient être assez rapides pour répondre aux demandes des utilisateurs dans un laps de temps courts. Ils devraient aider la tâche d'analyse en fournissant la souplesse dans l'utilisation des outils statistiques et toutes les logiques d'affaire. Ajoutant que le système doit créer un contexte où la confidentialité est préservée et doit gérer les cas où plusieurs utilisateurs ont des droits en écritures. Ce point est la plus grosse faiblesse des produits actuels. Ils devraient fournir une vue multidimensionnelle afin que la métaphore de cube de données peut être utilisée par les utilisateurs. Enfin, le system doit fournir toutes les données et les informations nécessaires pour un produit OLAP.

OLAP présente un certain nombre d'opérateurs, classés dans trois familles, les opérateurs ensemblistes (Slice et Dice), les opérateurs liés à la granularité (rollup et drill-down) et les opérateurs liés à la structure (rotation, switch, push et pull) (Leh.1998). Chaque opérateur prend en entrée un cube OLAP et fournit un autre cube en sortie. Un cube OLAP représente dans ses cellules des faits à analyser (matérialisés par des mesures numériques) en fonction de dimensions (axes d'analyse) décrites par des attributs membres et susceptibles d'être hiérarchisées (par exemple, une dimension géographique comporte les attributs, ville, région et pays).

Dans ce travail nous nous concentrons sur les opérateurs ensemblistes (Slice et Dice) et sur les opérateurs liés à la granularité (rollup et drill-down).

3.2 Extension de XQuery

Dans ce travail nous proposons d'étendre le langage de XQuery, tout en utilisant des fonctions utilisateur, par des opérateurs OLAP. En effet, nous avons proposés les syntaxes de quatre opérateurs OLAP, qui seront appelés par la requête XQuery. L'ajout des opérateurs OLAP modifie la syntaxe de l'expression FLWOR propre à XQuery. Les clauses (*rollup, drilldown, slice* et *dice*) se placent entre les clauses *for* ou *let* et *return*. Suite à la compilation de ces opérateurs, on obtient ainsi une expression XQOLAP prête à être exécutée et par la suite à produire des résultats. La *figure.1* illustre schématiquement les étapes de l'évaluation d'une requête XQOLAP.

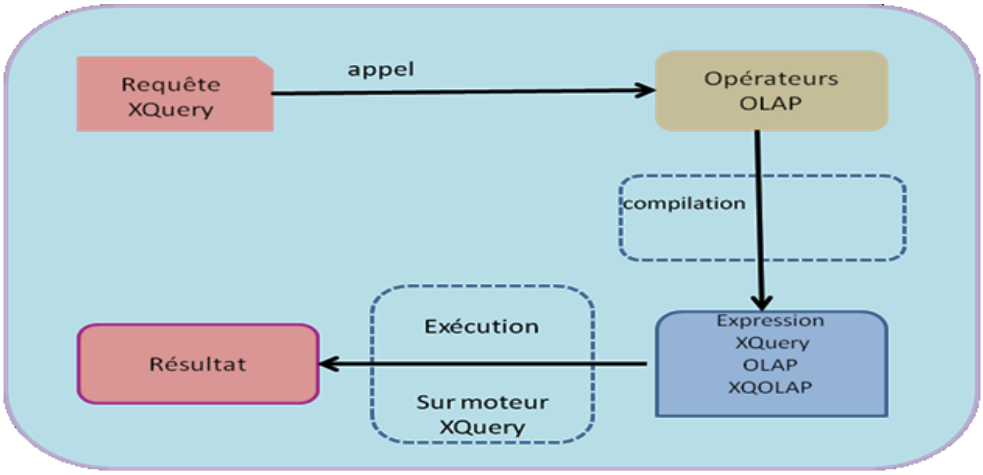


Fig.1 - Evaluateur d'une requête XQuery OLAP

3.3 Syntaxe des opérateurs OLAP

Notre approche consiste à écrire la syntaxe de quelques opérateurs OLAP et à les faire intégrer dans le moteur de XQuery à partir de notre propre interface. La syntaxe complète des clauses respectives rollup, drilldown, dice et slice sont présentés ci-dessous.

RollUpClause = " rollup“ “(“ cube “,“ dimension “,“ niveau “)“ “as“ RollVar

- *cube* est la collection des documents XML représentant les faits et leurs dimensions, déclarée précédemment comme une variable à l'aide d'une clause *let*;
- *dimension* représente l'élément sur lequel sera appliquée l'opération de forage vers le haut, déclaré précédemment comme une variable à l'aide d'une clause *for*;
- *niveau* représente le niveau de granularité à partir duquel commence le forage vers le haut, déclaré précédemment comme une variable à l'aide d'une clause *for*;
- *RollVar* est une nouvelle variable qui contiendra, l'ensemble des nœuds résultant de la clause *RollUp*.

DrillDownClause = “drilldown“ “(“ cube “,“ dimension “,“ niveau “)“ “as“ DrillVar

- *cube* est la collection des documents XML représentant les faits et leurs dimensions, déclarée précédemment comme une variable à l'aide d'une clause *let*;
- *dimension* représente l'élément qui possède un niveau d'agrégation inférieur à *niveau*, déclaré précédemment comme une variable à l'aide d'une clause *for* ;
- *niveau* représente le niveau de granularité à partir duquel commence le forage vers le bas, déclaré précédemment comme une variable à l'aide d'une clause *for* ;
- *DrillVar* est une nouvelle variable qui contiendra, l'ensemble des nœuds résultant par la clause *DrillDown*.

DiceClause=“Dice”“(“cube“,“attributdedimension1”,”valeur1”, ‘’attributdedimension2 ‘’,’’ valeur2“)“ “as“ DiceVar

- *cube* est la collection des documents XML représentant les faits et leurs dimensions, déclarée précédemment comme une variable à l'aide d'une clause *for*;

- *attributdedimension1* représente le nom d'attribut concernant la dimension 1 du cube *cube* ;
- *valeur1* représente la valeur d'attribut concernant la dimension 1 du cube *cube* ;
- *attributdedimension2* représente le nom d'attribut concernant la dimension 2 du cube *cube* ;
- *valeur2* représente la valeur d'attribut concernant la dimension 2 du cube *cube* ;
- *DiceVar* est une nouvelle variable qui contiendra, l'ensemble des nœuds résultant de la clause *Dice*.

SliceClause = "slice" "(" cube "," dimension "," attribut ", ''valeur '')" "as" Slice-Var

- *cube* est la collection des documents XML représentant les faits et leurs dimensions, déclarée précédemment comme une variable à l'aide d'une clause *for*;
- *dimension* représente l'élément sur lequel sera appliqué la coupe du cube *cube,* déclarée précédemment comme une variable à l'aide de la clause *for* ;
- *attribut* représente le nom ou le prédicat de la dimension ;
- *valeur* représente le nom ou la valeur de l'attribut de la dimension ;
- *SliceVar* est une nouvelle variable qui contiendra, l'ensemble des nœuds résultant de la clause *Slice ;*

La façon dont nous avons implémenté ces différentes clauses OLAP est basée sur des algorithmes différents selon leurs rôles. En effet, l'opération *RollUp* consiste à une association ou groupement des nœuds XML.

Par contre, *DrillDown* vise à les dissocier. D'autre part, la clause *dice* permet d'extraire un sous ensemble des nœuds XML. L'opération *slice* consiste à filtrer les nœuds selon la ou les valeurs des dimensions. Une illustration du fonctionnement des clauses OLAP pour XQuery est représentée par la *figure.2*.

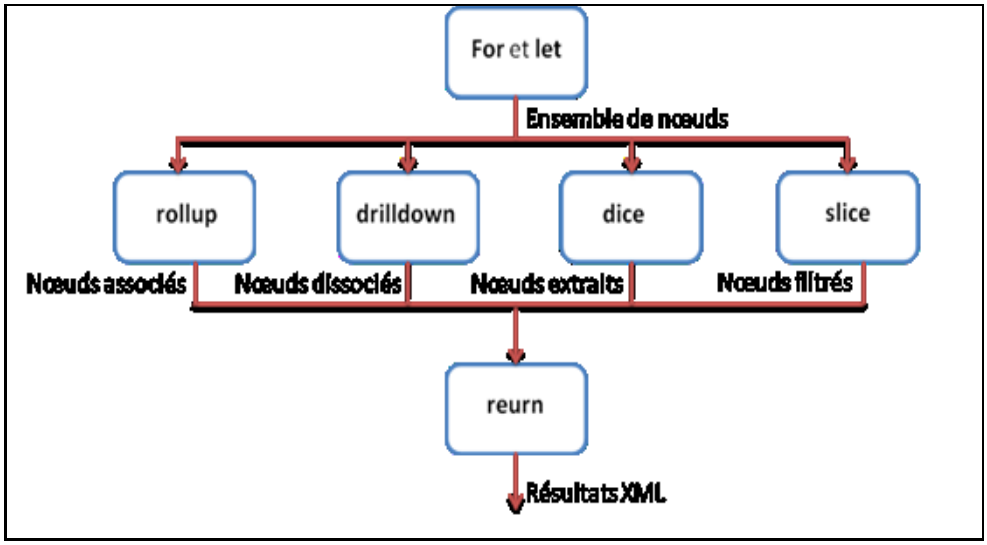


Fig.2 - Représentation schématique du fonctionnement des opérateurs OLAP pour XQuery.

4 Implémentation et Evaluation

Afin de valider notre proposition théorique, nous l'avons implémenté au sein du notre prototype intitulé XqueryPlus v1.0-Gestion des XML. Cette section présentera ce prototype qui implémente les opérateurs XML-OLAP dans le moteur XQuery. Nous présentons aussi notre expérimentation qui évalue et analyse les performances de ces opérateurs dans une base XML native sous le système eXist.

Dans le but d'analyser les performances des opérateurs proposés, nous utiliserons le même exemple tout au long de cet article. Il s'agit d'un scénario d'entrepôt de données comprenant un ensemble de faits et trois dimensions hiérarchisées. Les faits sont les ventes (ventes) des produits aux clients en fonction du temps. Les mesures des faits sont le prix (prix) et le nombre des produits vendus (quantité). Chaque fait possède trois dimensions : le produit commandé (produit), le temps (temps) et le client (client). Ces trois hiérarchies de dimensions possèdent trois niveaux de granularité et sont strictes, c'est-à-dire qu'il existe une relation 1-n entre un niveau de granularité et le niveau inférieur. Dans cet exemple, un produit (produit) fera donc partie d'une et une seule catégorie (catégorie).

4.1 Modélisation XML-OLAP

Nous avons donc optés pour un modèle qui respecte la représentation multidimensionnelle classique des entrepôts de données et modélise explicitement les dimensions. Cette famille s'avère donc la plus adaptée dans notre contexte à savoir la variante X-Warehousing (Boussaid et al., 2006).

En effet, ce modèle considère l'entrepôt de données XML comme étant une collection de documents XML, où chaque document stocke un fait et les instances des dimensions correspondantes. Notons que la modélisation X-Warehousing utilise adéquatement les imbrications d'éléments XML pour modéliser les dimensions.

Grace à cela, on peut connaitre la profondeur des niveaux des hiérarchies et leurs imbrications, ce qui nous servira surtout pour les opérations liées à la granularité. On gagne donc en sémantique en utilisant ce modèle.

```
<?xml version="1.0" encoding="UTF-8"?>
<vente prix="34" quantite="5">

  <produit nomp= "VTT">
    <categorie nomg="velo">
      <famille nomf="transport"/>
    </categorie>
  </produit>

  <client nomc= "Mohamed Ali">
    <pays nomy= "Tunisie">
      <continent nomn= "Afrique"/>
    </pays>
  </client>

  <temps jour="23">
    <mois nomm= "09">
      <annee nome="2006"/>
    </mois>
  </temps>

</vente>
```

Fig.3 - Modélisation X-Warehousing d'un fait de l'exemple de travail.

Nous avons opté pour un stockage natif des données de l'entrepôt XML. En effet, ce type de stockage permet d'enregistrer des données XML sans modifier leur structure et offre ainsi la possibilité d'exprimer des requêtes XML d'analyse complexes. Ce type de stockage permet aussi de meilleures performances lors d'une interrogation des données, comparée à un stockage relationnel.

En effet, nous avons travaillé sur eXist qui est une base de données native XML open-source. Elle est à ce jour le produit le plus évolué et le plus flexible que l'on puisse trouver. Dans notre approche full XML, et grâce à la gestion par collection de documents, nous estimons que le stockage de données est convaincant et tout à fait adapté à des besoins d'analyse OLAP XML.

4.2 Performance

Pour mener à bien notre évaluation, nous avons procédé en différentes étapes. Tout d'abord nous avons augmenté le nombre de faits de l'entrepôt à 15 soit 3 fois plus que dans l'exemple précédant, puis à 50 soit 10 fois plus que l'exemple précédant. Ensuite, nous avons exécuté les requêtes utilisant les opérateurs OLAP proposés. La *figure.4* présente l'interface d'interrogation des entrepôts des données XML avec le prototype proposé alors que la *figure.5* présente l'interface d'interrogation des entrepôts des données XML sans le prototype proposé c'est-à-dire à l'aide des requêtes XQuery classique.

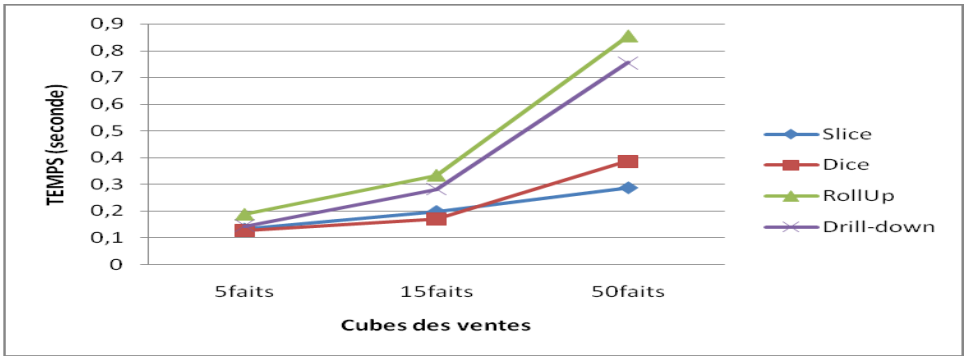


FIG.4_Temps d'exécution des requêtes OLAP en fonction de la taille du cube X

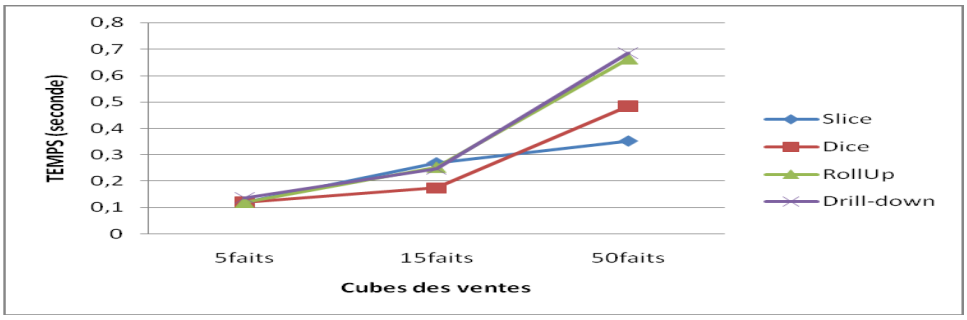


FIG.5-Temps d'exécution des requêtes XQuery classique en fonction de la taille du cube XML.

Nous remarquons rapidement que l'augmentation de temps de traitement lors de l'exécution des opérateurs liés à la granularité (rollup et drilldown) parait plus importante que celle de l'opérateur ensemblistes (slice et dice). Ceci est dû aux opérations de forages qui nécessitent le parcoure des nœuds de chaque fait dans le bute de représenter les données du cube à un niveau de granularité inférieur ou supérieur.

Un tel point pourrait être optimisé avec un travail de refactorisation de l'architecture d'indexation des donnes XML d'eXist. Il s'agit ici d'un point important si l'on compte gérer une base de données importante avec eXist et que l'on a donc besoin de performance. D'autre part, les clauses OLAP que nous avons proposé ont permis d'améliorer la lisibilité des requêtes ainsi que leur complexité.

Ajoutant également que le temps d'exécution des requêtes avec le prototype proposé est inférieur à celui des requêtes XQuery classique surtout l'orsqu'on augmente le taille de l'entrepôt. En effet, il été difficile de réaliser des opérations d'analyse multidimensionnelles avec les requêtes XQuery classiques. Car il fallait écrire des requêtes imbriqués et complexes.

5 Conclusion

Nous avons étudié les différentes formes de modélisation multidimensionnelles des données XML. Nous avons également analysés les travaux visant à optimiser l'analyse en ligne des données XML.

Ainsi nous avons constaté que des extensions du langage XQuery sont nécessaires. Nous avons étendu le langage XQuery par des opérateurs OLAP. Ces opérateurs ont permis d'améliorer la lisibilité des requêtes ainsi que leurs complexités.

Afin de pouvoir mettre en ouvre ces opérateurs, nous avons exécuté une série de requêtes typique d'analyse sur un entrepôt de données XML au sein de la base de données XML native eXist. Nous avons constaté pour celui là l'intérêt des opérateurs OLAP pour XQuery.

Pour finir, nous avons réalisé un test de performance de ces opérateurs sur des cubes XML. Nous avons ainsi constaté que les opérateurs OLAP sont indispensables pour l'exécution des requêtes d'analyse multidimensionnelles.

De nombreux aspects pourraient encore être analysés. En effet, l'aspect d'indexation présente un grand impact sur la performance de l'interrogation. Il serait intéressant d'utiliser des indexes spécifiques ou de chercher d'autres algorithmes. D'autre part, il existe d'autres langages spécifiés pour l'analyse comme MDX *(Multidimensional Expressions)* de Microsoft. Un tel langage compatible avec XML serait intéressant pour faciliter encore l'écriture des requêtes d'analyse multidimensionnelle. Enfin, l'analyse de performance que nous avons réalisée pourrait être plus complète. En effet, les données analysées pourraient être modélisées autrement, en constellation par exemple, et les requêtes pourraient être plus complexes et sur des sous-cubes plus grands. Il serait dés lors utile d'explorer ces voix.

Références

Beyer, K. S., D. D. Chamberlin, L. S. Colby, F. Özcan, H. Pirahesh, et Y. Xu (2005). Extending XQuery for Analytics. *In ACM SIGMOD 24th International Conference on Management of Data (SIGMOD 05), Baltimore, USA*, pp. 503–514. ACM.

Boussaïd, O., R. BenMessaoud, R. Choquet, et S. Anthoard (2006). X-Warehousing : An XML-Based Approach for Warehousing Complex Data. *In 10th East European Conference on Advances in Databases and Information Systems (ADBIS 06), Thessaloniki, Greece, Volume 4152 of LNCS*, pp. 39–54. Springer.

Borkar Vinayak R. and Michael J. Carey. Extending XQuery for Grouping, Duplicate Elimination, and Outer Joins. *In XML 2004, Washington DC, USA,* pages 1-11,November 2004.

BenMessaoud, R., S. Rabaséda, et O. Boussaïd (2006). A Data Mining-Based OLAP Aggregation of Complex Data : *Application on XML Document. International Journal of DataWarehousing & Mining* 2(4), 1–26.

Bremer, J. and M.Gertz(2002) XQuery/IR: *Integrating XML Document and Data Retrieval Department of Computer Science University of California*, Davis, CA.

Boussaid, O., BenMessaoud, R., Choquet, R., & Anthoard, S. (2006). X-Warehousing: An XML-Based Approach for Warehousing Complex Data. *10thEast-European Conference on Advances in Databases and Information Systems(ADBIS 06)*, Thessaloniki, Greece, Vol. 4152 of Lecture Notes in Computer Science (pp. 39-54). Springer.

Codd. E.F., Codd. S.B., Salley. C.T., Providing OLAP (On Line AnalyticalProcessing) to user analyst: an IT mandate, rapport technique, E.F. Codd and associates, *(white paper de Hyperion Solutions Corporation), 1993*.

Darmont, J., O. Boussaïd, J.-C. Ralaivao, et K. Aouiche (2005). An Architecture Framework for Complex Data Warehouses. *In 7th International Conference on Enterprise Information Systems (ICEIS 05), Miami, USA*, pp. 370–373.

Deutsch. A, Papakonstantinou.Y, and Y. Xu. Minimization and group-by detection for nested xqueries. *In Proceedings of the ICDI, pages 839–854*, 2004.
Hachicha.M, Darmont.J. Modèles d'arbre pour XOLAP. *In 6èmes Journées francophones sur les Entrepôts de Données et l'Analyse en ligne EDA 2010.Djerba Tunisia*. B-6 (2010) p.97-106.
Jagadish. H. V, Laks.V. S. Lakshmanan, Divesh Srivastava, and Keith Thompson.TAX: A Tree Algebra for XML. *In 8th International Workshop on Database Programming Languages (DBPL 01), Frascati, Italy, volume 2397 of Lecture Notes in Computer Science, pages 149-164*. Springer, 2001.
Lehner.W, "Modelling Large Scale OLAP Scenarios", *6th Intl. Conf. on Extending Database Technology - Advances in Database Technology (EDBT),LNCS 1377, Springer, p. 153–167*, 1998.
Nuwee Wiwatwattana H. V. Jagadish Laks V. S. Lakshmanan Divesh Srivastava Xˆ 3: A Cube Operator for XML OLAP 1-4244-0803-2/07/ ©2007 IEEE.
Chang.T.K, Hwang.G.H. Developing an efficient query system for encrypted XML documents. *The Journal of Systems and Software 84 (2011) 1292–1305*. © 2011 Elsevier.
Park, B.-K., Han, H., & Song, I.-Y. (2005). XML-OLAP: A Multidimensional Ana-lysis Framework for XML Warehouses. *7th International Conference on DataWarehousing and Knowledge Discovery (DaWaK 05), Vol. 3589 of Lecture Notes in Computer Science (pp. 32-42)*. Springer.
Pedersen.D, Pedersen.J, and Pedersen.T.B. Integrating XML Datain the TARGIT OLAP System. *In 20th International Conference on Data Engineering(ICDE 2004), Boston, USA, pages 778-781*. IEEE Computer Society, 2004.
Tan.Z, Liu.C, Wang.W, Shi.B. Consistent query answers from virtually integrated XML data. *The Journal of Systems and Software 83 (2010) 2566–2578*. © 2010 Elsevier.
Wang.W, Zhang.J, Liu.X, and Zhang.S. X-warehouse: building querypattern-driven data. *In 14th International Conference on World Wide Web (WWW 05), Chiba, Japan, pages 896-897*. ACM, 2005.

Summary

The OLAP (On-Line Analytical Processing) technology allows easy access to detailed and aggregated decision-making data and interactive exploration for decision support. These systems allow the analysis of the data of the company. As data warehouses and OLAP systems develop, complex data are and more use. XML (eXtensible Markup Language) is a flexible text format for the Exchange and the representation of the data. In fact, the decisional documents are more and more represented in the XML environment. In order to perform queries on semi structured data, several query languages have been proposed. XQuery (XML Query Language), can develop a complex queries. The objective of our work is to extend the language of the XQuery by different operators OLAP (slice, dice, rollup and drill down) to query XML-OLAP data warehouses.

Approche de sécurisation des communications d'un webhouse

Salma DAMMAK KTARI*,Faiza GHOZZI JEDIDI**
*damak.salma@gmail.com
**faiza_jedidi@yahoo.fr

Laboratoire Mir@cl: Multimedia, InfoRmation systems and Advanced Computing Laboratory
Université de Sfax ISIMS, BP 2042, CP 3021, Sfax, Tunisie

Résumé. L'avènement du web donne naissance à une nouvelle génération des systèmes décisionnels: le Webhouse. L'utilisation des webhouse et la diversification et la multiplicité des utilisateurs et des communications augmentent les risques d'intrusion. De ce fait, il devient primordial d'offrir les mesures de sécurité pour la protection des données et un mécanisme de contrôle d'accès des utilisateurs via des chemins de transmission sécurisés.
Les besoins de sécurité des communications sont définis généralement par un ingénieur réseau suite à la construction du webhouse. Cette définition tardive de ce type de besoin de sécurité augmente les risques d'intrusion aux données. En outre, la définition de ces besoins doit être faite par un responsable métier dès le début de la démarche de conception du webhouse. Dans ce papier, nous présentons une approche de sécurisation des webhouse et spécifiquement des communications en suivant l'architecture MDA. Pour les transformations entre les modèles MDA, nous appliquons les règles du langage de transformation QVT.

1 Introduction

Un entrepôt de données est une vision centralisée et intégrées de toutes les informations de l'entreprise. C'est une structure qui a pour but, contrairement aux bases de données, de regrouper les données de l'entreprise pour des fins analytiques et pour aider à la prise de décision stratégique. Les données d'un entrepôt sont épurées, organisées, historisées et provenant de plusieurs sources de données.

Les actions des internautes dans un site web présentent une nouvelle source de donnée née avec l'avènement du web. Ces données sont nommées Clickstreams d'après Kimball et Merz (2000) et ont donné naissance à une nouvelle génération de système décisionnel appelée Webhouse. Un webhouse est une nouvelle vision des systèmes décisionnels contenant des données du web organisées, historisées et stockées sous format XML.

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. Ces inconvénients nous amènent à considérer que la sécurité est un facteur déterminant pour un webhouse.
En plus, les besoins de sécurité ne sont pas pris en considération lors de la conception des entrepôts de données, la plupart des méthodes de conception ne les intègrent pas. Ces aspects seront 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.

L'objectif de cet article est de définir une approche de sécurisation des communications d'un webhouse. Et devant la complexité des outils et des techniques de sécurité, nous adoptons sur l'architecture MDA, nous proposons de concevoir un Webhouse sécurisé et de

définir des contraintes de sécurité au niveau métier sans engager le concepteur à chercher quel mécanisme de contrôle d'accès à choisir ? Et comment doit il protéger les chemins de transmission ?

Notre proposition prend en considération la protection des communications entre les utilisateurs et le webhouse en intégrant la spécification des besoins de sécurité des réseaux. Il est à noter que les contraintes de sécurité des communications ne sont pas traitées dans la littérature au niveau conceptuel. En plus, nous proposons une classification des utilisateurs selon des compartiments et des rôles afin d'empêcher toutes tentatives de violation des droits d'accès.

Ce papier est composé, outre l'introduction, d'une première partie présentant l'état de l'art des travaux relatifs à la sécurité des systèmes décisionnels. La deuxième partie présente notre approche de sécurisation des webhouses en se basant sur une démarche MDA. Et nous clôturons par une conclusion présentant les avantages et les perspectives de nos recherches.

2 Sécurité des entrepôts et le web :

Un objectif de sécurité exprime l'intention de contrer des menaces identifiées et/ou de satisfaire à des hypothèses, de ce fait la sécurité qui en découle n'est pas définie dans l'absolu mais bien relative à un but que l'on veut atteindre. .....

Dans cette partie, nous présentons les travaux relatifs à la sécurité des entrepôts de données et des webhouse.

2.1 Sécurité des entrepôts:

La sécurisation des systèmes décisionnels peut être abordée à deux niveaux Triki et al. (2010): (i) niveau conception qui vise à concevoir un webhouse sécurisé; et (ii) niveau exploitation qui vise à renforcer les droits et à interdire tout utilisateur malicieux d'inférer des données interdites à partir des données auxquelles il a accès. Triki et al. (2010) traitent la sécurisation des entrepôts de données contre les inférences à travers une approche basée sur les réseaux Bayésiens.

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

ROSENTHAL et Sciore. (2000) étendent la politique d'accès SQL (Structured Query Language). Ils définissent la permission d'accès d'un utilisateur comme un quadruplet (sujet, opération, objet, mode) avec 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 et permission physique.

En se basant sur les modèles de l'architecture MDA, Blanco et al. (2008) développent des entrepôts de données sécurisées. MDA propose de définir un modèle métier indépendant de toute plate-forme technique et de générer automatiquement du code vers la plateforme choisie. Dans le même cadre de travaux, 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 les éléments de l’entrepôt de données et les exigences de QoS qui analysent les problèmes de sécurité via le modèle SR de la stratégie i* YU (1997). VILLORROEL et al. (2006) définissent le modèle sécurisé de l’entrepôt de donnée SECDW du niveau **PIM** (Platform Independent Model) qui analyse les exigences définies précédemment en se basant sur une extension du profil UML appelée SECure Data Warehouse pour résoudre les problèmes de confidentialité de modélisation conceptuelle des DW et une extension OCL WARMER et KLEPPE. (2003) qui permet de spécifier les contraintes de sécurité des éléments de l’entrepôt de donnée. MEDINA et al. (2008) utilisent, au niveau **PSM** (Platform Specific Model), des mécanismes d’extension propres fournis par le CWM (Common Warehouse Metamodel) OMG. (2003) et étendent le paquet relationnel pour obtenir le modèle SECRDW qui est défini par un schéma en étoile représentant toutes les mesures de sécurité et d’audit capturées pendant la phase de modélisation conceptuelle de l’entrepôt de donnée.

À notre connaissance, seul le travail de Kimball et Merz (2000) a traité les aspects de sécurité dans les Webhouse (système décisionnel web). Dans ce travail, l’assurance de sécurité est basée sur 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.

2.2 La Sécurité et le web :

Devant l’augmentation de l’utilisation des technologies de communication Web, les risques concernant la sécurité d’information augmentent aussi. Afin de répondre à ces risques et assurer une gestion des informations sécurisées, l’analyse de la sécurité doit être intégrer dans les premières phases de conception du système.

Best el al. (2007) proposent une conception de l’aspect de sécurité à l’aide d’UmlSec qui présente une extension d'UML permettant au développeur d'applications d'intégrer des informations liées à la sécurité dans la conception du système, ainsi que d'effectuer des analyses de sécurité sur la couche modèle.

D’autre part, Juan et al (2010) présentent un travail qui vise à intégrer les aspects non-fonctionnels dans le développement de services web, spécifiquement la sécurité. Il présente un méta modèle pour le contrôle d’accès aux services web.

Dans le même contexte de la sécurité web, nous avons étudié le travail de Pozo el al. (2009) qui traite spécifiquement la sécurité des réseaux de communications. Pozo el al. (2009) proposent un framework MDA pour les firewalls avec la définition d’un nouveau langage DSL pour les firewalls.

Löf et al. (2010) présentent une nouvelle approche pour la construction de systèmes sécurisés. Dans cette approche, appelée Model Driven Security, les concepteurs doivent spécifier des modèles de systèmes ainsi que leurs exigences de sécurité et utiliser des outils pour générer automatiquement des architectures de système. Ce travail décrit une méthode pour évaluer la sécurité du réseau basé sur le modèle probabiliste relationnelle (PRM) appliqué sur les réseaux de communication, qui est une combinaison de réseaux bayésiens, des graphiques d'attaque, et des modèles d'architecture. La méthode proposée est basée sur un méta-modèle décrivant l'architecture du système, en termes de ses composantes et de leurs attributs, ainsi que les attaques possibles, sous la forme de graphes d'attaque conceptuels.

3 Problématique

Dans la section précédente, nous avons étudié plusieurs travaux présentant des solutions pour la sécurité des entrepôts de données. Dans la plupart des travaux, la sécurité de l'entrepôt est traitée qu'au niveau physique.
Aussi, nous remarquons que la sécurisation des réseaux de communications des données est rarement traitée. Pozo et al (2009) ont présenté une solution de sécurité des communications relative uniquement aux firewalls qui représentent l'un des composants du réseaux.
Les besoins de sécurité de ces aspects de communication sont définis généralement par un ingénieur réseau suite à la construction du système. Cette définition tardive de ce type de besoin de sécurité peut entrainer des risques d'intrusion aux données. En outre, la définition de ces besoins doit être faite par un responsable métier dès le début de la démarche de conception du webhouse. Nous considérons qu'il est vital d'incorporer les exigences de sécurité et de les mettre en application dès la phase de conception.
Dans les bibliographies étudiées, nous rencontrons uniquement le travail Kimball et Merz (2000) qui traite la sécurité des webhouse. Ce travail présente une solution pour le problème de sécurité des webhouse mais il n'intègre pas les exigences de sécurité dans la conception. Kimball et Merz (2000) ont proposé une solution qui offre des techniques de sécurité à intégrer après le développement du webhouse.
Devant les insuffisances rencontrées, nous proposons une approche de sécurisation du webhouse et spécifiquement des communications en tenant compte des besoins au niveau métier en se basant sur l'architecture MDA. Notre démarche propose dans une première étape la définition des exigences métiers de sécurité au niveau CIM. La transformation de ces modèles vers les modèles de la couche PIM est basée sur le standard QVT.

4 Démarche de sécurisation d'un webhouse

Notre démarche de sécurisation d'un webhouse adopte l'architecture MDA pour la spécification des besoins de sécurité. MDA préconise l'élaboration de modèle d'exigences (CIM), d'analyse et de conception (PIM) et de code (PSM) BLANC et SALVATORI (2005). Nous appliquons cette démarche sur un exemple d'étude de cas présentant un webhouse d'un système bancaire

4.1 Définition des exigences de sécurité (CIM)

Il est à noter que le modèle CIM permet la vision du système dans l'environnement où il opèrera, mais sans rentrer dans le détail de la structure du système, ni de son implémentation. L'indépendance technique de ce modèle lui permet de garder tout son intérêt au cours du temps et il est modifié uniquement si les connaissances ou les besoins métier changent.
Nous avons deux types d'exigences celles relatives aux données du webhouse répondant aux besoins utilisateurs et celles définissant les besoins de sécurité des utilisateurs (relatives au chef de sécurité). Nous avons choisi, pour la représentation de ces exigences, le modèle SR (Strategic Rationnel) de la structure de modélisation i* Yu (1997), SOLER et al. (2008) qui offre des modèles permettant de représenter des objectifs associés à chaque acteur et leurs dépendances.

Nous nous concentrons sur les exigences de sécurité qui sont définies par les chefs de sécurité du système.

Pour l'expression de notre modèle de sécurité SR, nous exploitons les avantages de la méthode de sécurité **EBIOS** Poggi, S (2005) (Expression des Besoins et Identification des Objectifs de Sécurité). Cette méthode permet de déterminer d'une manière exhaustives des objectifs de sécurité en suivant cinq étapes : étude de contexte, expression des besoins de sécurité, étude de menace, identification des besoins et expression des exigences de sécurité.

Nous présentons dans la figure ci-dessous le méta modèle des exigences de sécurité contenant la méta classe **MODELE SR** composée des métas classes **Exigence** et **Acteur**.
Exigence admet les attributs identifiant, type représentant une énumération pouvant avoir les valeurs donnée, utilisateur ou communication et l'attribut prédicat. les exigences peuvent être classifiées en deux types : **Besoin** et **Menace**. **Besoin** présente le besoin de sécuriser notre système contre un risque. **Menace** est la possibilité qu'une **vulnérabilité** soit exploitée. Une vulnérabilité est un défaut ou une faiblesse d'un système pouvant mener à une faille de sécurité ou à la violation de sa politique de sécurité.
La méta classe **Acteur** est caractérisée par un rôle et un identifiant. Une relation de dépendance existe entre les exigences d'un même modèle.

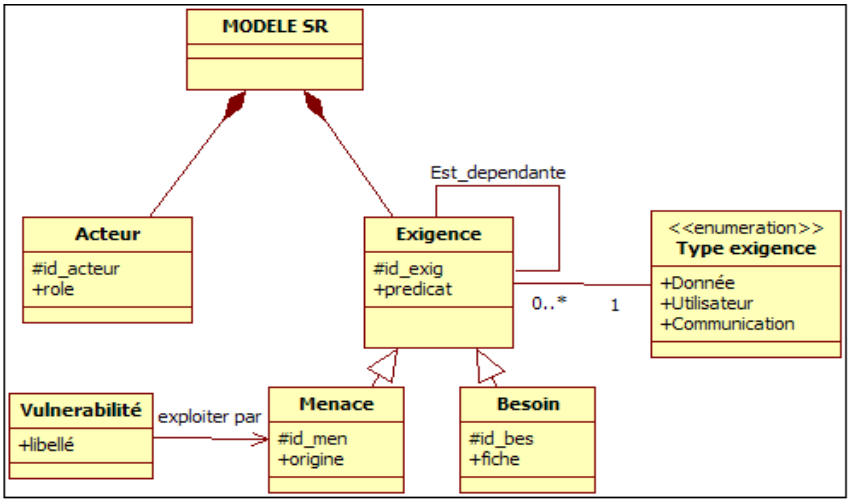


Fig. 1 – Le méta modèle en UML du modèle SR de i de définition des exigences de sécurité (CIM)*

L'expression des exigences de sécurité concerne trois axes: la sécurité des profils décideurs, la sécurité des données et la sécurité des chemins de transmission webhouse-décideur. Dans cet article nous nous intéressons à la sécurité des chemins de transmission. Ce type d'exigence possède une structure réutilisable d'un chef de sécurité à un autre. Nous profitons de cet aspect pour définir un modèle générique d'expression des exigences de sécurité.
Ce modèle offre une représentation claire et générique des exigences de sécurité. Les exigences sont présentées sous format de softgoal. Pour la sécurité des chemins de transmission nous devons : fournir un environnement sécurisé de transmission, contrôler le

trafic, ne pas transmettre les données en clairs et faire recours à des techniques de sécurisation de la donnée telle que le cryptage.
La figure2 présente les exigences définies par le chef de sécurité pour un décideur, du webhouse du système bancaire, appartenant au service Marketing ayant le grade d'un Chef service. Dans ce modèle, nous définissons les exigences de sécurité des communications: Le contrôle de trafic, le cryptage des données et la non transmission des données en clair (données des clients, données des opérations de la banque).

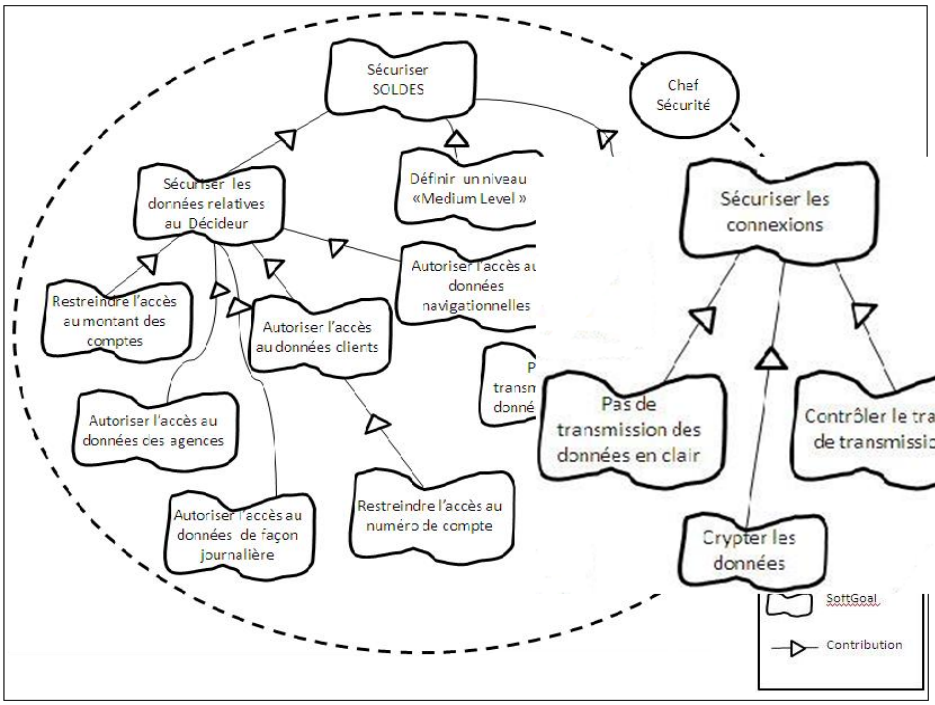


Fig. 2 – Le modèle SR de i de définition des exigences de sécurité (CIM)*

4.2 Analyse des exigences de sécurité (PIM)

Dans l'architecture MDA, le modèle PIM présente une vue partielle d'un CIM GRARI (2007) et il doit être fidèle aux exigences métiers de sécurité et de données exprimées et ne présente pas d'informations sur les technologies qui seront utilisées pour déployer un data webhouse sécurisé. Notre modèle PIM comportera trois niveaux dont le premier traite les profils des décideurs, le deuxième décrit les données sécurisées du webhouse et finalement un troisième étudie les connexions webhouse-décideur.

4.2.1 Le modèle de sécurité des utilisateurs:

Le modèle de sécurité des utilisateurs consiste à classifier les utilisateurs d'un webhouse en des décideurs et des administrateurs. Pour les décideurs, nous les classifions selon les

compartiments auxquels ils appartiennent (secteur de travail : Service Marketing, Service GRH, …) et selon leurs rôles dans le compartiment pour sécuriser l'accès aux données.
Cette classification aide à définir les droits d'accès de chaque décideur et à limiter les accès non désirés.

4.2.2 Le modèle de sécurité des données du webhouse

Selon Kimball un webhouse est un data warehouse qui stocke les clickstreams (les traces de l'utilisateur) et les données métier de l'application afin de suivre le comportement de l'utilisateur. Ainsi, les données d'un Webhouse peuvent être classifiées en deux catégories :

- Des données de base qui sont les données classiques qu'on rencontre dans un entrepôt de données telles que les informations concernant les ventes, les produits, les magasins,...
- Des données propres aux comportements du client à travers ses actions de navigation (clickstreams) telles que les pages les plus consultés, le temps que passe l'utilisateur à lire l'information, ...

Pour la sécurisation des données stockées dans le webhouse, nous spécifions un métamodèle SecDataWebhouse qui contient des stéréotypes nécessaires pour la définition des propriétés de sécurité dans le webhouse. Ce méta modèle décrit les caractéristiques principales d'un modèle multidimensionnel et ses aspects de sécurité Villarroel et al. (2006), LUJÁN et al. (2002). L'instanciation de ce méta permet de concevoir un modèle conceptuel multidimensionnel intégrant les règles de sécurisé et classifiant les informations selon leur niveau de sécurité afin de n'autoriser que les utilisateurs appropriés à les accéder.

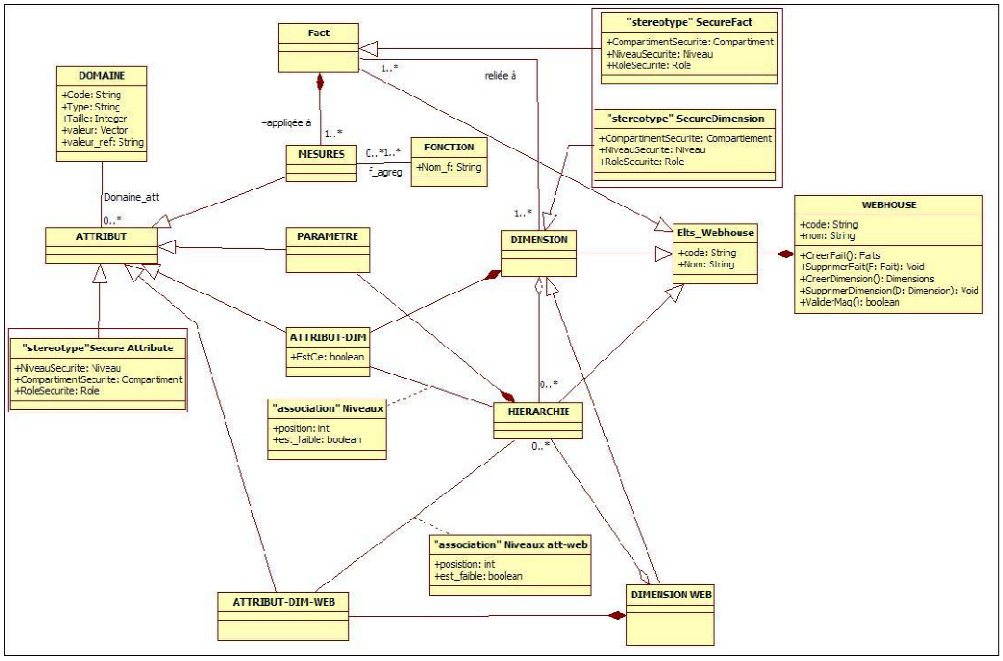


Fig. 3 Couche PIM : Méta modèle du Webhouse sécurisé

La figure 3 présente le méta modèle de notre webhouse. Nous enrichissons chaque élément du méta modèle (Fait, dimension, attribut) par des étiquettes de sécurité représentant les caractéristiques des utilisateurs autorisés. Une étiquette est sous la forme de (SL, SR, SC) représentant respectivement le niveau de sécurité, le rôle et le compartiment. Ces étiquettes limitent l'accès aux données du webhouse en fixant les niveaux de sécurité et les rôles autorisés à consulter les données. Nous définissons les contraintes nécessaires à la restriction d'accès aux données.
Ce méta modèle présente deux parties ; une parties décrivant les données multidimensionnelles spécifiques au domaine d'analyse et une deuxième partie générique relatives aux actions Web. Cette dernière partie est intégrée à tout modèle multidimensionnel comportant des données web extraites des clickstreams.
Exemple : A partir du modèle SR (figure2), nous analysons les exigences de données définies et nous déterminons pour la dimension CLIENT et AGENCE cette étiquette SC="Service Marketing", SR="ChefService", SL="Medium Level", nous définissons pour l'attribut Jour de la dimension DATE la même étiquette pour répondre à l'exigence accès au données de façon journalière.
Ainsi, le décideur : Chef service Marketing ayant le niveau Medium Level ; accédera aux données :

- CLIENT : code_c, nom, prenom, age, sexe, profession, ville, departement, region
- AGENCE : code_a, rais_soc, chiffre_a_, ville_a, departement_a, region_a
- DATE : code_d, num_jour

4.2.3 Le modèle de sécurité des chemins de transmission

Cette partie représente la partie principale de notre approche, et pour définir le modèle de sécurité des communications, nous avons appliqué un ensemble de règles de transformation en nous basant sur le langage QVT. QVT est un Langage de transformation et de manipulation de modèles normalisés par l'OMG. La technique utilisée par QVT consiste à sélectionner des éléments d'un modèle sous la forme d'une sous-partie ou une vue pour la transformer vers un autre modèle.

Lors de la définition des règles de transformation, nous considérons le méta modèle d'exigence comme modèle source pour obtenir le méta modèle de sécurité des communications (modèle destination) présenté par la figure 4.

```
Transformation CIMModelToPIMMOdel (CIMModel : Modele SR, PIMModel : Politique)
{
    Top relation ModelSRToPolitique
    {
        Nom-m : String ;
        Checkonly domain CIMModel ms: ModelSR {name=nom_m};
        Enforce domain PIMModel p:politique {name=nom_m} ;
    }
    Top relation ExigenceToSolution
    {
```

```
        Pred, niv : String ;
        Checkonly domain CIMModel e: exigence {namespace= ms:modelSR};
        Enforce domain PIMModel s: solution {namespace= p:politique} ;
        When { ModelSRToPolitique{ms ,p}}
        Where{ PredicatToOutil{pred, o};}
    }
    Relation BesoinToMécanismedeSurete
    {
        Checkonly domain CIMModel be: Besoin{SuperClass=e:exigence};
        Enforce domain PIMModel mse:MecanismedeSurete{SuperClass= s :solution};
        When { ExigenceToSolution {e,s}}
          }
Relation MenaceToMécanismedeProtection
    {
        Checkonly domain CIMModel me: Menace{SuperClass=e:exigence};
        Enforce domain PIMModel mp:MecanismedeProtection{} ;
                When { ExigenceToSolution {e,s}}
          }
          Top relation ActeurToTypePolitique
          {
            Checkonly domain CIMModel a:acteur {namespace=ms:modelSR,};
            Enforce domain PIMModel tp:TypePolitique {namespace=p:politique} ;
        When {ModelSRToPolitique{ms ,p};}
          }
}
```

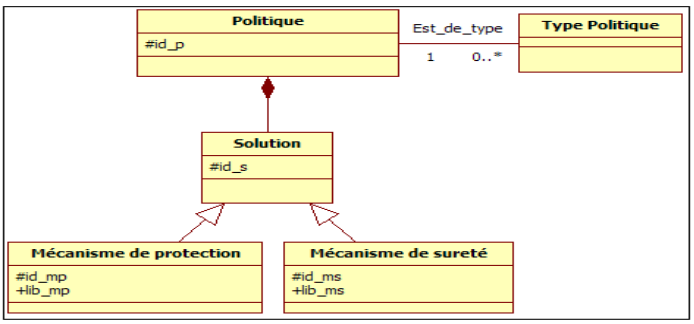


Fig.4 – Méta Modèle des communications Webhouse-Décideur

Ce méta modèle consiste à définir pour chaque Politique un type dépendant du niveau de sécurité attribué au décideur. Cette politique est composée de solutione qui offrent des mécanismes de sureté et d'autres de protection.

Par l'opérationnalisation des softgoals (NFRs exigence non fonctionnelles) d'un graphe orienté but, nous instancions ces mécanismes de sécurité. Notre tâche consiste à choisir l'opérationnalisation qui satisfait le besoin de communication de chaque décideur. Cette étape est nommée Opérationnalisation des buts, pour laquelle, il existe dans la littérature une diversité de mécanismes. Notre choix se base sur la sensibilité des données accédées par un décideur car comme nous l'avons mentionné dans les parties précédentes, nous attribuons des étiquettes de sécurité à chaque donnée du webhouse. Ces étiquettes précisent quels sont les décideurs autorisés à consulter une telle donnée.

Dans la figure suivante, nous présentons un graphe orienté but offrant les mécanismes nécessaires pour sécuriser les communications d'un décideur (ayant un niveau de sécurité Top Level) autorisé à consulter des données du webhouse très sensibles et importantes.

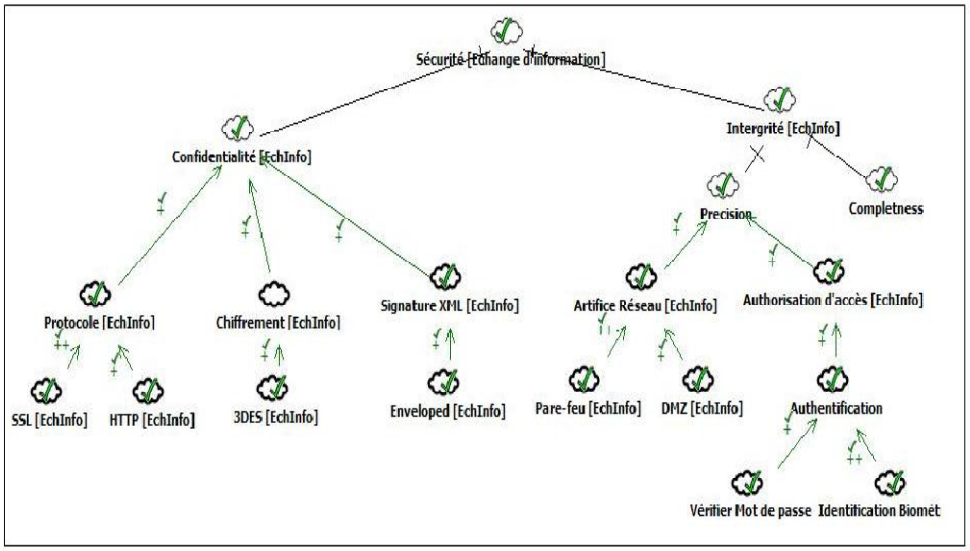


Fig.5 – Graphe orienté but de sécurisation des communications des données TopLevel

4.3 Définition des modèles de code des Communications (PSM)

Après l'analyse des exigences et la définition du modèle d'analyse, nous passons à la définition des modèles de code correspondant au modèle PSM de l'architecture MDA. Le modèle PSM sert essentiellement de base à la génération de code exécutable vers la ou les plates-formes techniques. L'output de la transformation du modèle de sécurité des communications webhouse-Décideur défini au niveau PIM, est un fichier XACML (Oasis, 2005) contenant la politique de sécurité.

XACML (Extended Access Control Markup Language) est un standard OASIS dédié à la définition des politiques de contrôle d'accès dans des fichiers XML. Il existe un framework (SUN) qui soutient et aide à l'accès, à écrire et analyser le fichier XACML. Cette partie est en cours de traitement.

5 Conclusion

Dans cet article, nous proposons une approche de sécurisation des communications webhouse-décideur en tenant compte des exigences de sécurité dès le niveau CIM. A l'aide du langage de transformation QVT, nous générons le modèle de sécurité du niveau PIM.

Ce travail présente une étape dans la définition d'une démarche complète de conception de webhouse sécurisé. Nos futures recherches vont traiter plus largement ce sujet et proposer une approche qui couvre la sécurité des données, des utilisateurs et des communications notamment en considérant l'aspect web, d'un coté, et les normes et les standard de sécurisation des systèmes, de l'autre.

Références

André, S. (2004). *Mda (model driven architecture) principes et états de l'art*. Examen probatoire du diplôme d'ingénieur C.N.A.M en INFORMATIQUE, Conservatoire national des arts et métiers centre d'enseignement de LYON.

Baril, X. et Z. Bellahsène, (1996). *Designing and Managing an XML Warehouse*. Extrait d'un Chapitre de livre XML Data Management : Native XML and XML-Enabled Database Systems, écrit par Akmal B. Chaudhri, Awais Rashi et Roberto Zicari.

Best. B., J. Jan, N. Bashar, (2007), Model-based Security Engineering of Distributed Information Systems using UMLsec.

Blanc, X. avec la contribution de O. Salvatori (2005), extrait du livre *MDA en Action*, Éditions Eyrolles-Collection : Architecte logiciel, 1ère édition.

Blanco, C., F-M. Eduardo, J. Trujillo et M. Piattini (2008*), Implementing Multidimensional Security into OLAP Tools*, The Third International Conference on Availability, Reliability and Security, Barcelona, pp.1248-1253.

Chang D.T (2001). *CWM Enablement Showcase.* IBM Database Technology Institute, 2001.

GRARI, M. (2007). *Principes et états de l'art de l'approche MDA et applications pour des plates-formes PHP orienté 3-tiers*. Diplôme des Etudes Supérieures Approfondies (DESA), Université Mohammed Premier Faculté des sciences OUJDA.

Juan, P., G. Silva, A. Miguel, B. Javier Fernández et A. Alejandro, (2010), Model-Driven Development of a Web Service-Oriented Architecture and Security Policies.

Katic, N., G. Quirchmayr, J. Schiefer1 M. Stolba et A.M. Tjoa, (1998), *A prototype model for data warehouse security based on metadata*, Proceedings of the 9th International Workshop on Database and Expert Systems Applications (DEXA'98), Vienna, Austria, pp.300-309.

Kimball. R et R. Merz, (2000). *Le DATA WEBHOUSE : analyser les comportements client sur le Web*", Eyrolles Edition , ISBN 2212091648.

Löf, F., J. Stomberg, T. Sommestad, M. Ekstedt, J. Hallberg, (2010), *An Approach to Network Security Assessment based on Probabilistic on Probalistic Relational Models*, First workshop on secure Control System sockholm Sweden.

Lonjon. A., J-J. Thomasson et L. Maesano, (2006). *Modèlisation XML*, EYROLLES.

LUJÁN, S.M., J.TRUJILLO et I.Y.SONG, "Extending the UML for multidimensional modeling", 5th International Conference on the Unidfied Modeling Language, Dresden, Germany, pp.265-276.

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

OASIS *"eXtensible Access Control Markup Language (XACML) Versión 2.0"*, IN Moses, T. & Godik, S. (Eds.), OASIS, February 2005.

OMG, (2003). *CWM:Common Warehouse Metamodel(CWM) Specification*.

Poggi, S (2005), *Rapport de veille sur les standards et méthodes en matière de sécurité informatique* , http://www.cases.lu.

Pozo, S., A.J. Varela-Vaca et M.G. Rafael (2009), *MDA-Based Framework for Automatic Generation of Consistent Firewall ACLs with NAT*.

Rosenthal. A. et E.Sciore, (2008), *View security as the basic for data warehouse security*. Workshop on Design and Management of Data Warehouse (DMDW'00), Sweden, pp.1-8.

Soler. E., V. Stefanov, J-N. Mazon, J. Trujillo, F-M. Eduardo et M. Piattini, (2008), *Towards Comprehensive Requirement Analysis for Data Warehouses: Considering Security Requirements*, The Third International Conference on Availability, Reliability and Security.

Triki. S., H. B. Abdallah, J. Feki, et N. Harbi, (2010), Sécurisation des entrepôts de données contre les inférences en utilisant les réseaux bayésiens, EDA, Tunisie, pp.35-47.

Summary

The advent of the Web gives rise to a new generation of decision support systems: Webhouse. The use of WEBHOUSE and diversification and multiple users and communications increase the risk of intrusion. Therefore, it becomes essential to treat the subject of safety in the design of WEBHOUSE providing security measures for data protection and a mechanism to control user access to data via secure transmission paths. The definition of these needs must be made by a business manager from the beginning of the design process of WEBHOUSE. In this paper, we present an approach for securing WEBHOUSE and specifically communications byfollowing the MDA architecture. For transformations between models of MDA, we apply the rules of the transformation language QVT.

Agrégation sémantique du texte

Meriem Bouslah[(1)] et Nadjia Benblidia [(1,2)]
(1) Laboratoire de Recherche pour le Développement des Systèmes Informatisés
(2) Laboratoire de Traitement du Signal et de l'Image
Université Saad Dahlab de Blida –BP 270, Route de Soumâa ; 9000 Blida
bouslah.meriem@yahoo.com, benblidia@univ-blida.dz

Résumé. Les technologies entrepôt de données et OLAP actuelles permettent d'analyser et d'interroger les données structurées mais demeurent inefficaces pour l'analyse des données textuelles non structurées faute d'opérateurs adéquats. En effet, un enrichissement des opérateurs OLAP classiques s'avère plus que nécessaire. Dans cet article, nous nous intéressons à une nouvelle fonction d'agrégation permettant d'agréger un contenu textuel non structuré par ses mots les plus représentatifs en utilisant la fonction de pondération de termes TF-IDF tout en préservant la sémantique du contenu textuel lors de l'agrégation.

1 Introduction

Les technologies entrepôt de données et OLAP (on-line analytical processing) actuelles permettent d'analyser et d'interroger les données numériques que les entreprises stockent dans leurs bases de données. Cela s'effectue grâce à des bases multidimensionnelles pour le stockage des données et grâce à des opérateurs OLAP pour la manipulation et l'interrogation de ces données. Ces opérateurs permettent différentes opérations d'analyses telles que le groupement ou l'agrégation. En effet, l'un des avantages les plus importants de ces approches est notamment de pouvoir utiliser des opérateurs d'agrégation telle que Roll-Up ou Drill-Down pour naviguer au travers des dimensions et ainsi agréger les données en fonction des requêtes utilisateurs. Cependant, ces entrepôts de données, aussi efficaces dans l'analyse des données numériques, ne s'appliquent pas malheureusement sur les données textuelles.

Or selon Tseng et Chou (2006), les données numériques exploitées par l'entrepôt de données ne représentent que 20% des données du système d'information de l'entreprise. Les 80% restantes, généralement contenues dans des documents électroniques, restent hors de portée de la technologie OLAP par manque de méthodes de conception adaptées et d'opérateurs adéquats capable d'analyser du contenu textuel. L'omission des données textuelles, situées essentiellement au niveau des documents, de la procédure d'analyse va priver l'utilisateur d'une quantité d'information assez importante. Afin de pallier à ce manque, et de permettre l'analyse des données textuelles, nous pouvons faire appel aux différents domaines qui s'intéressent au traitement de texte tel que : la recherche d'information, la fouille de données, …etc. Dans cet article, nous nous intéressons à la problématique d'agrégation du contenu textuel dans un environnement OLAP en utilisant des techniques issues de la re-

cherche d'information. Cette dernière a connu le développement de nombreux outils robustes et validés pour le traitement du texte. D'autre part, les ressources sémantiques (thésaurus, ontologies, etc.) ont un apport considérable dans le traitement du texte, ils permettent de ne pas considérer les termes dans un texte comme de simples chaines de caractère mais plutôt comme des concepts où la signification du terme et le rapport sémantique entre termes est pris en compte. Dans ce cadre, nous proposons une nouvelle fonction pour l'agrégation du contenu textuel en se basant sur des techniques de TAL (Traitement Automatique des Langues) et de la recherche d'information. Cette fonction est aussi guidée par une ontologie afin de préserver la sémantique du texte à agréger.

Le reste de l'article est organisé de la manière suivante. Dans la section 2, nous décrivons quelques travaux antérieurs liés à notre problématique. La section 3 détaille les différentes parties de notre proposition. Enfin nous concluons cet article en présentant quelques perspectives.

2 Travaux antérieurs

Pour que la donnée textuelle soit exploitable au niveau de l'entrepôt de données, il faut absolument étendre les approches classiques d'entrepôts de données et les technologies OLAP adaptés jusqu'à présent qu'aux données numériques. Dans ce cadre, il y a eu récemment plusieurs propositions pour l'analyse multidimensionnelle de contenu textuel au sein d'un environnement OLAP. Dans Park et al. (2005), les auteurs proposent un Framework d'analyse multidimensionnel du contenu textuel, plusieurs fonctions d'analyse sont proposées : summary, top_keywords, topic et clustering, toutefois, ces fonctions ne sont ni détaillées, ni formalisées, ni implantées.

D'autres travaux ont apporté des fonctions d'analyses plus détaillés en s'inspirant des techniques de la recherche d'information. Selon les auteurs de Pérez et al. (2006), nous pouvons classer les approches combinant recherche d'information et analyse OLAP selon deux catégories : Les approches utilisant les bases multidimensionnelles pour l'implémentation des systèmes de recherche d'information et les approches utilisant la recherche d'information pour la manipulation de contenu textuel au sein des bases de données multidimensionnelles.

2.1 Systèmes de recherche d'informations multidimensionnels

L'implémentation des systèmes de recherche d'information en utilisant les bases multidimensionnelles vise essentiellement d'améliorer la pertinence des réponses offertes aux utilisateurs par ces systèmes. Le fait de stocker les documents dans des entrepôts de données selon des thématiques (sujets) bien déterminés peut énormément faciliter la navigation au sein du corpus des documents. Toutefois, les travaux qui utilisent les bases de données multidimensionnelles de texte pour l'implémentation des systèmes de recherche d'information, ne vont pas jusqu'à la réalisation effective des opérations OLAP sur les documents textuels. C'est plutôt les avantages de l'organisation multidimensionnelle des documents qui justifie le recours à des bases multidimensionnelles pour l'implémentation de systèmes de recherche d'information.

Parmi ces travaux, nous pouvons citer :

2.1.1 Les travaux de McCabe et al

Selon McCabe et al. 2000, la donnée textuelle est souvent convertible en une donnée multidimensionnelle ce qui permet son intégration dans une base de données multidimensionnelle. Les auteurs proposent une approche pour l'exploitation d'un entrepôt de documents. L'interrogation de ce dernier s'effectue par la soumission d'une requête composée de mots clés. Une nouvelle dimension appelée *queryterm* pour abriter cette requête est crée au niveau de l'entrepôt de données. *Querryterm* est intégrée au niveau de la requête MDX (MultiDimensional eXpressions) afin de pouvoir sélectionner l'ensemble des documents satisfaisants à la requête *queryterm* et la requête MDX. La sélection des documents satisfaisants se base sur le calcul de la pertinence d'un document en utilisant TF et IDF comme mesures au sein de l'entrepôt de données.

2.1.2 Les travaux de Lee et al

Dans Lee et al. (2002), les auteurs proposent un nouveau moteur de recherche multidimensionnel qui est capable de prendre en charge la nature hiérarchique des données et qui permet de naviguer selon les hiérarchies. Pour ce faire, ils adoptent une modélisation multidimensionnelle des données structurées en utilisant un schéma en étoile. Quant au texte, il est intégré en utilisant la structure de donnée « index inversé » bien connue en recherche d'information. Par l'adoption d'une modélisation multidimensionnelle des données au niveau du moteur de recherche, Lee et al estiment ainsi permettre la possibilité d'effectuer des opérations OLAP tel que la navigation à travers la granularité (Roll-up et drill-down).

2.1.3 Les travaux de Lin et al

Les auteurs de Lin et al. (2008) proposent un système de recherche d'information sous forme d'un cube de données qu'ils appellent « *text cube* »basé sur une base multidimensionnelle de documents. Ce cube de données supporte deux mesures : une mesure textuelle qui représente l'ensemble de mots clés qui indexent le document et une mesure numérique élaborée calculée sur la base des deux fonctions de pondération : TF et IDF. Ce modèle comporte aussi une nouvelle dimension appelée hiérarchie des termes, elle représente l'ensemble des liens sémantiques qui relient tous les mots clés utilisées pour indexer les documents . L'interrogation du *text cube* se fait de la manière suivante : une requête composée de mots clés et des contraintes sur les dimensions est soumise. Une reformulation de la requête ainsi qu'une agrégation des documents sont effectuées. La reformulation consiste à remplacer chaque mot clé de la requête par ses descendants dans l'hiérarchie des termes. L'évaluation par le *text cube* de la requête reformulée nous donne des documents agrégés en un ensemble de mots clés combinés à des valeurs de TF et IDF mises à jour.

2.2 Entrepôt de données textuelles

Les approches que nous allons présenter dans cette section ont l'objectif commun de permettre la manipulation du contenu textuel non structuré au sein d'un environnement OLAP. Un critère important qui peut différencier ces approches est le type de l'intégration du contenu textuel appliquée au niveau de l'entrepôt de donnée. Ainsi, nous pouvons distinguer deux types d'intégration (Tournier (2007)) :

2.2.1 Intégration logique du contenu textuel au niveau des entrepôts de données

Le but de ce type d'intégration est d'essayer d'exploiter le contenu textuel qui est au niveau des documents et qui n'est pas pris en charge par les entrepôts de données classiques. Parmi les travaux qui ont adopté ce type d'intégration, nous citons Les travaux de Pérez et al. (2005) qui proposent un Framework pour l'exploitation des données documentaires situées au niveau des documents XML. Le travail présenté consiste à construire un nouveau entrepôt de données appelé *R-Cube* composé d'un entrepôt de données classique et d'un entrepôt de documents contenant les documents XML. Le *R-Cube* est aussi composé de deux nouvelles dimensions : contexte et pertinence. L'interrogation de l'entrepôt de données classique se fait par une requête MDX relative à un contexte donné tandis que l'interrogation de l'entrepôt de document se fait grâce à un ensemble de mots clés (une requête RI) liés au même contexte. La réponse du *R-Cube* serait un ensemble de faits et de documents. Ces documents représentent une information complémentaire à l'information offerte par l'entrepôt de donnée classique pour un contexte d'analyse donné ce qui aide dans la compréhension des faits observés. L'entrepôt de documents sert aussi à extraire de nouveaux faits liées au contexte et qui ne sont pas déjà présents dans l'entrepôt de données classique ce qui constitue un enrichissement de ce dernier. Néanmoins, le *R-Cube* ne permet pas une interrogation du contenu textuel des documents à travers des opérations d'analyse OLAP tel que le groupement ou l'agrégation. Le *R-Cube*, et pour un contexte d'analyse donnée, permet l'enrichissement de l'entrepôt de données par de nouveaux faits extraits depuis les documents ainsi que la sélection d'un ensemble de documents ou de fragments de documents intéressants à lire par rapport à un fait donné.

2.2.2 Intégration physique des documents au niveau des entrepôts de données

Les travaux de Ravat et al. Ce travail Ravat et al. (2008) s'articule autour de la proposition d'une fonction d'agréation TOP-KEYWORD du contenu textuel issu de documents XML. Cette fonction agrège ou résume un ensemble de mots en se basant sur la fonction de pondération TF-IDF. Deux nouvelles mesures ont été introduites par les auteurs : la mesure textuelle brute et la mesure textuelle élaborée. L'opération d'agrégation, proposée dans ce travail est composée de trois étapes: prétraitement, ordonnancement des termes et choix des k les plus termes représentatifs. Le prétraitement consiste à éliminer les termes susceptibles de biaiser les calculs et de parasiter les résultats, les termes résultants de cette étape seront ordonnés selon un poids calculé sur la base de la formule de pondération TF-IDF, les K premiers termes de cette liste ordonnées seront considérés comme l'agréation du contenu textuel.

3 Contribution

3.1 Cas d'étude

Afin de mieux illustrer la problématique posée, nous allons nous intéresser à un cas d'étude bien précis, l'analyse de l'activité de recherche scientifique à l'aide d'un entrepôt d'articles scientifiques. En effet, nous pouvons analyser les avancées de la recherche scientifique dans un domaine donné par la consultation et l'évaluation des articles scientifiques

publiés relativement à ce domaine, mais ce travail constitue une tâche très laborieuse vu le nombre important d'articles scientifiques publiés chaque année. Un entrepôt d'articles scientifiques serait un moyen très efficace pour réaliser cette tâche sauf que cet entrepôt doit être capable de manipuler le contenu textuel des articles scientifiques ce que les entrepôts de données classiques demeurent incapables de réaliser.

Le modèle multidimensionnel de données représente le cœur d'un entrepôt de données. Pour l'entrepôt d'articles scientifiques, nous avons adopté le modèle multidimensionnel en étoile suivant(FIG.1) :

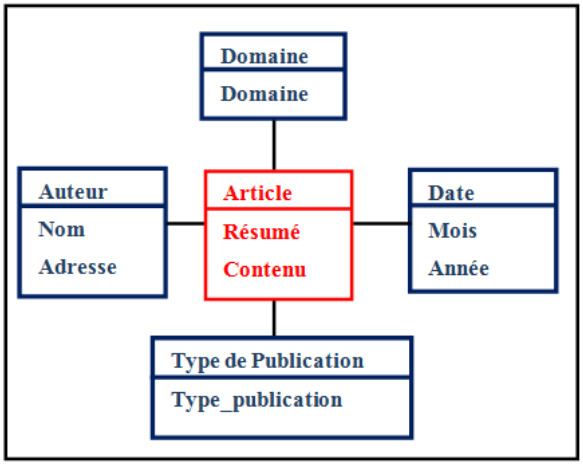


FIG.1- *Modèle multidimensionnel en étoile-entrepôt d'articles scientifiques*

Les articles scientifiques sont organisés selon quatre dimensions, à savoir : l'auteur, le domaine scientifique concerné, la date de publication (le mois et l'année) et le type de publication dont l'article scientifique a fait objet. La table de fait Article contient deux mesures textuelles : *résumé* et *contenu*. *Résumé* n'est que le résumé sous forme d'un paragraphe qu'ajoute la majorité des auteurs à leurs articles, quant à la mesure *contenu*, elle englobe le contenu textuel que contient l'article délimité par l'introduction et la conclusion jusqu'à sa fin. Des parties telles que la bibliographie ou l'annexe ne font pas partie de la mesure *contenu*. En effet, *résumé* représente l'idée générale de l'article du point de vue de l'auteur et *contenu* représente l'ensemble des détails qu'a apporté l'auteur sur le sujet abordé, selon le type d'analyse voulue : détaillée ou non, nous choisissons la mesure *résumé* ou *contenu*.

3.2 Alimentation des mesures :

Nous avons précisé dans la section précédente la localisation du contenu des deux mesures : *résumé* et *contenu* par rapport à l'article scientifique. Néanmoins, il ne s'agit pas d'un copiage fidèle du contenu, les mesures *résumé* et *contenu*, ne contiendront pas le contenu textuel intégral correspondant dans l'article scientifique mais un ensemble de termes choisis à partir de ce contenu selon leurs degrés d'importance. En effet, nous allons identifier le contenu textuel correspondant à chacune des deux mesures par rapport au contenu de l'article. Par la suite, nous allons appliquer un ensemble de techniques issues du domaine de

TAL en vue de ne garder que les termes susceptibles d'être représentatifs. Cette étape aide à normaliser ce contenu initial de mesures et de lever certaines ambigüités. Elle consiste en trois opérations élémentaires qui sont :

3.2.1 Tokenisation:

D'abord, nous devons reconnaitre les termes composant le contenu textuel.

3.2.2 Suppression de mots vides à partir d'une « Stoplist » :

Une fois on a la liste de termes après Tokenisation, nous devons en retirer tous les termes non porteurs de sens grâce à une stop liste. Plusieurs stop liste ont été proposées, nous avons choisi de travailler avec la stop liste de Fox (Fox, C. (1990)). Comme il s'agit d'une stop liste générale, un travail de personnalisation au domaine des articles scientifiques a été effectué.

3.2.3 Lemmatisation :

Nous allons réaliser cette opération, en utilisant l'étiqueteur morphosyntaxique « TreeTagger » (Schmid, H. (2011)) qui génère pour chaque terme son lemme associé et sa catégorie grammaticale. Remplacer chaque terme par son lemme associé va nous permettre de lever de nombreuses ambigüités liée aux flexions quant à la catégorie grammaticale, elle va nous aider à réaliser un filtrage sur les termes où nous gardons que les trois catégories grammaticales suivantes : nom, verbe et adjectif.

A la fin de cette première étape, nous pouvons alimenter les deux mesures *résumé* et *contenu* par la liste de termes retenus correspondante.

3.3 Approche d'agrégation sémantique

Permettre l'agrégation du contenu textuel dans un environnement OLAP revient à la création de nouvelles fonctions pour l'opérateur d'agrégation OLAP usuel afin que ce dernier soit capable d'agréger du texte. Par ailleurs, il ne faut pas négliger l'aspect sémantique du contenu textuel. En effet, chaque terme est porteur d'un sens qu'il faut prendre en considération. Dans ce cadre, nous proposons une fonction pour l'agrégation du texte tout en préservant sa sémantique. Le texte à agréger est une donnée textuelle non structurée (elle n'obéit à aucune forme ou structure), il peut être une phrase, un fragment de texte, ou tout simplement un ensemble de termes. Dans notre cas, nous avons à agréger le contenu textuel d'un article scientifique à travers les deux mesures *résumé* et *contenu*. L'agrégation consiste à résumer le texte de chacune des deux mesures en une liste de *m* termes les plus représentatifs. Pour arriver à repérer ces m termes, nous avons trois étapes à accomplir: pondération des termes choisis, repérage des liens sémantiques entre termes et finalement choix des *m* termes à retenir.

L'agrégation se déroule comme suit :

3.3.1 Pondération des termes retenus

Il s'agit de calculer la représentativité (l'importance) de chacun des termes du texte à agréger. A l'issue de cette étape, nous aurons une liste de termes pondérés. Le calcul de représentativité se fait en utilisant la formule de pondération TF-IDF [Salton et al. (1975)]. La formule TF-IDF a pour objectif de sélectionner les termes clés les plus significatifs à partir d'un ensemble de termes. Elle permet de calculer le poids d'un terme dans un document par rapport à une collection de documents. Une adaptation de la formule à notre problématique consiste à choisir le paragraphe comme une unité d'agrégation au lieu du document. Le texte à agréger doit être réorganisé en paragraphes.

Le calcul de TF-IDF nécessite le calcul des deux fonctions TF et IDF selon les formules adaptées suivantes :

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

Où $TF_{i,j}$ est la fréquence du terme i par rapport au paragraphe j, $n_{i,j}$ est le nombre d'occurrence du terme t_i dans le paragraphe p_j et $\sum_k n_{k,j}$ e st le nombre d'occurrence de tous les termes dans le paragraphe p_j .

$$IDF_i = log_{10}\left(\frac{|P|}{|\{p_j : t_i \in p_j\}|}\right)$$

IDF_i est la fréquence inverse du paragraphe liée au terme t_i, |P| est le nombre de paragraphes dans le texte et $|\{p_j : t_i \in p_j\}|$ est le nombre de paragraphe où le terme t_i apparait
Finalement, le poids s'obtient en multipliant les deux fonctions :

$$TF - IDF = TF * IDF$$

A l'issue de cette étape, à chaque terme est associé un poids représentant l'importance de ce dernier.

3.3.2 Repérage des liens sémantiques entre termes

Ce traitement a pour but d'identifier les liens sémantiques qui lient les termes pondérés résultants de l'étape précédente afin d'élire les termes dont la sémantique est la plus représentative. En effet, TF-IDF associe au terme un poids numérique de représentativité par rapport au texte et à l'unité d'agrégation (paragraphe) ou se trouve le terme, mais la sémantique du terme n'est nullement prise en compte dans le calcul de ce poids ce qui peut générer une grande perte de la sémantique du texte lors de son agrégation.

Pour pallier à ce problème, nous proposons un ajustement du calcul de ce poids en prenant en considération les différents liens sémantiques qui existent entre les termes tel que la synonymie. Pour ce faire, nous allons calculer la distance sémantique entre tous les termes pondérés résultants de l'étape précédente deux à deux, en utilisant l'ontologie WordNet 2.1 (Chaumartin,F,R. (2007)). Cette distance sera comparée à un seuil de similarité α. Si la distance résultante est inférieure à α, nous considérons les deux termes comme similaires. Pour

chaque deux termes dit similaires, nous ne gardons qu'un seul et nous éliminons l'autre. Le choix du terme à garder se fait par rapport aux poids des deux termes, nous gardons le terme qui a le poids le plus fort, une fois le terme retenu, son poids va changer, il correspondra à la somme de son ancien poids et le poids du terme similaire qui a été éliminé. Dans la littérature, plusieurs travaux sur la mesure de similarité sémantique utilisant une ontologie ont été développés [Thabet et al, 2007], nous avons choisi de travailler avec la mesure Resnik(1995).

A la fin de cette étape, nous aurons une liste de termes pondérés. L'agrégation consiste à choisir les *m* termes de la liste des termes ayant les poids les plus forts avec *m* un paramètre que l'utilisateur fixe selon la taille d'agrégat voulue.

Conclusion

Dans ce travail, nous avons proposé une approche pour l'agrégation du contenu textuel en un ensemble de termes représentatifs avec prise en considération de la sémantique du texte. Cette approche se base sur un calcul de pondération des termes. Ce calcul de pondération se fait en deux parties : d'abord un calcul de représentativité basée sur la formule TF-IDF en divisant le texte en paragraphes. Puis, un ajustement du poids déjà calculé par un calcul de similarité sémantique à l'aide de la mesure Resnik et de l'ontologie WordNet 2.1 dans le but de comptabiliser les liens sémantiques qui existent entre les termes.

Dans le future, nous envisageons de développer l'approche sur le volet sémantique pour une meilleure qualité de l'agrégat et de meilleures performances.

Références

Chaumartin, F, R. (2007). WordNet et son écosystème : un ensemble de ressources linguistiques de large couverture. Université Paris 7.

Fox, C. (1990). A Stop List for General Text. Laboratories Lincroft New Jersey.

Schmid, H. (2011). TreeTagger – a language independent part–of–speech tagger. Institute for Computational Linguistics of the University of Stuttgart.

Lee,J, D. Grossman, et R. Orlandic (2002). MIRE: A Multidimensional Information Retrieval Engine for Structured Data and Text. In *Proceedings of the International Conference on Information Technology: Coding and Computing*, pages 224–229. IEEE Computer Society, Washington, DC.

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

McCabe, C. Lee, J. Chowdhury,A. A. Grossman, D et Frieder, O (2000). On the design and evaluation of a multi-dimensional approach to information retrieval. 23rd Intl. ACM Conf. on Research and Development in Information Retrieval (SIGIR), ACM Press, p. 363–365.

Park, B-K. H, Han et I-Y Song (2005). XML-OLAP: A Multidimensional Analysis Framework for XML Warehouses, 7th Intl. Conf. on Data Warehousing and Knowledge Discovery (DaWaK), LNCS 3589, Springer, p. 32–42.

Pérez, J-M. R, Berlanga. M-J, Aramburu et T- B Pedersen (2006). Integrating Data Warehouses with Web Data: A Survey. DB technical report, TR-18.

Pérez, J-M. R, Berlanga. M-J, Aramburu et T- B Pedersen (2005). A Relevance-Extended Multi-dimensional Model for a Data Warehouse Contextualized with Documents. In *Proceedings of the 8th ACM international workshop on Data warehousing and OLAP*, pages 19–28. ACM Press, New York.

Ravat, F. Teste, O. Tournier, R. et Zurfluh, G (2008). Top-keyword: An aggregation function for textual document OLAP. Lecture Notes in Computer Science, 5182, 55-64.

Resnik, P (1995). Using information content to evaluate semantic similarity in taxonomy, In Proceedings of 14th International Joint Conference on Artificial Intelligence, Montreal.

Salton, G., A. Wong, et C. S. Yang (1975). A vector space model for automatic indexing *Commun ACM 18*(11), 613–620.

Tournier, R (2007). Analyse en ligne (OLAP) de documents. Université Toulouse 3, france.

Tseng F.S.C., A.Y.H Chou (2006). The concept of document warehousing for multidimensional modeling of textual-based business intelligence. *Journal of Decision Support Systems* (*DSS*), vol.42(2), Elsevier, p. 727–744.

Summary

Current data warehouse and OLAP allow the query and the analysis of the structured data but they remain ineffective in the analysis of the unstructured textual data in the absence of suitable operators. Indeed, an enrichment of traditional operators is more than necessary. In this paper, we focus on a new aggregate function to aggregate an unstructured text content with its most representative words using the terms weighting function "TF-IDF" while preserving the semantics of the textual content in the aggregation. Experiments on data from databases of scientific papers are presented.

Summarization des Documents par Catégorisation dans les *Text Cubes*

Aicha Lababou*, Nadjia Benblidia*,**

* Laboratoire de Recherche pour le Développement des Systèmes Informatisés
** Laboratoire de Traitement du Signal et de l'Image
Université Saad Dahlab Blida –BP 270, Route de Soumâa ; 9000 Blida
alababou@live.fr, benblidia@univ-blida.dz

Résumé. Avec la croissance explosive des données textuelles dans les organisations ainsi que sur le web, il devient nécessaire d'analyser aussi bien les données structurées que les données texte non structurées. Les processus OLAP ont prouvé leur efficacité dans l'analyse des données structurées. Cependant, ils sont inadaptés à l'analyse des données texte. Un des défis que doivent relever les processus OLAP (*On-Line Analytical Processing*) est l'agrégation des données texte. Dans cet article, nous proposons une approche de *summarization*, basée sur la catégorisation de texte. Une évaluation de l'approche a été menée à travers une étude expérimentale sur un corpus de documents issu du domaine des ressources humaines, les CVs.

1 Introduction

L'entreposage de données et l'analyse en ligne (OLAP) gagnent de plus en plus en popularité au sein des organisations. Ces dernières ont réalisé les avantages des analyses multidimensionnelles sur de grands volumes de données. Lorsque ces dernières sont historisées, et organisées par sujet dans le but d'une prise de décision. Cependant, seuls 20% des données d'un système d'information d'une entreprise sont stockés dans les bases de données relationnelles et peuvent être traités par un système OLAP (Tseng et Chou (2006)). Pour Sullivan (2001), la plupart des efforts n'ont touché que la partie visible de l'iceberg des flux informationnels. En effet, les 80% restants de l'information sont contenus dans des documents non structurés ou semi-structurés appelés *données ou objets complexes* (Darmont et al. (2005)). Ces documents sont principalement des documents texte, car ce type de données est le plus répandu pour exprimer les informations et les connaissances.

Avec la croissance explosive des données textuelles, aussi bien dans les organisations que sur le web, il devient nécessaire d'aller au-delà de l'analyse en ligne des données structurées pour prendre en charge celle des données texte, non structurées, et couvrir ainsi les 100% des données d'un système d'information. Cependant, aller au-delà des nombres constitue un

challenge pour les processus OLAP. Deux principaux problèmes sont à considérer : Le premier est relatif à l'intégration et au stockage des informations issues de documents hétérogènes. Cette hétérogénéité est aussi bien structurelle que sémantique. De plus, le contenu de documents textuels étant peu structuré, ce qui explique la difficulté de leur intégration dans les systèmes décisionnels. La tendance actuelle pour résoudre ce problème consiste à utiliser le standard XML[1] (eXtensible Markup Language). Des travaux ont montré que XML peut être une solution permettant de représenter et d'intégrer des documents peu ou pas structurés au sein des systèmes décisionnels (Boussaïd et al. (2006) ; Pérez et al. (2007)). XML, une norme du W3C[2], est considéré comme un standard dans la description et l'échange des données. Il représente les données de façon semi-structurée. Aussi, sa capacité d'auto-description et sa structure arborescente donnent à ce formalisme une grande flexibilité et une puissance suffisante pour décrire des données complexes, hétérogènes et provenant de sources éparpillées.

Le second problème consiste à déterminer, les informations à extraire des documents textuels pour servir aux différents processus de restitution, notamment l'analyse OLAP. Se pose alors les questions suivantes. Est-il possible d'analyser des données textuelles représentées à l'aide d'un schéma multidimensionnel ? Doit-on faire évoluer les modèles de schémas en étoiles existants ? Comment peut-on agréger les données textuelles ? Ce dernier problème constitue l'un des challenges que doit relever les processus OLAP. En effet, avec les outils OLAP classiques, il est impossible d'agréger des données textuelles selon des fonctions telles que la somme ou la moyenne. L'environnement OLAP de données textuelles, appelé aussi *Text OLAP* (Park et Song (2011)) a besoin d'outils appropriés et de nouvelles techniques d'agrégation pour ce type de données. La tendance actuelle, vise à combiner l'OLAP avec une des principales technologies manipulant les données textuelles, la Recherche d'Information (RI), et la Fouille de Texte (Text mining, TM). Notre approche s'inscrit dans la tendance de celles qui ont pour but de combiner l'OLAP et la fouille de texte. Nous pensons que ces deux techniques se complémentent et leur association serait une solution envisageable pour une analyse plus élaborée et sémantiquement plus riche des données textuelles. Nous proposons dans cet article la catégorisation comme technique pour résumer les documents textuels dans les *Text cubes*. Pour mettre en évidence le principe de notre approche nous avons pris un exemple concret, celui de l'analyse et le traitement automatique des candidatures (CVs) dans le domaine des ressources humaines.

La suite de cet article est organisée comme suit. Dans la section 2, nous passons en revue les principaux travaux qui ont abordé l'analyse en ligne de données textuelles. La section 3 aborde le principe de la *summarization* par catégorisation des documents textuels. La section 4 est consacrée à la partie catégorisation réalisée sur des documents CVs, ainsi qu'à une expérimentation. Enfin, nous terminons par une conclusion et nous dressons quelques perspectives.

2 Approches *Text OLAP* : état de l'art

L'analyse multidimensionnelle repose sur la capacité à résumer et à synthétiser des données très volumineuses. Le problème majeur quant à l'analyse OLAP des documents texte

[1] Extensible Markup Language (http://www.w3.org/XML/)
[2] http://www.w3.org/

est de pouvoir agréger les données textuelles. L'environnement OLAP ne fournit pas d'outils à l'agrégation de ces dernières. Des travaux récents se sont intéressés à intégrer les données textuelles dans les modèles multidimensionnels. Des dimensions spécifiques aux données textuelles, ainsi que des mesures et méthodes d'agrégation adaptées à ces données sont proposées. Dans Zhang et al. (2009) par exemple, les auteurs proposent une dimension *Topic* construite à partir d'une arborescence hiérarchique de thèmes (*Topics*). Ces derniers sont extraits des documents en utilisant l'approche PLSA. Cette dimension va permettre à l'utilisateur de procéder à des forages le long de la hiérarchie pour explorer le contenu des documents textuels selon différents niveaux de granularité. Deux mesures sont utilisées : la distribution des mots d'un thème dans le document et la couverture du thème par le document. Dans Park et al. (2005), les auteurs proposent des fonctions inspirées de la fouille de texte pour l'analyse du contenu de documents textuels : *SUMMARY*, permet la génération d'un résumé du texte à agréger ; *TOP-KEYWORDS*, sélectionne les n principaux mots-clés du texte à agréger ; *TOPIC*, extrait le thème d'un bloc de texte et *CLUSTER* partitionne le texte en fonction du contenu. Les travaux de Ravat et al. (2007, 2008) proposent deux fonctions d'agrégations. *TOP_KW* qui fournit les n principaux mots-clés en utilisant la fonction de poids des termes *tf.idf* issue de la recherche d'information et *AVG_KW* qui combine plusieurs mots clés en un mot-clé plus général en utilisant une ontologie de domaine. Ces auteurs proposent aussi une dimension documentaire construite sur la base du contenu et de la structure logique des documents. La hiérarchisation de ce type de dimension représente la structure logique générique d'une collection de documents. Dans Zhang et al. (2011), un résumé est fourni à l'utilisateur, sous forme de documents les plus représentatifs de la cellule du cube de texte, en utilisant la technique du *clustering*. Si l'utilisateur désire un résumé formé de *k* documents les plus représentatifs, *k* clusters seront créés et le document centroide de chaque cluster est choisi pour former le résumé. Lin et al. (2008) proposent une nouvelle dimension, une hiérarchie sémantique de mots avec deux nouveaux opérateurs OLAP *pull-up* et *push-down* et des mesures issues de la RI (*term frequency* et *inverted index*). Dans Simitsis et al. (2008), les auteurs proposent MCX (*Multidimensional Content eXploration*) un framework dans lequel les concepts : document, contenu, métadonnées, liens entre documents, sont formellement mappés sur un schéma multidimensionnel : table de faits, dimensions statique ou dynamique, mesures. Dans un premier temps les documents les plus pertinents, par rapport à une requête utilisateur, sont extraits en utilisant un système classique de recherche d'information. Dans un deuxième temps, les métadonnées ainsi que les thèmes, extraits de l'analyse du contenu des documents, sont utilisés comme dimensions pour construire le cube de texte. Dans Pérez et al. (2007), les auteurs proposent de contextualiser un entrepôt de données avec des documents. Ces derniers vont constituer un contexte pour l'analyse des données structurées. A cet effet deux dimensions sont proposées, une dimension contexte et une dimension pertinence pour construire le cube contextualisé par les documents textes, appelé *R-Cube*.

Dans tous ces travaux, deux technologies majeures, permettant la manipulation des données textuelles sont utilisées : la Recherche d'Information (RI) et la Fouille de Texte (Text mining). Nous pensons que celles-ci permettent d'apporter des solutions réelles permettant d'étendre la portée des systèmes décisionnels pour prendre en charge les données texte. Dans Park et Song (2011), les auteurs proposent d'allier à ces deux technologies celle de l'extraction d'information (EI) pour un système décisionnel complet prenant en charge, aussi bien les données structurées et non structurées.

Notre approche s'inscrit dans la tendance des travaux qui combinent OLAP et fouille de texte pour l'analyse en ligne de données textuelles. Nous pensons que des techniques telles que la catégorisation et le *clustering* permettent la *summarization* des données textuelles et offre des agrégats sémantiquement pertinents pour une meilleure prise de décision.

3 *Summarization* par catégorisation

Les technologies OLAP reposent sur des outils pour la visualisation, la structuration et l'exploration des cubes de données. Un cube de données est constitué d'un ensemble de cellules où chacune représente un sujet d'analyse, appelé *fait*. Ce dernier est observé par un ou plusieurs indicateurs d'analyse appelés *mesures*, selon plusieurs axes d'analyse, appelés *dimensions*. Ces dernières sont composées d'attributs et peuvent être agencées de manière hiérarchique. De telles hiérarchies permettent d'observer des indicateurs selon plusieurs niveaux de granularité et de construire des agrégats à partir des faits. Cette structure multidimensionnelle permet de construire des cubes qui supportent des opérations d'analyse en ligne (dit OLAP). Des opérateurs fournissent aux utilisateurs des moyens pour naviguer dans les données multidimensionnelles afin d'y découvrir des informations pertinentes. Pour intégrer les données textuelles dans les systèmes OLAP, trois aspects doivent être pris en charge. La spécification de *mesures* appropriées aux données textuelles, des *hiérarchies de dimensions* spécifiques aux documents textuels pour permettre des analyses selon différents niveaux de granularités ainsi que l'*agrégation* des données textuelles. Ces choix dépendent essentiellement du problème à analyser.

Dans le domaine des ressources humaines par exemple, les réponses des candidats à une offre d'emploi représentent une grande quantité d'information difficile à gérer rapidement et efficacement par les recruteurs. Il devient nécessaire de traiter cette masse d'information de façon automatique ou assistée pour une meilleure prise de décision. Par rapport à une offre d'emploi et sur la base d'un document CV du candidat, la démarche habituelle d'un recruteur consiste à faire une présélection des candidatures en les classant en trois catégories. Les candidatures pertinentes (classe A), celles susceptibles d'être pertinentes (classe B) et enfin celles qui sont non pertinentes (classe C). Ce classement se fait sur la base d'informations telles que : la localité du candidat, son niveau d'étude, sa spécialité, son expérience professionnelle, ainsi que les compétences recherchées.

Nous proposons au recruteur un modèle multidimensionnel. Ainsi l'indicateur à observer est une mesure textuelle, le CV. Trois axes d'analyses sont proposés, le lieu du poste, le niveau d'étude et le secteur d'activité. Par exemple, si le recruteur soumet une requête (Lieu = "Alger", secteur d'activité = "Informatique", niveau d'étude = "Ingénieur"). Un ensemble de documents est renvoyé dont le nombre peut être très grand. Il serait souhaitable de pouvoir fournir au recruteur un résumé de cet ensemble de documents sous forme de trois classes. La figure 1, illustre le principe de la *summarization* par catégorisation sur l'exemple des CVs.

L'idée de base de la *summarization* par catégorisation est d'exploiter la mesure textuelle contenue dans le cube pour présenter au décideur un résumé sous forme de classes.

D'autre part, et comme le CV est un document qui présente une structure conventionnelle, nous souhaitons offrir au recruteur la possibilité de visualiser tout ou partie de ce document à travers une dimension structure. Il peut ainsi mieux apprécier la pertinence du profil du candidat par le biais des parties de son CV, comme l'expérience professionnelle par

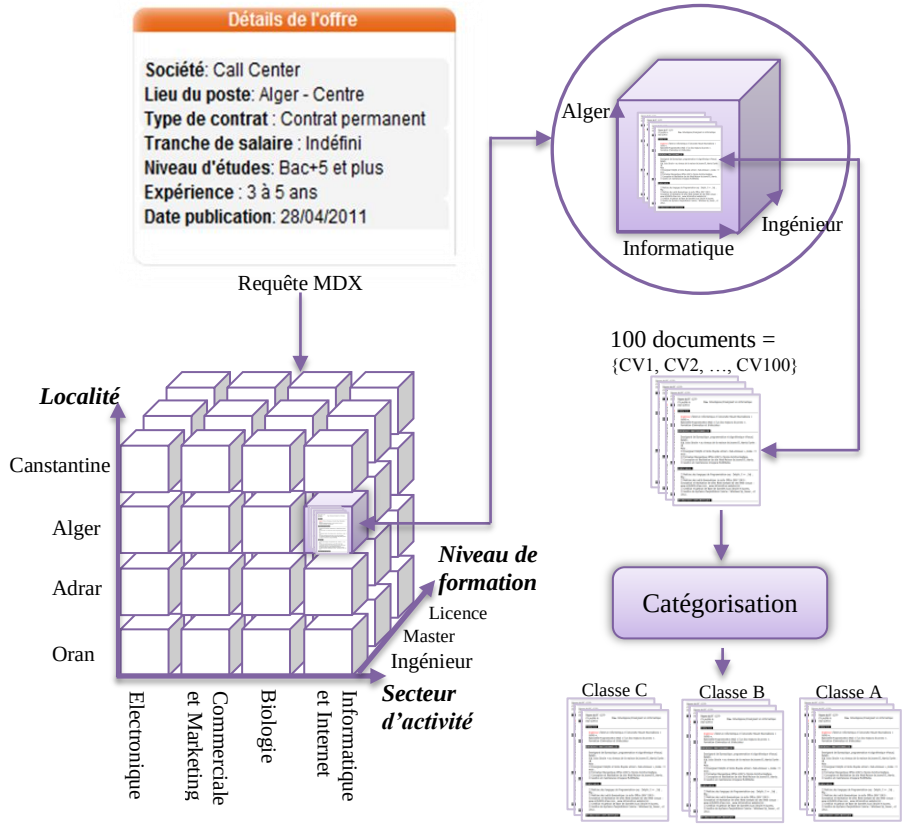


FIG. 1 – *Principe de la summarization par catégorisation appliqué aux CVs*

exemple. La section 4 présente la partie catégorisation. Des expérimentations ont été menées dans le but de voir l'influence des différentes sections du CV sur les résultats de la catégorisation.

4 Catégorisation des documents CVs

Le classement, appelé aussi catégorisation automatique de documents, est défini comme étant la tâche d'assigner un document texte à une ou plusieurs classes prédéfinies (Sebastiani (2002)). Cette tâche est à la croisée des chemins entre le domaine de la recherche d'information et celui de l'apprentissage automatique. En effet, un processus général inductif construit automatiquement un classifieur, en apprenant depuis un ensemble de documents pré classés, il s'agit des caractéristiques de la catégorie (Sebastiani (2002)). Plusieurs méthodes de catégorisation de texte existent, parmi lesquelles nous distinguons: les méthodes à base d'instances comme *k-NN* (*k Nearest Neighbors*) ou la méthode du plus proche voisin, les méthodes probabiliste comme le classifieur bayésien naïf, la méthode des SVM (*Support Vector Machine*), les arbres de décision, les réseaux de neurones, etc. (Sebastiani (2002)).

Il n'y a pas de méthode générique ayant donné la preuve de sa supériorité dans tous les cas de catégorisation de textes (Torres-Moreno (2007)). Néanmoins, la classe d'algorithmes basés sur la similarité, a montré qu'ils produisent de bonnes performances pour les données textuelles (Shankar et Karypis (2000)). Cette classe contient les algorithmes des classifieurs tels que *k-NN*, ou *"basé centroide"* et leurs variantes. Pour ces classifieurs, la classe d'un nouveau document est déterminée en calculant la similarité entre le document test et les instances individuel (comme dans le cas de *k-NN*) ou agrégats (comme dans le cas du *"basé centroide"*) de l'ensemble d'apprentissage. La classe est déterminée en se basant sur la distribution de la classe des plus proches instances ou agrégats.

Notre choix s'est porté sur une méthode simple, efficace et peu couteuse. C'est la méthode *"basé centroide"* dont les performances ont surpassé celles des algorithmes, *k-NN*, *Bayésien naïf* et *C4.5*, sur un large éventail de jeux de données (Han et Karypis (2000)). Avant de voir le modèle de classement utilisé dans la section 4.4., nous allons tout d'abord aborder la préparation des données dans la section 4.1. La section 4.2 illustrera la représentation choisie pour les documents CVs ; la mesure de similarité sera abordée dans la section 4.3.

4.1 Prétraitements

Dans un premier temps, nous avons effectué un prétraitement du contenu des documents CVs afin de filtrer les termes pertinents par élimination des segments de texte non pertinents. Les principales tâches effectuées concernent (1) la *Tokenisation* ou segmentation du texte ; elle consiste à parcourir le texte en vue de récupérer les termes et supprimer les caractères spéciaux et la ponctuation. (2) La *Normalisation* qui réalise la conversion du texte en minuscule. (3) L'*Elimination des mots vides***,** consiste à éliminer les mots outils de la langue en utilisant une liste de mots vides pour le français (ex : de, en, dans, etc.). Nous avons choisi également d'éliminer les accents. Car une partie considérable de la source des erreurs d'orthographe, réside dans l'oubli des accents. En effet, les mots « ingénieur » et « ingenieur » sont considérés comme étant deux termes différents, ce qui augmente la dimension de l'espace des caractéristiques. (4) La *Racinisation* ou *stemming*, consiste à trouver la racine des verbes fléchis et à ramener les mots pluriels et/ou féminins au masculin singulier[3].

4.2 La représentation des documents CVs

Nous avons choisi de représenter les documents CVs en adoptant l'approche «*bag of words*». L'unité textuelle choisie est donc le terme ou groupe de termes, en utilisant le modèle vectoriel (*Vector-Space Model* ou VSM). Dans ce dernier, chaque document d_j est un vecteur des poids des termes dans l'espace des termes.

$$\vec{d_j} = (w_{1j}, w_{2j}, w_{3j} \dots, w_{nj}) \quad (1)$$

où : w_{ij} est le poids du terme t_i dans le document d_j. Nous avons choisi le schéma de pondération : $tf.idf$ pour la pertinence du terme dans le document (Salton et Buckley (1988)) :

[3] Ainsi, les mots : développe, développé, développeront, développement, seront ramenés à la même forme, « développ ».

$$w_{ij} = tf_{ij} * idf_{ij} = tf_{i,j}.log\frac{N}{n_i} \quad (2)$$

où, tf_{ij} est la fréquence du terme t_i dans le document d_j ; idf_{ij} est l'inverse de la fréquence des documents dans lesquels le terme t_i apparaît ; N est le nombre total de documents dans le corpus ; n_i le nombre de documents dans lesquels apparaît le terme t_i. Pour pallier au problème de la longueur des documents et ne pas pénaliser les documents courts au dépend des documents longs, chaque vecteur document a été normalisé de sorte que sa gueur $\|\vec{d}_j\| = 1$. Après prétraitements, l'ensemble des CVs utilisés dans cette expérimentation ont été transformés en des vecteurs de poids des termes en utilisant le schéma de poids $tf.idf$. Ils ont été normalisés de sorte que leur longueur devienne unitaire.

4.3 Mesure de similarité

Dans le modèle vectoriel, la similarité entre deux documents d_i et d_j est communément mesurée en utilisant la fonction cosinus donnée par la formule (3).

$$\cos(\vec{d}_i, \vec{d}_j) = \frac{\vec{d}_i.\vec{d}_j}{\|\vec{d}_i\| * \|\vec{d}_j\|} \quad (3)$$

où $\vec{d}_i.\vec{d}_j$ est le produit vectoriel des vecteurs des documents d_i et d_j. Comme les vecteurs sont normalisés, la formule (3) devient :

$$\cos(\vec{d}_i, \vec{d}_j) = \vec{d}_i.\vec{d}_j \quad (3)'$$

4.4 Principe de l'algorithme *"basé centroide"*

L'algorithme *"basé centroide"* a montré des performances ayant surpassé celles des algorithmes, *k-NN*, *Bayésien naïf* et *C4.5*, sur un large éventail de jeux de données, d'après Han et Karypis (2000). Basé sur le modèle vectoriel, le principe de ce classifieur est le suivant :

- Pour chaque ensemble de documents appartenant à la même classe, un vecteur *centroide* est calculé ; il est appelé prototype de la classe. Si k classes existent dans la base d'apprentissage, k vecteurs *centroides* sont calculés $\{\vec{C}_1, \vec{C}_2, \vec{C}_3, \dots, \vec{C}_k\}$, où $\vec{C}_i$ est le *centroide* de la $i^{ème}$ classe. Le calcul du *centroide* est donné par la formule (4).
- Les similarités entre un nouveau document x et les k centroides sont calculées, en utilisant la mesure du cosinus. Le calcul de ces similarités est donné par formule (5).
- A partir de ces similarités, x est assigné à la classe dont le vecteur *centroide* est le plus similaire au vecteur document. La classe du document x est donnée par la formule (6).

Pour une classe i, étant donnée un ensemble S de documents et leurs représentation vectorielle, le vecteur *centroide* $\vec{C}_i$ est calculé par :

$$\vec{C}_i = \frac{1}{|S|}\sum_{d \in S} \vec{d} \quad (4)$$

où $|S|$ représente le nombre de documents de l'ensemble S.

La similarité entre le vecteur du document d et le *centroide* $\overrightarrow{C}$ d'une classe est calculée par la formule ci-dessous :

$$\cos(\vec{d}, \vec{C}) = \frac{\vec{d}.\vec{C}}{\|\vec{d}\| * \|\vec{C}\|} = \frac{\vec{d}.\vec{C}}{\|\vec{C}\|} \quad (5)$$

La classe d'un nouveau document x est définie par :

$$arg\ max_{j=1,\dots,k}(\cos(\vec{x}, \vec{C}_i)) \quad (6)$$

4.5 Expérimentations

Nous avons effectué nos expérimentations sur un ensemble de 141 CVs constitué pour la catégorie « Ingénieur en Informatique ». Chaque candidature est identifiée comme pertinente, pouvant être pertinente ou non pertinente par rapport à une offre d'emploi donnée. Le tableau 1 résume les informations sur le corpus utilisé et son étiquetage.

Nombre de CVs	Pertinence de la candidature		
	Oui	Peut être	Non
141	35	33	73

TAB. 1 - *Statistiques du corpus en fonction de l'étiquetage.*

4.5.1 Protocole expérimental

Afin d'éviter de faire l'expérimentation sur un seul ensemble d'apprentissage et un seul ensemble de test[4], nous avons choisi la stratégie de la validation croisée ou «*k-fold cross validation*» (Torres-moreno (2007)). Dans cette approche, les données sont réparties aléatoirement (mais de façon équilibrée) en *k* ensembles. *K* exécutions sont réalisées, où pour chacune un des *k* ensembles est utilisé pour la validation du modèle et les *(k-1)* ensembles restants sont utilisés pour l'apprentissage. Les valeurs de *k* les plus utilisées sont celles comprises entre 5 et 10.

Nous avons choisi *5-fold cross validation* ; les 141 CVs ont été scindés en cinq sous-ensembles approximativement de même taille $F_i; i = 1, \dots, 5$. Avec une répartition aléatoire mais équilibrée des candidatures dans chaque sous-ensemble. Cinq exécutions ont été réalisées $E_i; i = 1, \dots, 5$ où pour chaque exécution E_i, quatre des cinq sous-ensembles sont concaténés pour constituer l'ensemble d'apprentissage et le cinquième est utilisé pour la validation (ex : les sous-ensembles $F1, F2, F4, F5$ sont utilisés comme ensemble d'apprentissage et le sous-ensemble $F3$ pour le test). Nous avons utilisé la mesure $F - score$ (formule 7) des documents bien classés, moyennée sur toutes les classes.

$$F - score = \frac{2 * \langle Précision \rangle * \langle Rappel \rangle}{\langle Précision \rangle + \langle Rappel \rangle} \quad (7)$$

[4] Ce qui pourrait conduire au phénomène de sur-apprentissage ou apprentissage par cœur.

4.5.2 Résultats et discussion

Pour voir l'impact de la structure du CV sur les résultats de la catégorisation et identifier ainsi les parties du CV contenant les informations les plus pertinentes, nous avons mené trois expérimentations différentes. Dans la première expérimentation les parties formations, expériences professionnelles et compétences (*CV_FEC*) sont utilisées. Dans la deuxième expérimentation la partie expériences professionnelles (*CV_E*) est utilisée. Dans la troisième, les parties expériences professionnelles et compétences (*CV_EC*) sont utilisées. Chacune de ces expérimentations a été faite avec le protocole décrit dans la section précédente. Les résultats obtenus, par l'algorithme *"basé centroide"* sur l'ensemble des 141 CVs pour les trois expérimentations, sont résumés dans le tableau 2, ci-dessous :

	Précision			Rappel			F-score		
	CV_FEC	CV_E	CV_EC	CV_FEC	CV_E	CV_EC	CV_FEC	CV_E	CV_EC
Classe A	0,634	0,418	0,500	0,743	0,685	0,629	**0,684**	0,519	0,557
Classe B	0,392	0,297	0,422	0,267	0,276	0,395	0,318	0,286	**0,408**
Classe C	0,703	0,688	0,719	0,726	0,469	0,631	**0,714**	0,558	0,672
Toutes classes	0,576	0,468	0,547	0,579	0,477	0,552	**0,577**	0,472	0,549

TAB. 2 – *Précision, Rappel et F-score obtenus pour les trois expérimentations.*

Les résultats obtenus dépendent, d'une part de la qualité des données et de l'expertise humaine. D'autre part, d'un point de vue ressources humaines, une candidature appartenant à la classe B est une bonne candidature mais ne correspond pas forcément au profil recherché ; ou une bonne candidature mais pas la meilleure (Kessler et al. (2008)). Cette dernière affirmation rend difficile l'initialisation de cette classe. Néanmoins, la démarche adoptée dans cette section, montre qu'il est possible de récupérer les documents d'un *Text Cube*, en réponse à une requête multidimensionnelle pour appliquer la catégorisation de texte. Les classes obtenues représentent les valeurs de la mesure textuelle OLAP, le CV dans notre cas. Ces classes représentent aussi, pour le décideur, des agrégats en rapport avec la nature des données textuelles et la spécificité du domaine.

Par ailleurs, ces résultats montrent aussi que les meilleurs scores ont été obtenus en utilisant les parties formations, expériences professionnelles et compétences (l'expérimentation CV_FEC). Nous pouvons donc conclure que les informations pertinentes permettant de déterminer la candidature adaptée à une offre d'emploi sont contenues dans ces parties. La figure 2 donne les résultats graphiques du tableau 2.

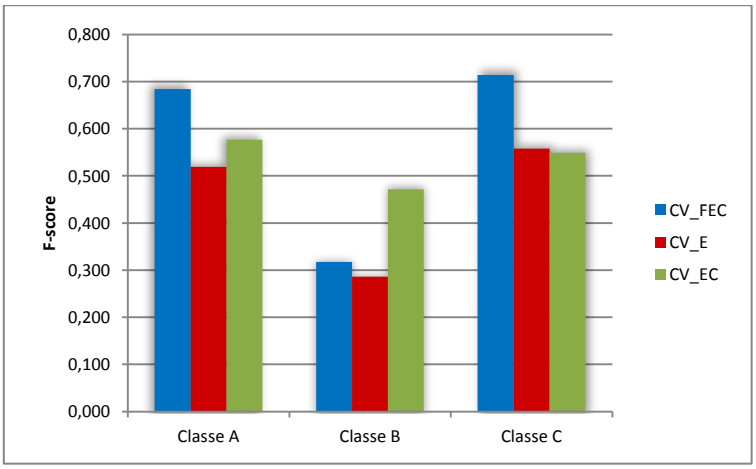


FIG. 2 – *F-score obtenus pour toutes les classes, en utilisant différentes parties du CV.*

4.5.3 Performances en temps d'exécution

En plus de sa simplicité, l'algorithme *"basé centroide"* a une complexité computationnelle linéaire. En effet, le temps nécessaire pour classer un nouveau document x est au plus $O(km)$, où m est le nombre de termes présents dans x et k le nombre de classes. Dans Shankar et Karypis (2000), les auteurs ont montré que *"basé centroide"* est de 10 à 20 plus rapide que le classifieur *SVM*. Cette efficacité computationnelle est très importante pour les systèmes OLAP qui nécessitent des temps de réponses très courts. Le tableau 3 donne les temps de calcul pour les trois expérimentations menées.

Expérimentations	**Taille de l'espace des caractéristiques**	**Temps de calcul (en secondes)**
CV_FEC	2747	6,79
CV_E	1761	3,26
CV_EC	2461	5,50

TAB. 3 - *Temps de calcul obtenus pour les trois expérimentations.*

Ces résultats ont été obtenus sur un PC portable, Intel Core™2 Duo 2.0 GHz et 3,00 Go de RAM.

5 Conclusion et perspectives

Dans cet article nous avons proposé la catégorisation de texte comme approche pour la *summarization* de documents dans les *Text cubes*. Notre objectif est de pouvoir offrir au décideur des agrégats significatifs pour une meilleure prise de décision. L'exemple concret que nous avons pris, celui de l'analyse et du traitement des candidatures dans le domaine des

ressources humaines, confirme la pertinence de nos propositions. Nous avons présenté les résultats de la catégorisation menée sur un corpus de documents CVs en utilisant un algorithme dont les performances computationnelle sont très efficaces.

Les perspectives associées à notre approche sont nombreuses. Tout d'abord, il y a lieu d'améliorer les performances de l'algorithme de catégorisation choisi. Nous comptons inclure une phase de sélection d'attributs en amont du processus de catégorisation, afin d'éliminer les termes non pertinents présents dans les documents CVs. Par ailleurs, Nous prévoyons de permettre au recruteur de ne visualiser que des parties du CV. En effet, le CV a une structure visible composée de parties telles que : formations, expériences, compétences et divers. Pour cela, nous prévoyons d'inclure dans notre modèle une dimension structure à travers laquelle le recruteur peut visualiser le contenu textuel du CV selon différentes granularités. Nous prévoyons aussi d'inclure un modèle permettant d'apprécier la qualité des agrégats (classes) obtenus.

Références

Boussaid, O., R. Ben Messaoud, R. Choquet, S. Anthoard (2006). X-Warehousing: An XML-Based Approach for Warehousing Complex Data. In *Proceedings of the 10th East-European Conference on Advances in Databases and Information Systems (ADBIS'06), LNCS 4152*, pp. 39–54.

Darmont, J., O. Boussaïd, J.C Ralaivao, K. Aouiche (2005). An architecture framework for complex data warehouses. *7th International Conference on Enterprise Information Systems (ICEIS)*,pp. 370-373.

Han, E.H., G. Karypis (2000). Centroid-based Document Classification: Analysis and Experimental Results. *Principles of Data Mining and Knowledge Discovery*, pp. 424–431.

Kessler, R., J.M. Torres-Moreno, M. El-Bèze (2008). E-Gen : Profilage automatique de candidatures. *Traitement Automatique des Langues Naturelles (TALN)*, pp. 370-379.

Lin, C.X., B. Ding, J. Han, F. Zhu, B. Zhao (2008). Text Cube: Computing IR Measures for Multidimensional Text Database Analysis. *International Conference on Data Mining (ICDM)*, pp. 905-910.

Manning, C.D., P. Raghavan, H. Schütze (2008). An Introduction to Information Retrieval. *Cambridge University Press*.

Park, B-K, I-Y Song (2011). Toward Total Business Intelligence Incorporating Structured and Unstructured Data. *BEWEB'11 Proceedings of the 2nd International Workshop on Business intelligencE and the WEB*.

Park, B-K., H. Han, I-Y Song (2005). XML-OLAP : A Multidimensional Analysis Framework for XML Warehouses. In *7th International Conference on Data Warehousing and Knowledge Discovery (DaWaK), LNCS 3589, Springer*, pp. 32–42.

Pérez, J.M., R. Berlanga, M.J Aramburu, T.B Pedersen (2007). R-cubes : OLAP cubes Contextualized with documents. *ICDE 2007, IEEE 23rd International Conference*, pp.1477-1478 .

Ravat, F., O. Teste, R. Tournier (2007). OLAP Aggregation Function for Textual Data Warehouse. In *9th International Conference on Entreprise Information System (ICEIS)*, pp. 151-156.

Ravat, F., O. Teste, R. Tournier, G. Zurfluh (2008). Top_Keyword: an Aggregation Function for Textual Document OLAP. *10th International Conference on Data Warehousing and Knowledge Discovery*, pp.55-64.

Salton, G., C. Buckley (1988). Term-weighting approaches in automatic text retrieval. *information processing and management. volume 24, N°5*.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM computing surveys, 34(1)*, pp.1-47.

Shankar, S., G. Karypis (2000). Weight adjustment schemes for a centroid based classifier. *Principles of Data Mining and Knowledge Discovery*.

Simitsis, A., A. Baid, Y. Sismanis, B. Reinwald (2008). Multidimensional Content eXploration. In *Proceedings of the VLDB Endowment. 1(1)*, pp. 660–671.

Sullivan, D. (2001). Document warehousing and text mining : Techniques for Improving Business Operations, Marketing, and Sales. *John Wiley & Sons*.

Torres-Moreno, J. M. (2007). Du textuel au numérique : analyse et classification Automatiques. *HDR, université d'Avignon*.

Tseng, F.S.C., A.Y.H. Chou (2006). The concept of document warehousing for multi-dimensional modeling of textual-based business intelligence. *Science Direct, Decision Support Systems 42* p. 727–744.

Zhang, D., C. Zhai, J. Han (2011). MiTexCube: MicroTextCluster Cube for Online analysis of Text Cells. *Conference on Intelligent Data Understanding (CIDU)*, pp. 204-218.

Zhang, D., C. Zhai, J. Hany, A. Srivastavaz, N. Ozaz (2009). Topic Modeling for OLAP on Multidimensional Text Databases: Topic Cube and its Applications. *Statistical Analysis and Data Mining*, pp. 378-395.

Summary

With the explosive growth of textual data as well in the organizations as on the Web, it becomes necessary to analyze not only structured data but unstructured text data as well. The OLAP (*On-Line Analytical Processing*) Processes have proved their efficiency in the structured data analysis. However, they are unsuited to the text data analysis. One of the challenges the OLAP processes must raise is the aggregation of the text data. In this paper we propose a summarization approach, based on text categorization. An evaluation of the approach has been carried out through an experimental study on a corpus of documents resulting from the field of human resources, the CVs.

Recherche d'information contextuelle par segmentation thématique de documents: Application au corpus 20 Newsgroups

Rachid Aknouche

Laboratoire ERIC, Université Lumière Lyon2
5 avenue Pierre Mendès-France, 69676 Bron Cedex, France

Rachid.Aknouche@univ-lyon2.fr

Résumé. L'accès à une information pertinente, adaptée aux besoins et au contexte de l'utilisateur, est un véritable défi pour la communauté de la recherche d'information (RI). Les systèmes de recherche d'information (SRI) classiques ne prennent pas en compte les facteurs contextuels dans le processus d'appariement et/ou dans la phase de classement. Ils s'intéressent plutôt aux métriques de pondération et aux formules de calcul de similarité entre les termes de la requête utilisateur et ceux des documents du corpus. Dans cet article, nous proposons une approche qui prend davantage en compte les paramètres contextuels utilisés dans le domaine de la RI et les centres d'intérêts de l'utilisateur obtenus par le processus de segmentation thématique des documents. Pour considérer les unités thématiques, représentées sous forme de fragments de texte, nous proposons lors de la phase de représentation et de classification des documents une adaptation de la formule *TF-IDF*. Nos expérimentations faites sur le corpus 20 Newsgroups montrent que notre approche améliore considérablement l'efficacité de la recherche par l'amélioration des taux de la précision par rapport aux SRI classiques qui ne considèrent pas les facteurs contextuels.

1 Introduction

Les systèmes de recherche d'informations classiques utilisent souvent des méthodes de pondération et des mesures de similarités pour retrouver des textes ou des extraits de textes pertinents par rapport à une requête utilisateur. La pondération *tf-idf*[1] est l'une des techniques les plus utilisées dans ces systèmes. Elle permet d'évaluer l'importance d'un terme dans un document par rapport à une collection ou à un corpus. Cependant cette formule, telle qu'elle est souvent présentée dans la littérature, ne tient pas compte du contexte de la recherche d'information (RI). On entend par contexte l'endroit d'apparition des termes recherchés dans un document. Il s'agit notamment des fragments de texte qui sont représentés sous forme de sections

1. désigne un ensemble de schémas de pondération de termes. *tf* signifie *(Term Frequency)* qui désigne le nombre d'occurrence du terme dans le document et *idf (Inverted Document Frequency)* qui est la valeur inverse du nombre de documents dans lesquels le terme est présent ou le pouvoir de discrimination de ce terme.

et de paragraphes dans un document. Par contre, une recherche pertinente devrait considérer ces éléments lors de la phase de reformulation de la requête, dans le processus d'appariement et/ou dans la phase de classement des résultats. Par ailleurs, la volumétrie des collections ne cesse d'accroître ; ce qui rend la phase de recherche lente et fastidieuse. Cette problématique est déjà connue dans la littérature et elle est à la base de la plupart des travaux réalisés par les chercheurs dans la communauté de la RI. Plusieurs approches ont donc été proposées pour faciliter la RI, pour augmenter son degré de pertinence et pour réduire son temps d'exécution. Elles utilisent de nouvelles techniques d'organisation et d'analyse de larges collections de documents pour accélérer le processus de recherche d'information. Ces approches sont souvent organisées autour des méthodes de classification, de catégorisation et de segmentation de documents. Ces termes ont des origines différentes et ils sont utilisés indifféremment dans les publications. Cependant, ils désignent des méthodes permettant de positionner et/ou classifier un document dans une structure, de l'étiqueter dans une, plusieurs ou aucune catégorie et enfin d'analyser le contenu des documents et de les regrouper au sein de différents thèmes extraits.

Parmi les approches proposées pour faciliter le processus de la recherche de l'information, on peut citer les premières techniques, dites traditionnelles, qui sont basées sur la représentation linéaire des documents. D'après (Zargayouna, 2004), ces techniques procèdent à des requêtes plates (recherche par mots clés) et ignorent, par conséquent, la structure du document. De leurs côtés, (Pinel-Sauvagnat et Boughanem, 2005) stipulent que la granularité des réponses renvoyées aux utilisateurs est restreinte au document tout entier. Or, un document possède souvent des contenus complexes et une recherche pertinente devrait, par contre, se faire au milieu des autres thèmes liés à ce document. D'autres approches plus complexes sont aussi apparues pour pallier aux problèmes liés à la recherche sémantique de l'information. Elles sont organisés autour d'une source de connaissances qui est souvent modélisée au moyen d'une ontologie pour approximer les concepts et les relations de la requête utilisateur.

De nombreux travaux proposent par contre des méthodes de classification et de catégorisation pour organiser les collections. Qu'elles soient supervisées ou non supervisées, ces méthodes permettent de réduire davantage le temps d'accès et de recherche dans ces collections. La classification supervisée suppose qu'il existe déjà une classification de documents et le but alors est de classer automatiquement un nouveau document. Les approches utilisées dans ce mode de classification sont : *K-plus proches voisins*, *Arbre de décision*, *naïve bayésienne*, *Machine à vecteur de support (SVM)*... La classification non-supervisée, quant à elle, est utilisée lorsque que l'on possède des documents qui ne sont pas classés et dont on ne connaît pas de classification. Les deux approches principales de cette méthode sont : les méthodes hiérarchiques et les méthodes non hiérarchiques. Parmi les travaux importants qui se sont intéressés à ces méthodes, on peut citer ceux de (Denoyer et al., 2003) qui combinent plusieurs fonctions d'affectation afin de classer des documents XML multimédia. Dans (Despeyroux et al., 2005) les auteurs identifient, pour une collection homogène donnée, les types d'éléments XML les plus pertinents pour un objectif de classification. Les travaux de (Doucet et Myka, 2002) et de (Yi et Sundaresan, 2000), quant à eux, étendent le modèle vectoriel pour définir la similarité entre documents. Ils utilisent pour cela, soit la technique *tf-idf* pour normaliser les éléments du vecteur soit l'algorithme de *K-means* pour la classification.

2 Contexte et motivations

Le but principal des systèmes de recherche d'informations classiques est de retrouver dans un corpus de documents l'information considérée comme pertinente pour une requête utilisateur. Cette pertinence, dans ces systèmes, est souvent liée à la fréquence d'apparition des mots dans le texte par rapport au corpus, mais sans pour autant tenir compte du contexte de la recherche et ni du fragment de texte où se trouve cette information.

Pour illustrer ce constat prenons l'exemple d'une collection de Curriculum Vitæ (CV) et considérons la requête "*Quels sont les candidats ayant le profil* ***ingénieur*** *?*" pour laquelle le système préférera ressortir tous les CVs dans lesquels le mot ***ingénieur*** (cf fig.2). Alors que, ce mot est parfois utilisé pour désigner une structure, le cas d'une école d'*ingénieurs* dans laquelle le candidat a déjà travaillé mais sans pour autant qu'il soit de ce profil. Il peut également indiquer un lieu s'il est utilisé dans la rubrique adresse du candidat exemple "Rue de l'***ingénieur*** Robert Keller". Le même constat est valable également pour la requête "*Quels sont les candidats ayant une expérience dans l'****enseignement*** *?*" et pour laquelle les résultats obtenus par les systèmes classiques concerneront aussi les candidats ayant retracés dans leur CV leur parcours scolaire dans la rubrique par exemple "***Enseignement*** supérieure".

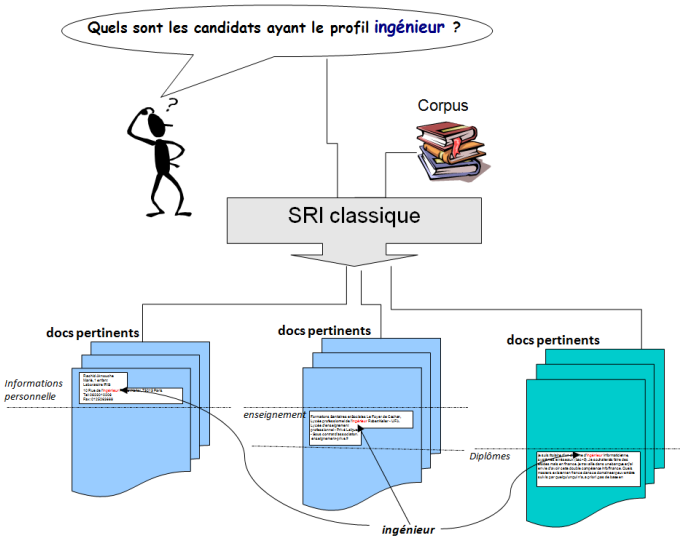


FIG. 1: Contexte et motivations

Les SRI classiques ne tiennent donc pas compte du contexte de la RI. Ils partent de l'hypothèse qu'un utilisateur a besoin rapidement d'une liste de réponses, quitte à passer du temps pour retrouver l'information pertinente. De ce fait, une recherche d'une information précise dans ces système devient une tâche complexe. Dans cet article nous détaillons de façon précise notre approche de recherche d'information contextuelle par segmentation thématique de documents , baptisée RISCH (une version plus synthétique a récemment fait l'objet d'une publication sous forme d'un papier court (Aknouche et al., 2012)). De plus, nous avons validé

notre approche sur a un nouveau corpus plus important "20 Newsgroups [2]" qui est constitué de 19997 documents, issus de 20 forums différents et décrits par 145980 descripteurs. Ce corpus est devenu une référence sur laquelle des techniques de Text Mining telles que la catégorisation ou la classification non supervisée sont testées et comparées. Sa caractéristique essentielle est son hétérogénéité en termes de taille des documents, en termes de thématiques et en termes de style.

L'approche RICSH est composée d'une phase de prétraitement de corpus visant l'identification et l'extraction des entités thématiques pour chaque document. Puis d'une phase de représentation et de classification des résultats obtenus lors d'exécution de la recherche. Cette approche s'intéresse, lors de la phase d'indexation du corpus, à la méthode de pondération *tf-idf* que nous avons adapté pour notre étude de cas. Le reste du document est organisé comme suit. Dans la section 2 nous présentons notre approche RICSH. La section 3 illustre la mise en oeuvre et l'efficacité de notre approche le corpus 20 Newsgroups. Enfin, la section 4 conclut ce papier et présente les perspectives de recherches associées.

3 Approche RICSH : Recherche d'information contextuelle par segmentation thématique de documents

L'approche RICSH comprend deux phases pour le processus de recherche d'information dans une collection de documents (cf fig.2) et chacune des phases est constituée de plusieurs étapes de traitement. La première phase est celle du prétraitement du corpus qui consiste, dans un premier temps, à extraire toutes les unités textuelles (mots) contenues dans ces documents, élaguer les mots fonctionnels [3] pour ensuite repérer les unités thématiques de chaque document. La deuxième phase, quant à elle, concerne le processus d'exécution de la recherche qui est basé sur les mécanismes de représentation des documents, sur les techniques utilisées pour la comparaison des documents par rapport à une requête utilisateur, et enfin sur la classification des résultats obtenus en des classes distinctes.

3.1 Architecture fonctionnelle de l'approche

3.1.1 Phase de prétraitement

C'est une phase de préparation et de nettoyage des données texte. Elle consiste d'abord à parser les documents pour en extraire les unités textuelles. Ensuite, celles-ci sont prétraitées selon une source de connaissances et enfin passées de leur forme linéaire à une représentation hiérarchique grâce au processus d'identification des différentes unités thématiques des documents. La source de connaissance que nous avons utilisé pour traiter la langue anglaise est la base de données lexicale Wordnet [4]. Par contre, pour ce qui des mots français, nous avons conçu notre propre thésaurus que nous avons alimenté à partir d'un dictionnaire.

2. http ://people.csail.mit.edu/jrennie/20Newsgroups/

3. ce sont les mots vide de sens ou *stopwords* en anglais qui n'intéressent pas la recherche d'information.

4. WordNet (Miller, 1995) est une base de données lexicale développée depuis 1985 par des linguistes du laboratoire des sciences cognitives de l'université de Princeton. C'est un réseau sémantique de la langue anglaise, qui se fonde sur une théorie psychologique du langage. La première version diffusée remonte à juin 1991.

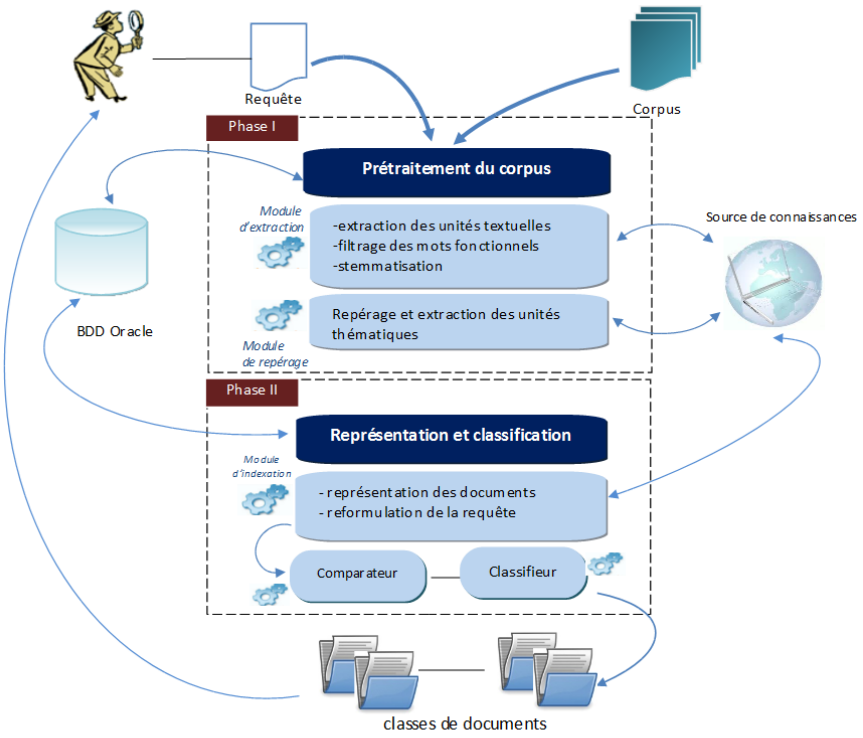


FIG. 2: Architecture fonctionnelle de l'approche RICSH

Extraction des structures Les documents peuvent se présenter sous des formats variés (Doc, Pdf, Xml, Html...), dans plusieurs langues et peuvent souvent contenir des images. Dans notre démarche ce n'est pas toute l'information structurelle de ces documents qui est utilisée dans la classification. Mais, on s'intéresse plutôt aux données texte que véhiculent ces documents écrits en anglais ou en français.

Filtrage des mots fonctionnels C'est un traitement qui permet d'éliminer les mots vides de sens et qui n'intéressent surtout pas la RI. On peut citer l'exemple des propositions, des pronoms, certains adverbes ou même adjectifs. Certains d'entre eux n'apparaissent pas souvent dans les documents et donc ne seront pas éliminés par la formule de calcul de la valeur de discrimination où même par la pondération *tf-idf*. Ainsi, selon le domaine de l'application, il est créé une liste appelée *Stoplist* (*Stopwords*) ou anti-dictionnaire, qui regroupe tous les mots vides de sens et ceux qui sont peu importants.Pour une requête utilisateur, la fréquence élevée d'apparition de ces mots dans plusieurs documents de la collection ne permet pas de discriminer ni de partitionner les textes pertinents de ceux non pertinents. De ce fait, le traitement réservé a une *Stoplist* consiste à ignorer tout mot du document existant dans cette liste.

Stemmatisation ou racinisation Après les mots vides de sens, il y a lieu de citer ceux qui ont des formes légèrement différentes, mais qui gardent souvent leur sens ou leur similarité. C'est le cas des mots conjugués par exemple. Cette différence de forme n'intéresse pas autant le domaine de la RI du fait que le résultat pour une requête de recherche doit retourner tous les documents traitant les mots de la même famille. La méthode généralement utilisée pour aboutir à cette mise en forme est d'éliminer les terminaisons de mots pour les faire ramener à leur forme canonique appelée *lemme* ou *stemme* (racine ou parfois radical). Ce traitement est basé sur un dictionnaire de suffixes qui permet d'extraire le radical du mot grâce à l'étude morphologique des mots. Dans notre approche nous avons utilisé l'algorithme de Porter pour l'anglais et l'algorithme proposé dans le projet *CLEF* pour le français (Rizoiu et al., 2010)

Identification et extraction des unités thématiques Il s'agira de faire collaborer des outils de TAL (Traitement Automatique de la Langue) traditionnellement utilisés en RI avec la prise en compte des indicateurs lexicaux définis au préalable pour le découpage d'un document en unités thématiques. *Une unité thématique désigne les fragments de nature textuelle d'un document faisant référence à un seul thème*. Notre démarche s'appuie sur un thésaurus qui englobe l'ensemble des indicateurs lexicaux jugés importants pour ce processus de découpage. Ils sont crées, dans le cadre de cette étude, d'une façon manuelle, par contre pour élargir la couverture des unités thématiques dans un document, nous avons rajouté les synonymes et les différentes formes lexicales que peuvent avoir ces unités dans le corpus. Les fragments de texte obtenus sont ensuite regroupés par unité thématique. La RI liée au contexte d'apparition des termes s'effectuera, par contre, par rapport à ces fragments. Cette façon de faire permettrait un gain en temps d'exécution et une amélioration dans la pertinence des résultats.

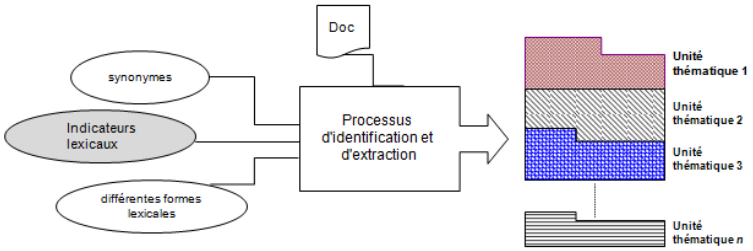


FIG. 3: Identification et extraction des unités thématiques

3.1.2 Phase de de représentation et de classification des documents

Phase d'indexation Les *stemmes* obtenus lors de la phase de prétraitement constitueraient des éléments essentiels pour indexer le contenu de notre corpus. Ils sont représentés comme des vecteurs dans un espace vectoriel à n dimensions où n est le nombre total de *stemmes*. La matrice générée dans notre étude de cas permet de ressortir une relation "*stemme-fragments*". En effet, les lignes de cette matrice représentent les fragments par unité thématique et les colonnes désignent les *stemmes* extraits de ces fragments. Chaque valeur de la matrice détermine la pertinence du terme par rapport au fragment de texte. Elle est obtenue par l'adaptation de la

formule *tf-idf*. Les formules *tf* et *idf* les plus utilisées se résument comme suit (Jeh et Widom, 2002) : *tf* = fréquence d'occurrences du terme dans un document *f(t, d)* ;
tf = f(t,d)/Max[f(t,d)] ou *Max[f(t,d)]* : fréquence maximale des termes dans *d* ;
tf = log(f(t, d)) ; *tf = log(f(t,d) + 1)* ;
idf = log(N/n) où *N* est le nombre de documents, et *n* ceux qui contient le terme.

Adaptation de la mesure *tf-idf* pour l'indexation de corpus Pour considérer les unités thématiques lors de la phase d'indexation du corpus, nous calculons la fréquence d'occurrences $TF_{(t,frag_i)}$ d'un terme t par rapport au fragment i de l'unité thématique (U_i). Cependant, la formule *tf-idf* adaptée à ce besoin combinera les deux critères : l'importance du terme dans un fragment de texte et son pouvoir de discrimination par rapport à l'ensemble des fragments. Ainsi, un terme d'une valeur de *tf-idf* adaptée élevée est considéré important dans le fragment, et il doit peu apparaitre dans les autres fragments. La valeur de fréquence d'un terme est :

$$\boldsymbol{TF}_{(t,frag_i)} = \frac{\boldsymbol{F_{t,frag_i}}}{\boldsymbol{N_{frag_i}}}$$

$F_{t,frag_i}$: désigne la fréquence d'apparition du terme t dans le fragment $frag_i$
N_{frag_i} : le nombre total des termes existants dans le fragment $frag_i$

Le calcul de la valeur de discrimination d'un terme par rapport aux fragments est :

$$\boldsymbol{IDF}_{(t,U_i)} = \boldsymbol{log}\ (\frac{\boldsymbol{N}_{U_i}}{\boldsymbol{n}_{t,U_i}})$$

N_{U_i} : le nombre total de fragments dans l'unité thématique i
n_{t,U_i} : le nombre de fragments dans l'unité thématique i contenant le terme t

Ainsi, nous obtenons la formule suivante pour chaque terme contenu dans le fragment :

$$\boldsymbol{TF_{(t,frag_i)}} \cdot \boldsymbol{IDF_{(t,U_i)}} = \frac{\boldsymbol{F_{t,frag_i}}}{\boldsymbol{N_{frag_i}}} \cdot \boldsymbol{log}\ (\frac{\boldsymbol{N}_{U_i}}{\boldsymbol{n}_{t,U_i}})$$

Cette phase d'indexation permettrait ainsi de représenter dans un espace vectoriel à *n* dimensions la pondération des *stemmes* par rapport aux fragments de l'unité thématique. Pour chaque stemme t du fragment i de l'unité thématique j, désignée par $frag_i^{U_j}$, on calcule sa valeur de pondération $TF_{(t,frag_i)}.IDF_{(t,U_j)}$ présentée dans les tableaux suivants par $V_{(t_i,frag_i)}$.

Unité thématique : U_1				
Fragments	t_1	t_2	...	t_n
$frag_1^{U_1}$	$v_{1,1}$	$v_{1,2}$	...	$v_{1,n}$
:	:	:	...	:
$frag_n^{U_1}$	$v_{n,1}$	$v_{n,2}$	...	$v_{n,n}$

TAB. 1: Indexation de l'unité thématique U_1

....

Unité thématique : U_n				
fragments	t'_1	t'_2	...	t'_n
$frag_1^{U_n}$	$v'_{1,1}$	$v'_{1,2}$	...	$v'_{1,n}$
:	:	:	...	:
$frag_n^{U_n}$	$v'_{n,1}$	$v'_{n,2}$	...	$v'_{n,n}$

TAB. 2: Indexation de l'unité thématique U_n

Phase de reformulation de la requête L'étape de reformulation de la requête consiste à enrichir la requête utilisateur avec des informations liées au contexte de la recherche avant le lancement du processus d'appariement et d'indexation. Les informations du contexte sont

fournies grâce : (1) au processus d'expansion de la requête généré en fonction des synonymes en communs des termes de la requête initiale ; (2) au choix de l'utilisateur de l'unité thématique correspondant à ses critères de recherche. Pour traiter la synonymie, nous avons intégré dans *RICSH* un module permettant de retrouver, pour chaque mot de la requête utilisateur, la liste des synonymes lui correspondant dans le dictionnaire et de ne garder que ceux qui sont en communs. Nous utilisons *WordNet* pour retrouver les synonymes des mots en anglais. Par contre pour les mots en français nous avons conçu un thésaurus des synonymes qui regroupe 36200 mots recensés dans le dictionnaire de la langue française. La requête est ensuite racinisée à son tour, pour qu'elle soit représentée par un vecteur de stemmes selon la composition de la matrice générée pour le corpus lors de l'indexation. Cette opération permet d'augmenter davantage les chances de retrouver soit les documents qui correspondent le mieux aux mots de la requête utilisateur, soit ceux qui se rapprochent le plus sémantiquement de ces critères.

Phase de comparaison Une fois la matrice construite, la similitude entre fragments et requête peut être calculée selon différentes méthodes. Dans *RICSH*, nous utilisons une métrique à base de cosinus. Cette mesure a l'avantage d'être simple et d'avoir de bonnes performances. Elle permet de calculer l'angle entre deux vecteurs (**?**). La valeur du cosinus est normée (entre 0 et 1). Si le cosinus tend vers 1 alors les deux documents sont proches, sinon s'il tend vers 0 alors ils sont éloignés. Un document est représenté par un vecteur $\vec{d} = (t_1, t_2, ...t_i, ...t_n)$, où $t_i \in [0, 1]$ est le poids d'un terme *i* dans le document. Une requête est également représentée par un vecteur $\vec{q} = (q_1, q_2, ...q_i, ...q_n)$, où $q_i \in [0, 1]$ est le poids du terme *i* dans la requête.

La fonction de correspondance mesure donc la similarité entre le vecteur requête et les vecteurs correspondant aux fragments de texte. Elle est définie comme suit :

$$\mathbf{Cos}(\vec{\mathbf{d}}, \vec{\mathbf{q}}) = \frac{\vec{\mathbf{d}}.\vec{\mathbf{q}}}{\| \vec{\mathbf{d}} \| . \| \vec{\mathbf{q}} \|} = \frac{\sum_{i=1}^{n} \mathbf{t}_i.\mathbf{q}_i}{\sqrt{\sum_{i=1}^{n} \mathbf{t}_i^2 + \sum_{i=1}^{n} \mathbf{q}_i^2}}$$

4 Expérimentations

4.1 Implémentation de l'approche RICSH

Pour apprécier les résultats de notre approche RICSH, nous avons implémenté en Java, sous l'environnement Eclipse, des programmes illustrant chaque étape de notre démarche. Notre prototype s'appuie sur JWNL [5] (Java WordNet Library), une API permettant un accès facile au thésaurus WordNet pour retrouver les synonymes d'un mot et sur SAX (Simple API for XML) pour parser un document XML. Pour faciliter le processus d'évaluation des performances de notre système, nous avons développé une interface permettant à l'utilisateur d'introduire ses requêtes, de combiner ses critères de recherche et enfin d'afficher les résultats selon leur degrés de pertinence par rapport à sa requête. Nous avons également implémenté sous le SGBDR Oracle une base de données qui héberge à la fois le thésaurus des synonymes de mots de la langue française, les unités thématiques générées lors du processus d'extraction et les scores de pertinence obtenus après la phase de comparaison.

5. http ://jwordnet.sourceforge.net/handbook.html

4.2 20 Newsgroups Corpus

Le 20 Newsgroups est une collection d'environ 19997 documents. Il s'agit d'échanges entre personnes dans le cadre d'un forum de discussions autour de 20 thématiques différentes (groupes de discussions). Il est devenu l'un des corpus les plus populaires pour évaluer les techniques de classification et/ou de catégorisation des textes. Sa caractéristique essentielle est son hétérogénéité en termes de taille des documents, en termes de thématiques et en termes de style. Le tableau 3 montre la liste des thématiques du 20 Newsgroups (sport, religion, politique, musique...).

comp.graphics	rec.sport.baseball	talk.politics.misc	comp.os.ms-windows.misc
rec.sport.hockey	talk.politics.guns	misc.forsale	comp.sys.ibm.pc.hardware
talk.politics.mideast	sci.crypt	talk.religion.misc	comp.sys.mac.hardware
comp.windows.x	sci.electronics	alt.atheism	rec.autos
sci.med	soc.religion.christia	rec.motorcycles	sci.space

TAB. 3: la liste des 20 Newsgroups

Pour effectuer la segmentation thématique des documents, nous avons d'abord prétraité les documents en supprimant les mots outils (stopswords) et toutes les en-têtes des articles de groupes de discussions (Usenet). Nous avons ensuite représenté dans un espace vectoriel chaque mot par son radical (stemme ou stems) grâce au processus de stemming. Les résultats de cette phase sont présentés dans le tableau 4.

corpus	**docs**	**stems**	**corpus**	**docs**	**stems**
20 newsgroups	20017	192375	alt.atheism	1001	15618
comp.graphics	1001	17731	comp.os.ms-windows.misc	1001	54511
comp.sys.ibm.pc.hardware	1001	16575	comp.sys.mac.hardware	1001	15011
comp.windows.x	1001	24915	misc.forsale	1001	17518
rec.autos	1001	15415	rec.motorcycles	1001	15108
ec.sport.baseball	1001	14000	rec.sport.hockey	1001	15610
sci.crypt	1001	17436	sci.electronics	1001	15622
sci.med	1001	19963	sci.space	1001	18432
soc.religion.christian	1001	13915	talk.politics.guns	1001	20258
talk.politics.mideast	1001	20546	talk.politics.misc	1001	17782

TAB. 4: Prétraitement du corpus

4.3 Évaluation des performances et résultats

Pour tester la performance et la fiabilité des résultats obtenus, nous avons effectué des expérimentations sur un ordinateur, Intel Core i5 de 2,8 GHz avec 3 Go de RAM. Ces expérimentations comparent les résultats obtenus par un système classique de la RI sans prise en compte des unités thématiques avec ceux obtenus après leur repérage et leur extraction.

4.3.1 Baseline

Pour évaluer la pertinence des résultats fournis par notre système RICSH, nous avons utilisé les mesures Rappel et Précision habituelles dans l'évaluation des systèmes de la RI. Pour calculer ces métriques nous avons utilisé le Toolkit Lemur [6], qui est un standard pour mener des expériences en recherche d'information. A l'aide de son outil d'indexation *Lemur's indexer* (cf. fig.4), un index des documents est créée pour chaque thématique du 20 Newgroup. Pour évaluer la pertinence de la recherche sur un SRI classique et sur notre approche, nous avons pris 20 requêtes, variant de 1 à 7 termes comme base de test. Les résultats de la recherche dans un système classique sont obtenues grâce au moteur de recherche *Lemur's retrieval* intégré dans le toolkit. En effet, il permet de retourner pour chaque requête une liste de documents classés par ordre de pertinence. La figure 5 montre, par exemple, le classement des documents pour la requête "athlétisme jouer".

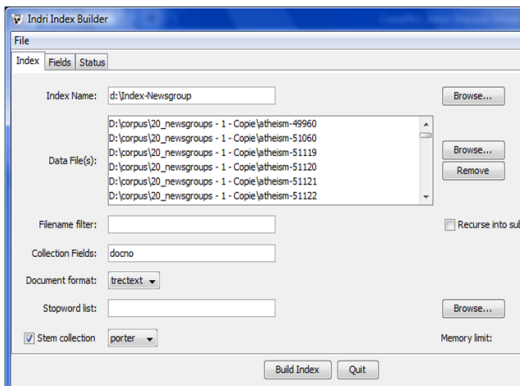


FIG. 4: *Lemur's indexer*

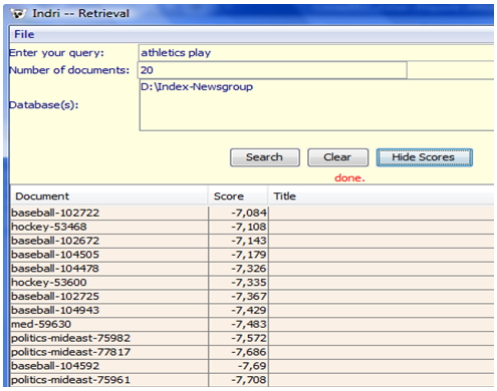


FIG. 5: *Lemur's retrieval*

4.3.2 Résultats

Nous avons utilisé l'outil d'évaluation TREC [7] pour calculer les valeurs de précision et de rappel (le script Perl "*ireval.pl*" livré avec le toolkit Lemur pour interpréter les résultats du programme "*trec-eval*"). La figure 6 montre les courbes de ces métriques obtenues par un système de recherche classique et par notre approche.

Les résultats de notre approche montrent une nette amélioration dans les mesures de rappel et de précision par rapport à ceux des méthodes classiques. Cette amélioration réside particulièrement dans les taux de précision. Elle est obtenue grâce à la nouvelle technique d'indexation du corpus issue de l'adaptation de la mesure tf-idf et au processus d'expansion de la requête généré en fonction des synonymes en communs des termes de requêtes. Ces résultats expliquent notamment notre première motivation de considérer le contexte de la recherche dans notre approche RICSH.

6. http ://www.lemurproject.com
7. http ://trec.nist.gov/

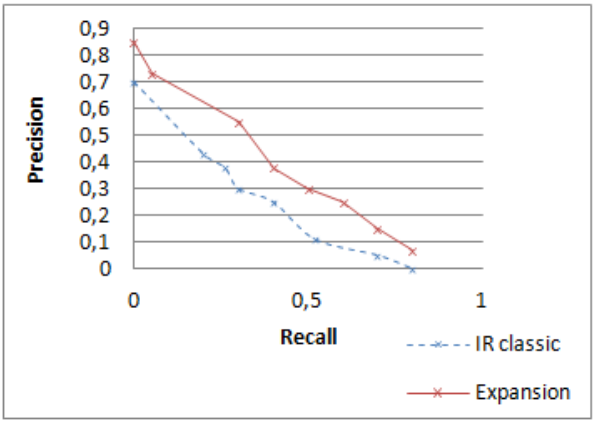


FIG. 6: Courbes de précision et rappel pour 20 Newsgroups

5 Conclusion

Dans cet article, nous avons présenté notre approche RICSH (recherche d'information contextuelle par segmentation thématique de documents) qui prend en compte deux paramètres contextuels pour améliorer le processus de recherche d'information : (1) Le contexte de l'utilisateur qui est assimilé à tous les facteurs qui peuvent décrire ses intentions et/ou son domaine d'intérêt. (2) Le contexte de la requête qui comprend, quant à lui, les connaissances linguistiques et sémantiques de la requête pour une meilleure compréhension des besoins de l'utilisateur. Notre approche repose sur huit étapes de traitement regroupées dans deux phases principales : Une de prétraitement du corpus et une de représentation et de classification des documents. L'originalité de notre démarche est donc double. D'une part, nous avons défini des techniques de repérage et d'extraction des unités thématiques des documents. D'autre part, nous avons adapté la mesure *tf-idf* pour évaluer la pertinence d'apparition des termes dans ces documents. Les travaux d'expérimentation ont été réalisés sur le corpus 20 Newsgroups. Les résultats que nous avons obtenus montrent que notre approche améliore considérablement la pertinence de la recherche par rapport aux SRI classiques.Dans les travaux futurs, nous comptons approfondir notre approche par l'introduction d'autres métriques telle que la divergence de Kullback-Leibler prenant davantage en compte la notion de contexte dans la RI. Un deuxieme objectif serait d'intégrer les techniques développées dans notre approche avec celles utilisées par les systèmes décisionnels *OLAP* (*On Line Analytical Processing*).

Références

Aknouche, R., O. Boussaid, et F. Bentayeb (2012). RICSH : Recherche d'information contextuelle par segmentation thématique de documents. In *Actes de la 12e Conférence Internationale Francophone (EGC 2012)*, Bordeaux, France, pp. 327.

Denoyer, L., J.-N. Vittaut, P. Gallinari, S. Brunessaux, et S. Brunessaux (2003). Structured multimedia document classification. In *Proceedings of the 2003 ACM symposium on Document engineering*, DocEng '03, New York, NY, USA, pp. 153–160. ACM.

Despeyroux, T., Y. Lechevallier, B. Trousse, et A.-M. Vercoustre (2005). Experiments in clustering homogeneous xml documents to validate an existing typology. *CoRR abs/cs/0507024*.

Doucet, A. et H. A. Myka (2002). Naïve clustering of a large XML document collection. In *INEX*, pp. 81–87.

Jeh, G. et J. Widom (2002). Simrank : A measure of structural-context similarity. In *In KDD*, pp. 538–543.

Pinel-Sauvagnat, K. et M. Boughanem (2005). A la recherche de noeuds informatifs dans des corpus de documents XML. In *Conférence francophone en Recherche d'Information et Applications (CORIA), Grenoble, 09/03/2005-11/03/2005*, pp. 119–134. IMAG.

Rizoiu, M.-A., J. Velcin, et J.-H. Chauchat (2010). Regrouper les données textuelles et nommer les groupes à l'aide des classes recouvrantes. In *10ème Conférence (EGC 2010), Hammamet, Tunisie*, Volume E-19 of *Revue des Nouvelles Technologies de l'Information*.

Yi, J. et N. Sundaresan (2000). A classifier for semi-structured documents. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '00, New York, NY, USA, pp. 340–344. ACM.

Zargayouna, H. (2004). Contexte et sémantique pour une indexation de documents semi-structurés. In *CORIA*, pp. 161–178.

Summary

Access to relevant information adapted to the needs and the context of the user is a real challenge in information retrieval systems IRS. The user context can be assimilated to all factors that can describe his intentions and perceptions of his surroundings. The classic information retrieval systems do not take into account these contextual factors in the similarity function or in the results ranking. They are often interested to calculate the similarity metric weights between the user query and the documents of the corpus based on the terms occurrence frequency in the text and ignore the user context. In this paper, we propose an approach that takes into account the contextual parameters in information retrieval IR and the user's domain of interest obtained by the thematic segmentation of the documents. To consider the thematic units, which are represented as text fragments, we propose, in the representation and classification of documents phases, an adaptation of the TF-IDF formula. Our experiments on the 20 Newsgroups corpus show that our approach improves significantly the retrieval effectiveness by improving the precision rate compared to the classic information retrieval systems, which do not consider contextual factors.

Unification de DTDs : Une étape vers la construction d'entrepôts de documents XML

Haithem Aouabed*, Ines Ben Messaoud*, Jamel Feki*
Gilles Zurfluh**

* Université de Sfax, Laboratoire Mir@cl, Faculté des Sciences Economiques et de Gestion,
Route de l'Aéroport Km 4, 3018 Sfax, BP. 1088 - Tunisie
haithem.abdi@gmail.com, {ines.benmessaoud ; jamel.feki}@fsegs.rnu.tn
** Université Toulouse 1, IRIT, Institut de Recherche en Informatique de Toulouse,
2 Rue du Doyen Gabriel Marty, 31042 Toulouse Cedex 9 – France
zurfluh@univ-tlse1.fr

Résumé. Les documents constituent une source capitale d'information pour les analyses décisionnelles. Ils aident les décideurs à mieux comprendre et expliquer certaines activités de leur organisation. Généralement, ces documents existent en format XML et sont structurellement hétérogènes. En conséquence, leur exploitation est délicate par un décideur. Cet article présente une méthode d'unification des structures des documents XML appartenant à un même domaine, et un outil la supportant. La méthode vise à produire une vue globale et générique de ces documents afin de faciliter leur interrogation dans un processus de prise de décisions. Elle comporte trois étapes : i) Représentation des structures des documents XML sous forme d'arbres, ii) Unification de ces arbres, et iii) Validation des arbres résultats.

1 Introduction

Afin d'améliorer le processus de prise de décisions, les décideurs exploitent des systèmes d'information décisionnels (SID) dont l'entrepôt de données (ED) constitue le noyau dur. En effet, ces ED permettent l'analyse d'énormes quantités de données numériques en un temps opportun. Néanmoins, avec l'avènement des nouvelles technologies de communication et plus précisément l'ouverture des organisations sur Internet, les documents représentent une capitalisation importante des connaissances du système d'information de production (SI) et deviennent de plus en plus utiles pour le système de pilotage Tournier (2007). En fait, le texte représente le moyen le plus répandu pour exprimer l'information et les connaissances sous-jacantes. Des études récentes comme celles de Tseng et Chou (2006) affirment que seuls 20% des données d'un SI sont transactionnelles et peuvent être traitées par un système de traitements analytiques en ligne « *OLAP : On Line Analytical Processing* », alors que les 80% restantes sont constituées de données non numériques (i.e., des documents).

Cependant, ces documents ont des contenus souvent peu structurés ; en conséquence, il est difficile de les intégrer dans les SIDs. Face à cette situation, les analystes-décideurs n'arrivent pas à explorer facilement, rapidement et efficacement ces documents ce qui risque de conduire

à des décisions contestables Tseng et Chou (2006). De ce fait, nombreux chercheurs recommandent d'entreposer les documents McCabe et al. (2000) Sullivan (2001).

Nous nous intéressons particulièrement aux documents XML « *Extensible Markup Language* » du fait que ce format est le plus utilisé pour la représentation et l'échange de données sur le Web W3C-XML (2008). De plus, ils sont produits conformément à une grammaire exprimée sous forme de DTD « *Document Type Definition* » ou de XSchema « *XML Schema Document* ». Toutefois, ces grammaires sont généralement hétérogènes même pour les documents d'un même domaine et d'une même organisation. Par conséquent, l'exploitation de ces documents nécessite une étape d'unification de leur structure. Elle génère une structure unifiée jouant le rôle d'un schéma global facilitant l'interrogation des documents. Cette étape d'unification est suivie par une étape de modélisation permettant la traduction de la structure unifiée sous forme d'un schéma multidimensionnel qui sera utilisé plus tard dans l'interrogation. Dans Ben Messaoud et al. (2010), les auteurs proposent une méthode de construction d'un entrepôt de documents XML, comportant deux processus : *unification des structures des documents XML* Ben Messaoud et al. (2011a) et *modélisation multidimensionnel* Ben Messaoud et al. (2011b). Dans cet article, nous présentons l'outil qui supporte le processus d'unification des structures des documents XML.

Cet article est structuré comme suit : la section 2 étudie les travaux les plus pertinents relatifs à l'unification des structures des documents XML. La section 3 donne un aperçu de la méthode d'unification des structures des documents XML. Quant à la section 4, elle présente notre outil appelé USD « *Unification of Structures of XML Documents* » qui implante la méthode d'unification. Finalement, la section 5 synthétise ce travail et envisage ses extensions futures.

2 Etat de l'art

Le format XML joue un rôle important dans l'échange de données sur le Web W3C-XML (2008). Il est caractérisé par une grande flexibilité lui permettant le stockage de données de diverses natures. Il existe deux types de documents XML : *Document XML orienté données* et *Document XML orienté documents*. Le premier type est très structuré tandis que le deuxième est principalement composé de textes et est décrit par des structures hétérogènes. Deux formalismes permettent de décrire la structure d'un document XML : la *DTD* et le *XSchema*. Ces deux formalismes jouent le même rôle. Toutefois, ils diffèrent sur quelques aspects. Par exemple, un XSchema peut définir des domaines de validité pour la valeur d'un champ, alors que ce n'est pas le cas pour une DTD. Généralement, les structures des documents XML sont hétérogènes même pour un ensemble de documents appartenant à un même domaine et à une même organisation; d'où la nécessité de leur déterminer une description unifiée.

Dans cette section, nous présentons les travaux les plus pertinents relatifs à l'unification des deux structures des documents XML : DTD et XSchema.

Dans Lee et al. (2002), les auteurs proposent une méthode d'intégration des DTDs appelée « *XClust* ». Initialement, les degrés de similarité entre les DTDs sont calculés en tenant compte des informations linguistiques, structurelles et sémantiques. Pour gérer les acronymes existants dans différentes DTDs, les auteurs utilisent une table spécifique contenant des acronymes et leurs formes complètes. Aussi, ils interrogent *Wordnet* pour déterminer les synonymes des balises. Ensuite, les clusters de DTDs similaires sont formés en se basant sur les degrés de similarité calculés précédemment. Finalement, une DTD est générée à partir de

chaque cluster de DTDs. Nous constatons que, malgré que la DTD générée donne une description globale du cluster, cependant, les auteurs ne détaillent pas les règles d'intégration qui permettent de générer ces DTDs résultats.

Par ailleurs, Yoo et al. (2005) proposent un algorithme pour l'unification des DTDs des documents XML appartenant à un même domaine. Leur algorithme reçoit en entrée un ensemble de DTDs ayant des structures similaires et génère une DTD unifiée. Une telle DTD joue le rôle d'un schéma conceptuel global dans le contexte d'un domaine donné. L'algorithme proposé comporte quatre étapes : *Prétraitement*, *Représentation des DTDs* sous forme d'arbres et d'automates finis, *Génération d'une DTD unifiée*, et *Post-traitement*. Tout d'abord, le prétraitement résout les ambiguïtés des noms des éléments des DTDs. Ensuite, les DTDs sont représentées sous forme d'arbres et d'automates finis. Puis, ces deux structures sont fusionnées pour créer une DTD unifiée. Enfin, la DTD résultat est syntaxiquement vérifiée en utilisant un parseur de DTD. Néanmoins, lors de l'étape de prétraitement les auteurs utilisent une table nommée « *Element Name Resolution Table* » pour résoudre les ambigüités des noms des éléments des DTDs. En fait, cette table doit être soigneusement préparée à l'avance, son contenu est dépendant du domaine comme il risque d'être incomplet; ces insuffisances pourraient affecter la qualité du résultat d'unification.

Quant aux travaux de Zhang et Liu (2002), les auteurs présentent un processus pour intégrer des XSchema. Ce processus génère un modèle conceptuel global et comprend trois étapes : *Classification des concepts*, *Unification des concepts*, et *Restructuration des relations* entre concepts. Pour exécuter ce processus, chaque XSchema est converti en un diagramme de classes UML étendu « EUML : *Extended UML class diagram* ». Ce diagramme définit l'ensemble des cardinalités entre les éléments de chaque schéma, les relations d'agrégation, etc. Tout d'abord, l'étape de classification utilise l'ontologie *Wordnet* pour résoudre les conflits de nommage. Ensuite, les conflits de typage et de structure sont résolus dans l'étape d'unification. Enfin, la troisième étape, restructure les relations entre les différents concepts et supprime celles qui sont redondantes. Nous constatons que le processus proposé nécessite la traduction des XSchema en diagrammes de classes UML étendu. Cependant, lorsque le nombre de XSchema en entrée est important, leur traduction en diagrammes EUML peut consommer beaucoup de temps. De plus, ces diagrammes peuvent appartenir à des domaines différents, ce qui pourrait affecter la qualité du modèle conceptuel résultat.

Dans Júnior et Mello (2008), les auteurs proposent une approche pour l'intégration des documents XML. Cette approche se compose de deux processus : *Définition de similarité* et *Unification*. La définition de similarité compare chaque paire d'instances de documents et génère un score de similarité. Elle génère des ensembles d'instances de documents XML sémantiquement similaires. Alors que, le deuxième processus (i.e., unification) génère une instance de document XML unifiée pour chaque ensemble. Il utilise une ontologie de domaine et un dictionnaire pour nommer et structurer les représentations unifiées. Toutefois, la comparaison des paires d'instances de documents ne peut être que gourmande en temps vu le nombre élevé de documents à comparer ; aussi certaines comparaisons doivent être évitées lorsque deux documents possèdent une même structure.

D'autres travaux tels que Khrouf (2004) et Djemal (2010) proposent des processus d'intégration de documents dans un entrepôt. Ces processus sont basés sur un calcul de similarité entre les structures des documents.

Dans Khrouf (2004), l'auteur propose un processus permettant l'intégration de documents dans un entrepôt de documents. Tout d'abord, la structure logique[1] et le texte du document à intégrer dans l'entrepôt sont identifiés. Ensuite, les structures logiques contenues dans l'entrepôt et ressemblant à celle du document sont sélectionnées afin d'être comparées. Cette comparaison est basée sur le calcul d'un degré de similarité pour déterminer les structures qui peuvent être fusionnées avec le document. Finalement, le contenu du document est inséré dans l'entrepôt. Néanmoins, cette approche est conçue pour comparer deux structures. Par conséquent, lorsque nous voulons comparer plusieurs documents, nous sommes obligés d'appliquer ce processus sur chaque paire de documents. Ce processus peut être amélioré par l'utilisation d'une matrice qui englobe tous les degrés de similarité des documents à intégrer.

Dans Djemal (2010), l'auteur propose un modèle nommé MVDM « *MultiView Document Model* » pour les documents multi-structurés. Ce modèle décrit le niveau spécifique[2] et le niveau générique[3] des documents à travers respectivement une vue spécifique et une vue générique. L'auteur définit un processus comportant deux étapes : *Extraction de structure* et *Calcul de similarité* afin d'intégrer un document dans ce modèle. La première étape extrait la structure et le contenu du document. Ensuite, un degré de similarité est calculé afin de déterminer la vue générique qui sera associée à la vue spécifique du document ; ce degré mesure la similarité entre ces deux vues. Toutefois, cette comparaison peut consommer du temps du fait que chaque vue sera traduite en une matrice pour déterminer le degré de similarité.

Dans cette section, nous avons présenté les travaux qui nous semblent les plus pertinents relatifs à l'unification des structures de documents XML ; certains travaux proposent des méthodes pour l'unification des DTDs et d'autres présentent des méthodes pour l'unification des XSchema. Aussi, nous notons que certains travaux traduisent la structure du document XML en un modèle conceptuel. D'autres travaux définissent un degré de similarité pour mesurer la pertinence d'intégration des structures XML. En outre, nous soulignons que souvent l'utilisateur ne participe pas dans le processus d'unification. Dans ce qui suit, nous présentons une méthode et un outil d'unification des structures des documents XML, appartenant à un même domaine, basée sur un calcul de similarité. Dans cette méthode nous réservons place à l'utilisateur pour intervenir et valider le résultat d'unification.

3 Aperçu de la méthode d'unification des structures des documents XML

Rappelons que les documents XML appartenant à un même domaine sont décrits par des structures pouvant être hétérogènes. Ainsi, lors d'une analyse OLAP le décideur sera contraint de prendre en considération l'hétérogénéité structurelle de ces documents. De ce fait, il se trouve obligé d'écrire de multiples requêtes, soit autant de requêtes que de structures distinctes. De plus, il doit rassembler les résultats partiels pour obtenir le résultat final. Pour pallier à ces inconvénients, le décideur a besoin de travailler sur une structure globale, c'est-à-dire commune à l'ensemble des documents XML. Cette structure joue un rôle similaire à celui d'un schéma d'une base de données.

[1] La structure logique décrit l'ensemble des informations structurées contenues dans le document.

[2] Le niveau spécifique décrit un document à travers ses entités.

[3] Le niveau générique définit les types de documents à travers des groupements de structures similaires (par exemple DTD et XSchema).

Dans cette section, nous présentons un aperçu de notre approche d'unification des structures des documents XML publiée dans Ben Messaoud et al. (2011a). Cette approche permet de définir une structure générale pour un ensemble de documents XML ; cette structure se présente sous forme d'un ou plusieurs *arbres unifiés*. L'approche comporte trois étapes :

- représentation arborescente,
- génération des arbres unifiés, et
- validation des arbres unifiés.

La figure 1 illustre l'enchaînement de ces étapes que nous détaillons dans les sous sections suivantes.

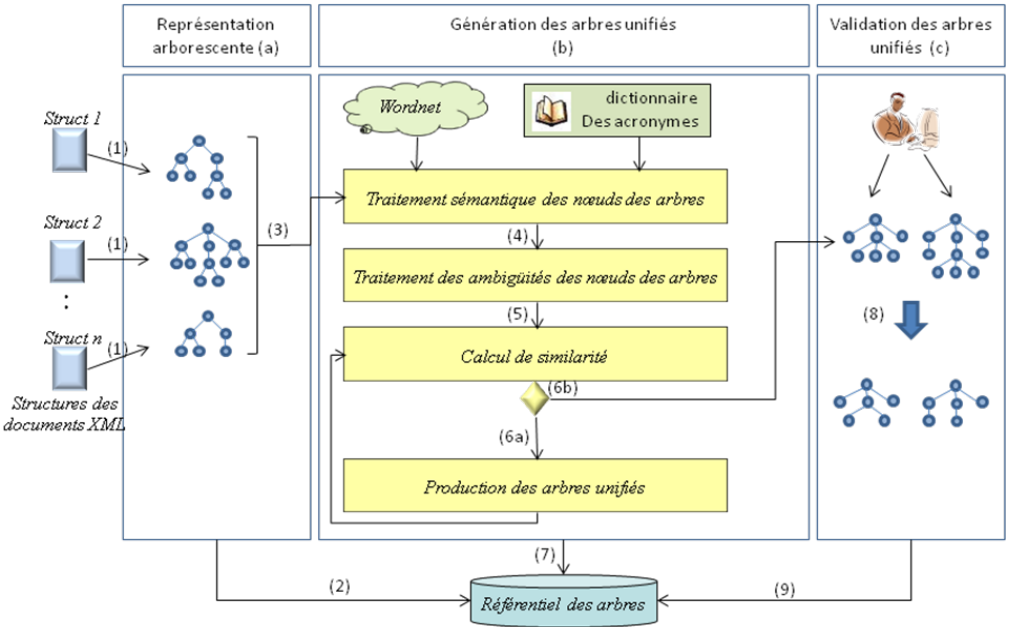


Fig. 1 – Etapes d'unification des structures des documents XML.

Les résultats intermédiaires et le résultat final du processus d'unification sont stockés dans le référentiel décrit par le diagramme de classes de la figure 2.

3.1 Représentation arborescente

Cette étape consiste à traduire chaque structure XML (DTD ou XSchema) sous forme d'un arbre comme dans les travaux de Lee et al. (2002) et Yoo et al. (2005). Chaque nœud de l'arbre résultat représente un élément de la structure. Il est annoté par une cardinalité. Alors que, les arcs indiquent les relations existantes entre les éléments de la structure. Nous choisissons le formalisme de l'arbre puisque il est facile à comprendre par les décideurs ; ce qui les motive à participer dans l'étape de validation.

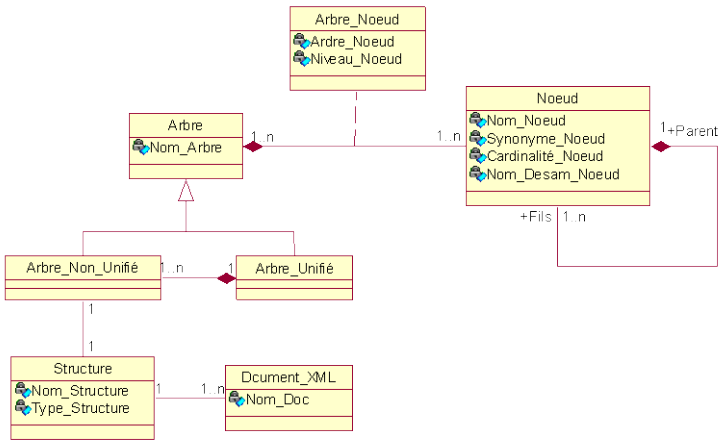


Fig. 2 – Diagramme de classes du référentiel des arbres.

3.2 Génération des arbres unifiés

La génération des arbres unifiés reçoit en entrée un ensemble d'arbres résultat de l'étape précédente et génère un ou plusieurs arbres unifiés. Cette génération comporte quatre sous-étapes :

- traitement sémantique des nœuds des arbres,
- traitement des ambiguïtés des nœuds des arbres,
- calcul de similarité, et
- production des arbres unifiés.

La figure 1.b illustre l'enchaînement de ces sous-étapes.

Traitement sémantique des nœuds des arbres. Cette sous-étape vise à résoudre les ambiguïtés sémantiques des nœuds des arbres ; elle utilise d'une part l'ontologie *Wordnet* et, d'autre part, un *dictionnaire des acronymes*. *Wordnet* permet de déterminer l'ensemble des synonymes d'un mot, ensuite, les nœuds ayant des noms de même sens sont détectés et leur nom seront remplacés par un nom commun, c'est le nom le plus fréquent dans l'ensemble des noms des nœuds synonymes qui sera retenu. Tandis que, le dictionnaire des acronymes permet de remplacer l'acronyme d'un nom de nœud par sa forme complète. Par exemple, le dictionnaire des acronymes permettra d'identifier que *Par* est une abréviation de *Paragraphe* et *Aut* signifie *Auteur*.

Traitement des ambigüités des nœuds des arbres. Il vérifie l'unicité des noms des nœuds d'un même arbre. En effet, il renomme les nœuds ayant le même nom en précédant le nom du nœud par le nom de son nœud père. En sortie de cette seconde étape, l'approche produit un ensemble d'arbres ayant des nœuds étiquetés par des noms uniques et standards ce qui permet de calculer des degrés de similarité plus précis que dans les approches de la littérature.

Calcul de similarité. Cette sous-étape permet de déterminer les arbres les plus prioritaires à être fusionnées. Pour ce faire, elle utilise une matrice triangulaire avec *n* arbres en lignes et *n* arbres en colonnes. Cette matrice est inspirée de la matrice définie dans Feki (2004) où elle a

été utilisée pour l'intégration des schémas multidimensionnels. Chaque cellule (*i*, *j*) de la matrice représente le facteur de similarité entre l'arbre de la ligne *i* et celui de la colonne *j*. Cette matrice a l'avantage de faciliter la recherche des arbres à unifier (i.e., les arbres ayant le plus grand facteur de similarité). Nous notons que la fusion de deux arbres est significative lorsque le facteur de similarité calculé est supérieur à un seuil ; ce seuil peut être déterminé expérimentalement.

Production des arbres unifiés. Elle construit des arbres unifiés à partir d'un ensemble d'arbres en entrée. Pour ce faire, elle fusionne chaque paire d'arbres ayant un degré de similarité supérieur au seuil et ceci en se basant sur trois opérateurs : *Fusion par inclusion*, *Fusion par union des sous-arbres* et *Fusion par union des nœuds*. La fusion par inclusion est réalisée lorsque l'un des deux arbres en entrée est inclus dans l'autre ; dans ce cas, l'arbre résultat est celui qui couvre tous les nœuds. La fusion par union de sous-arbres est opérée lorsque les nœuds communs de deux arbres en entrée ne partagent pas les mêmes nœuds fils. Dans ce dernier cas, l'arbre résultat est composé de l'union des sous-arbres des arbres en entrée. Alors que, la fusion par union des nœuds est utile quand deux sous-arbres identiques possèdent des nœuds pères différents. L'arbre fusionné est caractérisé par le nœud spécifique *ou*. Ce nœud substitue les nœuds parents distincts des arbres en entrée et relie les nœuds communs. Notons que lors de la fusion des arbres les cardinalités des nœuds sont pris en considération et sont traitées selon les règles présentées dans Hachaichi et al. (2010). Pour plus de détails concernant les opérateurs, nous dirigeons le lecteur vers Ben Messaoud et al. (2011a).

3.3 Validation des arbres unifiés

Dans cette étape, les arbres résultats de l'étape précédente sont présentés au décideur sous forme graphique afin d'être validés. Ainsi, le décideur/concepteur ajuste ces arbres selon ses besoins analytiques : il peut supprimer et/ou renommer les nœuds des arbres. Toute modification est enregistrée dans le référentiel afin de pouvoir retrouver le lien entre un nœud renommé et son origine dans la structure (DTD ou XSchema). Cette correspondance aidera ultérieurement l'interrogation des documents sources correspondant aux arbres unifiées.

4 L'outil USD « Unification of Structures of XML Documents »

Pour valider notre approche, nous avons implanté un outil nommé USD *« **U**nification of **S**tructures of XML **D**ocuments »*. Cet outil génère une ou plusieurs structures unifiées (i.e., arbres unifiés) à partir d'un ensemble de structures des documents XML (DTD ou XSchema) appartenant à un même domaine.

Premièrement, USD utilise le logiciel *XMLSpy*[4] pour générer pour chaque document XML une DTD ou un XShema. Ensuite, ces structures générées sont représentées sous forme d'arbres et les éléments de chaque arbre sont stockés dans un référentiel (figure 2).

Afin de construire des arbres syntaxiquement corrects, nous avons défini un ensemble de quatre contraintes ; elles garantissent la bonne forme des arbres (non unifiés ou unifiés). Elles aident à dériver, plus tard, des schémas multidimensionnels corrects et, par conséquent facili-

[4] http://www.altova.com/xml-editor/

tent l'interrogation des documents XML décrits par les structures en entrée. Ces contraintes sont :

- *C1 :* Connexité,
- *C2 :* Hiérarchie,
- *C3 :* Existence du nœud racine, et
- *C4 :* Acyclicité.

Définition 1 : La *Connexité* exige que chaque nœud appartenant à un arbre soit obligatoirement lié à au moins un nœud (i.e., nœud père).

Cette propriété garantie que chaque arbre présenté graphiquement ne comportera aucun nœud isolé. Par conséquent, lors de l'interrogation, le décideur peut utiliser tous les éléments multidimensionnels dans l'expression de son besoin analytique. Il est à noter que chaque nœud sera présenté comme étant un élément multidimensionnel (i.e., dimension, paramètre) dans le processus de modélisation multidimensionnelle.

Définition 2 : La *hiérarchie* des nœuds exige que chaque nœud soit lié à un et un seul nœud père.

Cette contrainte contrôle l'absence de nœud ayant plus d'un nœud père. Elle garantit l'obtention de structures purement arborescentes.

Définition 3 : L'*Existence du nœud racine* garantit que chaque arbre comporte un et un seul nœud racine.

Nous exploitons cette contrainte pour construire un arbre n'ayant pas des sous-arbres déconnectés. Lors de la modélisation multidimensionnelle de ces arbres, les éléments du schéma produit seront tous connectés ce qui permettra d'obtenir des schémas multidimensionnels valides.

Définition 4 : L'*Acyclicité* contrôle l'absence de cycles dans un arbre. Elle exige qu'un nœud ne puisse être père et fils du même nœud par transitivité.

Cette propriété assure que chaque arbre présenté ne comporte pas de cycles. S'appuyant sur un arbre vérifiant cette propriété, nous générons des schémas multidimensionnels ne comportant pas de cycles et, par conséquent les opérations de forage seront réalisables.

Dans le reste de cet article, nous détaillons les fonctionnalités de l'outil USD à travers l'unification de trois DTDs d'articles scientifiques, présentées dans la figure 3.

Le processus d'unification débute par la sélection de l'ensemble des structures (DTD ou XSchema) à unifier ; l'utilisateur sélectionne au moins deux structures. En fait, s'il choisit un document XML, l'outil USD appelle *XMLSpy* et génère une structure pour le document. Ensuite, les structures choisies seront représentées graphiquement sous forme d'arbres. Nous avons choisi cette forme de représentation puisqu'elle est simple et facile à manipuler par le décideur. Les figures 4, 5 et 6 présentent respectivement l'arbre 1, 2 et 3 correspondant aux trois DTDs de la figure 3. La validité syntaxique de ces arbres est vérifiée selon les contraintes définies précédemment.

DTD1

```
<!ELEMENT Article (Title, Writer+, Section+)>
<!ELEMENT Title (#PCDATA)>
<!ELEMENT Writer (Name, Affiliation)>
<!ELEMENT Name (#PCDATA)>
<!ELEMENT Affiliation (#PCDATA)>
<!ELEMENT Section (Title?, Subsection+)>
<!ELEMENT Subsection (Para)>
<!ELEMENT Para (#PCDATA)>
```

DTD2

```
<!ELEMENT Article (Title, Aut, Body, References)>
<!ELEMENT Title (#PCDATA)>
<!ELEMENT References (#PCDATA)>
<!ELEMENT Aut (Name, Institute)>
<!ELEMENT Name (#PCDATA)>
<!ELEMENT Institute (#PCDATA)>
<!ELEMENT Body (Section, Section_Number)>
<!ELEMENT Section (Paragraph)>
<!ELEMENT Paragraph (#PCDATA)>
<!ELEMENT Section_Number (#PCDATA)>
```

DTD3

```
<!ELEMENT Article (Title, Author, Summary, Section+)>
<!ELEMENT Title (#PCDATA)>
<!ELEMENT Summary (#PCDATA)>
<!ELEMENT Author (Name, University)>
<!ELEMENT Name (#PCDATA)>
<!ELEMENT University (#PCDATA)>
<!ELEMENT Section (Section_Number, Title, References, Subsection+)>
<!ELEMENT Section_Number (#PCDATA)>
<!ELEMENT References (#PCDATA)>
<!ELEMENT Subsection (Subsection_Number, Table?, Figure?)>
<!ELEMENT Subsection_Number (#PCDATA)>
<!ELEMENT Table (#PCDATA)>
<!ELEMENT Figure (#PCDATA)>
```

Fig. 3 – Exemple de trois DTDs.

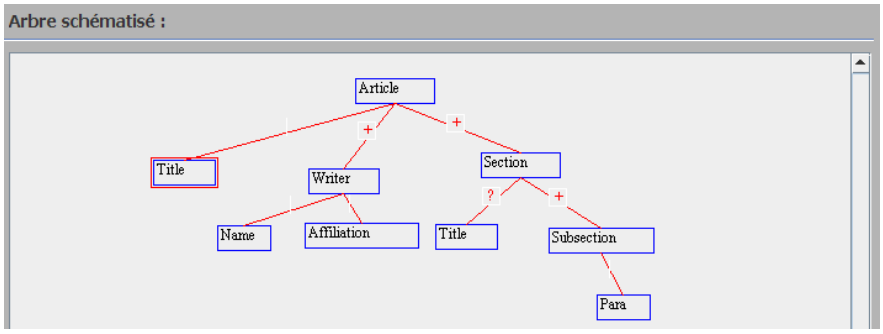


Fig. 4 – Arbre1 : représentation arborescente de la DTD1.

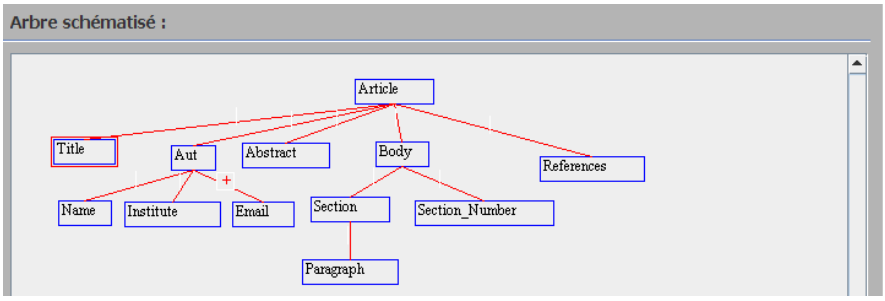


Fig. 5 – Arbre2 : représentation arborescente de la DTD2.

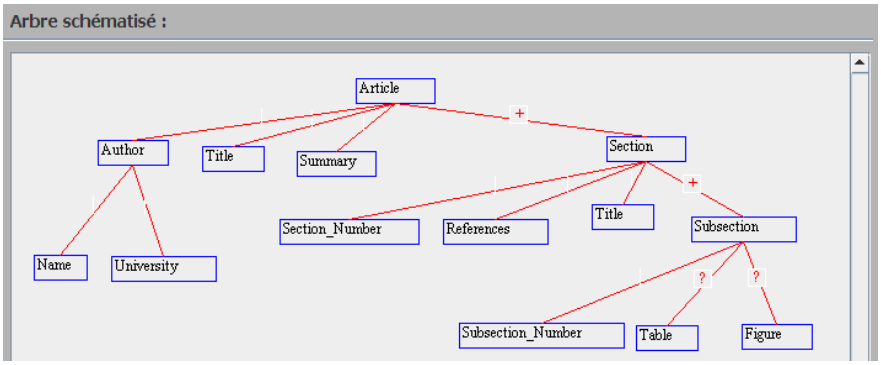


Fig. 6 – Arbre3 : représentation arborescente de la DTD3.

L'outil USD donne la main au décideur d'adapter les arbres présentés selon ses besoins analytiques ; par exemple pour supprimer les nœuds jugés inutiles.

Puis, les nœuds des arbres subissent un traitement sémantique pour résoudre les acronymes et les synonymes. Ce traitement prépare à l'étape de calcul de similarité, il permet d'obtenir de bonnes valeurs de similarité. A titre d'exemple, en utilisant le dictionnaire des acronymes, le nœud *Para* de *l'arbre 1 sera* remplacé par son nom complet *Paragraph*. De même, dans *l'arbre 2* , le nœud *Aut* sera remplacé par *Author*. La figure 7 décrit l'interface résultat du traitement des acronymes des trois arbres présentés précédemment.Tandis que, les synonymes sont résolus en accédant à *Wordnet* ; le nœud *Writer* de *l'arbre 1* sera ainsi remplacé par son synonyme *Author* qui représente le terme le plus fréquent dans l'ensemble des arbres à unifier.

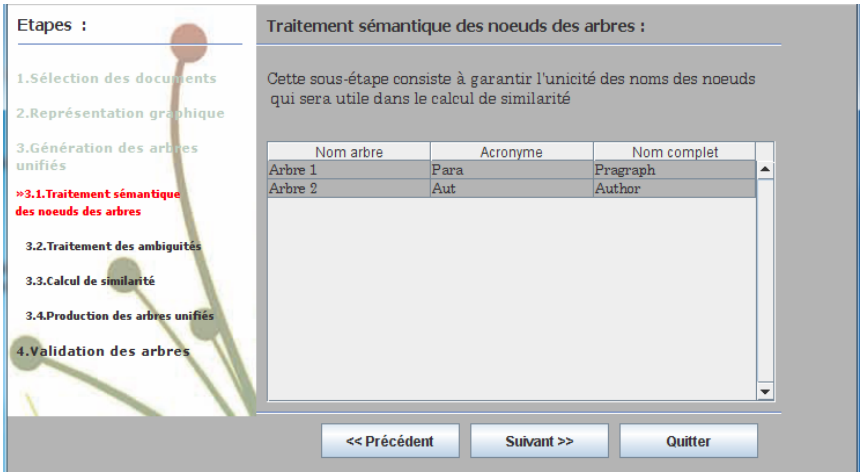


Fig. 7 – Traitement des acronymes des DTDs 1, 2 et 3.

Après le traitement sémantique des noms des nœuds, USD renomme les nœuds ayant le même nom et appartenant à un même arbre et ceci en préfixant le nom d'un nœud par le nom de son nœud père. Par exemple, dans *l'arbre 3*, il existe deux nœuds de même nom *Title*. Les noms de ces deux nœuds sont remplacés respectivement par *Article_Title* et *Section_Title*. Ce traitement permet d'aboutir à des arbres ayant des nœuds étiquetés par des noms uniques.

En outre, la matrice de similarité est calculée pour déterminer les arbres qui méritent d'être fusionnés. Notons que la fusion de deux arbres n'est réalisée que lorsque le degré de similarité entre ces deux arbres est supérieur ou égal à un seuil déterminé par expérimentation. USD donne la main au décideur de fixer son propre seuil.

Dans notre exemple, le décideur a choisi un seuil égal à 0.4. La première itération produit la matrice présentée dans la figure 8.

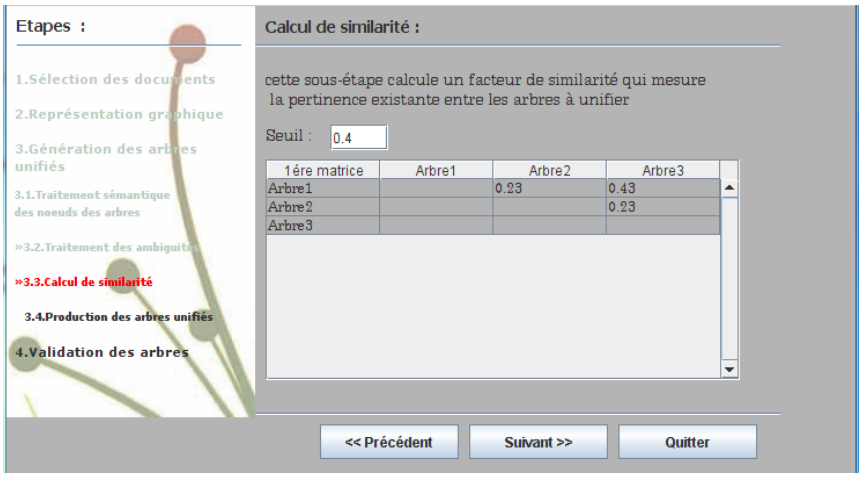


Fig. 8 – Première itération du processus de calcul de similarité.

Nous remarquons que le degré de similarité entre les arbres 1 et 3 dépasse le seuil retenu 0.4. Alors, ces arbres seront fusionnés par le biais de l'opérateur *Fusion par union des sous-arbres*. La figure 9 illustre l'arbre : Arbre' résultat de cette unification.

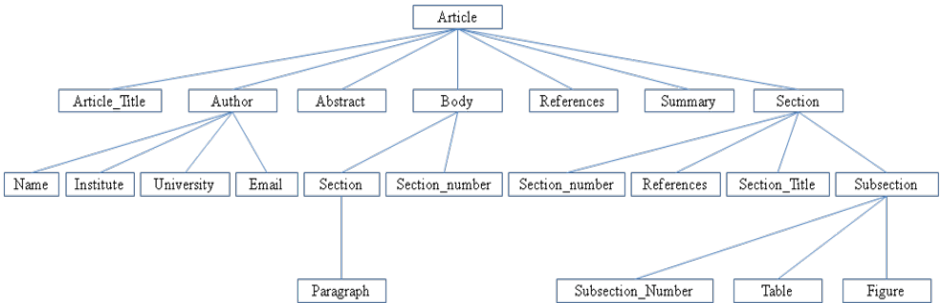


Fig. 9 – Arbre' : résultat de fusion de l'arbre 1 et 3.

La deuxième itération du calcul de similarité produit la matrice suivante :

2ème Matrice	Arbre 1	Arbre'
Arbre 1		0.35
Arbre'		

La similarité, entre les arbres Arbre 1 et Arbre', égale à 0.35 est inférieure à 0.4. D'où l'arrêt du processus d'unification. Le résultat final d'unification produit les deux arbres : *Arbre 1* et *Arbre'*.

Finalement, ces arbres sont fournis au décideur pour ajustement selon ses besoins analytiques ; il peut y supprimer et/ou renommer des nœuds. Les arbres résultats (i.e., unifiés et validés) sont traduits, plus tard, sous forme de schéma multidimensionnel par le biais des règles décrites dans Ben Messaoud et al. (2011b). C'est sur ces schémas que se fonderont les requêtes OLAP du l'entrepôt de documents.

5 Conclusion

Les documents aident les décideurs à mieux comprendre l'évolution des activités de leur organisation. Ils représentent une source principale dans le processus d'analyse décisionnelle et méritent d'être intégrés dans l'entrepôt. L'extension des entrepôts de données par des données textuelles améliore les résultats des analyses. Dans ce papier, nous avons présenté une méthode et un outil, appelé USD (Unification of Structures of XML Document), d'unification des structures des documents XML hétérogènes appartenant à un même domaine. Cette méthode est articulée autour de trois étapes que nous avons expliquées et illustrées à travers USD que nous avons développé pour supporter les étapes de notre approche.

Cette méthode consiste à unifier les structures des documents XML. Initialement, les structures des documents XML sont présentées sous forme d'arbres. Ensuite, les ambigüités des noms des nœuds sont résolues en utilisant d'une part l'ontologie *Wordnet* et, d'autre part, un dictionnaire d'acronymes. Puis, une matrice de similarité est calculée pour déterminer les arbres à unifier. L'unification est réalisée en se basant sur un ensemble d'opérateurs. Finalement, les arbres résultats d'unification sont retournés aux décideurs pour les valider selon leurs besoins analytiques.

Actuellement, nous continuons à tester notre méthode sur d'autres cas afin d'évaluer le processus d'unification. Egalement, nous sommes en phase de finaliser le développement de notre outil d'unification USD.

Références

Ben Messaoud, I., J. Feki, et G. Zurfluh (2010). *Unification des structures des documents XML pour l'entreposage de documents.* Cinquième Atelier sur les Systèmes Décisionnels (ASD'10), pp. 1-12, ISBN 9973-9900-2-0, Sfax, Tunisie.

Ben Messaoud, I., J. Feki, K. Khrouf, et G. Zurfluh (2011a). *Unification of XML document structures for Document Warehouse (DocW).* 13th International Conference on Enterprise Information Systems (ICEIS'11), pp. 85-94, Beijing, China.

Ben Messaoud, I., Feki, J., Zurfluh, G., (2011b). *Modélisation multidimensionnelle des documents XML*. 7ème journée francophones sur les Entrepôts de Données et d'Analyse en ligne (EDA'11), RNTI, Vol. B-7, Hermann, pp. 55-70, Clermont Ferrand, France.

Djemal K. (2010). *De la modélisation à l'exploitation des documents à structures multiples*. Thèse de doctorat en Informatique, Université Toulouse III – Paul Sabatier.

Feki, J., (2004). *Vers une conception automatisé des entrepôts de données : Modélisation des besoins OLAP et génération de schémas multidimensionnels*. 8th Maghrebian Conference on Software Engineering and Artificial Intelligence (MCSEAI'04), pp. 473-485, ISBN 9973-37-193-3, Sousse, Tunisie.

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

Júnior, C. A. S. et R. S. Mello (2008). *An ontology-driven process for unification of XML instances.* Brazilian Symposium on Multimedia and the Web, Vila Velha, Brazil, 242-249.

Khrouf K. (2004). *Entrepôts de documents : De l'alimentation à l'exploitation*. Thèse de doctorat en informatique, Université Paul Sabatier, Toulouse (France).

Lee, M. L., L. H. Yang, W. Hsu, et X. Yang (2002). *XClust: clustering XML schemas for effective integration*. Proc. of the ACM International Conference on Information and Knowledge Management (CIKM'02), pp. 292–299, McLean, Virginia.

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*. Proceedings of the 23th Annual International ACM SIGIR Conference, pp. 363–365.

Sullivan, D. (2001). *Document Warehousing and Text Mining: Techniques for Improving Business Operations*. Marketing and Sales. John Wiley & Sons, Inc.

Tseng F. S. C. et A. Y. H. Chou (2006). *The concept of document warehousing for multi-dimensional modeling of textual-based business intelligence*. Decision Support Systems (DSS), vol 42, Elsevier, pp. 727– 744.

Tournier R. (2007). *Analyse en ligne (OLAP) de documents*. Thèse de doctorat en informatique, Université Paul Sabatier, Toulouse (France).

W3C-XML (2008). *Extensible Markup Language (XML) 1.0*, http://www.w3.org/xml/.

Yoo, C. S., S. M. Woo, et Y. S. Kim (2005). *Unification of XML DTD for xml Documents with Similar Structure*. Computational Science and its Applications – ICCSA, LNCS 3482, pp. 954-963.

Zhang Y. F. et Wei-YI Liu. *Semantic integration of XML schema*. Proceedings of the first International Conference on machine Learning and Cybernetics, Beijing, 4-5 November 2002, PP 1085-1061.

Summary

Documents present an important source of information for decision analyses. They help decision makers to better understand and explain the activities of their organization. Usually, these documents are available in XML format and described by multiple and different structures. As a result, the use of these documents in a decisional process is difficult for a decision maker. This paper deals with a method and a tool to unify structures of XML documents belonging to the same domain. This method aims to produce a generic overview of these documents and to facilitate their querying in a decision making process. It consists of three steps: (i) tree representation, (ii) generation of unified trees, and (iii) validation of unified trees.

Mesure de la qualité de la vaccination guidée par les données

Laid AMAMRA*, Mostéfa MOKADDEM*, Baghdad ATMANI*
*Equipe de recherche SIF «Simulation, Intégration et Fouille de données »
Laboratoire d'Informatique d'Oran (LIO), Département Informatique Université d'Oran
sir_laid_m@hotmail.fr,{mokaddem.mustapha,atmani.baghdad}@univ-oran.dz

Résumé - Mesurer la qualité d'une vaccination repose sur une gestion avancée des données de référence, la possibilité de réutiliser des règles de qualité pour la gouvernance des données et la mise en place d'alertes proactives sur la qualité de cette vaccination.L'augmentation massive des volumes de données transactionnelles et l'explosion des volumes de données d'interaction générées par les différents services de vaccination rendent complexes cette mesure de qualité. Il faut donc fournir une plateforme d'intégration de données unifiée pour inscrire le concept d'entreprise data-centrique dans le processus de vaccination etoffrir la possibilité de mettre en place des solutionsen libre-service.La plateforme doit intégrer les données suffisamment vite pour assurer une qualité de vaccination, détecter et résoudre efficacement les problèmes de qualité des données, maitriser l'ensemble de ces données et en tirer des informations pertinentes, fiables et exploitables.
Nous proposons, dans cet article, une plateforme orientée services munie d'outils de fouille de données pour arriver à cette fin.

1 Introduction

Les SEMEP[1] doivent utiliser et faire coexister plusieurs systèmes d'information dont les systèmes de captation des naissances, les systèmes d'approvisionnement en vaccin, les services de contrôle et de supervision du processus de vaccination, les systèmes d'administration de vaccin, etc. En règle générale, l'intégration de plusieurs sources d'information ou de plusieurs systèmes mène à combiner ces différentes sources ou ces différents systèmes de manière à ce qu'ils forment une vue uniforme pour les utilisateurs, leur donnant l'illusion de n'interagir qu'avec un seul système. Pouvoir continuer à progresser vers une qualité de vaccination de plus en plus performante et compétitive,ilest devenu nécessaire de faire coopérerces processus. Dans ce contexte, l'interopérabilité est devenue de plus en plus importante afin de pouvoir satisfaire à la fois les besoins en vaccins, tout en rentabilisant leurs investissements relatifs et offrir ainsi une qualité de vaccination meilleure.L'interopérabilité représente la capacité qu'ont deux ou plusieurs composants (applications, sources de données, services mais aussi processus métier) de communiquer et de coopérer en dépit du modèle d'abstraction choisi (niveau d'abstraction choisi pour représenter une information par exemple). Il s'agit donc d'un objectif à atteindre afin de bénéficier d'un ensemble d'applications interopérables, de bases de données interopérables ou de services interopérables. Les trois propriétés principales de systèmes interopérables sont

[1]Services d'Épidémiologie et de Médecine Préventive

la prise en compte de la distribution, de l'hétérogénéité, et du respect de l'autonomie de chaque composant.

Mesurer la qualité d'une vaccination repose sur une gestion avancée des données de référence, la possibilité de réutiliser des règles de qualité pour la gouvernance de ces données et la mise en place d'alertes proactives sur la qualité de cette vaccination.L'augmentation massive des volumes de données transactionnelles et l'explosion des volumes de données d'interaction générées par les différents SEMEP rendent complexes cette mesure de qualité.Aussi et au fil du temps, le volume de ces données de transaction va augmenterpour dépasser les capacités de l'informatique.Il faut donc fournir une plateforme d'intégration de données unifiée pour inscrire le concept d'entreprise data-centrique dans le processus de vaccination etoffrir la possibilité de mettre en place des solutionsen libre-service.

La qualité de données doit, donc, être fournie sous forme de service, via de multiples protocoles facilement utilisables par des systèmes en aval aussi divers que nombreux. Et ces services de données doivent s'adapter de façon dynamique, pour répondre aux divers besoins.Les services destinés à l'amélioration de la qualité des données doivent assurer l'analyse et le nettoyage des stocks de données actuels tout en empêchant la saisie de données erronées et la création de doublons avant l'insertion des données dans la base de données. Dans ce contexte, les architectures orientées services (SOA) offrent des opportunités majeures pour améliorer l'efficacité des services de qualité des données et simplifier l'intégration dans l'infrastructure informatique de l'entreprise. On doit permettre,donc, de déployer des fonctions de qualité de données sous forme de services Web, capitalisant ainsi sur tous les avantages qu'offre une SOA, notamment la flexibilité, la réutilisabilité et l'évolutivité.

L'Analyse en Composantes Principales (ACP) est une méthode d'analyse de données.Elle cherche à synthétiser l'information contenue dans un tableau croisant des individuset des variables quantitatives. Produire un résumé d'information au sens de l'ACP c'est établir une similarité entre les individus, chercher des groupes d'individus homogènes,mettre en évidence une typologie d'individus. Quant aux variables c'est mettre en évidence des bilans de liaisons entre elles, moyennant des variables synthétiques et mettreen évidence une typologie de variables. L'ACP cherche d'une façon générale àétablir desliaisons entre ces deux typologies(Kaouani et al, 2007). L'intégration de l'ACP sous forme de services permettra, certainement, à la plateforme d'intégrer les données suffisamment vite pour assurer une qualité de vaccination, détecter et résoudre efficacement les problèmes de qualité des données, maitriser l'ensemble de ces données et en tirer des informations pertinentes, fiables et exploitables.

L'intégration d'applications est une tâche difficile, plutôt, c'est l'un des plus difficiles problèmes faisant face au développement d'applications d'entreprisedata-centrique. Prendre état de l'usage d'un vaccin, ne plus s'approvisionner en vaccins devenus inutiles et lier la vaccination au suivi des épidémies et la situation de santé de la population tels sont les valeurs ou résultats attendus de cette initiative.

La plateforme fera usage de plusieurs bases de données Oracle XE dispersées (une base par SEMEP du pays, une base pour chaque service de captation de naissance (APC) du pays, etc.) et de plusieurs Serveurs d'application WebLogic10.3.4 (respectifs aux bases). Sur chaque serveur, il est déployé plusieurs Business Components et/ou Services Web appelés à coordonner leurs activités. Les JSF de présentation aux différents niveaux de l'application communiquent avec les Business Components, les Services Web ou les Beans Session conformément à la Fusion Middleware et la SOA suite d'Oracle 11g.

Ce papier qui s'inscrit dans le cadre du projet « Architecture Orientée Service pour le Programme Elargie de Vaccination » (PNR, 2011) s'articule en 3 sections. La première section concernel'intégration et la qualité des données en vaccination avec principes et issues. La section 2 présente l'architecture du système proposé. La section 3 couvre l'implémentation et quelques expérimentations. Enfin une conclusion relève les points clé du projet et propose des perspectives.

2 La vaccination

La vaccination a pour définition traditionnelle la lutte contre les maladies. Elle assure un bénéfice individuel grâce à l'élimination de maladies infectieuseset un bénéfice collectif, de santé publique, en limitant la circulation et la transmission de ces maladies.

Le PEV (2010) décrit le processus complet du programme élargi de vaccination, jusqu'à 2001, en Algérie. Il cite les principales dates historiques, le calendrier vaccinal, quelques situations épidémiologiques, l'évolution de la couverture vaccinale etquelques indicateurs de performance de la surveillance de certaines maladies. Cependant le calendrier vaccinal a subi plusieurs amélioration jusqu'en 2010.On peut prévenir de nombreuses maladies infantiles grâce aux vaccins généralement recommandés pour les enfants. Depuis l'introduction généralisée de ces vaccins en Algérie, des maladies telles que la poliomyélite, la rougeole, la diphtérie, la coqueluche ont diminué de 85 à 95%. Avant l'avènement de la vaccination et la mise en œuvre effective du PEV(PEV,2010) et des campagnes de consolidation et de rattrapage, des dizaines de milliers d'enfants étaient infectés et un grand nombre d'entre eux mourraient en Algérie chaque année de ces maladies. La Cellule de Communication, en s'inspirant largement et librement des dossiers de l'UNICEF et de l'OMS, a proposé des réponses aux questions qu'on peut se poser au sujet de la vaccination et a montré que les avantages et les bénéfices liés à la vaccination l'emportent largement sur les risques infimes qui peuvent se produire dans n'importe quel pays du monde.

Toutefois, la qualité d'une vaccination ne peut être considérée comme acquise, et des enfants souffrent et meurent à cause d'erreurs de vaccination (efficacité du vaccin, mauvaise ventilation des vaccins, erreur de dose, perdus de vue…). Le fameux rapport,publié en 1999,*To ErrisHuman*de l'Institute of Medicine (IOM) a mis en lumière une statistique inquiétante selon laquelle, aux États-Unis, les erreurs médicales faisaient probablement plus de morts que les accidents de la circulation. Une mauvaise qualité de vaccination a des répercussions partout, même si ce sont surtout les enfants quien souffrent. Par ailleurs, les coûts de vaccination sont plus élevés qu'ils ne devraient l'être. De plus, le rapport « *Crossing the qualitychasm*» publié en 2001 par la même source définit les soins axés sur le patient comme l'un des principaux domaines de la qualité. Il propose un système de santé qui respecte les valeurs et les préférences des patients, assure la coordination et l'intégration décloisonnées des soins, informe, communique et éduque, et enfin garantit le bien-être physique, le soutien psychologique et l'implication de l'entourage dans les soins à dispenser. Un système de vaccination doit faire autant et une question très importante se pose: comment mesurer la qualité d'une vaccination et sa performance à partir des données?

Comme le souligneHarding(2005), dans un contexte de prise de décision, l'utilisation des résultats fiablesest en partie dépendante de l'utilisationprévue des données ainsi que de leur interopérabilité avec d'autres sources de données. La disponibilité des données n'assure pas forcément que celles-ci soient compatiblesavec l'environnement technique, ni qu'elles soient

conformes à la qualitéattendue. Il est évident que l'usage de données de mauvaise qualitéconduit presque toujours à des résultats incohérents voire erronés, ce qui est préjudiciable dansun processus de prise de décision.
Dans ce contexte Gutiérrez et Servigne(2009) définissent la qualité comme étant la proximitéentre les caractéristiques des données et les besoins d'un utilisateur pour une applicationdonnée à un instant donné.

Avec le développement des connaissances médicales, les nouvelles possibilités technologiques et la fragmentation des prestations de soins, l'évaluation de la qualité des processus et des résultats sur le plan de la santé revêt une importance croissante. Néanmoins, il est plus difficile,que jamais, de dispenser des soins et d'évaluer leur qualité.
Depuis sa création en 2001, le projet de l'OCDE sur les indicateurs de qualité des soins de santé (HCQI), mené en partenariat avec des organisations et des pays à la pointe dans ce domaine, a permis de mettre en place un cadre conceptuel et une base méthodologique pour fournir les informations nécessaires sur la qualité.Ce programme collecte les indicateurs aisément accessibles sur les processus et résultats en matière de soins, et entreprend des activités de recherche et développement coopératifs dans les domaines prioritaires des indicateurs (notamment les soins primaires, la santé mentale, la sécurité des patients et les expériences des patients) tout en encourageant l'amélioration de l'homogénéité des systèmes d'information et des indicateurs à l'échelle internationale. Même si des données manquent, ce projet a produit des statistiques utiles portant sur les aspects suivants : efficacité clinique, sécurité des patients et expérience des patients.Actuellement, environ 40 indicateurs de la qualité des soins de santé sont jugés appropriés pour la collecte de données internationales et ont été publiés dans des documents de travail et dans les éditions 2007 et 2009 de la publication biennale de l'OCDE intitulée Panorama de la santé (site Internet du projet HCQI, 2010). Le projet HCQI a progressé dans l'amélioration de la qualité et de la comparabilité des données provenant de différentes sources(OCDE, 2010). On ne peut guère disséquer la vaccination de l'état de santé d'une population. De la même façon, il faut dégager les indicateurs de la qualité de vaccination tels qu'il a été fait pour la qualité de soins. Pour élaborer un système qui prend en compte lamesure de la qualité de la vaccination quelques questions surgissent :

- Pourquoi avons-nous besoin d'informations sur la qualité de la vaccination?
- Quelles sont les données comparables à l'échelle internationale sur la qualité de vaccination ?
- Comment utiliser les informations relatives à la qualité pour améliorer le processus de la vaccination?

Face à ces questions les responsables de la santé ont besoin de mesurer, d'évaluer et de comparer la qualité de la vaccination pour trois raisons principales : responsabiliser les prestataires du processus de la vaccination, élaborer des politiques mieux adaptées et permettre aux prestataires et aux autres parties prenantes d'échanger leurs connaissances (et leur savoir-faire). De ce fait, il n'y a guère aujourd'hui de politique qui ne cherche à améliorer la qualité de la vaccination et/ou qui ne dépende de l'aptitude à mesurer cette qualité(OCDE,2010).

Et par contre les établissements qui mesurent de manière systématiques la qualité de leurs données, disposent d'une base solide pour l'améliorer efficacement, ainsi que pour comprendre l'utilisation réelle des données à travers l'entreprise. Normaliser et systématiser

la manière de mesure de la qualité de données est la première étape d'une stratégie globale et efficace pour la représentation des connaissances.

Pour aider les établissements à mettre en place rapidement un système de mesure de qualité des données fiable, le logiciel Infomatica9.1 a identifié six grandes dimensionspermettant de juger, mesurer et améliorer la qualité de chaque ensemble de données. Ces dimensions sont :

- La complétude des données
- Leur conformité aux règles internes et externes
- Leur cohérence
- Leur exactitude
- Leur intégrité et le taux de données dupliquées

La majorité des initiatives liées à la qualité des données concernent les données opérationnelles, autrement dit les données utilisées dans les systèmes métiers assurant la bonne marche de l'entreprise au quotidien. Aussi, une solution de qualité de données doit pouvoir prendre en charge les opérations cruciales, en temps réel, et inter-opérer en toute transparence avec l'ensemble de l'infrastructure informatique.

3 L'architecture du système proposé

A l'origine, on a d'abord vu apparaître de nombreux travaux autour de l'intégration de données, avec en particulier tous les travaux sur l'intégration de données dans les SGBD. Sont venus ensuite les travaux autour de l'interopérabilité autour du contrôle. En effet, les activités d'un système sont basées sur les règles explicites de fonctionnement du système c'est ce que l'on appelle les processus. Enfin, le découpage des activités en unités plus petites, et censées être plus modulables a permis de parler de services (ou de composants suivant le contexte d'utilisation à l'intérieur d'une organisation ou au-delà d'une organisation). Permettre à plusieurs services de fonctionner en orchestration est devenu un problème important.

Ainsi, lorsque l'on passe dans une PMI[2] pour vacciner son enfant, nous ne percevons pas nécessairement les différentes applications utilisées pour satisfaire notre requête : système d'identification de l'enfant, système de gestion des vaccins auprès des fournisseurs et depuis les SEMEP, système de gestion de stock, système de gestion de la livraison des vaccins, etc. Chaque accès sur un SEMEP, un centre de vaccination ou une PMI peut mettre en jeu une multitude de services pour collecter les données et les afficher sur une page.

Nous sommes dans le cas où le PEV[3] sollicite plusieurs entités qui sont, à la fois, distribuées à travers le temps, l'espace et que leurs processus peuvent être distribués dans plusieurs organisations voir la figure 1. Cela signifie que le PEV peut être scindé en plusieurs parties (un processus coopératif dépasse les frontières d'une PMI) et que chaque entité participante implante une partie seulement du processus (sous-traitance). Le service captation des naissances traite les données administratives de l'enfant, le SEMEP traite ses données de vaccination, la disponibilité des vaccins et leur livraison, la PMI traite l'administration du vaccin à l'enfant. On a aussi le scénario dans lequel l'administration du vaccin peut être implantée par plusieurs PMI parallèles, et il faut alors gérer la synchronisation des données

[2]Protection Maternelle et Infantile
[3]Processus élargi de vaccination

produites (Co-développement). On distingue alors une alliance temporaire amenée à composer certains services, à partager certaines ressources et connaissances provenant de services divers sans que ces derniers perdent leur autonomie. Ceci nécessite donc des mécanismes d'intégration (dans le sens composition) relativement flexibles (modification de sous-traitant, arrivée de nouveaux enfants, modification des participants à une activité donnée, capitalisation de projets précédents avec certaines modifications mineures ou majeures, etc.) et il est maintenant établi que les architectures centralisées issues du monde des systèmes d'information ne sont pas adaptées au support de ce modèle d'entreprise.

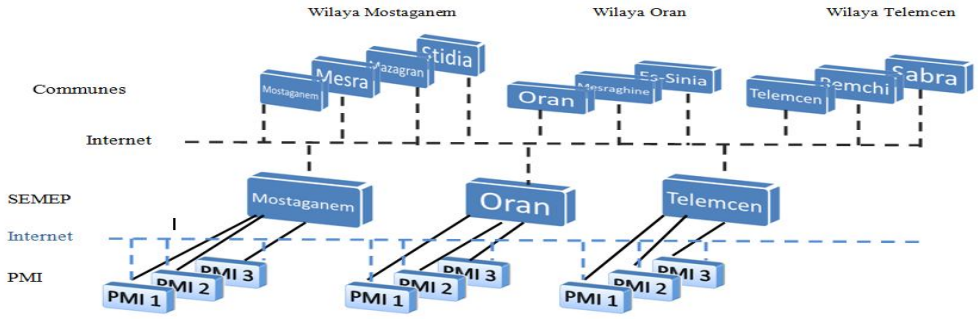


FIG. 1-*Architecture du processus de vaccination.*

Le PEV est donc un processus complexe qui fait partie intégrante de plusieurs responsabilités entre autres le SEMEP. Celui-ci doit être compétent depuis la captation de naissance d'un enfant jusqu'au suivi de la prise de tous les vaccins. D'où la phrase célèbre: « le PEV a pour objectif d'administrer le bon vaccin sous la bonne forme avec la bonne dose au bon enfant et au bon moment ». L'objectif du processus étant centré sur l'enfant, tout doit concourir à une meilleure prise en charge de son traitement, tant en terme de qualité et de fiabilité, que de sécurité et de délai (SAS, 2010).

Si le personnel du SEMEP se trouve contraint à remplir à la main les bons de commande et de livraison de vaccins ainsi que les formulaires relatifs, on ne peut prétendre à une qualité de vaccination meilleure puisque cette façon de gérer manuellement le circuit des vaccins a créé avec le temps une multitude de problèmes tels que :

- Un nombre important d'archives qui engendre une difficulté de stockage.
- Les difficultés de trouver les dossiers des enfants.
- Les difficultés de gérer les dossiers.
- Une mauvaise codification sur quelques objets dans la gestion de l'information
- La possibilité d'erreur dans les calculs des statistiques.
- Le temps à perdre pour remplir les formulaires.
- Une recherche difficile sur les registres qui engendre une perte de temps.

Le PEV est une composante essentielle de la mesure de la qualité de vaccination et représente aujourd'hui une préoccupation majeure pour les établissements de santé. Ainsi l'objectif de la plateforme est double, d'abord améliorer la qualité de vaccination, et ensuite valoriser les informations par leur réutilisation dans les différents sous-produits relatifs à l'enfant tels que l'historique de vaccination, le calendrier vaccinal, les perdu de vue, le

changement de résidence, le Dossier Vaccinal Personnel(DVP), et à la gestion tels que l'établissement des tableaux de bords, de pôles de gouvernance, etc.

La plateforme doit permettre :

- la réalisation du calendrier vaccinal et à l'évaluation des protocoles du PEV.
- d'assurer la sécurité et la qualité de la vaccination.
- d'acquérir des informations pour la valorisation de l'analyse de la vaccination.
- d'assurer l'interopérabilité avec différents logiciels touchant le PEV.
- le gain de temps, ou gain de productivité.
- la diminution des dépensesde vaccins.
- les conséquences liées à la meilleure qualité de vaccination apportée à l'enfant.

Pour mettre, plus facilement, en œuvre le PEV au sein du SEMEP, nous optons, pour développer un ensemble d'outils en Java et XML (technologie Ajax) permettant d'assurer tous les objectifs précités avec une prise en charge en temps réel du PEV et ainsi perfectionner le SEMEP devant les situations critiques, et améliorer par effet de simulation, de fouille de données et d'intelligence artificielle le PEV. La plateforme devra être suffisamment performante pour effectuer l'optimisation des divers processus et éviter les erreurs relatives à la qualité de vaccination.

Les données relatives au PEV proprement dites sont étroitement liées aux informations de captation de naissance à caractère administratif et, par ailleurs, les responsables de ces services souhaitent disposer d'indicateurs d'activité du PEV et de la mesure de la qualité de vaccination afin d'optimiser le rendement et minimiser la perte en vaccins sujets de péremption. C'est pourquoi les technologies Web et la gestion électronique des documents du PEV font actuellement l'objet du développement de Systèmes d'Information Vaccinal Décisionnel.

Les nouveaux modes de gestion des connaissances, notamment ceux qui font appel aux technologies de l'information et de la communication, peuvent améliorer l'efficacité en permettant de mieux gérer le temps, d'offrir des services de meilleure qualité, de favoriser l'innovation et de réduire les coûts.

Nous souhaitons pour cela développer un système générique flexible, extensible, visuel et compatible avec les formats de données les plus courants.Nous allons tenter de synchroniser des services web et implémenter le modèle MVC (Model-View-Controller)(Bean, 2010)(Martin et Marlies, 2008) très populaire et très adéquat pour ce genre d'applications. Nous allons aussi expérimenter des techniques plus complexes telles que la fusion middlewares d'Oracle 10g Norbert et al (2008), et enfin intégrer le tout grâce à JDeveloper 11g (Robin et al., 2009).

La distribution engendre des problèmes de localisation en temps réel des données qui sont utilisées au niveau des acteurs (SEMP, PMI et l'état civil) du processus vaccinal. Le volume important d'information disponible sur ces sources de données distribuées nécessite une stratégie de recherche et de sélection de plus en plus performante pour localiser et extraire l'information désirée. Pour y remédier, nous avons conclu que l'utilisation d'une SOA est plus appropriée pour notre démarche car elle présente des avantages multiples quant à la localisation en temps réel des données. Pour se faire le système d'intégration que nous proposons est un système basé sur l'invocation dynamique de services web. L'invocation dynamique consiste à construire,en run-time, une requête non statique, à découvrir dynamiquement le service auquel elle sera envoyée et à décortiquer,enfin,la réponse obtenue. Ce type d'invocation est à la base de l'orchestration de services dans notre cas. Nous

disposons alors de toute la liberté souhaitée pour avoir ce que nous voulons comme information. La figure2montre l'architecture de la plateforme.

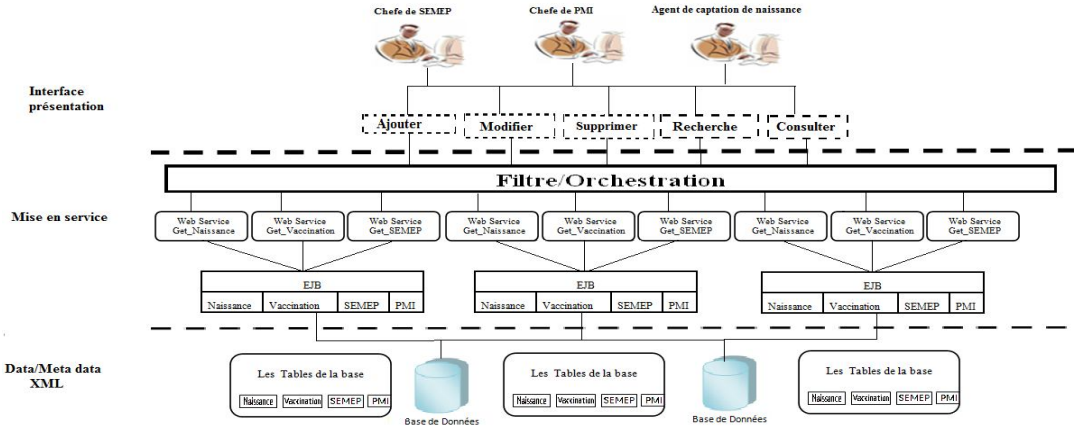


FIG.2 : *Architecture Orienté Service du système d'information Vaccinal.*

A présent, nous allons présenter et définir les trois groupes de processus (SEMEP, PMI, captation des naissances) en interaction dans un PEV schématisée par la figure 3.

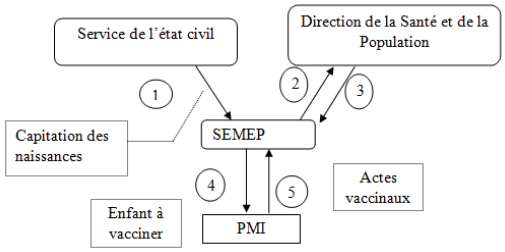


FIG.3 - *l'interaction dans le processus de vaccination.*

Processus'Captation des Naissances' :

La source de la captation des naissances est le service de l'état civil de la commune. Cependant, la commune où est né l'enfant n'est pas, nécessairement celle où résident ses parents. Le service de l'état civil et les structures d'accouchement (cliniques, maternités, hôpitaux, etc..) doivent être en mesure d'obtenir des familles l'adresse précise et la commune de la résidence familiale. L'enregistrement d'un nouveau-né nécessite certainement la validation de certaines informations primordiales comme l'existence des parents, leurs résidences etc. Or il s'avère qu'une naissance peut avoir lieu dans une commune où le nouveau-né sera enregistré alors que les parents sont déclarés dans des communes différentes entre elles et entre la commune où la naissance a eu lieu. L'adresse de résidence de la naissance n'est pas forcément dans la commune de naissance.

Exemple : une femme enceinte se rend à Oran pour une visite familiale. Elle met au monde son nouveau-né à Oran. Le nouveau-né sera enregistré dans la commune

d'Oran. Son père est, quant à lui, né à Alger et sa maman à Constantine. Les parents sont mariés à Mostaganem et leur lieu de résidence est Tlemcen.

Le système doit pendant la saisie de l'enfant vérifier qu'effectivement le père existe et est enregistré normalement à Alger et que la maman est normalement enregistrée à Constantine.

Pour cela, figure 4, le système doit se connecter aux systèmes distants (APC d'Alger et de Constantine) pour valider ces informations. Si le père existe effectivement, on doit vérifier que le nom du père est le même que le nom saisi pour la nouvelle naissance. Le système doit aussi valider la résidence des parents en accédant ausystème de l'APC de Tlemcen pour porter que le nouveau-né, à sa naissance, réside à Tlemcen le lieu de résidence des parents. Enfin, une dernière validation exige que le père et la mère soient bien mariés dans la légalité et déclarés à l'APC de Mostaganem. Ceci concerne la naissance, un processus similaire se déclenche dans le cas de décès qui consiste à valider qu'une personne décédée à Oran, née à Alger et résidant à Mostaganem est bien mise à jour sur les différents systèmes mis en œuvre.

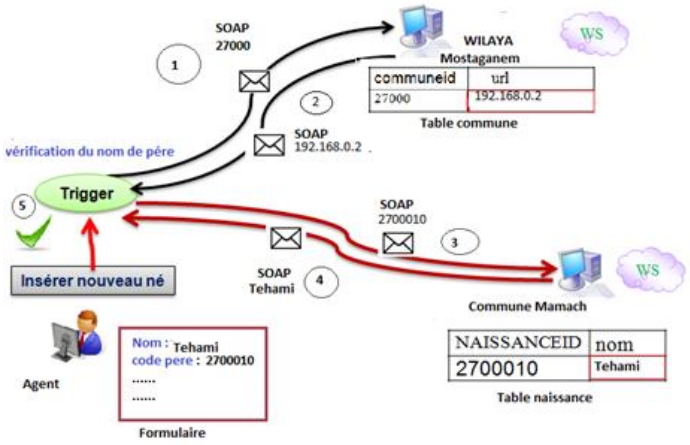


FIG.4 – *L'envoi et la réception des messages SOAP avec l'invocation dynamique des Web service.*

Un ensemble de services Web, en orchestration, permettant la découverte des perdus de vue, de changement de résidence, etc. est également en œuvre à ce niveau. La qualité des données est assurée par cet ensemble. L'état du lieu où l'enfant réside, l'état social de l'enfant, l'environnement sont aussi considérés. Ce sous-système assure que les données sont mises à jour en temps-réel. Il peut être soutenu par une ontologie de la captation des naissances pour donner plus de sémantique aux données et aux processus en interaction.

Processus 'PMI' :

Les PMI et les points de vaccination sont destinés à vacciner les enfants et enregistrer la consommation en vaccin pour améliorer leur productivité. Un enfant né dans une commune est censé posséder un carnet de santé délivré par le SEMEP lié à sa commune. Une PMI, dans l'état actuel, vaccine tous les enfants selon un calendrier

vaccinal et ne prend en charge que les enfants dont la liste a été fournie par l'agent responsable du SEMEP. Or on peut très simplement se présenter à n'importe quelle PMI et vacciner son enfant. Il suffit que la PMI puisse se connecter ausystème d'origine, où l'enfant a été saisi initialement, pour obtenir les informations qui le concerne et procéder à sa vaccination. Elle porte sa consommation, sa gestion de stock, sur son système local et personnel. Un schéma, similaire à celui de la figure 4, prend en compte ces opérations.

Processus 'SEMEP' :

Le SEMEP est l'organe usant, le plus, du décisionnel. Ces processus sont constamment en interaction avec les processus des autres sous-systèmes. Il doit faire état des enfants perdus de vue et lier ceci aux conditions sociales et environnementales où vit l'enfant, étudier les courbes des maladies liées à la vaccination, se prononcer sur la couverture vaccinale, étudier la qualité des vaccins, etc. enfin, il doit exprimer son expertise sous forme de règles adoptée par la population. Il est concerné, à cet effet,par la sélection des données à partir du SIVD. Les tâches de cette sélection consistent en la préparation et le filtrage des données pour pouvoir appliquer la méthode de projection (ACP), ces résultats seront traduits dans un graphique visuel.L'intérêt de cette représentation visuelle est de rendre le processus de prise de décision plus facile et abordable par le décideur.

La partie décisionnelle se compose de trois phases majeures :

- Sélection : L'expert se connecte au SIVD via un compte où se trouvent les données demandées, et exécute sa requête pour sélectionner ces données.
- Préparation : l'expert doit charger ces données pour construire la table individus/variables. Il pourra dynamiquement modifier, ajouter des attributs ou/etdes individus.
- Evaluation : fin de la préparation de données, et applicationde l'ACP avec une représentation visuelle des résultats.

Plusieurs expérimentations et la combinaison avec d'autres outils de fouilles de données, on pourra aboutir à un ensemble de données de qualité qui toucheront de près le PEV.

4 Conclusion

Les travaux de recherches menés dans le cadre de ce papier avaient un double objectif : d'une part développer une plateforme interopérable pour automatiser le PEV et avoir en permanence disponible toutes les données de vaccination pour une meilleure exploitationet d'autre part implémenter une méthode de projection visuelle pour la classification afin de découvrir des méthodes faisant preuve de qualité de données.

La démarche expérimentée adopte le principe de disponibilité et accessibilité temps réel à l'information distante.Nous avons proposé une architecture orientée service permettant d'agir sur dessystèmes distants comme un système local et unique en utilisant l'orchestration dynamique de services Web. La qualité de données est assurée par la prise en charge des opérations cruciales, en temps réel, et l'interopérabilité en toute transparence avec l'ensemble du SIVD. La qualité de données est fournie sous forme de services adaptés dynamiquement, pour répondre aux divers besoins.Les services destinés à l'amélioration de la qualité des données prennent en charge l'analyse et le nettoyage des stocks de données actuels tout en empêchant la saisie de données erronées (noms père et enfant, mariage des parents,

validation des adresses, etc.) et la création de doublons avant l'insertion des données dans la base de données. Dans ce contexte, la SOA adoptée a offert des opportunités majeures pour améliorer l'efficacité des services de qualité des données et simplifier l'intégration dans le SIVD.Ce qui capitalise ainsi sur tous les avantages qu'une SOA offre, notamment la flexibilité, la réutilisabilité et l'évolutivité.

Le décisionnel implémenté sous forme de services est, en réalité, une simple expérimentation des outils de fouille et d'analyse de données. Le travail étant ancré sur la SOA, l'ACP permet juste l'exploration et la visualisation de l'information sous forme graphiqueen vue d'une interprétation rapide pour l'aide à la décision sur la qualité des données et du PEV.

Décider sur la qualité des données ou d'un processus métier est certainement bien soutenu avec une expertise exprimée sous forme de connaissance. Ainsi, compléter les données par l'aspect sémantique exprimé par des ontologies de données et de processus constitue une extension à ce travail. L'extraction des connaissances à partir des données est une première initiative qui débordera sur des résultats à ne pas négliger.

En perspective à ce travail, nous allons orienter nos efforts vers l'intégration des ontologies à la mesure de la qualité et expérimenter les outils d'extraction de connaissance en vue d'enrichir ces ontologies. La découverte de connaissance à partir du Web (réseaux sociaux, forums, laboratoires, blogs, etc.) via du TextMining conduit également à l'extension de l'ontologie envisagée.

5 Références

ANDS, (2010) *: Module d'épidémiologie et médecine préventive.www.adns.com.*

Kouani A., S. El Jamali et M.Talbi (2007) : *Analyse en composantes principales Une méthode factorielle pour traiter les données* didactiques,Radisma, numéro 2, 2007.

Bean J., (2010): *SOA and Web Service Interface Design*, Morgan Kaufman, 2010 ISBN 978-0123748911.

Christophe L., (2008) : *Retour d'expérience et perspectives de mutualisation en matière de Système d'Information Décisionnel (S.I.D.,*Projet Sis-If – Groupe S.I.D.

CISSEM., (2009) : *Manuel de formation sur le système d'information sanitaire du Burkina Faso,* Direction générale de l'information et des statistiques sanitaires.

Daniel A., (2004): *Transactions On Visualization And Computer Graphics Vol. 7*, Information Visualization Journal.

Deepak V., (2009):*Processing XML documents with Oracle JDeveloper 11g, Creating, validating, and transforming XML documents with Oracle's IDE.*

Duncan M., Peter K., Avrom R., (2010) : *Oracle JDeveloper 11g Handbook A Guide to Oracle Fusion Web Development*, Oracle Team.

GHELLAB A., (2007) : Conceptiond'une Basede Données Décisionnelle. Mémoire de magister.

Gutiérrez, C. et S. Servigne (2009).*Métadonnées et qualité pour les systémes de surveillance en temps-réel.* Revue Internationale de Géomantique 19/2, pp. 151–168.

Harding, J. (2005) :*Qualité des données vectorielles : perspective d'un producteur de données. In Qualité de l'information géographique*, pp. 171–192. Traités IGAT, Hermés Sciences, Lavoisier. ISBN 2-7462-1097-5.

Lamirel J. C., (2006) : *Combinaison de méthodes avancées de visualisation et de sélection d'information pour la fouille et l'analyse de données.*

Martin V., Marlies S., (2008): *Building an Enterprise Architecture Practice*, Kluwer Academic Publishers.

Menouer T., Dermouche M., (2010) : *Application de techniques de datamining pour la classification automatique des données et la recherche d'associations,*Ecole nationale supérieur d'informatique.

Norbert B., Robert G., Keith J., Tilak M., (2008): *A Practical Guide for the Service-Oriented Architect,* Pearson Education.

OCDE, (2010) *: Améliorer la performance des soins de santé : comment mesurer leur qualité,* Forum sur la qualité des soins, Paris, 7-8 octobre 2010

Peguiron F., (2006) : *Application de l'Intelligence Economique dans un Système d'Information Stratégique universitaire : les apports de la modélisation des acteurs*, Thèse présentée et soutenue publiquement le 16 novembre 2006.

PNR., (2011): Architecture Orientée service pour le programme élargi de vaccination. http://www.nasr.dz/pnr2011/cerist_12.html

Robin W. & al, (2009): *Oracle Fusion Middleware Web User Interface Developer's guide for Oracle Application Development Framework 11g, book*, oracle Corporation.

SAS, (2010) :*Qualité des données La matière première de la décision mérite votre plus grand soin.* http://www.sas.com.

Un modèle basé agent pour l'aide à la décision coopérative dans une chaîne logistique

Souheila BOUDOUDA*, Mahmoud BOUFAIDA**

Labouratoire LIRE, 25000 Constantine, Algerie
*boudoudasouha@yahoo.fr
** mboufaida@umc.edu.dz

Résumé. Dans cet article nous présentons un modèle distribué d'une chaîne logistique. Ce dernier utilise le paradigme agent pour modéliser les différents acteurs de la chaîne. Chaque acteur de la chaîne est représenté par un sous système constitué de trois types d'agents (agent approvisionneur, agent de gestion de stock, agent livreur). La coopération et la négociation entre les différents agents permettent la synchronisation des décisions prises par les différents acteurs de la chaîne. Les agents seront donc capables de s'alimenter en informations auprès des systèmes d'information des différents acteurs ensuite interagir et négocier afin de proposer des solutions pour l'aide à la prise des décisions. Les agents des sous systèmes de la chaîne peuvent communiquer selon deux types d'interaction : une interaction entre les différents sous systèmes appartenant à des niveaux différents liés par la relation de type demandeurs de services ou de biens et fournisseurs, ce type génère des décisions concernant les différents types des flux dans la chaîne. Une deuxième interaction entre les sous systèmes appartenant au même niveau afin de collaborer, échanger et négocier les stocks, rendre le système plus réactif et de diminuer la taille des commandes.

1 Introduction

Aujourd'hui, la performance des entreprises ne suffit plus pour assurer leur pérennité. En effet, les exigences des clients sont si multiples et les intervenants dans la chaîne (chaîne d'entreprises) si nombreux, que les entreprises se trouvent donc en situation d'interdépendance très forte avec ses partenaires. Elles se constituent ainsi en réseaux, dans le but de mieux prendre en compte les besoins du client final à moindre coût. Au sein de ces réseaux, des mécanismes de collaboration sont apparus pour faciliter les relations entre un donneur d'ordres et ses sous-traitants ou fournisseurs, cet ensemble de mécanismes sont appelés « Chaînes Logistiques (CL)[1] ». Une chaîne logistique peut prendre l'allure d'un modèle très simple constitué d'un nombre réduit d'acteurs liés linéairement, comme elle peut être extrêmement complexe avec un nombre élevé de sous traitants, de fournisseurs, de centres d'assemblage, de distributeurs et des détaillants, interagissant ensemble et simultanément avec différents outils et techniques de gestion (Simchi et al, 2000). La gestion de

[1] **CL :** Traduction française de la notion de **Supply Chain (SC).**

cette chaîne (GCL)[2] se matérialise par un partage d'information et un redéploiement des activités entre les différents maillons qui la composent (Mentezer et al, 2001).

Beaucoup de problèmes sont apparus avec l'évolution de la dynamique des interactions de la CL, tels que l'augmentation des coûts de production et de stockage, l'amplification de la demande (coup de fouet)[3] ou la complexité de flux d'information. La plupart des travaux de recherche de la GCL concernent la modélisation de la CL et les méthodes permettant l'amélioration des performances. Nous pouvons citer, par exemple, les méthodes de programmation mathématique linéaire, intégrée avec ou sans contraintes (Rota et al, 2000) et la recherche opérationnelle avancée (Hong et al, 2000).

La modélisation d'une chaîne logistique ayant un nombre d'acteurs variable avec de multiples relations est une tâche difficile. Cette chaîne comporte souvent des phénomènes qui ne peuvent être modélisés par des méthodes classiques (réseau de Petri, approches probabilistes, recherche opérationnelle, etc.). Vu la dynamique et la diversité des flux entre ces acteurs, cette organisation industrielle nécessite des méthodes de modélisation avancées pour donner une description complète de ses mécanismes de fonctionnement. Dans ce contexte, notre contribution se situe parmi les travaux qui s'intéressent à la problématique de modélisation de la CL dans le but d'obtenir un modèle générique suffisamment proche de la réalité industrielle et capable de répondre aux différentes difficultés de sa gestion. L'utilisation des modèles issus de l'IAD[4] et plus particulièrement des SMA[5] dans les outils de gestion des entreprises s'avère être efficace pour simuler et reproduire les comportements collaboratifs et adaptatifs tels qu'ils apparaissent actuellement dans les entreprises. Les SMA constituent donc un outil puissant pour comprendre la dynamique des systèmes complexes car elle permet de représenter de façon explicite les phénomènes d'interaction et de collaboration entre les entités intervenant dans le système. Notre ambition est de proposer un modèle distribué à base d'agents capables de coopérer afin d'étudier la prise de décision décentralisée entre les différents acteurs de la chaîne dans le but d'améliorer globalement les performances de la CL.

Cet article est organisé comme suit : La section 2 présente une vue générale sur les différentes approches de modélisation d'une CL. Ensuite nous présentons dans la section 3 les SMA et leurs rôles pour le pilotage et la gestion des CL ainsi que les motivations qui nous ont amené à les employer pour répondre à notre problématique. Dans la section 4 nous donnons un aperçu général de notre proposition. La section 5 est consacrée à la description et le rôle des différents agents de chaque sous système d'une chaîne logistique, dans la section 6 nous détaillons les deux types d'interaction proposés dans notre modèle. Enfin, une conclusion résume notre travail, en citant l'apport des SMA dans la modélisation, la communication et la collaboration entre les systèmes de GCL et en donnant quelques perspectives de travail.

2 Complexité des CLs

Un système de GCL peut être vu comme un concept développé par les entreprises pour apporter une réponse à une demande client personnalisée en termes de qualité de ser-

[2] **GCL :** Traduction française de la notion de **Supply Chain Management (SCM).**
[3] **Coup de fouet:** Traduction française de la notion de **Bulwip effect,** il décrit comment les petites fluctuations de la demande au niveau du client sont amplifiées (Geune et al, 2003).
[4] **IAD: I**ntelligence **A**rtificielle **D**istribuée**.**
[5] **SMA : S**ystème **M**ulti **A**gents**.**

vice (Muller et al, 2002). Ainsi, le système de GCL a pour premier objectif d'éliminer les barrières qui limitent la communication et la coopération des différents membres d'une chaîne (figure 1) (Cutting et al, 2006). La GCL implique donc des entités à la fois autonomes mais aussi en forte interaction dans un environnement de plus en plus dynamique et donc leur gestion nécessitera une coordination minutieuse et une synchronisation parfaite si l'on veut réussir une optimisation des chaînes au sein d'une entreprise (Giard et al, 2006).

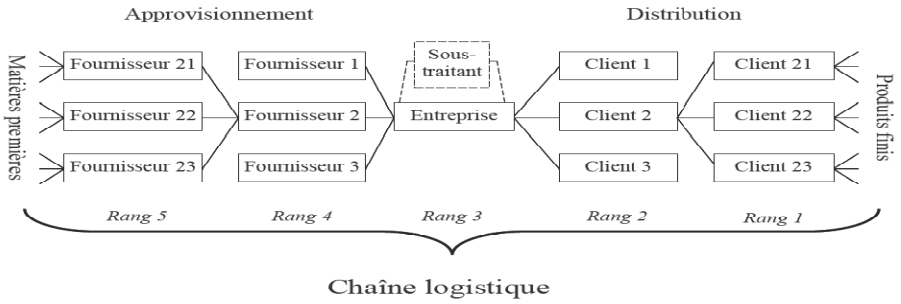


FIG. 1 – *Exemple d'une chaîne logistique.*

Afin d'appréhender et de réduire la complexité du système réel, la modélisation est une étape essentielle pour l'étude des systèmes complexes et dynamiques. La mise en œuvre d'un projet de chaîne logistique, oblige les entreprises à maîtriser leurs processus en interne afin d'appréhender leurs interactions en externe. Dans ce cadre, de multiples approches de modélisation sont utilisées afin d'en décrire et analyser l'organisation. (Anane et al, 2009) ont dressé un panorama de ces approches. Dans son étude, l'auteur distingue 4 catégories de modèles, à savoir :

- Les modèles analytiques déterministes : à titre d'exemple, le modèle déterministe et non linéaire de programmation mathématique en nombre entiers qui a été étendu par un autre modèle d'optimisation sous contraintes ;
- Les modèles analytiques stochastiques : à titre d'exemple, les modèles stochastiques pour calculer la valeur des variables aléatoires introduites dans le modèle mathématique ;
- Les modèles économiques : à titre d'exemple, les modèles de la théorie des jeux pour analyser les relations fournisseurs-clients ;
- Les modèles de simulation : à titre d'exemple, les modèles de simulation pour évaluer les effets des stratégies sur l'amplification de la demande.

Plusieurs autres travaux de recherche peuvent être cités (Dupont et al, 2008), (Ghedira et al, 2008), (Karam et al, 2010). Ils traitent de la performance d'une chaîne logistique équitable, constituée des donneurs d'ordres et des sous-traitants dans un contexte contractuel particulier qui est le contrat réservation de capacité. L'objectif est de maximiser la marge de la chaîne logistique.

Cependant ces différents travaux présentent plusieurs limites. L'aspect interaction entre les différents acteurs n'est pas pris en compte. Ces travaux ne permettent pas la modélisation du comportement des acteurs, notamment leur caractère réactif, autonome et proactif. L'approche agent apparaît comme une technologie intéressante pour modéliser la complexité de la chaîne logistique et de ses comportements. La section suivante explicite la manière dont ces systèmes sont susceptibles de modéliser et d'implémenter des systèmes décentralisés, communicants et collaboratifs.

3 Apport des SMA dans la modélisation des CLs

Un intérêt majeur de la modélisation orientée agents est qu'elle permet en général une simulation autorisant une étude de la dynamique du système considéré. L'approche de modélisation orientée agents permet de représenter le système à travers l'identification et la spécification du comportement des individus en interactions qui le composent. Ces individus peuvent évoluer et agir sans contrôle ni intervention extérieure (autonomie). Ils peuvent percevoir leurs environnements et répondre à ces modifications (réactivité), initier des comportements dirigés par des buts internes (pro-activité), et interagir avec d'autres individus (habilité sociale). Selon (Parunak et al, 2000), les organisations en réseau de production et de distribution possèdent les mêmes caractéristiques que les agents : autonomie, capacité sociale, réactivité et pro-activité. De ce fait les SMA s'adaptent bien aux systèmes complexes et ouverts (comme les chaînes logistiques) où il est difficile de tout décrire à l'avance. Mark Fox a été l'un des premiers à proposer d'organiser la chaîne logistique comme un réseau d'agents intelligents (Dodd et al, 2001). Dans son projet, les chaînes logistiques sont composées de sous-systèmes de production hétérogènes qui se regroupent en vastes coalitions dynamiques et virtuelles. Les SMA permettaient la représentation de l'autonomie de chaque membre de ce réseau d'entreprises.

D'autres projets se sont confrontés à des problématiques liées à la prise de décision dans les CL et ont employé des technologies à base d'agents. (Moyaux et al, 2008) utilisent le paradigme agent pour modéliser la variabilité de la demande des chaînes logistiques (coup de fouet) en fonction de la disponibilité de l'information. Plus précisément, ils modélisent un nouveau schéma pour passer les commandes (moins de stocks mais plus d'information).
(Brahimi et al, 2009) ont présenté un autre modèle basé essentiellement sur deux concepts clés : l'agent et le web service. La partie agents est le noyau de ce modèle, elle est représentée par un ensemble de groupes identiques. Chaque groupe possède un ensemble d'agents d'application dirigés par un agent coordinateur.

Plusieurs autres travaux de recherche (Montreuil et al, 2008), (D'amours et al, 2006) (Nfaoui et al, 2008), (Zhang et al, 2009), (Anane et al, 2009) sont souvent finalisés par une simulation pour mettre en œuvre le modèle développé. Nous nous situons dans cette catégorie de travaux de recherche, et nous présentons dans les paragraphes suivants notre proposition.

4 Aperçu du modèle proposé

La complexité de la modélisation des systèmes de gestion des chaînes logistiques ainsi que le support de mise en œuvre, nous amènent à proposer une architecture de modélisation basée agents. Notre modèle (figure 2) est constituée d'un ensemble des sous systèmes multi agents constitués de trois types d'agents cognitifs (agent approvisionneur, agent de gestion de stock, agent livreur), chaque sous système représente un acteur de la chaîne, ces sous systèmes sont regroupées en plusieurs niveaux. Ils cherchent à coordonner et à synchroniser des activités très différentes. Ils peuvent communiquer entre eux par des interfaces de communication.

Le modèle proposé est dynamique, permettant d'avoir différents formes et tailles d'une chaîne logistique. Les sous systèmes gèrent non seulement les activités traditionnelles de la

chaîne logistique (production, planification, ordonnancement, stockage), mais manipulent aussi les flux de communication entre les différents acteurs.

Notre modèle permet la coopération et la négociation entre les différents maillons de la chaîne dans le cas des situations de planification opérationnelle (par exemple commande urgente) ou bien l'apparition d'un problème (par exemple problème de production) conduisant à une prise des décisions collaboratives. Pour cela les acteurs peuvent communiquer selon deux types d'interaction: une interaction entre les différents sous systèmes appartenant à des niveaux différents, liées par la relation de type demandeurs de services ou de biens et fournisseurs (Chaque acteur peut communiquer avec l'acteur qui est en amont ou en aval de lui même), dans ce type : l'interaction entre acteurs génère des décisions concernant les différents types des flux dans la chaîne (matériels, informationnels, monétaires). Une deuxième interaction entre les sous systèmes du même niveau afin de collaborer, échanger les différentes informations et négocier les différents nivaux des stocks.

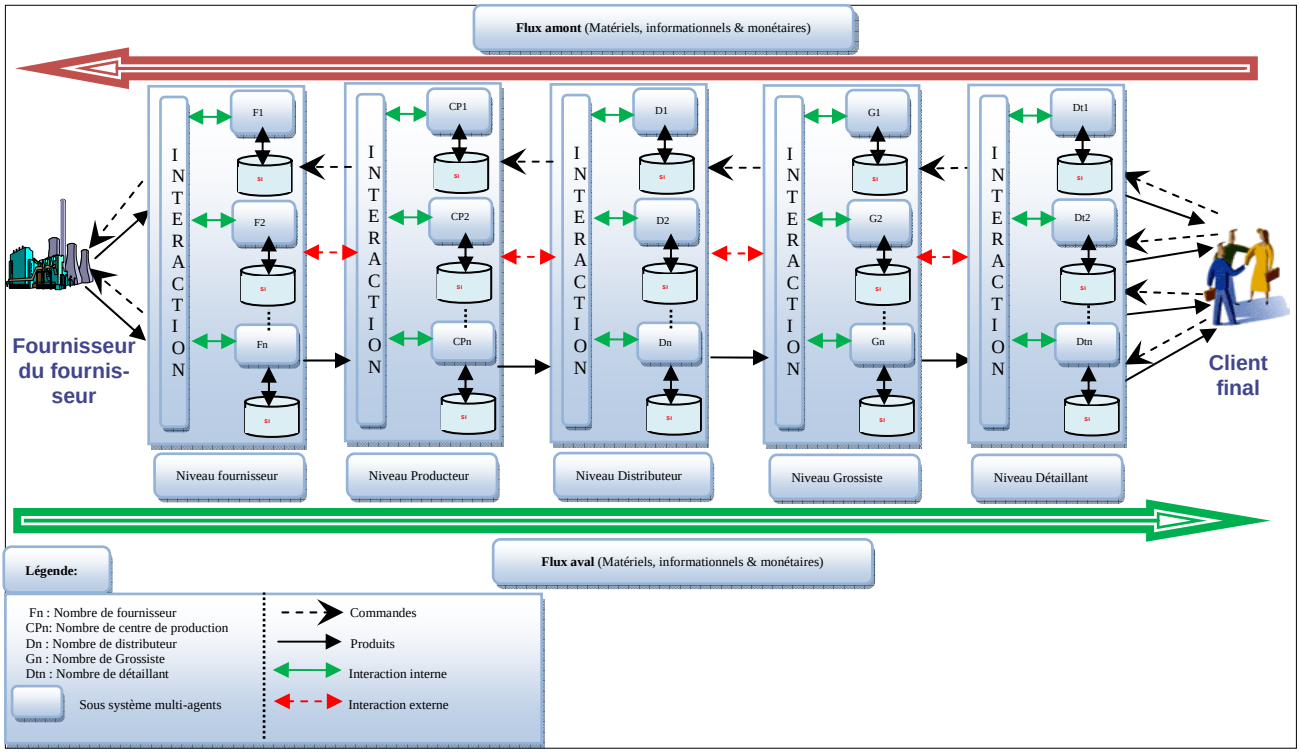


Fig. 2 – *Présentation du modèle multi agents multi niveaux d'une CL.*

Le but de notre modèle est d'étudier la prise de décision décentralisée et la coordination d'entités autonomes en vue de réduire l'effet dit « coup de fouet ». Elle permet également la flexibilité, c'est-à-dire que les sous systèmes qui modélisent la chaîne peuvent être étendus par l'introduction de nouveaux sous systèmes sans avoir à modifier la structure existante.

5 Architecture globale du sous système

Chaque acteur de la chaîne logistique n'a qu'une vision partielle de toute la chaine, il est uniquement en contact avec ses clients (acheteurs) et ses fournisseurs directs (vendeurs). La coopération inter-entreprise nécessite des processus de négociation et de renégociation, des propositions et des contre-propositions et finalement une acceptation ou refus de solutions partenariales. Pour cela nous modélisons chaque sous structure par trois agents de type cognitif (figure 3) (agent approvisionneur, agent de gestion de stock, agent livreur), Ces derniers interagissent entre eux pour accomplir leurs taches internes de chaque sous système et peuvent communiquer, négocier et collaborer avec les autres sous systèmes à travers une interface de communication.

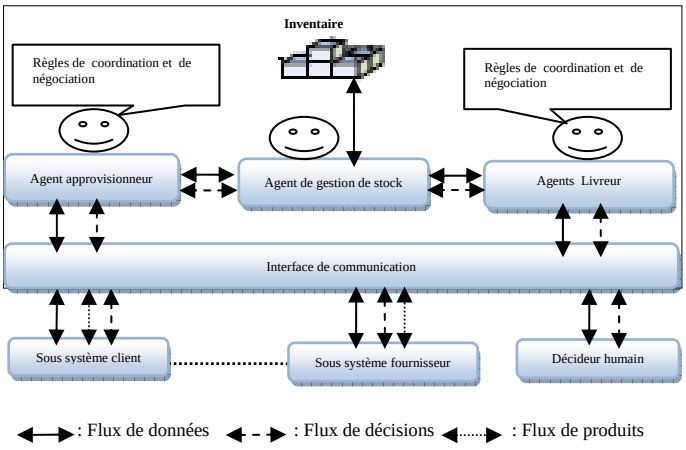


FIG. 3 – *Architecture globale du sous système.*

- Agent approvisionneur : Le rôle principal de ce type d'agent est la prise en charge de la négociation des achats et des approvisionnements du sous système. Il est en contact avec les autres types d'agents ainsi qu'avec l'ensemble des fournisseurs de l'entreprise. L'activation de ce type d'agent se produit lorsqu'un flux d'information lui est adressé.
- Agent de gestion de stock : Cet agent doit déterminer l'état du stock de l'entreprise, suivant son état, il négocie la quantité à vendre ou à acheter pour compenser et équilibrer son stock. Cet agent consulte et mit à jour donc la base de donnée d'une façon continue et périodique et guette les cas qui peuvent provoquer des situations anormales afin de prévenir et envoyer des messages aux autres agents. L'activation de cet agent est réalisée par des flux d'informations ainsi que par l'intervention d'un décideur humain.
- Agent livreur: Cet agent prend en charge les procédures de ventes (lié au processus de livraison et au processus relatifs aux clients). Il peut alors soit répondre par lui-même à une demande client soit s'appuyer sur les autres agents de la structure pour lui apporter les éléments nécessaires à la prise de décision.

Les agents des sous systèmes doivent communiquer, négocier et chercher les données exactes pour une meilleure coordination et une meilleure prise de décision. Ils peuvent générer donc des décisions concernant les différents flux à faire transiter d'une entité de la chaîne à une autre.

6 Interaction entre les sous systèmes de la CL

Dans notre proposition les sous systèmes de la chaîne peuvent collaborer et négocier selon deux types d'interaction. Le premier type concerne les acteurs appartenant au même niveau, ces derniers peuvent par exemple négocier et rééquilibrer les différents nivaux des stocks par la négociation entre les différents agents des sous systèmes. Le deuxième type qui peut avoir lieu à des sous systèmes de deux niveaux successifs, les agents de ces derniers génèrent des décisions concernant les différents types des flux. Nous donnons dans la section suivante un exemple explicatif pour détailler ces deux types d'interaction complémentaires.

6.1 Exemple d'interaction entre sous systèmes

Lorsqu'un acteur par exemple de la CL veut réapprovisionner son stock de composants par planification ou pour le besoin d'une commande ferme, il sollicite, généralement ses fournisseurs habituels (disponibles dans son carnet d'adresses). Dans le cas ou aucun ne peut y répondre totalement, l'acteur ne peut pas réaliser sa production du fait du manque de composants. Pour cela, il doit consulter les autres acteurs du même niveau afin de rééquilibrer son stock. Si un des acteurs accepte, un processus de négociation entre l'agent approvisionneur de sous système client et l'agent livreur de sous système fournisseur s'établit. Pour bien comprendre ce type d'interaction, nous donnons ci dessous un exemple explicatif :

Supposons qu'un client particulier avait placé une inattendue commande urgente qui se caractérise par deux critères :

- une quantité (Q), qui doit être livrée complètement.
- un délai de livraison (T_L), qu'on ne doit pas dépasser.

Deux cas sont alors possibles :

1- le fournisseur n'a pas complètement la quantité commandée.

2- il a une partie de la quantité commandée (Q_i) et doit compléter le reste (Q_k).

Où Q_i : Quantité disponible pouvant être délivré par le client ;

Q_k : Quantité restante qui doit être livré en respectant le délai de livraison.

Une des pratiques des fournisseurs consiste à rechercher la quantité restante d'un autre fournisseur. Dans la pratique, afin de satisfaire les besoins des clients, trois conditions doivent être vérifiées:

- Une solution rapide et automatique ;
- La confidentialité des données et l'autonomie de chaque participant doit être garantie ;
- Les coûts de transport doivent être minimisés.

Le non respect de la quantité commandée (Q) occasionne une série de coûts logistiques et administratifs supplémentaires, ruptures de stock, ventes perdues et perte de crédibilité auprès des clients. En ce moment, en profitant des caractéristiques des SMA, les trois agents de chaque sous système peuvent effectuer la recherche de la quantité commandée pour résoudre le problème.

```
// Agent de gestion de stock
Début
Si "je reçois un message de commande de l'agent livreur pour un article A (quantité Q)"
Alors "CONSULTER le niveau du stock"
   Si "état = sur-stock " Alors " ACCEPTER la commande " Fin si
   Si "état = sous-stock" Alors "ENVOYER un message à l'agent approvisionneur" Fin si
Fin si
      ETUDIER commande ( )
      PASSER commande ( )
      ANNULER commande ( )
      METTRE A JOUR stock ( )
      CALCULER prévision ( )
Fin
```

Fig. 4– *Algorithme de l'agent de gestion de stock.*

L'agent de gestion de stock consulte donc la base de donnée d'une façon continue et périodique et guette les cas qui peuvent provoquer des situations anormales afin de prévenir et envoyer des messages aux autres agents de son sous système. Il envoie des messages d'urgence à l'agent approvisionneur ou à l'agent livreur, pour un type d'article, s'il y a une situation de sur-stockage ou sous-stockage.

```
// Agent Approvisionneur
Début
Si "je reçois un message de sous-stockage de l'agent de gestion de stock  pour un article k" Alors
   "ENTRER en négociation avec les sous systèmes pour acheter la quantité demandée"
    RECEPTIONNER les offres d'appels ( )
    OPTIMISER les  coûts ( )
    CLASSER  les offres ( )
    REPONDRE aux requêtes émanant des sous systèmes selon le meilleur coût et délais ()
Fin si
Fin
```

FIG. 5 – *Algorithme de l'agent approvisionneur*.

L'agent Approvisionneur communique les quantités entrantes d'articles, C'est lui qui lance les requêtes d'acquisition des articles de rechange. Il envoie les messages aux autres sites dans le cas de sous-stockage et entre en négociation avec eux pour arriver à une meilleure solution.

```
// Agent Livreur
Début
Si " je reçois un message de sur-stockage de l'agent de gestion de stock  pour un article k" Alors
"ENVOYER un message d'offre de l'article k à mon environnement" Fin si
Si " je reçois un message de demande d'un sous système pour un article A"
Alors "ENVOYER un message à l'agent de gestion de stock"
        Si " je reçois un message d'acceptation de l'agent de gestion de stock"
       Alors " ENTRER en négociation avec le sous système concerné" Fin si
Fin si
Fin
```

Fig. 6 – *Algorithme de l'agent Livreur*.

L'agent livreur communique les quantités sortantes des différents articles. Il peut recevoir des messages de sur-stockage de l'agent de gestion de stock ou des commandes envoyées par

les agents « Approvisionneurs » des autres sous systèmes. Il peut entrer donc en négociation avec eux pour une modification d'un paramètre d'une commande (par exemple une quantité commandée ou un délai de livraison).

Nous nous sommes basés sur le fait que les trois agents de notre architecture sont coopératifs de sorte qu'ils cherchent un but global : trouver un scénario acceptable pour résoudre une situation d'urgence (dans notre exemple une commande urgente).

Cette recherche peut être résumée dans les étapes suivantes:

- Rechercher la quantité restante. Elle peut être livrée entièrement par un seul fournisseur ou recueillie à partir de plusieurs fournisseurs.
- Trouver et classer une série de chemins possibles dans un ordre croissant en fonction de coûts de transport (qui dépend étroitement de la distance) pour le transport de la quantité commandée tout en respectant le délai de livraison.

Le but principal de notre modèle est de coordonner et faire collaborer tous les agents des sous systèmes participants au même niveau afin de trouver la quantité restante et minimiser le coût total de transport quand une inattendue commande urgente est présentée.

Pour modéliser les négociations entre les agents qui composent notre exemple, nous considérons les aspects suivants:

- L'objet de négociations: Dans notre exemple, plusieurs objets peuvent être négociés : la commande et ses attributs (la quantité et la date de livraison), le contrat de livraison et ses attributs (quantité et plan de la livraison), les prévisions et leurs attributs (quantités, les dates et les exceptions).
- Le processus de prise de décision: La partie la plus importante de la prise de décision est la stratégie de la négociation qui permet à l'agent de communiquer à un certain moment, les messages pouvant être échangés entre agents sont de type : demander, refuser, informer, proposer, accepter-proposition, etc. Ci-dessous (figure 7) un exemple d'échange de messages lors du processus de négociation entre agents en se basant sur le langage de modélisation AUML (Agent UML).

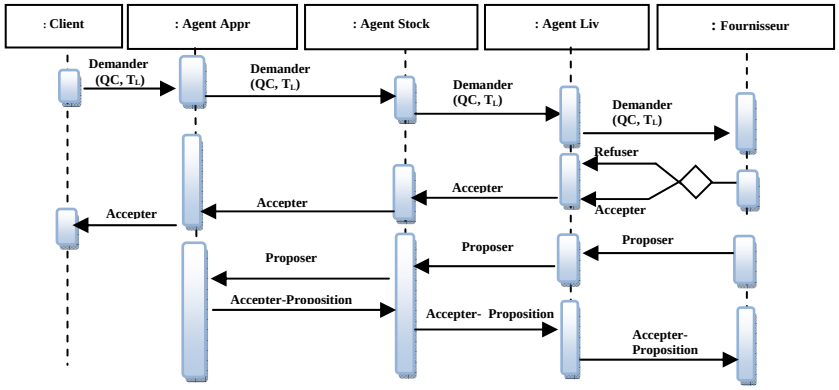


Fig. 7 – *Diagramme de séquence illustrant un exemple d'interaction.*

Comme nous avons indiqué dans notre proposition, le deuxième type d'interaction concerne les acteurs de la CL qui ont des positions différentes (relation de type client /fournisseur). Les agents des sous systèmes de deux nivaux successifs peuvent aussi générer des décisions concernant les flux de produits à faire transiter d'un acteur de la chaîne à un autre, ils peuvent communiquer leurs décisions lors du passage des commandes pour faciliter

les prévisions aux autres acteurs. Ils peuvent donc négocier et collaborer pour trouver un arrangement satisfaisant, dans ce cas, les négociateurs vont vers un compromis (négociation par exemple sur le prix d'achat entre un client et ses fournisseurs).

Notre modèle permet d'accroitre le niveau de collaboration et de négociation entre les différents acteurs de la chaîne logistique avec une diminution de l'effet de l'amplification de la commande (coup de fouet). Il ne remplace pas les outils existants et les stratégies de la CL, mais il peut être utilisé comme un complément qui les améliore en cas de la présence d'une commande urgente imprévue.

7 Conclusion et perspectives

Afin d'améliorer les performances des chaînes logistiques, les entreprises sont dans l'obligation de remettre en question leurs processus de collaboration et de partage d'information avec leurs différents partenaires. Cette problématique est l'une des plus étudiées dans le domaine des systèmes d'information et des outils d'aide à la décision appliqués à la gestion des chaînes logistiques. Notre modèle utilise le paradigme SMA pour modéliser les acteurs de la chaîne logistique, afin de bénéficier de ses caractéristiques importantes d'autonomie et de facilité la coordination. Les acteurs de la chaîne logistique dans le modèle proposé sont représentés par des sous systèmes constitués de trois types d'agents cognitifs (agent approvisionneur, agent de gestion de stock, agent livreur) qui assurent leurs fonctions premières comme le stockage, l'ordonnancement (autonomie). De plus, ces agents permettent la gestion des flux inter-acteurs (habilité sociale), cela permettrait de réduire les stocks, de rendre le système plus réactif, de diminuer la taille des commandes. Deux types d'interaction entre sous systèmes sont définies : une interaction interne qui concerne les acteurs appartenant au même niveau, une deuxième interaction externe qui peut avoir lieu à des acteurs de différents niveaux.

Notre modèle implique une plus grande coopération entre les différents agents des sous systèmes. Il permet le partage d'informations et de connaissances comme: les données sur les ventes, les prévisions de la demande, l'état des stocks. Les partenaires de la chaîne peuvent prendre des décisions conjointes basées sur l'information et les connaissances combinées. Sans ce partage d'informations, les commandes aux fournisseurs ont tendance à avoir plus de variance que les ventes à l'acheteur, et une anormalité en amont, dans une forme amplifiée «coup de fouet ». Le modèle proposé peut être dimensionné et redimensionné à tout instant selon le nombre, la localisation et les relations des acteurs, il permet de structurer et de faciliter les échanges entre les partenaires de la chaine logistique et d'avoir une vision claire de l'ensemble de la chaine.

Afin de montrer la validité de notre proposition, nous avons choisi de l'appliquer sur une étude de cas, dont des travaux sont en cours, ce qui va permettre la simulation de notre modèle. Nous envisageons dans des futurs travaux, d'enrichir notre proposition, en focalisant sur les protocoles de communication entre les différents agents des sous systèmes (pour les deux types d'interactions proposées).

Références

Anane, D., S. Pinson, S. Aknine (2009), "Les approches agents pour la coordination d'activités dans les chaînes logistiques". Cahier du lamsade, CNRS FRE 3234, Labora-

toire d'Analyse et Modélisation de Systèmes pour l'Aide à la Décision, CNRS UMR 7024, Université Paris-Dauphine, F-75016.

Brahimi, M., L. Seinturier, M. Boufaida (2009), “A Multi-Agent Architecture for Developing Cooperative E-Business Applications”. In the International Journal of Information System and Supply Chain Management, pp 43-62.

Cutting, A., B. Das, R. Young,, C.Rahimifard, , C. Anuba, N.Bouchlaghem (2006), “Supply Chaine Communication system: a review of methods and techniques”, Data Science Journal, Volume 5.

D’Amours, S., B. Chaib-Draa, T. Moyaux, (2006), “Supply Chain Management and Multi Agent System”: an Overview. In MultiAgent-Based Supply Chain Management, ISBN: 3-540-33875-6 (Springer).

Dodd, C., S. Kumara (2001),“A distributed multi-agent model for value nets”. IEA/AIE 2001, p 718-727.

Dupont, L., K.Ghedira, O. Kallel, I.Benjaafar (2008), “Résolution d’un problème de contrat réservation de capacité dans une chaîne logistique équitable”, MOSIM’08. – Paris-France. Pp : 1484-1490(Volume 2), Editions Tec&Doc – Lavoisier ISBN : 978-2-7430-1057-7.

Geune, J., M. Panos Pardalos and H. Edwin Romejin (2002), “Supply Chain management”: models, applications, and research directions, Netherlands.

Ghedira, K., L. Dupont, O. Kallel, I. Benjaafar (2008), “Multi-agent negotiation in a supply chain: case of the wholesale price contract”, 10th International Conference on Enterprise Information Systems.

Giard, V., G. Mendy (2006), “Amélioration de la synchronisation de la production sur une chaîne logistique”, Revue Française de Gestion Industrielle, vol 25, n°1, p.63-82.

Hong, Y., Y. Zhenxin, T.C. Edwin (2003), “A strategic Model for Supply Chain Design with Logical Constraints”: Computer & Operations Research, Vol.30, pp 2135-2155.

Karam, M. E., B. Tranvouez, A. Espinasse, A. Ferrarini (2010), “Agent-Based Supply Chain Simulation: Towards an Organization-Oriented Methodological Framework”. MOSIM’10 Hammamet.

Mentzer, J., W.Dewitt, S .Keebler (2001), “Définir le Supply Chain Management: Logistique & Management”, Journal of Business logistics, vol 22, n°2.

Montreuil, B., O. Labarthe, S. D’amours, D. Roy, A. Ferrarini, B. Espinasse, T. Monteiro, D. Anciaux (2008) “Simulation à base d’agents des systèmes de coordination et de planification des réseaux d’entreprises ”, dans LAVOISIER – HERMES, pp. 227 – 260.

Moyaux, T., B.Chaib-draa S.D'Amours (2008), “Spreadsheet vs. Multiagent based simulations: The case of supply chains”, International Journal of Simulation and Process Modelling . Vol 4 n°2, p 89-105.

Muller, M. (2003), “The Use of Information Technologies in Supply Chains” – A Transaction Cost Analysis, in Strategy and Organization in Supply Chains.

Nfaoui, E.H., Y. Ouzrout, O. ElBeqqali, A. Bouras **(**2008**),** "Architecture Distribuée à base d'Agents pour la Simulation Proactive et l'Aide à la Décision dans la Chaîne Logistique", MOSIM'08, 2008 – Paris- France. Pp : 1476-1484 (Volume 2), Editions Tec&Doc – Lavoisier ISBN : 978-2-7430-1057-7.

Parunak, H.V.D., S. Brueckner, J. Sauter, R. Matthews (2000), « Distinguishing Control and Plant Dynamics in Enterprise Modeling ", Proceedings of the 2nd DARPA-JFACC Symposium on Advances in Enterprise Control.

Rota, K., "Coordination temporelle de centres gerant de façon autonome des ressources. Application aux Chaînes Logistiques Intégrés en Aéronautique", Thèse de doctorat, Université de Toulouse, 2000.

Simchi-Levi, D., P.Kaminiski, E.Simchi-Levi (2000), "Desingning and Managing the Supply Chain: Concepts, Strategies and Case Strudies", Ed. McGraw Hill.

Zhang**,** Q., H. xue1, C. chen (2009), "An Adaptive Planning Model Based On Multi-Agent". Proceedings of the Seventh International Conference on Machine Learning and Cybernetics, IEEE ISBN : 978-1-4244-2096-4.

Summary

In this paper we present a distributed supply chain model. This model uses the multi agent system paradigm for modeling the various actors. Each actor in the supply chain is represented by a subsystem consists of three types of agents (Purchaser agent, manager stock agent Agent, delivery agent). The cooperation and the negotiation among these agents can synchronize the decisions taken by the different actors of a SC. The agents are therefore able to get the information from the different actors then interact and negotiate in order to propose solutions to aid decision making. Two types of interaction between subsystems are defined: an external interaction that involves coordination between the various subsystems belonging to different levels and internal interaction between the subsystems of the same level in order to collaborate, discuss and negotiate the stocks, make the system more responsive and reduce the size of orders.

Déploiement d'une Méta-modélisation Holistique pour l'aide à la prise de décision

Noureddine Falih*, Azedine Boulmakoul*
Rabia Marghoubi**

*FST Mohammedia, Département informatique, B.P. 146 Mohammedia Maroc
** INPT – 2, AV Allal EL Fasse - Madinat AL Irfane - Rabat – Maroc

Résumé. Faisant suite à notre démarche basée sur le paradigme structurel et systémique de l'entreprise pour l'alignement stratégique du Système d'Information, nous proposons, dans ce travail, le déploiement de notre Méta-modélisation holistique auprès d'une société spécialisée dans le domaine du transport au Maroc. Nous étudions en particulier la matrice structurelle Processus/Axes stratégiques où l'on va dresser les Treillis de Galois y associés en vue de dégager toute connaissance pertinente susceptible d'aider les managers à la prise de décision. Nous tirons profit essentiellement des scalogrammes de Gutman pour analyser les fermés représentant toutes les intersections possibles ainsi générées. Cette démarche pragmatique s'inscrit dans le cadre de l'alignement stratégique, partie intégrante du Management des Systèmes d'Information, qui constitue un levier stratégique décisionnel permettant à l'entreprise de mieux voir son existant pour mieux prévoir son futur.

1 Introduction

L'entreprise moderne est fortement structurée par des processus informatiques répondant aux différents processus métiers au service des objectifs stratégiques Leader et Sethi (1992). Le Système d'Information (SI) garantit la communication entre le système opérant et le système décisionnel ainsi que l'échange avec l'environnement. Les dirigeants d'entreprises, ainsi que les professionnels des SI, sont confrontés, en permanence, à la problématique de l'alignement stratégique des SI. La résolution de cette problématique est un facteur essentiel et incontournable pour prévoir et organiser les synergies SI/Métiers conformément aux orientations stratégiques et intelligences intégrées dans la gouvernance globale de l'entreprise. En vue de concrétiser notre approche de Méta-modélisation holistique évoquée lors de nos dernières recherches en la matière Boulmakoul et al. (2009), nous déployons cette démarche auprès d'une entreprise de transport au Maroc en vue de détecter les failles de synchronisation entre les processus métiers et la stratégie de l'entreprise. En effet, nous étudions en particulier la matrice structurelle Processus/Axes stratégiques où l'on va dresser les treillis de Galois y afférents pour apporter une meilleure visibilité à la question de l'alignement stratégique. L'analyse des fermés basée sur les scalogrammes de Guttman s'inscrit dans le cadre du reengineering des processus qui permet aux décideurs d'avoir une vision informationnelle plus nette pour une meilleure prise de décision.

2 Etat de l'art

2.1 Alignement stratégique des SI

L'utilisation stratégique des technologies de l'information mieux connue sous le nom de « alignement stratégique » a considérablement augmenté en raison de l'extrême dépendance de l'organisation des activités avec les SI et leurs technologies supports. L'alignement stratégique est considéré comme un élément clé pour l'amélioration de la performance de l'organisation afin d'accroître l'efficacité et l'efficience et de permettre aux entreprises d'être plus compétitives dans leurs secteurs respectifs Jouirou et Kalika (2004). Le terme «alignement stratégique» exprime l'idée d'établir et de suivre un cap. C'est de coordonner la stratégie du SI avec la stratégie de l'entreprise sur les métiers CIGREF (2009). Lederer et Sethi (1992) définissent l'alignement stratégique des SI comme étant "The process of deciding the objectives of organizational computing and identifying potential computer applications which the organization should implement''. D'autres approches définissent l'alignement stratégique selon la citation suivante : "The alignment process refers to an organizational process where the mission, goals, objectives, and activities of the IS function change over time in parallel with changes in the organization." Henderson & Venkatraman (1993). Il y a quatre grands objectifs pour s'engager dans la formulation de la planification stratégique des SI Fimbel (2007):

- Alignement: identifiant les applications informatiques susceptibles d'aider l'entreprise à atteindre ses objectifs métiers.
- Impact: recherche des applications à impact important susceptibles d'aider l'organisation à obtenir un avantage compétitif sur le marché.
- Développement d'infrastructures technologiques flexibles et efficaces,
- Développement des ressources et compétences nécessaires pour déployer le système d'information avec succès dans toute l'organisation.

2.2 Modélisation de l'entreprise

Le concept d'entreprise, tel qu'il est entendu dans le cadre de la modélisation d'entreprise, se réfère à un ensemble organisé d'activités mises en œuvre par des ressources sociotechniques dans le cadre d'une finalité identifiée. Dans de tels systèmes, la dimension financière est généralement présente, que ce soit en termes de gain, ou plutôt la consommation de ressources financières. Nous considérons l'entreprise comme un système, dans le sens systémique du terme. L'entreprise est un système qui fonctionne dans son environnement, selon des objectifs (profit, puissance, durée de vie ...), et s'organise elle-même pour les atteindre (définition de plans d'action, budgets ...) avec structures de gestion et de contrôle. L'entreprise est aussi un ensemble de sous-systèmes en interaction les uns avec les autres. La modélisation d'entreprise est un procédé incontournable d'études des organisations en vue d'améliorer leurs performances, ça permet de représenter la société, selon une abstraction multi points de vue. C'est une pratique qui garantit à l'entreprise d'être en mesure de collecter l'information et poursuivre intelligemment ses objectifs. Les efforts de recherche des années 1990 ont conduit à un cadre normalisé pour répondre aux besoins d'une approche systémique de l'entreprise ISO 19440 (2007).

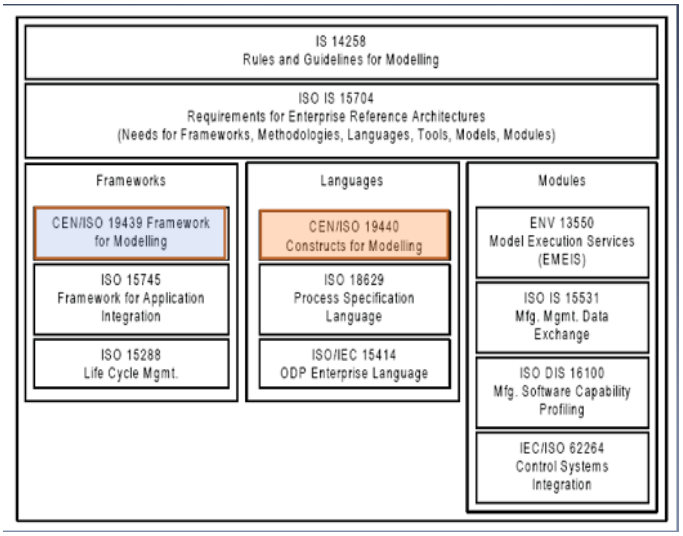


FIG. 1 – *Techniques de modélisation origines du Méta-modèle ISO/DIS 19440*
Source : www.omg.org

2.2.1 Le Méta-modèle ISO 19440

La norme ISO 19440:2007 spécifie les caractéristiques du noyau des construits nécessaires à la modélisation d'entreprise conformément à la norme ISO 19439. L'ISO 19440 définit sept phases dans le cycle de vie des modèles : la définition du domaine étudiée, la définition des concepts nécessaires, la définition des besoins de l'entreprise, la conception du modèle, la mise en œuvre du modèle, l'utilisation du modèle dans les opérations, le retrait ou l'arrêt des opérations. Elle propose quatre vue sur ces modèles, la vue organisationnelle, la vue informationnelle, la vue fonctionnelle et la vue des ressources ISO 19440 (2007). La vue informationnelle porte sur la représentation des données du SI. La vue organisationnelle est focalisée sur la stratégie de l'entreprise. La vue fonctionnelle vise les processus. La vue des ressources est liée aux ressources utilisées par les processus métiers de l'entreprise. La norme ISO/DIS 19440 propose un ensemble d'éléments de modélisation pour la représentation de l'entreprise. Il est orienté vers la modélisation par processus. Dans cette section, nous présentons le Méta-modèle proposé dans l'ISO/DSI 19440. Ce modèle est donné dans la figure 3, il intègre les quatre points de vue. Un domaine représente la frontière et les contenus d'une entreprise ou une partie d'entreprise. Un processus métier représente une certaine partie du comportement de l'entreprise. Un processus métier est une agrégation de processus métier et/ou activités de l'entreprise, ainsi que l'information décrite par les règles de gestion. L'activité d'entreprise est la réalisation des entrées aux sorties par des ressources spécifiques. L'activité d'entreprise et le processus métier sont appelés collectivement « Enterprise function ». Des règles de gestion sont utilisées pour définir le comportement d'un processus métier. Elles définissent les contraintes sur l'ordonnancement et les dépendances entre les processus métiers et/ou activités d'entreprise. Un événement lance l'exécution d'un processus métier ou activité de l'entreprise. Un type spécial de la classe événement est un ordre. Un ordre est une instruction pour l'exécution d'une activité. Ci-dessous, nous rappelons brièvement le diagramme UML du Méta-modèle ISO/DIS 19440.

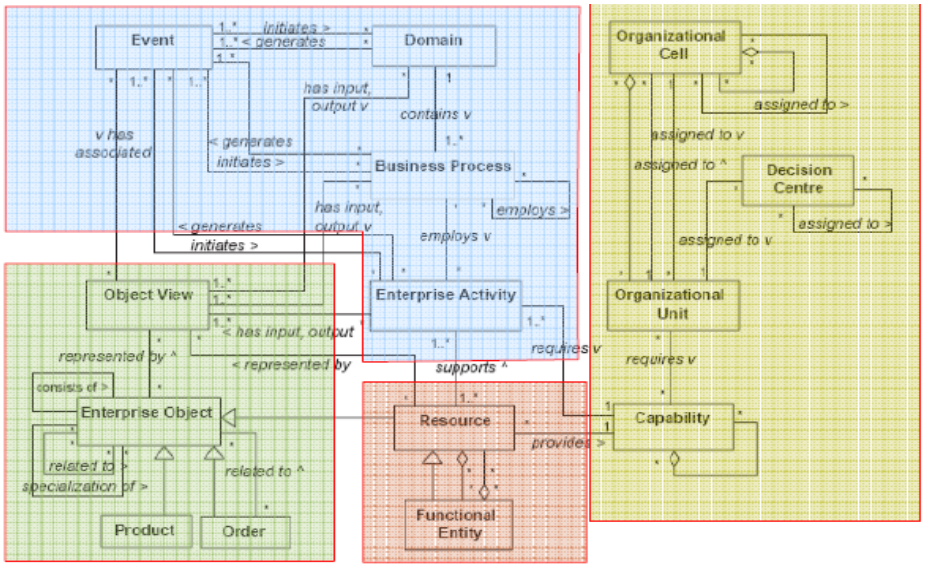


FIG. 2 – Le Méta-modèle de l'entreprise ISO 19440

Source : www.iso.org

2.2.2 Un Référentiel de bonnes pratiques appliqué au modèle de l'entreprise : COBIT

Le référentiel COBIT (Control Objectives for Information and Technology) est créé en 1996 par ISACA (Information Systems Audit and Control Association). Ce référentiel constitue un cadre de référence et un ensemble d'outils pour assurer le contrôle et le suivi de la gouvernance des SI ISACA (2008). COBIT est basé sur un ensemble de bonnes pratiques qui se propose d'établir un cadre de pilotage orienté processus afin de contribuer efficacement à l'alignement des technologies sur la stratégie de l'entreprise. COBIT est centré sur les métiers de l'entreprise, organisé par les processus et basé sur des contrôles et s'appuie systématiquement sur des mesures. Tous les composants COBIT sont reliés entre eux et visent à répondre aux besoins de gouvernance, de gestion, de contrôle et d'assurance des différents acteurs. La figure suivante illustre le degré de cohérence entre les différentes composantes de Cobit par rapport aux cinq domaines de gouvernance des TI dans l'entreprise.

	Goals	Metrics	Practices	Maturity Models
Strategic alignment	P	P		
Value delivery		P	S	P
Risk management		S	P	S
Resource management		S	P	P
Performance measurement	P	P		S

P: Primary enabler; S: secondary enabler

FIG. 3 – *Le cadre Cobit avec les domaines de « IT Governance »*

Source: www.isaca.org/cobit

Dans la suite de cet article, nous empruntons à COBIT les éléments de contrôle et de mesure des processus IT. Ces éléments sont utilisés pour l'extension de certains aspects du Méta-modèle ISO/DSI 19440 extrêmement utiles pour l'alignement stratégique des SI.

2.3 Treillis de Galois

Le treillis de concept (ou treillis de Galois) est une structure mathématique permettant de représenter les classes non disjointes sous-jacentes à un ensemble d'objets décrits à partir d'un ensemble d'attributs Birkhoff (1940). Ces classes non disjointes sont aussi appelées concepts, hyper-rectangles, ou ensembles fermés.
Soit deux ensembles finis, A et *B*, et une relation binaire, $R \subseteq A \times B'$ (Tableau 1), entre ces deux ensembles, on peut représenter par un treillis les regroupements naturels des éléments de *A* et de *B* par rapport à la relation R (Figure 4). Cette structure est appelée treillis de Galois ou treillis de concepts. Chaque élément du treillis est un couple noté (*X*,Y) composé d'un ensemble $X \in P(A)$ et d'un ensemble $Y \in P(B)$ satisfaisant aux deux propriétés suivantes:
1) $Y = f(X)$ où $f(X) = \{y \in Y \mid \forall x \in X, xRy\}$
2) $X = f'(Y)$ où $f'(Y) = \{x \in X \mid \forall y \in Y, xRy\}$.
Le *treillis de Galois* (*L*) d'une relation binaire *R* est l'ensemble de tous les couples complets dérivés de *R*. Une fermeture dans un ensemble ordonné $(A,\geq)$ est une application, $h: A \rightarrow A$, ayant les propriétés suivantes:
i) $\forall x \ \forall y, x \geq y \Rightarrow h(x) \geq h(y)$; ii) $\forall x, h(x) \geq x$; iii) $\forall x, h(h(x)) = h(x)$.
L'image $h(x)$ de x dans A est la *h-fermeture* de x; si $x = h(x)$, x est *h-fermé* ou un élément fermé pour h. Les applications $h = f$ o f' et $h' = f'$ o f sont respectivement des fermetures dans A et *B*. L'ensemble des éléments *h*-fermés de *A* correspond aux ensembles *X* des couples complets de *L* et forme un treillis isomorphe à *L*. C'est aussi vrais Symétriquement pour les éléments *h'*-fermé de *B*. Boulmakoul et al. (2007).

R	a	b	c	d	e	f	g	h
1	1	0	1	0	1	0	1	0
2	1	0	1	0	0	1	0	1
3	1	0	0	1	0	1	0	1
4	0	1	1	0	1	0	1	0
5	0	1	0	0	0	1	0	0

TAB. 1 – *Matrice binaire d'un contexte donné*

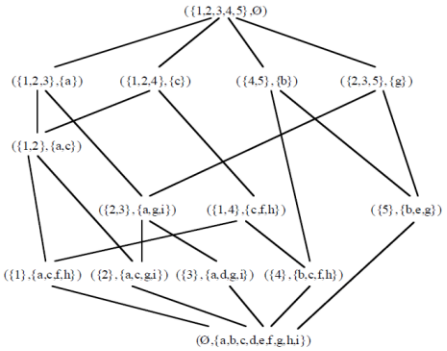


FIG. 4 – *Treillis de Galois y afférent*

Sur la base de ces concepts théoriques basés sur la systémique et le paradigme structural, nous pouvons monter divers types de couplages susceptibles d'être mesurés.

3 Vision étendue du Méta-modèle ISO/DIS 19440

Nous rappelons, dans cette section, notre approche holistique basée sur le Méta-modèle de l'entreprise ISO 19440. Nous proposons de construire une extension du Méta-modèle ISO 19440, visant à traduire explicitement la question de l'alignement des divers aspects du SI. Boulmakoul et al. (2009). Les frontières fondamentales de l'alignement se situent aux interactions et couplage des différents points de vue du Méta-modèle. Par exemple, l'interaction entre « enterprise activity » et « resource » illustre l'alignement <processus, activité | ressource>. L'interdépendance des entités resource et entreprise object situe l'alignement <ressource | information>; le couplage entre capability et resource qualifie l'alignement <organisation | ressource>. La structure du Méta-modèle de base permet donc l'expression de l'alignement du SI dans les formes décrites ci-dessus. Cependant, la formulation de l'alignement stratégique n'est pas explicite au niveau de la modélisation des quatre points de vue. Nous proposons d'utiliser les bonnes pratiques de COBIT pour le pilotage des processus IT. Ainsi, nous ajoutons le concept abstrait « objective » qui sera utilisé selon le point de vue. Le domaine d'activité de l'entreprise, les processus métiers, les activités, les centres de décisions sont contrôlés et pilotés par des objectifs (figure 6). Nous ajoutons aussi une spécialisation de Functional Entity pour modéliser les « IT processes » qui utilisent des ressources selon une connotation « IT resource » (figure 5).

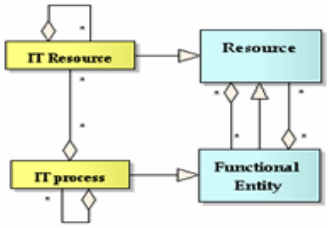


FIG. 5 – *Integration de "IT Resource"* et "IT process".

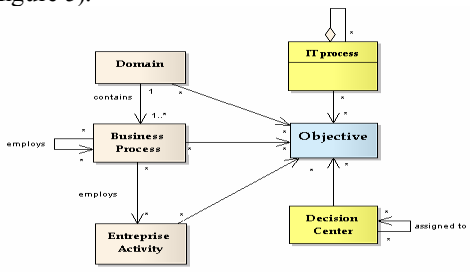


FIG. 6 – *Intégration du construit "Objective"*

Nous ajoutons aussi les construits indicators et metrics pour la mesure de la performance (figure 7).

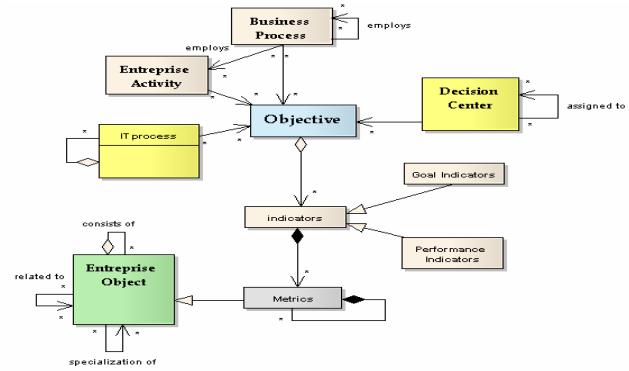


FIG. 7 – *Objectifs et indicateurs de mesure*

Par ailleurs, dans les bonnes pratiques de la systémique, les interactions entre les différentes composantes de l'entreprise peuvent être capturées par des matrices structurales appropriées, susceptibles d'apporter une vision détaillée sur l'arborescence des processus et leur agencement avec les autres constituants fondamentaux du Méta-modèle Roy (2005). Ainsi, pour les diverses problématiques d'alignement du SI, nous proposons la construction de matrices structurales permettant d'engager des analyses admises pour une meilleure vision de l'alignement stratégique du SI. Dans la figure 8, nous explicitons les construits Analyse structural et concepts dérivés pour l'évaluation de l'alignement avec des outils systémiques.

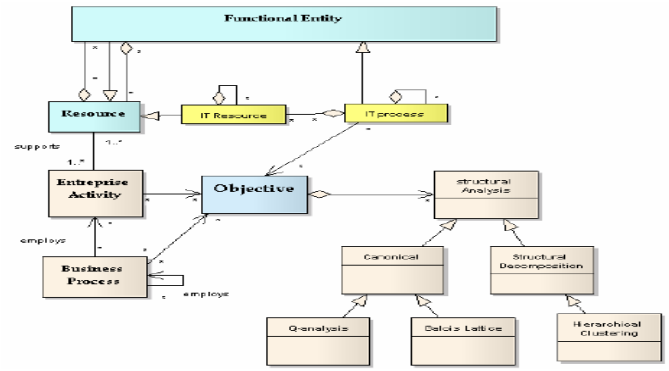


FIG. 8 – *Intégration de l'analyse structurale.*

4 Etude de cas

Dans cette partie, nous déployons notre démarche auprès d'une entreprise spécialisée dans le domaine du transport au Maroc, en vue d'un meilleur déploiement de ses processus au service des objectifs stratégiques. Les processus identifiés pour cette entreprise sont listés à titre indicatif comme suit :

	Projet
P1	Processus de Veille logistique
P2	Processus de Veille clientèle
P3	Processus de Veille Stratégique
P4	Simulation des caractéristiques des offres
P5	Prise en compte de l'historique
P6	Portail partenaires (e-partners)
P7	Suivi des moyens en adéquation p/r aux besoins
P8	Gestion de suivi du client (e-client)
P9	Gestion de la relation client voyageurs (CRM)
P10	Gestion de la relation client messagerie (CRM)
P11	Gestion de la relation client logistique (CRM)
P12	Gestion de suivi du client Logistique (e-logistique)
P13	Gestion de suivi du client messagerie (e-messagerie)
P14	Intégration de e-client, e-logistique et e-messagerie avec la circulation
P15	Activité multi-modale voyageurs (e-voyage)
P16	Echanges électroniques de documents conformes EDI
P17	Intégration du SI avec les SI des partenaires

FIG. 9 – *Liste des processus faisant marcher l'entreprise étudiée*

Les axes stratégiques relevés sont mentionnés dans la figure 10.

Axe1	Transporteur de référence au service de ses clients
Axe2	Entreprise performante et en croissance continue
Axe3	Modèle en matière de gestion de ressources humaines
Axe4	Entreprise au service de la collectivité
Axe5	Partenaire de référence pour ses fournisseurs
Axe6	Services intégrés à valeur ajoutée aux Clients
Axe7	Optimisation de la Planification et Circulation des moyens de transports
Axe8	Optimisation du Matériel et infrastructure
Axe9	Modernisation des processus de gestion
Axe10	Renforcement et modernisation du pilotage
Axe11	Renforcer l'ouverture et l'interopérabilité du SI
Axe12	Renforcer l'évolutivité, l'agilité et la sécurité du SI
Axe13	Renforcer l'urbanisation du SI (référentiels)
Axe14	Renforcer la transversalité du SI

FIG. 10 – *Liste des processus faisant marcher l'entreprise étudiée*

4.1 Choix de la matrice à étudier

Nous nous contentons d'étudier dans ce travail la matrice structurelle Processus/Axes stratégiques. Cette matrice se compose essentiellement d'un ensemble de processus qui répondent ou non à des objectifs stratégiques donnés.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15
p1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0
P2	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0
P3	1	1	0	0	0	1	1	1	0	0	1	0	0	0	0
P4	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
P5	1	0	0	0	0	1	0	0	0	0	1	1	0	0	0
P6	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0
P7	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
P8	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
P9	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0
P10	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
P11	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
P12	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
P13	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
P14	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0
P15	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0
P16	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0
P17	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0

FIG. 11 – *Matrice Processus/Axes stratégique de l'entreprise étudiée*

4.2 Analyse des concepts : Treillis de Galois

Nous intégrons notre matrice dans la solution Galicia qui est une plate-forme libre pour obtenir le treillis de Galois ci-après.

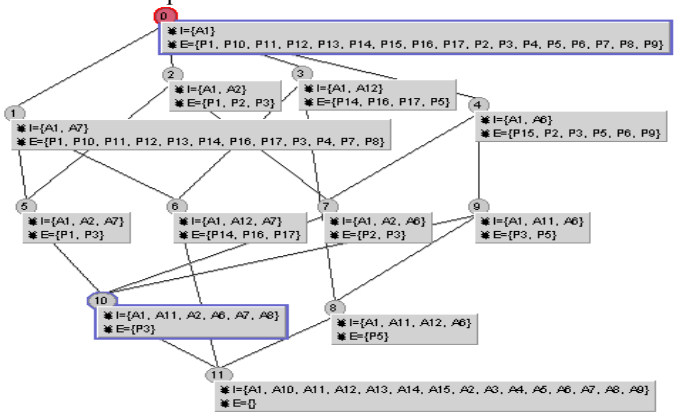


FIG. 12 – *Visualisation détaillée du treillis de Galois (processus/Axes stratégiques)*

L'analyse des fermés contribue en particulier à la Réingénierie des processus. Ses enjeux sont de booster la qualité de service pour une meilleure satisfaction clients et de réduire les coûts en améliorant la rentabilité et la productivité de l'entreprise.

4.3 Echelles de Guttman

Pendant la seconde guerre mondiale Louis Guttman, professeur de sociologie à l'université de Cornell aux USA, développe ce qu'il appelle l'analyse de scalogramme ou échelles de Guttman Guttman (1944). L'existence d'une échelle de Guttman révèle le degré d'importance des processus coïncidant avec les axes stratégiques majeurs de l'entreprise. Une chaîne maximale du treillis de Galois est une suite ordonnée de concepts allant du plus petit au plus grand concept Barbut (1965). La possibilité de situer tous les processus par coût de déploiement croissant et tous les axes stratégiques par degrés d'importance décroissant sur un même continuum orienté, constitue ce qu'on appelle, une échelle de Guttman ou un scalogramme (Figure 13).

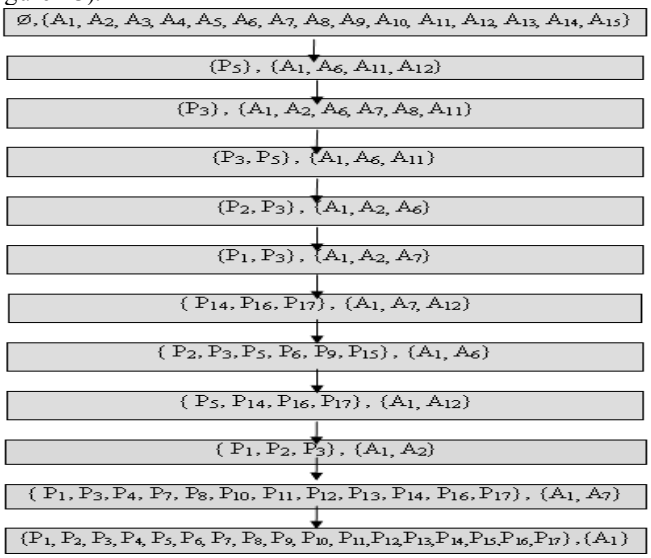


FIG. 13– *Echelle de Guttman obtenue pour une chaîne du treillis de Galois.*

4.4 Re-engineering des processus basé sur les fermés

Dans ce paragraphe nous décrivons une méthode contribuant à la Réingénierie des processus Falih et al. (2010). Cette démarche consiste à identifier les processus sans valeur ajoutée qui contribuent faiblement à la réalisation des objectifs stratégiques majeurs de l'entreprise, mais à coût de mise en œuvre très élevé. Nous reprenons la matrice binaire (Processus / Axe stratégique), mais annotée cette fois, des coût de déploiement de chaque processus par rapport au coût global de l'ensemble des processus mis en œuvre pour atteindre les objectifs stratégiques de l'entreprise.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15
p1	3%	2%	0	0	0	0	1%	0	0	0	0	0	0	0	0
P2	1%	3%	0	0	0	3%	0	0	0	0	0	0	0	0	0
P3	2%	5%	0	0	0	1%	1%	3%	0	0	2%	0	0	0	0
P4	5%	0	0	0	0	0	2%	0	0	0	0	0	0	0	0
P5	1%	0	0	0	0	2%	0	0	0	0	3%	5%	0	0	0
P6	5%	0	0	0	0	1%	0	0	0	0	0	0	0	0	0
P7	1%	0	0	0	0	0	2%	0	0	0	0	0	0	0	0
P8	2%	0	0	0	0	0	3%	0	0	0	0	0	0	0	0
P9	1%	0	0	0	0	1%	0	0	0	0	0	0	0	0	0
P10	2%	0	0	0	0	0	2%	0	0	0	0	0	0	0	0
P11	3%	0	0	0	0	0	1%	0	0	0	0	0	0	0	0
P12	2%	0	0	0	0	0	1%	0	0	0	0	0	0	0	0
P13	2%	0	0	0	0	0	2%	0	0	0	0	0	0	0	0
P14	1%	0	0	0	0	0	2%	0	0	0	0	2%	0	0	0
P15	5%	0	0	0	0	1%	0	0	0	0	0	0	0	0	0
P16	1%	0	0	0	0	0	1%	0	0	0	0	5%	0	0	0
P17	1%	0	0	0	0	0	2%	0	0	0	0	3%	0	0	0

FIG. 14 – Matrice annotée des coûts de déploiement par rapport au coût global des processus réalisant les objectifs stratégiques de l'entreprise

4.4.1 Notation :

Soient Π : ensemble des processus, Θ : ensemble des objectifs, λ l'application matérialisant le coût d'utilisation d'un processus par rapport au coût de mise en œuvre global pour réaliser l'ensemble des objectifs. $\lambda : \Pi \times \Theta \rightarrow R+$, ϖ une fonction d'agrégation, ϖ: $R+ \times R+ \times \ldots \times R+ \rightarrow R+$. Pour chaque processus Pi nous associons la mesure agrégée δ, relative au coût de déploiement par rapport au coût global de réalisation des objectifs.
$\delta(Pi) = \varpi((\lambda(Pi,R1), \ldots, \lambda(Pi,Rj), \ldots, \lambda(Pi,RN))$. La mesure normalisée μ est donnée par : $\mu(Pi) = \delta(Pi) / \Sigma(\delta(Pi))$.

4.4.2 Résultats

Calcul de $\mu(Pi) \ \forall \ Pi \in \Pi$
$\delta(P1) = \varpi((\lambda(P1,A1), \lambda(P1,A2), \lambda(P1,A7))$
Nous considérons que la fonction d'agrégation ϖ est relative à la valeur moyenne des $\lambda(Pi)$, on aura alors :
$\delta(P1) = \varpi(3\%, 2\%, 1\%) = 2\%$
De même, on calcule $\delta(P2)$, $\delta(P3)$,……,$\delta(P17)$
Par ailleurs, nous savons que $\mu(Pi) = \delta(Pi) / \Sigma(\delta(Pi)) \ \forall$ i
D'où : $\mu(P1) = \delta(P1) / \Sigma(\delta(Pi)) = 2/49 = 4\%$
Le tableau suivant récapitule tous les résultats obtenus:

Pi	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17
δ(Pi) (%)	2	3,5	7	3,5	5,5	3	1,5	2,5	1	2	2	1,5	2	2,5	3	3,5	3
μ(Pi) (%)	4%	7%	14%	7%	11%	6%	3%	5%	2%	4%	4%	3%	4%	5%	6%	7%	6%

TAB. 2 – *Valeurs de* δ et μ pour chaque processus Pi

Ordonnancement des processus selon la mesure μ
La mesure μ permet d'établir un tri décroissant des processus d'où l'ordre suivant :
P3>P5> P2≈P4≈P16> P6≈P15≈P17> P8≈P14> P1≈P10≈P1≈P13> P7≈P12> P9

Loi de Pareto

Rappelons que le fondement de la loi de Pareto revient à l'économiste italien Vilfredo Pareto qui a estimé que 80% de la richesse de ce monde était détenu par seulement 20% de la population. Dans un cadre commercial, par exemple, le principe de Pareto exprime le fait que 80% du chiffre d'affaire issu des activités commerciales est réalisé par seulement 20% de segments de clientèles.

Soit Δ la liste des processus dont les coûts sont classifiés selon la règle de PARETO.

Cet ensemble est constitué part la majorité des processus utilisant un minimum de coût pour la réalisation des objectifs assignés, conformément à la règle 80/20 de Pareto.

P3 : 14% ; P5 : 14% + 11% = 26% ; P2 ≈ P4 ≈ P16 : 14% + 11% + 7% = 33%

P6 ≈ P15 ≈ P17 : 14% + 11% + 7% + 6% = 39%

P8 ≈ P14 : 14% + 11% + 7% + 6% + 5% = 44%

P1≈P10≈P11≈P13 : 14% + 11% + 7% + 6% + 5% + 4% = 48%

P7≈P12 : 14% + 11% + 7% + 6% + 5% + 4% + 3% = 51%

P9 : 14% + 11% + 7% + 6% + 5% + 4% + 3% + 2% = 53%

Δ = {P5, P2, P4, P16, P6, P15, P17, P8, P14, P1, P10, P11, P13, P7, P12, P9} : ensemble des processus réalisant des objectifs assignés avec des coûts modérés.

Soit Λ = Π-Δ : l'ensemble des processus coûteux contribuant faiblement à la réalisation des objectifs. On a dans notre cas Λ = {P3}

Chaîne maximale de Guttman

Nous nous intéressons, en particuliers, aux fermés les plus consommateurs en termes de budget. Or, nous avons trouvé que Λ = {P3}, on en déduit le chemin des fermés contenant P3 selon une chaîne maximale de Guttman (Figure 15).

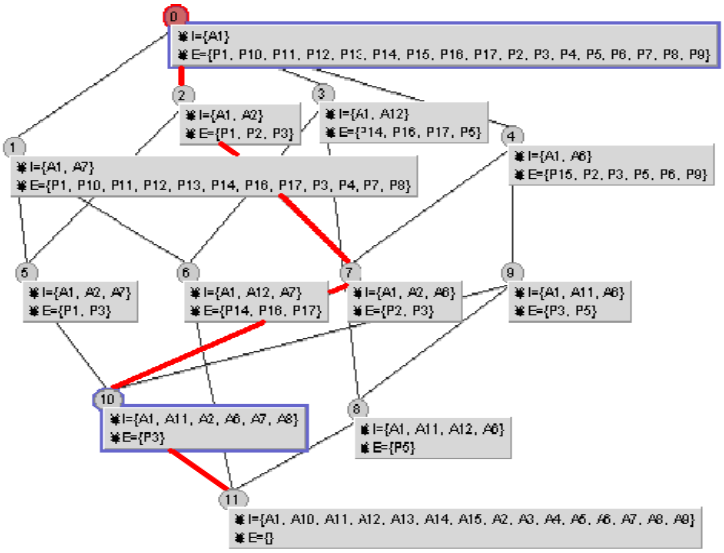


FIG. 15 – *Chaîne de Guttman relative au Processus P3*

4.4.3 Analyse

Pour le cas de notre entreprise, l'analyse des fermés nous mène à la conquête des entités organisationnelles déployant le processus Métier P3, en l'occurrence, le processus « Veille stratégique » rattaché à la direction Marketing. Pour cela on se propose de demander aux hautes instances de l'organisation de :

- Lister les principales rubriques du budget fonctionnel et d'investissement de la Direction Marketing;
- Définir la procédure suivie pour l'établissement du budget de chacun des processus ;
- Identifier, en termes de pourcentage par rapport au budget global, le budget annuel alloué au processus P3;
- Etudier la possibilité de fusionnement du processus « veille stratégique » avec d'autres processus, sans pour autant défavoriser la réalisation des objectifs assignés;
- Réviser les structures organisationnelles ;
- Rationaliser les coûts et optimiser les dépenses ;

4.1.1 Perspectives

Cette démarche est susceptible d'apporter une nouvelle vision de la Réingénierie des processus s'inscrivant dans le cadre de l'ingénierie des SI. L'analyse structurale des fermés aura pour objectif de revoir le Méta-modèle de l'entreprise sous sa forme fonctionnelle en vue de réviser la structuration des entités susceptibles de nuire à la performance globale de l'entreprise par une consommation gourmande en termes de budget. La méthodologie proposée a été mise en œuvre sur une grille d'impacts processus × objectifs stratégique. La démarche est fondée et reste à être validée par les acteurs du pilotage et du tableau de bord de l'organisation. D'autres matrices structurales pourront être étudiée dans des travaux futurs pour renforcer la bibliothèque des méta-connaissances utiles aux organisations sollicitant une aide à la prise de décision et un meilleur alignement de leurs SI en vue de conserver leur pertinence et leur statut dans un contexte de plus en plus concurrentiel.

Références

Barbut M. (1965), « Note sur l'algèbre des techniques d'analyse hiérarchique », B. Matalon (éd.), L'analyse hiérarchique, Paris, Gauthier- Villars

Birkhoff G. (1940), Lattice Theory (1e éd.), Providence (RI),

Boulmakoul A., Falih N. et Marghoubi R. (2009) ''Meta-Modelling and Structural Paradigm for Strategic Alignment of Information Systems'', In Information Society Research, Education, Policy and Practice in the Mediterranean Region, University of Economics and Business, Athens-Greece. ISBN : 978-9609-8566-7-6

Boulmakoul, A., Idri, A., Marghoubi, R. (2007), «Closed frequent itemsets mining and structuring association rules based on Q-analysis", in: Signal Processing and Information Technology, 2007 IEEE International Symposium on Publication Date: 15-18 Dec. 2007, page(s): 519-524, ISBN: 978-1-4244-1834-3

CIGREF (2009), "SI éco-responsables : L'usage des TIC au service de l'entreprise durable"

Falih, N., Boulmakoul, A., Marghoubi, R. (2010) ''Deploying Holistic Meta-modeling for Strategic Information System Alignment'', The International Arab Conference on Information Technology ACIT, University of Garyounis, Benghazi, Libya.
Fimbel E. (2007), "Alignement stratégique: Synchroniser les SI avec les trajectoires et manoeuvres des entreprises " Edition Pearson Education, Paris, ISBN : 978-2-7440-7226-0
Guttman, L., 1944. A basis for scaling qualitative data, American Sociological Review, 9: 139-150.
Henderson J.C., Venkatraman N., (1993), "Strategic Alignment: Leveraging IT for Transforming Organizations", IBM Systems Journal, Vol. 32 No. 1, pp. 4-16.
ISACA (2008), "Cobit 4.1", http://www.isaca.org/cobit
ISO 19440 (2007), "Enterprise integration -- Constructs for enterprise Modeling", Edition 1
Jouirou, N. et Kalika, M. (2004). L'alignement stratégique déterminant de la performance. In 9ème colloque de l'Association Informatique et Management – AIM2004.
Lederer, A.N. and Sethi, V, 1992. Root Causes of Strategic Information Systems Planning Implementation Problems. Journal of IS Management, Vol 9. N°1, pp: 25-45
OMG (2005), "BPDM: Business Process Definition Metamodel". http://www.omg.org
Roy B. (2005). Paradigms and Challenges dans Multiple criteria decision analysis: State of the Art Surveys Figueira, Greco et Ehrgott éditeurs,. Springer's international Series 3-24.

Summary

Following our approach based on the structural and systemic paradigm for strategic Information System alignment, we propose in this work, the deployment of our holistic meta-modeling in moroccan transport company. We study in particular the structural matrix Process / Strategic objectives in which we will develop the concept lattice associated to identify main knowledge that can assist managers in decision-making. We use Gutman scalograms to analyze the closed representing all possible intersections generated. This pragmatic approach is part of the strategic alignment, which is a main part of Information Systems Management. It´s considered as a strategic support for decision-making in the company.

Approche et Outil d'Aide à la Décision pour la Maintenance des Systèmes à Objets

M. Z. Dinedane*, M. K. Abdi**
Département d'informatique, Faculté de Sciences, Université d'Oran , 31000, Oran, Algérie
din_danos@hotmail.fr*, abdimk@yahoo.fr**

Résumé. Plusieurs travaux ont été proposés pour étudier la changeabilité ainsi que l'évolution des systèmes logiciels à objets. Cependant, peu de travaux qui ont traité la problématique d'analyse d'impact de changement, qui est en fait une technique parmi d'autres pour la maintenance de logiciels. Encore moins, sont les travaux qui ont essayé de porter de l'aide dans cette phase de maintenance.
L'objectif de ce travail est de proposer une approche et un outil d'aide pour la prise de décision lors de la phase de maintenance d'un système logiciel à objets. Pour cela, nous avons récolté des métriques de couplage sur un système test, puis nous avons porté des changements concrets afin de résoudre un problème au niveau de ce système. Ensuite, nous avons fait appel à la méthode ELECTRE III pour la prise de décision à propos de la solution à adopter parmi plusieurs alternatives. Les résultats de l'approche proposée ont confirmé certains résultats trouvés par d'autres approches.

Mots clés. Aide à la décision, analyse d'impact, analyse multicritère, maintenance, systèmes à objet, métriques de couplage.

1 Introduction

La maintenance est la dernière phase du cycle de vie d'un logiciel, elle est la phase la plus coûteuse. Selon Pfleeger et Shawn (1990), le coût de la maintenance dépend du degré de dépendance entre les entités d'une architecture logicielle. Un changement peut avoir des effets considérables et inattendus sur le reste du système. Le danger encouru lors d'une modification est la propagation du changement. De ce fait, il est préférable d'avoir une idée sur l'architecture du logiciel pour estimer l'impact de changement et ainsi réduire le coût de la maintenance. La modularité est considérée comme un critère important de la qualité du logiciel. Un produit logiciel est dit modulaire si ses composants présentent un faible degré de couplage. Dans le cadre des applications orientées objets (OO), il existe différents types de couplage entre classes.

Mesurer ces types de relations permet de mieux comprendre le lien qui existe entre le couplage des classes et les attributs de qualité. Nous définissons un changement dans un programme comme une modification apportée à un de ses éléments (classe, méthode ou variable). Ainsi, l'impact est vu dans notre contexte comme la conséquence d'un changement. L'analyse de l'impact est une activité dont l'objectif est de déterminer l'étendue d'une requête de changement. Elle estime les éléments affectés, au niveau du code source. Plus une classe est couplée avec d'autres classes, plus elle est sensible aux changements effectués dans ces classes et plus elle est susceptible de subir des erreurs.

1.1 Problématique et contribution

Avec l'évolution des systèmes logiciels, un flot important de changements doit être pris en considération ainsi que leur propagation sur le reste du système. Réussir à modifier un logiciel de façon disciplinée toute en maintenant son fonctionnement et son intégrité avec un coût raisonnable nécessite une analyse de l'impact du changement avant son implémentation. En effet, l'équipe de maintenance doit être en mesure de fournir des réponses aux questions suivantes : De quel type de changement s'agit-il ?, Quelle est l'étendue du changement ?

Dans la gestion des projets logiciels, plusieurs changements sont proposés pour résoudre le même problème au niveau d'un code et satisfaire le même besoin de l'utilisateur d'un système logiciel. L'utilisation d'analyse d'impact de changement permet de choisir le changement adéquat en prenant en considération deux aspects : la nature du changement, et son impact sur le reste du système. Vu que plus un changement se propage, plus le coût de sa maintenance augmente.

A cet effet, nous proposons dans cet article une approche d'aide à la décision pour les gestionnaires de maintenance des logiciels orientés objet dans leur choix de la meilleure solution parmi plusieurs proposées en utilisant la méthode ELECTREIII et en se basant sur certaines métriques de couplage comme indicateurs d'impact de changement.

La section 2 fait un tour d'horizon sur l'analyse d'impact de changement. La section 3 présente l'aide a la décision multicritère, et la quatrième est réservée à la mise en ouvre de l'outil. Les perspectives de notre travail sont discutées en conclusion.

2 Analyse d'impact de changement

Un impact, dans son sens le plus large, est l'effet qu'une chose a sur une autre. L'impact est souvent la conséquence d'un changement. L'analyse d'impact de changement est la détermination de la portée d'une modification, i.e., l'évaluation des éléments d'un système qui seront affectés par cette modification.

Le couplage mesure la force de l'interconnexion entre les modules d'un système. Durant le processus de maintenance de logiciel, le couplage prédit la difficulté de changement des modules du programme et les implications dans les autres modules (Lounis et Melo, 1997). Le couplage réfère au degré d'interdépendance entre les parties d'un programme. Plus une classe est couplée à d'autres classes, plus importante est sa sensibilité aux changements dans ces classes.

Un logiciel de bonne qualité doit obéir au principe de faible couplage (Chidamber et Kemerer, 1994), (Lounis et Melo, 1997). Un couplage faible facilite la maintenance vu que les dépendances entre les classes sont minimes.

3 Aide à la décision multicritère

L'objectif de l'Aide à la Décision Multicritère est d'aider un décideur à sélectionner une alternative parmi plusieurs sur la base de critères de décision. Dans l'approche dite constructive, une procédure interactive est alors utilisée pour aider le décideur à forger ses convictions quant à la façon de choisir la meilleure alternative. Une telle procédure repose

sur un modèle des préférences. La difficulté principale du choix du modèle réside dans le fait qu'il pourrait y avoir des contraintes contradictoires.

3.1 ELECTRE III

C'est une méthode de sur-classement, qui date de 1977, (Roy, 1977), et qui vise à résoudre la problématique de type gamma : classer les actions de la meilleure à la pire. Pour ce faire, ELECTREIII traite une matrice de performance contenant des actions et des critères. Les traitements de sur-classement munis sur cette matrice permettront d'établir une liste de pré-ordre partiel. ELECTREIII est composée de deux phases: la phase d'agrégation et la phase d'exploitation.

La phase d'agrégation : avant le début de traitement d'ELECTRE III, le décideur est invité à introduire les paramètres subjectifs de la méthode relativement à chaque critère (Roy, 1977): poids des actions, seuil de préférence, seuil d'indifférence, et le seuil de véto. Cette phase a pour objectif de construire les relations de sur-classements.

La phase d'exploitation : la méthode ELECTRE III propose de classer les actions selon un algorithme. Les actions sont comparées entre elles par paire selon tous les critères.

4 Approche proposée

L'approche que nous proposons dans le cadre de ce travail est divisée en deux parties, la première consiste à l'extraction de certaines métriques de couplage du code source Java du système en entrée, et la deuxième consiste à l'application de la méthode ELECTREIII.

L'identification et la compréhension des changements qui peuvent être apportés aux applications orientées objets s'avèrent importantes et fructueuses. Pour cela, on s'est inspiré du travail réalisé dans (Cheikhi, 2004). Nous avons repris un questionnaire visant à rassembler les perceptions des gens qui assurent la maintenance des logiciels à objets pour avoir une connaissance sur la nature des changements. Les changements sont groupés par classe, méthode, et attribut, et le résultat de ce questionnaire est soit le changement est souvent, rare, ou jamais.

Nous allons calculer les métriques de couplage concernées de toutes les classes à partir d'un code source Java, puis nous proposons quelques changements dont l'objectif de résoudre un problème au niveau du code. L'intersection entre les changements et les métriques calculées des classes portant le changement définit une matrice. Cette matrice va être utilisée comme entrée dans la méthode ELECTRE III. Le résultat final sera un rangement des ces changements du meilleur au moins bon.

4.1 Extraction de métriques

Notre objectif consiste à extraire des métriques capturant les caractéristiques importantes comme le couplage. Nous avons choisi la propriété de couplage pour deux raisons :

- Avoir une idée sur la qualité du système en termes de couplage, et cela permet d'estimer la propagation de changement dans le système.
- Exploiter ces métriques comme critères dans la matrice de performance.

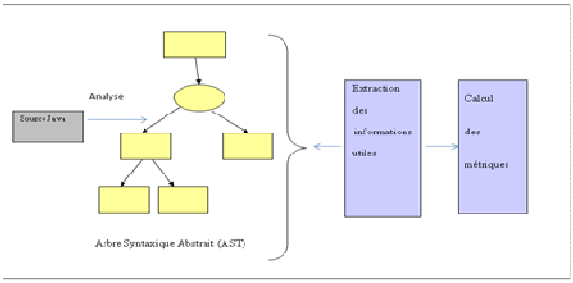


FIG. 1 – *Processus d'analyse et d'extraction de métriques*

Le processus décrit en figure1 permet d'analyser un fichier source Java et fournir son arbre syntaxique (AST) qui correspond à un ensemble de nœuds représentant tous les constituants de la classe jusqu'aux instructions; c'est une structure arborescente d'objet.

Nous avons développé un outil sous Eclipse qui permet d'analyser un code source java et d'extraire toutes les données importantes pour le calcul des métriques de couplage considérées pour chaque classe du système, cette analyse est assurée par plugin ASTParser. Les métriques de chaque classe portant le changement vont être exploitées comme critères dans la matrice de performance.

Pour un problème au niveau d'un logiciel écrit en Java, ou un nouveau besoin de l'utilisateur de ce dernier, plusieurs solutions (changements au niveau du code source) sont proposées afin de répondre au même besoin, l'analyse d'impact permet de choisir la solution adéquate, et la moins coûteuse, en prenant en considération deux aspects: la nature de changement et son impact sur le reste du système : vu que plus un changement se propage, plus le coût de la maintenance augmente. Donc l'analyse d'impact peut à notre avis être considérée comme un outil d'aide à la décision.

Dans notre travail, nous attribuons des indices à chaque changement, pour le changement jamais (indice 0), le changement rare (indice 1), et changement souvent (indice 2).

4.2 Application d'ELECTRE III

Notons qu'il est possible qu'un changement dont la nature est «souvent » soit porté sur une classe fortement couplée. Comme il peut y avoir un changement dont la nature est « rare » soit porté sur une classe faiblement couplée. D'où la nécessité de faire appel à la méthode ELECTRE III.

ELECTRE III traite une matrice d'évaluation contenant des actions et des pseudo critères. Les traitements de sur-classement munis sur cette matrice permettront d'établir un pré-ordre final partiel (Maystre, 1994).

La matrice de performance dans notre cas comprend dans les lignes les changements proposés ou les solutions pour la résolution d'un problème (les actions), et dans les colonnes on a les différentes métriques des classes portant le changement, plus la nature de changement (les critères). Par exemple si la solution1 (changement de signature d'une méthode de la classe4), les critères seront les différentes métriques de la classe4 plus la nature de ce changement.

Le décideur est invité ensuite à saisir ses préférences (les poids des critères et les différents seuils). Ces données vont servir aux calculs des deux phases d'ELECTRE III qui affichent à la fin un rangement de changements.

5 Expérimentation

Pour l'expérimentation de notre approche, nous avons pris un système réel de taille assez considérable, le système BOAP (EL Hachemi et Snoussi, 2002): il s'agit d'une application développée avec le langage Java, au laboratoire CRIM (Centre de Recherche en Informatique de Montréal). BOAP (Boite à Outils pour l'Analyse des Programmes), est un ensemble d'outils logiciels intégrés qui permet à un expert d'évaluer rapidement le niveau de qualité d'un logiciel. Nous avons choisi BOAP parce que tout d'abord, nous l'avons déjà utilisé comme système test dans nos travaux antérieurs (Abdi, 2007), et (Abdi et al., 2009) pour vérifier certaines hypothèses avec d'autres objectifs bien sure, et d'autre part, parce qu'il a été utilisé aussi dans (Cheikhi, 2004) afin d'estimer l'impact de changement en utilisant les algorithmes d'apprentissage automatique, et nous voulons dans le cadre de ce travail comparer nos résultats avec ceux obtenus dans (Cheikhi, 2004).

Les caractéristiques de BOAP sont présentées dans le tableau suivant :

Caractéristiques	**Système BOAP**
Nombre de fichiers	424
Nombre de modules	22
Nombre de classes indépendantes	103
Nombre de classes	394
Nombre de classes de base	117
Nombre de méthodes	3546
Nombre d'attributs	2247

TAB. 1 – *Caractéristiques du système étudié(BOAP)*

DAC	Data Abstraction Coupling: Cette métrique représente le nombre d'attributs non hérités, dont les types sont des classes définies dans l'application.
CBO	Coupling Between Object: Représente le nombre de classes avec lesquelles une classe est couplée.
RFC	Response For a Class : Représente le nombre de méthodes invoquées en réponse à un message.
MPC	Message Passing Coupling : Compte seulement les invocations de méthodes des autres classes.
AMMIC	Ancestors Method -Method Import Coupling: Correspond au nombre de classes parentes avec lesquelles une classe a une interaction de type méthode-méthode et un couplage de type IC.
OMMIC	Others Method -Method Import Coupling: Correspond au nombre de classes (autres que les super-classes et les sous-classes) avec lesquelles une classe a une interaction de type méthode-méthode et un couplage de type IC.
DMMEC	Descendants Method - Method Export Coupling: Correspond au nombre de sous-classes avec lesquelles une classe a une interaction de type méthode-méthode et un couplage de type EC.
OMMEC	Others Method - Method Export Coupling : Correspond au nombre de classes (autres que les super-classes et les sous-classes) avec lesquelles une classe a une interaction de type méthode-méthode et un couplage de type EC.
CBO'	Coupling Between Object: compte le nombre de classes avec lesquelles une classe c est couplée et n'ayant pas une relation d'héritage.
CBO-IUB	CBO Is Used By : Cette métrique consiste au nombre de classes utilisant la classe cible
CBO-U	CBO Using: Représente la partie du CBO qui s'intéresse aux classes utilisées par la classe cible c

TAB. 2 – *liste des métriques concernées*

5.1 Mise en œuvre de l'outil

L'option extraction (Figure 2) permet de parcourir le code source et donner son AST.

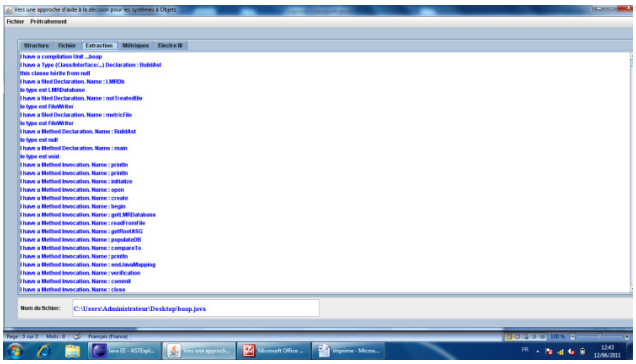


FIG. 2 – *Extraction de l'AST*

Après l'obtention de l'AST, on va procéder à la structuration en classes, méthodes, attributs, héritage, invocation de méthode,… etc.

L'option structure (figure3) permet d'afficher cette structuration du code source sous une forme exploitable afin de faciliter la manipulation.

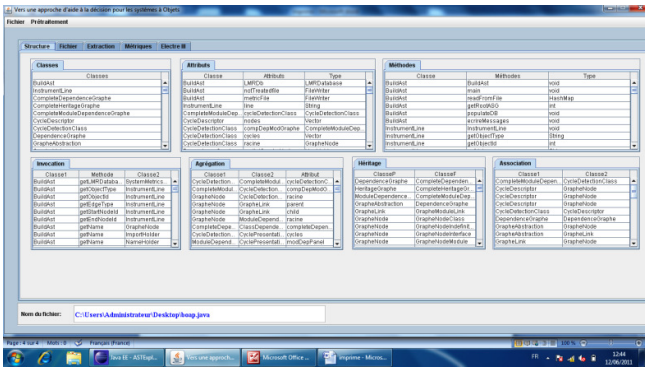


FIG. 3 – *Structure du code source*

Le calcul des métriques est fait sur la base des informations extraites dans la figure 3.

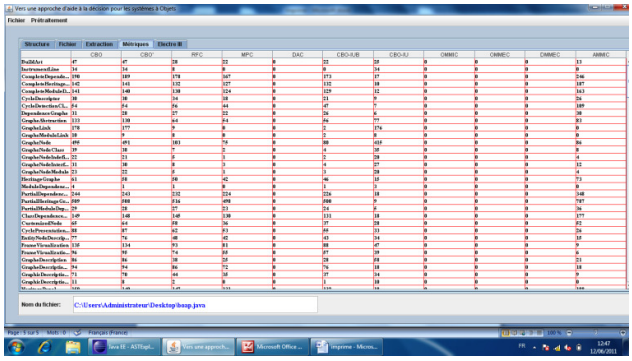


Fig. 4 – Calcul des différentes métriques

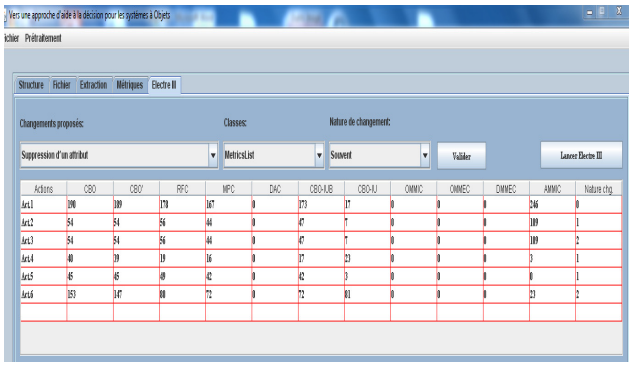


Fig. 5 – Matrice de performance

Les changements proposés pour notre expérimentation pour un besoin de maintenance du système sont :

- Ajout d'une méthode à la classe BuildAst.
- Ajout d'une superclasse à la classe DependancePanel.
- Changement de type d'un attribut de la classe GraphModuleLink.
- Changement de la portée d'une méthode de la classe – CompleteModule
- Changement d'implémentation d'une méthode de la classe laModuleDependance.
- Suppression de la classe partielHeritageGraph.

Nous avons supposé que ces six changements résoudront le même problème. Donc quel est le meilleur en tenant en compte de la nature du changement et de sa propagation sur le reste du système ?

Fig. 6 – Rangement des résultats

En comparant nos résultats avec ceux dans (Cheikhi, 2004), nous remarquons que les changements proposés par ELECTRE III comme meilleurs changements reflètent les règles obtenus dans (Cheikhi, 2004). Le rangement final de notre outil affiche que les changements 1, 3, et 5 sont les meilleurs changements ; et le changement 6 est le mauvais changement.

Nous avons repris quelques changements utilisés dans (Cheikhi, 2004) pour les comparer avec nos résultats.

Changement1 : changement d'implémentation de méthode
- Jeu de métriques : CBO, CBO', RFC, OMMEC, DAC
Nombre d'instances : 88, Algorithme : J48, Taux de succès : 0.56
Règle: (CBO >= 22) => chang1= impact grand (13.0/3.0) sinon chang1= impact faible.

Pour le changement 5 (changement d'implémentation d'une méthode de la classe ModuleDependance), nous remarquons que le CBO de la classe est 4, et que le changement est classé comme meilleure changement donc a un impact faible.

Changement2 : changement de la portée de la méthode de publique à privée
- Jeu de métriques: CBO, CBO', RFC, DIT, NOC
Nombre d'instances : 38, Algorithme : J48, Taux de succès : 0.75
Règles: CBO' <= 2 : impact faible (8.0)
CBO' > 2 : impact grand (30.0/11.0)

Pour le changement 4 (changement de la portée d'une méthode de la classe CompleteModule), nous remarquons que le CBO' de la classe est 141, donc ce changement a un impact grand, et par conséquent elle n'a pas été classée comme meilleure solution dans notre résultat.

Changement3 : changement du type d'attribut
- Jeu de métriques : OMMEC, DMMEC, OCMEC, OCAEC, DCMEC, DCAEC
Nombre d'instances : 21, Algorithme : J48, Taux de succès : 0.95
Règle : (DMMEC >= 19) => chang3= impact grand (2.0/0.0)

Pour le changement 3 (changement de type d'un attribut de la classe GraphModuleLink), le changement est classé parmi les meilleurs changements, et nous remarquons que le DMMEC de la classe est 0, donc ce changement a un impact faible.

6 Conclusion

L'analyse d'impact est une technique qui permet d'estimer les effets des changements afin de diminuer le coût croissant de la maintenance des systèmes logiciels. L'utilisation des métriques de couplage comme indicateur d'impact de changement a été validé dans plusieurs travaux (Chaumun et al., 2000), (Abdi, 2007), (Cheikhi, 2004), et (Abdi et al., 2009). Le problème auquel est confronté le chef de projet de maintenance réside dans le choix de la solution optimale parmi plusieurs solutions proposées en considérant plusieurs critères souvent contradictoires. Nous avons proposé dans cet article un outil qui permet de calculer certaines métriques de couplage à partir d'un système en entrée et d'aider le décideur dans son choix de la solution à adopter (ie, le changement optimum concernant ce système). Dans nos travaux futurs, nous envisageons faire d'autres expérimentations sur d'autres systèmes de natures différentes en considérant d'autres métriques (de cohésion, d'héritage,…).

Références

Abdi, M. K (2007) "*Analyse et Prédiction d'impact de Changement dans un système à objet*", Thèse de Doctorat d'Etat en Informatique, Université Es-Sénia d'Oran.

Abdi, M.K. H. Lounis, H. Sahraoui (2009) "*Predicting Change Impact in Object-Oriented Applications with Bayesian Networks*" in proceedings of the 33rd Annual IEEE International Computer Software and Applications Conference (COMPSAC2009), Seattle, Washington USA.

Chaumun, M. A. H. Kabaili, R. K. Keller et F. Lustman (1999) "*A Change Impact Model for Changeability Assessment in Object-Oriented Software Systems*". In Proceedings of the Third European Working Conference on Software Maintenance and Reengineering, Pages 130-138, Amsterdam, the Netherlands.

Chaumun, M. A. H. Kabaili, R. Keller, F. Lustman, et G. Denis (2000) "*Design Properties and Object-Oriented Software Changeability*". Fourth European Conference on Software Maintenance and Reengineering, Zurich, Switzerland, Pages 45-54.

Cheikhi, L (2004) "*Estimation de l'impact de changement dans les programmes a objet*". Thèse de Master, Département d'Informatique et de Recherche Opérationnelle, Montréal.

Chidamber, S. et C. Kemerer (1994) "*A Metrics Suite for Object- Oriented Design. IEEE Transaction on Software Engineering*", Vol. 20, No. 6, Pages 476-493.

(EL Hachemi et Snoussi, 2002) : Alikacem EL Hachemi, Hicham Snoussi, *"BOAP 1.1.0 : Manuel d'utilisation",* CRIM, Janvier 2002.

Li, W. et S. Henry (1993) "*Maintenance Metrics for the Object-Oriented Paradigm*". Proceedings of the First International Software Metrics Symposium, Pages 52–60.

Lounis, H. W.L.Melo (1997) "*Identifying and Measuring Coupling on Modular Systems*". 8th International Conferenceon Software Technology (ICST97).

Maystre, L. J. Pictet, J. Simos (1994) "*Méthodes multicritères Electre* ", Presses Polytechniques et universitaires Romandes, Lausanne, Suisse.

Pfleeger, S. L. et Shawn A. Bohner (1990) "*A Framework for Software Maintenance Metrics*". In proceedings of the Conference on Software Engineering, Pages 320-327.

Roy, B (1977) "*Electre III, un algorithme de classement fondé sur une représentation floue des préférences en présence de critères multiple*", rapport de recherche.

Summary

Several studies have been proposed to study the changeability and the evolution of objects software systems. However, few studies have addressed the problem of change impact analysis, which is actually a technique among others for the maintenance of software. Even fewer, are the work that has tried to bring help in the maintenance phase.

The objective of this work is to propose an approach and a tool for decision making during the maintenance phase of an objects software system. For this, we collected coupling metrics on a test system, and then we brought real change to solve a problem with this system. Next, we used the method ELECTRE III to make a decision about the best solution among several alternatives. The results of the proposed approach confirmed some results found by other approaches.

Keywords. Decision support, impact analysis, multicriteria analysis, maintenance, object oriented systems, coupling metrics.

A Multi-Agent Model for Web-based Collaborative Decision Support Systems

Abdelkader Adla, Bakhta Nachet

Department of Computer Science, University of Oran
Oran, Algeria
{adla.abdelkader, nachet.bakhta}@univ-oran.dz

Abstract. This article takes a multi-agent view of the web-based collaborative decision making process and examines the potential integration of agent technology into a distributed group decision support systems. We propose a Multi-agent model for web-based collaborative decision support system in which a facilitator and group decision makers are supported by agents. It considers group participants as multiple agents concerned with the quality of the collaborative decision. We define a facilitator agent as that agent responsible for the overall decision making process. This includes managing the complex negotiation processes that are required among those participants collaborating on decision making.
The use and the integration of software agents in the decision support systems provide cost-effective means for making decisions and automating more tasks for the decision maker, enabling more indirect management, and requiring less direct manipulation of the DSS. Specifically, agents were used to collect information and generate alternatives that would allow the user to focus on solutions that were found to be significant. The agents in the system autonomously plan and pursue their actions and sub-goals to cooperate, coordinate, and negotiate with others, and to respond flexibly and intelligently to dynamic and unpredictable situations.
The decision making process, applied to the boilers defects in an oil plant, relies on a cycle that includes recognition of the causes of a defect (diagnosis), plan actions to solve the incidences and, execution of the selected actions.

Keywords: Decision support, Collaborative decision making, Web-based decision support systems, Multi-agent systems.

1 Introduction

As organizations seek to adapt in a world of rapid change, decision making becomes increasingly dynamic and complex. Collaborative decision support systems provide a means by which a larger number of organizational stakeholders can efficiently and effectively participate in the decision making process. A greater number of organizational members participating in the decision making process logically leads to a better decision. The resulting decision should benefit by the richness of knowledge provided by the greater representation

of organizational members. A success factor critical to this involvement is the successful organization of massive amounts of information generated by such a group.

On the other hand, the Distributed Artificial Intelligence (DAI), which is commonly implemented in the form of intelligent agents, offers considerable potential for the development of information systems and in particular Decision Support Systems (DSS). Widely range applications domains, in which agent solution is suggested, are being applied or investigated [Cheung, 2005].This is because of the reason that intelligent agents have a high degree of self-determination capabilities, and they can decide for themselves when, where, and under what condition their action should be performed. Intelligent agents have the promise to provide timely assistance in various areas of such environments as information gathering, information dissemination, monitoring of team progress and alerting the team to various unexpected events.

This article takes a multi-agent view of the web-based collaborative decision making process and examines the potential integration of agent technology into a distributed group decision support systems. It considers group participants as multiple agents concerned with the quality of the collaborative decision. We define a facilitator agent as that agent responsible for the overall decision making process. This includes managing the complex negotiation processes that are required among those participants collaborating on decision making.

The use and the integration of software agents in the decision support systems provide an automated, cost-effective means for making decisions. The agents in the system autonomously plan and pursue their actions and sub-goals to cooperate, coordinate, and negotiate with others, and to respond flexibly and intelligently to dynamic and unpredictable situations.

We take first a literature survey of some related work in section 2 and 3. Then we propose a multi-agent architecture for web-based collaborative decision support systems in section 4. We also present some implementations issues in section 5. Finally, we conclude with future research direction in section 6.

2 Collaborative Decision Support Systems

2.1 Decision Support Systems

Decision support systems (DSS are designed to actively interact with an individual decision maker in order to assist him to make better decisions based on information obtained [Keen and Scott-Morton, 1978; Sprague and Carlson, 1982]. A number of frameworks or typologies have been proposed for organizing our knowledge about decision support systems [Power, 2000]. The two most widely implemented approaches for delivering decision-support are Data-Driven and Model-Driven DSS. Data-Driven DSS help managers organize, retrieve, and synthesize large volumes of relevant data using database queries, On-Line Access and Processing (OLAP) techniques, and data mining tools. Model-Driven DSS use formal representations of decision models and provide analytical support using the tools of decision analysis, optimisation, stochastic modelling, simulation, statistics, and logic modelling. Three other approaches have become more wide spread and sophisticated because of collaboration and web technologies: Knowledge-Driven DSS can suggest or recommend

actions to managers, Document-Driven DSS integrate a variety of storage and processing technologies to provide managers document retrieval and analysis, and Communication-Driven DSS rely on electronic communication technologies to link multiple decision makers who might be separated in space or time, or to link decision makers with relevant information and tools.

In the latter category, also known as Collaborative or Group Decision Support Systems (GDSS), which are closely related to DSS, facilitate the solution of unstructured and semi-structured problems by a group of decision makers working together as a team [Ribeiro, 2006; DeSanctis, and Gallup, 1997; Nunamaker, 1997]. Group Decision Support Systems (GDSS) are interactive computer-based environments which support concerted and coordinated team effort towards completion of joint tasks. DeSanctis and Gallup [1997] defined GDSS as a combination of computers, communications and decision technologies working in tandem to provide support for problem identification, formulation and solution generation during group meetings.

2.2 Collaborative Decision Making

Decision aid and decision making have greatly changed with the emergence of information and communication technology (ICT). Decision makers are now far less statically located; on the contrary they play the role in a distributed way. This fundamental methodological change creates a new set of requirements: web-based collaborative decisions are necessarily based on incomplete data. “web-based collaborative decision” means that several entities (humans and machines) cooperate to reach an acceptable decision, and that these entities are distributed and possibly mobile along networks. Distributed decision making must be possible at any moment. It might be necessary to interrupt a decision process and to provide another, more viable decision.

Research that studied group decision support systems in the existing literature used mainly face-to-face facilitated collaborative decision support systems. Some of its results may not apply to distributed teams that, it is difficult for distributed teams to arrange face-to-face meetings or to meet at the same time virtually. Furthermore, a review of the literature of group decision support systems (GSS) shows that the facilitator plays a major role in an electronic meeting’s success. The facilitator is also responsible for designing and managing the group processes, and for guiding groups in the decision process as well as for solving process problems such as cognitive overload. Namely, The facilitator guided the group through the agenda: first, to define a list of criteria the group wanted to use to evaluate the projects; second, to weight the criteria in order of importance; third, to evaluate the alternatives against the criteria; fourth, to calculate scoring based on the criteria weights and ratings; fifth, to allocate dollar amounts to the projects. Facilitated groups did achieve higher quality decisions, and group processes and cohesiveness are improved.

3 Multi-Agent Systems

In recent years, there has been considerable growth of interest in the design of a distributed, intelligent society of agents capable of dealing with complex problems and vast amounts of information collaboratively. Various researches have been conducted into

applying intelligent agent-based technology toward real-world problems. Furthermore, there has been a rapid growth in developing and deploying intelligent agent-based systems to deal with real-world problems by taking advantage of the intelligent, autonomous, and active nature of this technology. The main benefits of an agent-based approach come from its flexibility, adaptability, and decentralization.

There is a multitude of viewpoints on what exactly constitutes an agent, how they should be structured, and how collections of agents that are interacting with each other and the environment can be used to implement complex systems. Despite the lack of consensus, the benefits of implementing agent systems are little disputed, and several agent architectures have progressed to become usable technologies. The working definition of an agent is adapted from [Jennings, 1996]: an agent is an artificial, computational entity that can perform certain tasks with a certain degree of autonomy or initiative whilst intelligently adapting to its environment. Note that a human is not an agent in this definition. An agent, which is a part of a multi-agent system, is defined as being a part of a whole or of a larger organization.

The definition of multi-agent systems (MAS) is well known and accepted as a loosely coupled network of agents that work together to find answers to problems that are beyond the individual capabilities or knowledge of each agent and there is no global control system.

An agent's architecture is a particular design or methodology for constructing an agent. Wooldridge and Jennings refer to an agent's architecture as a software engineering model of an agent [Jennings, 1996]. Using these guidelines, agent architecture is a collection of software modules that implement the desired features of an agent in accordance with a theory of agency. This collection of software modules enable the agent to reason about or select actions and react to changes in its environment.

MAS are software systems composed of several autonomous software agents running in a distributed environment. Beside the local goals of each agent, global objectives are established committing all or some group of agents to their completion. Some advantages of this approach are: 1) it is a natural way for controlling the complexity of large and highly distributed systems; 2) it allows the construction of scalable systems since the addition of more agents become an easy task; 3) MAS are potentially more robust and fault-tolerant than centralised systems.

As is typical with an emerging technology, there has been much experimentation with the use of agents in DSS, but to date, there has been little discussion of a framework or methodological approach for using agents in DSS. In addition, because of the subjective nature of agency and the wide-range of contexts and disciplines in which agent-like programs have been deployed, a general definition or description of agents has been lacking within the DSS/MIS literature. This difficulty in describing what an agent is has resulted in overuse of the term agent and poor guidance for DSS developers seeking to agent-enable their applications. And while DSS researchers are discussing agents as a means for integrating various capabilities in DSS and for coordinating the effective use of information [Whinston, 1997], there has been little discussion about why these entities are fit for such tasks.

4 A Multi-Agent Architecture for Web-based Collaborative Decision Support Systems

We started our framework with the following fundamentals:

1. The first fundamental, in keeping with [Adla and Zaraté, 2006], was to segment web-based collaborative decision support systems into two components: Facilitator and participants (decision-makers)
2. The second fundamental we adopted was to include in each collaborative decision support system component an agent to oversee or manage the other agents within the component;

4.1 The Web-based Collaborative Decision Making

In [Adla and Zaraté, 2006; Adla et al., 2007] we consider the paradigm of web-based collaborative decision-support systems, in which several decision-makers who deal with partial, uncertain, and possibly exclusive information must reach a common decision. To this end, the use of a collaborative system makes possible the collaboration of distant decision makers. The cooperative work so initiated can be synchronous or asynchronous. A small group or a whole organization can be supported. The application can be carried in several sites over a common information base.

The networked decision makers work together to solve a particular problem although they might neither be present at the same time in the same place nor constitute a permanent organization. Thus, decision-makers can evaluate and rank alternatives, determine the implications of offers, maintain negotiation records, and concentrate on issues instead of personalities.

In the proposed framework (Figure 1) [Adla and Zaraté, 2006], the group is constituted of two or several decision-makers (participants) and a facilitator. Each participant interacts with individual DSS (C-DSS) integrating local expertise and allowing him to generate one or several alternatives of the problem submitted by the facilitator. The group (facilitator and participants) use the group toolkit for alternative generation, organization, and evaluation as well as for alternative choice which constitutes the collective decision. Therefore, we view the individual DSS as a set of computer based tools integrating expert knowledge and using collaboration technologies that provide decision-maker with interactive capabilities to enhance his understanding and information base about options through use of models and data processing, and collaborate with him. The main components of an individual DSS are:

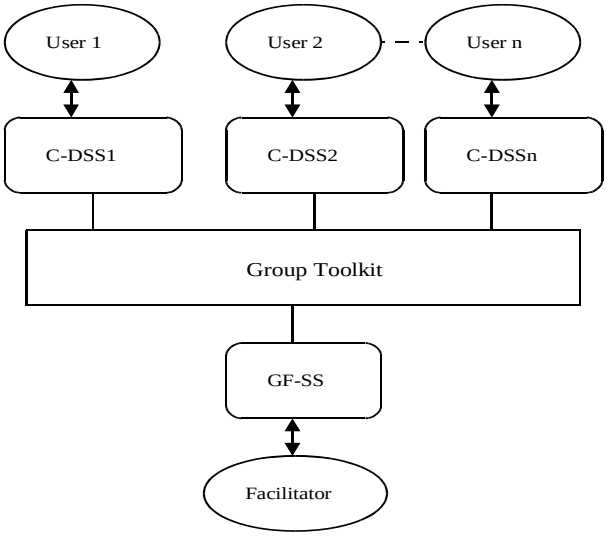


Figure 1: Web-based collaborative decision framework

The objective of this paper is to design a distributed GDSS based on a multi-agent architecture. Agents were integrated into the DSS for the purpose of automating more tasks for the user, enabling more indirect management, and requiring less direct manipulation of the collaborative decision support system. Specifically, agents were used to collect information outside of the organisation and to generate decision-making alternatives that would allow the user to focus on solutions that were found to be significant. A set of agents is integrated to the system and placed in the collaborative decision support system components, according to our architecture [Sprague and Carlson, 1982].

4.2 The Multi-Agent Collaborative Decision Support System Architecture

The goal of Distributed Group Decision making is to create a group of coarse-grained cooperating agents that act together to come to a collective decision. Participants in a collaborative decision making meeting are considered as a set of agents involved in creating a collective decision. These participant agents are involved with the content knowledge of the particular group problem at hand. The responsibility of managing any decision making process is typically put upon a supervisory agent. We call this agent the facilitator. We view the participants as multiple agents responsible for creating the *content* of the decision, and the facilitator as an outside agent responsible for managing the decision process that the participant agents use to come to common decision

For each participant (decision's maker), the following agents are defined:

• DA (Decision-maker Assistant): it's the interface between the participant and the system. It initiates the local decision making process, manages the private space of the participant and insures communication with the other participants and the facilitator. So

during idea (solution) generation stage, a decision maker can use its proper DSS (Decision Support System) through the DA.

• CA (Collaborator Assistant): The role of this agent is devoted exclusively to the collaboration of the decision maker in the process of decision making support. The only interaction it manages is with CRA of the facilitator and does not communicate directly with agents of other decision makers.

For the facilitator side, the following agents are defined:

• FA (Facilitator Assistant): it manages the interface between the system and the facilitator. It must bring ergonomic features that provide a work environment and comfortable communication, provides a private workspace for the facilitator and a public space for the group. It also allows the facilitator to communicate at any time with group members outside the decision making process, helps to establish communications with other system users through their assistants (DA). This last feature allows simplifying and reducing interactions within the MAS, since certain interactions of decision-making process happen between the coordinator agents (CAs), and others occur only between assistant agents (DAs). This reduces complexity and enables easy management of interactions. Also, in collaboration with the facilitator, FA solicits the right group of participants (depending on their profiles and the type of decisional problem), then sets the agenda for the meeting.

• CRA (CooRdinator Agent for the decision making process): It's the central agent of the decision making process. It remains in contact with the structure of the decision maker (CA). It is supervised by the facilitator via the FA. Its role is to ensure the rules checking and application during the various phases of the decision making process. FA starts the decision making session. The CRA takes in charge the following tasks of this

activity. It guides the group through the activity phases. To this end, it interacts with decision maker Collaborator Agents (CA). It displays the required information when needed, knows who proposed what, but it must also know how to preserve the anonymity of participants if necessary. Once the alternatives have been organised, the result (list of alternatives screened and cleaned) is sent by the Mediator Agent (MA) to the CRA, and according to the agenda initially established, the CRA sends to the participants the evaluation method with the organised list of alternatives.

• MA (Mediator Agent): is requested by the CRA during the alternatives organisation phase. The MA possesses knowledge on the field of decision making problem. Its role is to refine the alternatives (deletes or merges synonymous, redundant or inconsistent alternatives) and to classify the alternatives as well. The MA does not have a global vision on the group; it sends the results to the CRA.

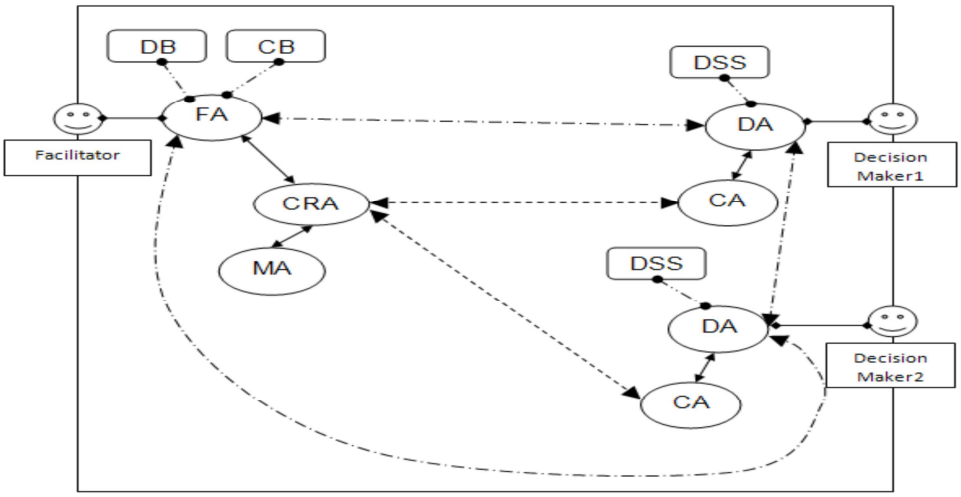


Figure 2: Distributed DSS-MAS logical Architecture

DA : Decision maker Assistant
CA : Coordinator Assistant (Decision Maker side)
DSS : Decision Support System
DB : Data Base
CB : Case Base.
FA : Facilitator assistant
CRA : CooRdinator Agent (Facilitator side)
AM : Mediator Agent
: User-system link
: Structure internal link (same structure)
: Inter-structure agents link
: Assistants link
: using link (DSS/DB/CB)

5 Implementation Issues

A prototype of the multi-agent architecture for distributed group decision support system is being implemented in order to generate results that can be analyzed and validate our work. To this end, we have used the FIPA compliant Multi-Agent System platform JADE multi-agent platform to implement our system. Some implementation details are given in the next section.

5.1 The Multi-Agent Platform

The FIPA compliant Multi-Agent System platform JADE is the most widely used platform in agent technology filed. It offers a broad range of functionalities, provides better support and availability of agent framework, where a lot of common tasks are already

implemented (i.e. agent communication at the syntax level, agent management, migration of agents etc.).

The main features of JADE platform are:

- Openness: agents may be written in Java language on operating systems. The language and the operating systems barriers are in this way reduced;
- Distributive: agents may be distributed on multiple networked machines. This is a very important advantage to increase simulation runs, especially in scenarios where a great number of agents are involved;
- Extensible: it is possible to add or remove agents at run-time allowing the creation of flexible and robust scenarios.

5.2 The Prototype

Currently, the various agents are created in java using JADE. The users' interfaces are separated from agents; they are linked to agents by means of message transmission.

The current prototype is based physically on hosts which can be on LAN or internet Network:

- Host-Jade Server: machine hosting JADE platform with the different agents of the system;
- Host-Facilitator: machine hosting the interface, different software used by the facilitator.
- Host-Decision's Maker: machine hosting the interface and can also hosts local DSS of the decision's maker.
- Host-Database Server: machine hosting various databases used by Facilitator and the proper database of our distributed model.

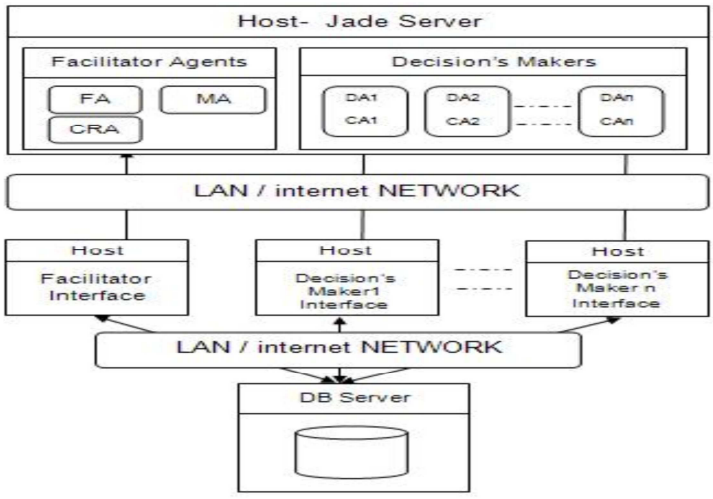


Figure 3: Distributed DSS-MAS Physical Architecture

As depicted in figure 4, a decision group composed of a facilitator and four decision makers collaborate and interact to solve a problem; the decision maker number three doesn't appear on the figure as it's disconnected and does not participate to the decision making session. A partial result of the interactions between agents (JADE's sniffer screen) is given Figure 4.

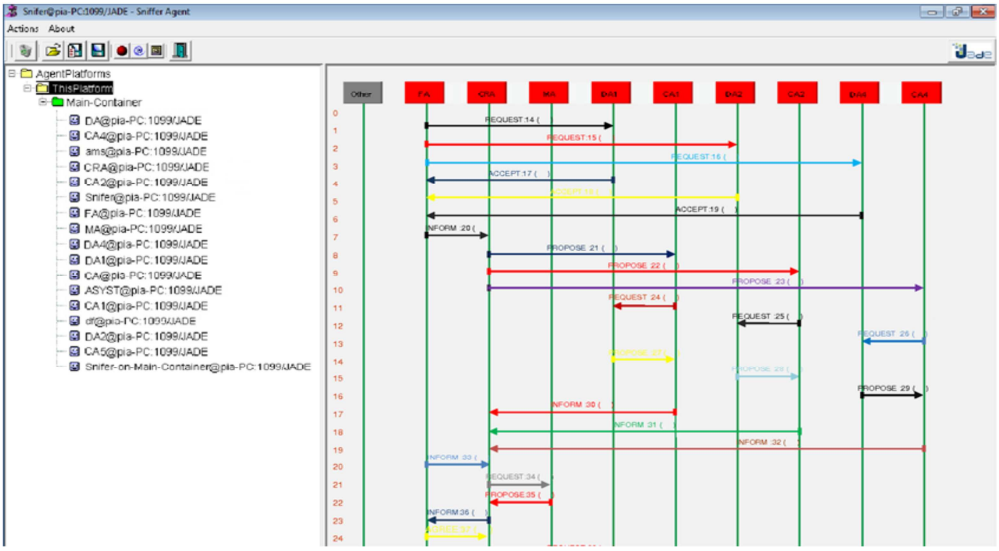


Figure 4: Partial result (sniffer screen)

6 Conclusion

The use of agent-based DSS to support decision making is important within the industry because they allow managers and stakeholders to quickly gather information and process it in various ways in order to assist with making diagnosis and treatment decisions.

In this paper we presented a distributed group decision making system based on a multi-agent architecture. We have integrated agents into a cooperative intelligent decision system for the purpose of automating more tasks for the decision maker, enabling more indirect management, and requiring less direct manipulation of the DSS. In particular, agents were used to collect information and generate alternatives that would allow the user to focus on solutions found to be significant. Agents are normally used to observe the current situation and knowledge base, and then make a decision on an action consistent with the domain they are in, an finally perform that action on the environment.

Based on this, and considering that communication capabilities play an essential role in DSS, further work based on coordination protocols between agents needs to be done. Particularly, the context information domain included in the software tool will be extended in order to improve the support for decision making and the coordination activities.

References

Adla, A., J-L, Soubie, and P. Zarate, "A Co-operative Intelligent Decision Support System for Boilers Combustion Management based on a Distributed Architecture", Journal of decision Systems, Lavoisier, 2007, Vol. 16, pp. 241-263. Systems, Lavoisier.

A. Adla, and P. Zarate, "A cooperative intelligent decision support system", *In Proceedings of International Conference on service systems and service management*, Troyes, France, October 25-27 2006.

Cheung, W. (2005): "An Intelligent decision support system for service network planning", Decision Support Systems, Lavoisier, 2005, Vol. 39, pp. 415- 428.

G. DeSanctis, and B. Gallup, "A foundation for the study of group decision support systems", *Management Science*, 1997, Vol. 13, pp. 1589-1609.

E. Jennings, "Using intelligent agents to manage business processes", In B. Crabtree and N. R. Jennings editors, In Proceedings of the 1st international conference on practical applications of intelligent agents and multi-agent technology (PAAM96), 1996, pp. 345-360.

P. Keen, and M. Scott-Morton *"Decision Support Systems: an organizational perspective"*, Addison-Wesley Publishing, 1978.

J. Nunamaker, "Lessons from a dozen years of group support systems research", *Journal of MIS*, 1997, Vol. 13, pp. 163-207.

D. J. Power, "Supporting Decision-Makers: An Expanded Framework", 2000.

R. Ribeiro, "Intelligent Decision Support Tool for Prioritizing Equipment Repairs in Critical/Disaster Situations", *In Proceedings of Workshop on Decision Support Systems*, 2006

R. Sprague, and D. Carlson, *"Building Effective Decision Support Systems"*, Prentice-Hall, Inc, Englewood Cliffs, 1982.

Whinston, A. (1997). Intelligent Agents as a Basis for Decision Support Systems. Decision Support Systems, 20(1).

Description et Classification de Services Web Sémantiques

Fatima Bedad* , Aek Haouas **
Djelloul Bouchiha ***

*Département informatique,Université d'USTO 31000,Algérie
e-mail bedad_fatima2006@yahoo.fr
www.univ-usto.dz
**Département informatique ,Université d'USTO 31000,Algérie
e-mail haouasab@yahoo.fr
www.univ-usto.dz
***Laboratoire EEDIS,Université d'UDL 22000,Algérie
e-mail bou_dje@yahoo.fr
www.univ-sba.dz

Résumé. Avec l'évolution du Web une nouvelle technologie a vu le jour ; il s'agit des services Web sémantiques (SWS). Cette nouvelle technologie permet d'automatiser la découverte, la composition et l'invocation des services à travers le Web. Pour profiter de ces avantages, un processus de réingénierie des applications Web réorienté vers les services Web sémantiques est nécessaire. La réingénierie permet de réutiliser les fonctionnalités des applications Web sous forme de services Web sémantiques sans reprendre l'écriture du code de ces applications. L'objectif de ce travail consiste à proposer une approche à base de réingénierie ontologique pour la description et la classification des applications SWS WSMO. L'approche consiste en deux grandes étapes : tout d'abord, l'extraction d'informations utiles à partir de codes sources de web services de SWS WSMO, ensuite l'analyse de l'information extraite à l'aide de l'ontologie du domaine en faisant appel à un critère de similarité.

1 Introduction

Le World Wide Web (WWW), ou tout simplement le Web, permet d'accéder à des sources de données hétérogènes résidant n'importe où sur Internet (Martin et al., 2004). Au cours des dernières années, le Web a rencontré deux changements révolutionnaires qui visent à le transformer d'une collection de documents statiques en un environnement intelligent et dynamique pour l'intégration de données.

La première technologie est celle des services Web. Ces services offrent des fonctionnalités accessibles via le Web en utilisant un ensemble de technologies courantes, telles que SOAP, WSDL, UDDI et XML.

La deuxième technologie est dite le Web sémantique (Roman et al.,2005). C'est un Web intelligent qui permet d'enrichir les données existantes sur le Web par des descriptions formelles indiquant leur signification. La technologie du Web sémantique est basée sur les ontologies. Selon Gruber, une ontologie est une spécification explicite d'une conceptualisation (Gruber,1993).

La limitation majeure de la technologie des services Web est que la découverte et la composition des services nécessitent encore une intervention humaine accrue. Ceci constitue un handicap, surtout avec les montées en charge des services Web. Pour résoudre ce problème, la communauté du Web sémantique a proposé d'enrichir les services Web avec un contenu sémantique de leurs fonctionnalités afin de faciliter la découverte et l'intégration de ces services. Cette technologie, qui combine les techniques des services Web et du Web sémantique est appelée les services Web sémantiques (SWS) (McIlraith et al.,2001).

Plusieurs propositions ont été soumises au W3C pour la conception des services Web sémantiques (OWL-S (Martin et al., 2004) , WSMO (Roman et al.,2005) , WSDL-S (Martin et al., 2004), etc.). L'un des plus importants challenges était de permettre la construction d'applications flexibles au niveau de plusieurs entreprises et de permettre ainsi leur composition dynamique pour offrir à leurs clients les meilleures solutions. Dans cette optique nous utilisons WSMO (Web Service Modeling Ontology) comme plateforme de modélisation des services Web sémantiques . WSMO possède quatre composantes principales : les ontologies, qui fournissent la terminologie ultérieurement, les services Web qui fournissent un accès aux offres, les buts qui représentent les désirs de l'utilisateur et les médiateurs qui traitent le problème d'interopérabilité entre ces différentes composantes.

Pour être en phase avec l'évolution de la technologie sans pour autant laisser tomber des années d'investissement et de savoir-faire, un nouveau domaine de recherche a vu le jour il y a des années dans le milieu académique, en l'occurrence la réingénierie, dans le but est justement de reconsidérer les applications existantes dans de nouvelles perspectives.

2 L'ingénierie des services web sémantiques

L'ingénierie des SWS consiste à générer des descriptions syntaxiques et sémantiques des services Web à partir de modèles conceptuels. La composante qui prend en charge les descriptions syntaxiques, peut directement intégrer les méthodologies classiques d'ingénierie des systèmes d'information. Par contre celle des descriptions sémantiques est manifestement plus complexe dans la mesure où on est confronté à des problématiques relevant de l'ingénierie ontologique.

En effet, plusieurs travaux menés dans ce dernier contexte d'ingénierie ontologique Sabou et al (Sabou et al., 2005), Buitelaar et Gmbh (Buitelaar et Gmbh 2003), Bell et Lycett (Bell et Lycett, 2005), (Bell et al., 2007).

Notre contribution est consacrée au problème de la description et de la classification de services web sémantique. Nous y proposons une approche qui se décline en deux grandes étapes: l'extraction d'informations utiles ou pertinentes et l'analyse de l'ensemble qui en résulte.

3 Approche proposée

Nous donnons tout d'abord une vue synthétique des étapes, leur organisation et les outils qu'elles font intervenir (FIG. 1).

La première phase est l'extraction d'informations utiles à partir du service web. L'élément de base de étant SWS WSMO. L'information pertinente est représentée dans un tableau. La deuxième phase est l'analyse qui commence par une comparaison réalisée entre les

informations extraites précédemment et les concepts de l'ontologie de domaine en faisant appel à des techniques de distance sémantique. Cette opération de comparaison sémantique permet le calcul des liens sémantiques. Il en résultera l'inférence de nouvelles ontologies importées et de même domaine de SWS.
Initialement, nous supposons que:

- A partir des services web, nous pouvons extraire des concepts ou bien des informations et le représenter dans un tableau.
- Un concept de l'ontologie de domaine est indiqué par un champ de tableau.
- Il existe au préalable une ontologie qui est construite par des experts et que cette ontologie est spécifique à un domaine.

Nous présentons ensuite, dans leur détails étapes du processus de description et de classification de SWS WSMO.

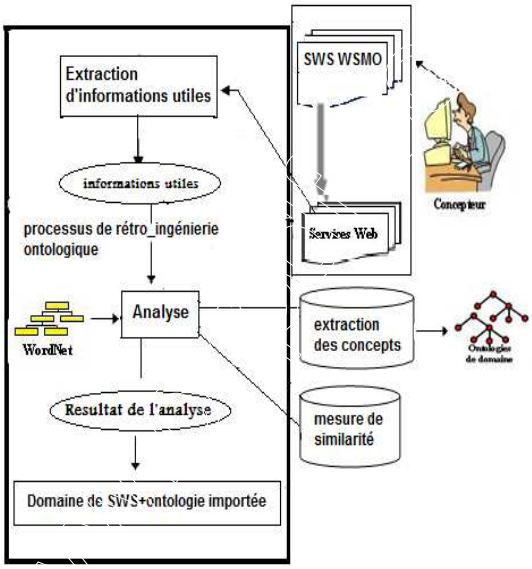


FIG. 1 – *Processus de description et de classification de SWS WSMO.*

3.1 Extraction des informations utiles

Cette opération commence par le filtrage de l'élément structurant ; le service web de SWS WSMO suivi de l'extraction des mots clés {'[' ,' memberOf'} qui se trouvent en variables partagées des services web et enfin l'extraction des informations utiles qui se trouvent après le mot clé ' memberOf' et avant le mot clé '['.

Le filtrage consiste à parcourir le code du service web, et garder uniquement les plus utiles, le résultat en est un ensemble de concepts. Ces concepts vont être représentés sous forme de tableau logique pour faciliter leur manipulation. Selon Lopresti et Nagy. les tableaux sont la manière la plus adéquate pour représenter des données structurées (Lopresti et Nagy 1999).

3.2 Analyse

Elle se constitue d'une série de sous-étapes pour le traitement de l'information utile résultante de la phase précédente. Le résultat en est le calcul des liens sémantiques.

3.2.1 Identification des concepts d'ontologie

Ce traitement permet de déterminer les concepts de l'ontologie de domaine qui appartiennent à l'application Web.

Il est appliqué aux ontologies de domaine, et consiste à enlever les espaces, les traits d'union, etc. et garder uniquement les racines des termes telles qu'elles apparaissent dans WordNet[1]. On en recueille un ensemble de termes qui peuvent être identifiés plus tard comme des concepts d'ontologie.

3.2.2 Calcul de la distance sémantique

Basée sur le calcul de la similarité, cette manipulation porte sur les informations retenues du service web qui se trouvent actuellement dans un tableau et les concepts des ontologies de domaine (cf. l'étape précédente).

Une mesure de similarité vise à quantifier la proximité sémantique de deux concepts. Pour calculer cette similarité sémantique, nous avons utilisé des approches qui se basent sur WordNet. Ces techniques peuvent être classifiées en trois catégories :

a) *les mesures de similarité basées sur la longueur du chemin entre les concepts :* la LCH (Leacock et Chodorow 1998), Le WUP (Wu et Palmer 1994) et La mesure path. le dernier est égal à l'inverse de la longueur du plus court chemin entre les deux concepts.

b) les mesures basées sur le contenu d'information : le RES (Resnik, 1995), le LIN (Lin ,1998) et la mesure JCN (Jiang et Conrath 1997).

c) les mesures basées sur le type des relations entre les concepts : le HSO (Hirst et Onge 1998) ,le LESK (Roman et al.,2005), et la mesure VECTOR (Patwardhan et al.,2003).

Une similarité entre deux concepts n'est tenue en compte que si elle dépasse un certain seuil. Le seuil est une valeur prise entre 0 et 1. La valeur 1 indique qu'il y a une équivalence sémantique totale entre deux concepts.

Les Méthodes et les stratégies décrites ci-dessus seront utilisées dans l'étape de calcul de la distance sémantique dans notre processus de rétro-ingénierie ontologique comme suit: Nous avons un tableau correspondant à un web service. Nous disposons aussi d'une ontologie de domaine contenant des concepts; chaque concept pouvant éventuellement avoir un ensemble d'attributs. Maintenant nous procédons à des calculs de distance sémantique pour identifier le domaine de SWS et importer de nouvelles ontologies.

Nous nous fixons au préalable un seuil pour la distance sémantique. Le seuil est une valeur comprise entre 0 et 1; la valeur 1 indiquant que les deux entités sont totalement similaires.

[1]WordNet : est une base de données lexicale qui organise les noms et les verbes dans des concepts (synset) en hiérarchies de relations is-a. Chaque concept est décrit par une brève glose.

En second lieu, nous devons choisir une méthode et finalement une stratégie pour calculer la distance sémantique comme décrit ci-dessus.

Si la distance sémantique entre un champ de table de web service et le champ de table d'un concept d'ontologie est inférieure ou égale au seuil fixé auparavant, alors on peut considérer que le SWS appartient au domaine de ce concept d'ontologie et que par conséquent cette ontologie peut être importée dans SWS.

4 Algorithme de l'approche proposée

Comme il est décrit, notre approche consiste à construire le domaine de service web de SWS WSMO et de plus établir un ensemble enrichi de concepts et de relations. Cet ensemble réfère en premier lieu aux concepts identifiés dans l'étape précédente. Il est en second lieu enrichi de relations de l'ontologie possédant comme extrémité l'un des concepts identifiés. Les concepts de l'autre bout de chaque relation sont aussi ajoutés à cet ensemble. A ce niveau, des groupes de concepts seront formés. Chaque groupe représente un domaine pour la classification de notre objectif (SWS WSMO) et même un ensemble des ontologies importées selon chaque concept Oj de la deuxième étape (cf. Fig. 2).

Algorithme Identification de domaine

Entrées
Les concepts d'Ontologie de domaine.
Les informations utiles de service web (champs de tableaux).
Une mesure de similarité.
Une stratégie de calcul de similarité entre ensembles.
Un seuil pour la mesure de similarité.
Sorties
Un ensemble de concepts.
Début-Algo
1- Créer un vecteur d'objets V1. Chaque objet possède comme attributs les champs des informations utiles de service web SWS WSMO .
2- Créer un vecteur d'objets V2. Chaque objet correspond à un concept de l'ontologie de domaine.
3- Pour chaque objet Oi du vecteur V1 ***Faire***
Pour chaque objet Oj du vecteur de concepts V2 ***Faire***
Calculer la similarité entre le concept de l'objet Oi et les concepts de l'objet Oj un par un ;
Si pour un attribut de O_i la similarité est égal au seuil, ***Alors*** marquer l'objet Oi correspondant au "domaine" de l'objet Oj .
Fin-Si
Fin-Pour
Fin-Pour
Fin-Algo

FIG. 2 – L'algorithme proposé

- *Complexité*: la complexité de cet algorithme est de l'ordre O(n*m), avec n le nombre de tableaux , et m le nombre de concepts de l'ontologie de domaine.

5 Expérimentation

Afin de mettre en œuvre notre approche et d'en évaluer les performances, nous considérons un exemple illustratif tiré du web service de SWS proposé par l'outil de WSMO Studio.

5.1 Résultats de l'extraction des informations utiles

Le filtrage du code de web service permet d'extraire tous les mots qui figurent après le mot-clé {memberOf} et avant le mot-clé {[}.

le filtrage de notre code donne lieu à trois thèmes remarquables qui forment l'ensemble {Child, location, parent}.

- {Child} : figure après le mot-clé 'memberOf'.
- {location, parent} figure avant celui '['.

On doit, selon ces trois thèmes, donner la classification de SWS WSMO et être à même de mener l'analyse de la phase suivante.

5.2 Résultats de la phase d'analyse

On fera l'analyse avec 3 ontologies de domaine qui sont : travel.owl, business.owl, Generations.owl.On fixe préalablement le seuil à 1.

5.2.1 Pour l'ontologie travel.owl

a) Identification des concepts d'ontologie

Accommodation,BedAndBreakfast,BudgetAccommodation,Campground,Hotel,LuxuryHotel,Activity,Adventure,BunjeeJumping,Safari,Relaxation,Sunbathing,Yoga,Sightseeing,Museums,Safari,Sports,Hiking,Surfing,Contact,Destination,BackpackersDestination,Beach,BudgetHotelDestination,FamilyDestination,QuietDestination,RetireeDestination,RuralArea, Farmland, NationalPark ,UrbanArea, City, Capital, Town, AccommodationRating.

b) Calcul de la distance sémantique

Les résultats de cette phase d'analyse sont mentionnés dans le tableau (cf. *TAB. 1*):

	child	Location	parent
Accommodation	-1.0	-1.0	-1.0
BedAndBreakfast	-1.0	-1.0	-1.0
BudgetAccommodation	-1.0	-1.0	-1.0
Campground	-1.0	-1.0	0.062
Hotel	-1.0	-1.0	0.076
LuxuryHotel	-1.0	-1.0	-1.0
Activity	-1.0	-1.0	-1.0
Adventure	-1.0	-1.0	-1.0
BunjeeJumping	-1.0	-1.0	-1.0
Safari	-1.0	-1.0	-1.0
Relaxation	-1.0	-1.0	-1.0
Sunbathing	-1.0	0.166	-1.0
Yoga	-1.0	-1.0	-1.0
Sightseeing	-1.0	-1.0	-1.0
Museums	-1.0	-1.0	0.076
Safari	-1.0	-1.0	-1.0
Sports	-1.0	-1.0	-1.0
Hiking	-1.0	-1.0	-1.0
Surfing	-1.0	-1.0	-1.0
Contact	-1.0	-1.0	-1.0
Destination	-1.0	-1.0	0.066
BackpackersDestination	-1.0	-1.0	-1.0
Beach	-1.0	-1.0	0.083
BudgetHotelDestination	-1.0	-1.0	-1.0
FamilyDestination	-1.0	-1.0	-1.0
QuietDestination	-1.0	-1.0	-1.0
RetireeDestination	-1.0	-1.0	-1.0
RuralArea	-1.0	-1.0	-1.0
Farmland	-1.0	-1.0	0.066
NationalPark	-1.0	-1.0	-1.0
UrbanArea	-1.0	-1.0	-1.0
City	-1.0	-1.0	0.062
Capital	-1.0	-1.0	-1.0
Town	-1.0	-1.0	0.062
AccommodationRting	-1.0	-1.0	-1.0

TAB. 1 –. *Mesure de similarité pour travel.owl*

Pour cette analyse aucun résultat de calcul de distance sémantique n'est égal à 1; on peut donc conclure que ce web service n'appartient pas au domaine de cette ontologie.

5.2.2 Pour l'ontologie business.owl

a) Identification des concepts d'ontologie

REA_Element,Exchange_Element,Agent,Agent_Type,Agreement,Contract,Agreement_Type,Association,Collaboration,Association_Type,Commitment,Commitment_Type,Event ,Event_Type ,Resource ,Resource_Type ,Stock_flow ,Recipe_Element ,Recipe ,Task ,Activity,Interface_Activity ,Begin ,End ,Fail ,Fork ,Join ,Transaction , Transition ,Script_Element ,Process ,Process_Stock_flow .

b) Calcul de la distance sémantique

De la même manière, le tableau (cf. TAB. 2) donne les résultats:

	child	Location	parent
REA_Element	-1.0	-1.0	-1.0
Exchange_Element	-1.0	-1.0	-1.0
Agent	0.111	0.25	0.083
Agent_Type	0.111	0.25	0.083
Agreement	-1.0	-1.0	-1.0
Contract	-1.0	-1.0	-1.0
Agreement_Type	-1.0	-1.0	-1.0
Association	-1.0	-1.0	-1.0
Collaboration	-1.0	-1.0	-1.0
Association_Type	-1.0	-1.0	-1.0
Commitment	-1.0	-1.0	-1.0
Commitment_Type	-1.0	-1.0	-1.0
Event_Type	-1.0	-1.0	-1.0
Resource	-1.0	-1.0	-1.0
Resource_Type	-1.0	-1.0	-1.0
Stock_flow	-1.0	-1.0	-1.0
Recipe_Element	-1.0	-1.0	-1.0
Recipe	-1.0	-1.0	-1.0
Task	-1.0	-1.0	-1.0
Activity	-1.0	-1.0	-1.0
Interface_Activity	-1.0	-1.0	-1.0
Begin	0.076	0.166	0.062
End	0.142	0.25	0.1
Fail	0.090	-1.0	0.071
Fork	-1.0	0.11	-1.0
Join	0.083	-1.0	0.066
Transaction	-1.0	.0	-1.0
Transition	-1.0	-1.0	-1.0
Script_Element	-1.0	-1.0	-1.0
Process	-1.0	-1.0	-1.0
Process_Stock_flow	-1.0	-1.0	-1.0

TAB. 2 – *Mesure de similarité pour business.owl*

Pour cette analyse non plus , aucun résultat des calculs de distance sémantique n'est égal à 1: il va donc de soi que ce web service est à exclure du domaine de cette ontologie.

5.2.3 Pour l'ontologie Generations.owl

a) Identification des concepts d'ontologie

Sex, Female, Person, Brother, Daughter, Father ,GrandFather ,GrandMother ,GrandParent Man ,Mother ,OffSpring ,Parent ,Sibling ,Sister ,Son ,Woman ,Male .

b) Calcul de la distance sémantique

Les résultats de la phase d'analyse figurent dans le tableau (cf. TAB.3) :

	child	Location	parent
Sex	-1.0	-1.0	-1.0
Female	0.166	0.142	0.111
Person	0.333	0.166	0.166
Brother	0.142	0.1	0.125
Daughter	0.125	0.090	0.111
Father	0.111	0.083	0.5
GrandFather	0.125	0.090	0.142
GrandMother	0.125	0.090	0.142
GrandParent	0.142	0.1	0.166
Man	0.2	0.125	0.125
Mother	0.111	0.083	0.5
OffSpring	0.2	0.125	0.166
Parent	0.125	0.090	**1.0**
Sibling	0.2	0.125	0.166
Sister	0.142	0.1	0.125
Son	0.125	0.090	0.111
Woman	0.2	0.125	0.125
Male	0.166	0.142	0.111

TAB. 3 – *Mesure de similarité pour Generations.owl*

Cette dernière analyse produit par contre un résultat où le calcul de distance sémantique donne 1, d'où la possibilité de conclure que ce service web sémantique appartient effectivement au domaine de l'ontologie Generations.owl, et que par conséquent celle-ci peut être importée à ce web service.

En résumé, comme exprimé auparavant, lors de la considération des 3 types d'ontologies, le résultat final est exposé dans le tableau (cf. TAB. 4) :

	Ontologie 1 (seuil=1)	Ontologie 2 (seuil=1)	Ontologie 3 (seuil=1)
Child	0	0	0
Location	0	0	0
Parent	0	0	1

TAB. 4 – *Mesure de similarité pour les 3 ontologies*

6 Discussion

L'ingénierie ontologique des services web s'applique à leurs codes sources qui sont décrits en WSDL WSMO. Elle fait intervenir une méthode qui consiste à utiliser la technique de mesures de similarité, en l'occurrence la méthode path pour identifier les relations entre tous les concepts. Comme résultat de ce processus, on aboutit à la classification de SWS WSMO et à l'importation de nouvelles ontologies.

L'approche proposée est supportée par un ensemble d'outils précisément WSMO Studio qui permet de générer les différentes composantes WSMO par des descriptions syntaxiques WSDL et Eclipse qui est un environnement de développement Java.

Le processus de réingénierie est un processus semi-automatique d'assistance du concepteur et de l'expert de domaine.

WordNet est utilisé ici comme base multilingue de données lexicales.

La fixation d'un seuil de distance sémantique signifie que le taux d'erreur est toléré. Nous admettons ce risque parce que le domaine de conception n'est pas déterministe, mais heuristique. Il n'y a pas un modèle unique correct pour une situation, mais seuls des modèles adéquats ou inadéquats (Rumbaugh et al., 1991).

En outre, l'implication de l'ontologie rend les résultats beaucoup plus adéquats pour le champ d'application. Pour éviter toute ambiguïté et obtenir des résultats plus réalistes, nous fixons le seuil à 1.

Le point fort de cette approche est qu'elle repose sur une référence sémantiquement très riche qui est en l'occurrence l'ontologie. Il est en outre possible d'affirmer qu'elle donne des résultats très satisfaisants qui permettent la réingénierie, la migration, la compréhension et l'évolution des applications Web.

Il est aussi à noter la possible de réutiliser et adapter ces procédés de l'approche WSMO à OWL-S, WSDL-S.

7 Conclusion

Le but de ce travail est de proposer une approche pour le service web sémantique en utilisant l'ingénierie ontologique de domaine pour générer les liens sémantiques.

Notre approche consiste en deux phases: l'extraction d'informations utiles, et l'analyse. L'approche permet une classification de SWS WSMO selon le domaine d'ontologie.

A cette fin, nous considérons l'ontologie de domaine comme la principale source sémantique qui permet l'identification de domaine de SWS.

Il serait utile de noter la spécificité de l'étude au SWS WSMO, les autres approches étant génériques et ne traitant pas les particularités de chaque ontologie de service à part.

Après plusieurs expérimentations, les résultats obtenus ont été très satisfaisants en termes de qualité et de quantité. Les modèles générés par le processus de rétro-ingénierie ont été très proches de ceux attendus par les concepteurs. Le taux de descriptions sémantiques générées par le processus d'ingénierie direct a atteint une moyenne de soixante pourcent.

8 Références

Bell, D., de Cesare, S., and Lycett, M. (2005). Semantic Transformation of Web Services. OnTheMove 2005 (SWWS 2005 Workshop), Springer LNCS 3762, pp: 856.

Bell, D., de Cesare, S., Iacovelli, N., Lycett, M., and Merico, A. (2007). A framework for deriving semantic web services. Information Systems Frontiers, Volume 9, Number 1, pp : 69-84, ISSN:1387-3326.

Buitelaar, P., and Gmbh, D. (2003). Ontology Learning for Semantic Web Services. In Proceedings of ONLINE2003, Düsseldorf, Germany.

Gruber, T-R. (1993). A translation approach to portable ontology specifications. Knowl. Acquis., 5(2): 199–220.

Hirst, G., and St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. In Fellbaum, C., ed., WordNet: An electronic lexical database. MIT Press. Pages: 305–332.

Jiang, J., and Conrath, D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings on International Conference on Research in Computational Linguistics, Pages: 19-33.

Leacock, C., and Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. In Fellbaum, C., ed., WordNet: An electronic lexical database. MIT Press. Pages: 265-283.

Lin, D. (1998). An information-theoretic definition of similarity. In Proceedings of the International Conference on Machine Learning.

Lopresti, D., and Nagy, G. (1999). Automated table processing : An (opinionated) survey. In Proceedings of the Third IAPR Workshop on Graphics Recognition, pages 109–134, Jaipur, India.

Martin, D., Paolucci, M., McIlraith, S., Burstein, M., McDermott, D., McGuinness, D., Parsia, B., Payne, T., Sabou, M., Solanki, M., Srinivasan, N., and Sycara K. (2004). Bringing Semantics to Web Services : The OWL-S Approach. In Proceedings of the First International Workshop on Semantic Web Services and Web Process Composition (SWSWPC 2004), San Diego, California, USA.

McIlraith, S-A., Son, T-C., and Zeng, H. (2001). Semantic Web services. IEEE Intelligent Systems, Special Issue on the Semantic Web, Volume 16, Number 2, pp: 46-53.

Patwardhan, S., Banerjee, S., and Pedersen, T. (2003). Using measures of semantic relatedness for word sense disambiguation. In Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics. Pages: 241–257.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In Proceedings of the 14th International Joint Conference on Artificial Intelligence. Pages: 448–453.

Roman, D., Keller, U., Lausen, H., de Bruijn, J., Lara, R., Stollberg, M., Polleres, A., Feier, C., Bussler, C., and Fensel, D. (2005). Web Service Modeling Ontology. Applied Ontology journal, Volume 1, Number 1, Pages : 77-106.

Rumbaugh, J., Blaha, M., Lorensen, W., Eddy, F., Premerlani, W., and Rumbaugh, J. (1991). *Object Oriented Modeling and Design*. Edition Prince Hall Inc. Englewood Cliffs.

Sabou, M., Wroe, C., Goble, C., and Stuckenschmidt, H. (2005). Learning domain ontologies for semantic Web service descriptions. Journal of Web Semantics, Volume 3, pages 340-365, Number 4.

Wu, Z., and Palmer, M. (1994). Verb semantics and lexical selection. In 32nd Annual Meeting of the Association for Computational Linguistics, Pages: 133–138.

Annexe

WSMO-Studio est un environnement de modélisation des services Web sémantiques WSMO, disponible en open source et développé en utilisant la plateforme Eclipse de JAVA.

WSMO-Studio est composé de plusieurs modules. Chacun est dédié à une tâche particulière. Nous avons utilisé plus particulièrement WSML-Validator , pour obtenir des exemples de web service pour la phase de filtrage de service web SWS WSMO.

Summary

With the evolution of a Web , a new technology has emerged; it is the Semantic Web Services (SWS). This new technology can automate the discovery, composition and invocation of services through the web. To enjoy these benefits, a re-engineering process of legacy Web applications to Semantic Web Services is required. The Reengineering allows reusing the Web application functionalities as Semantic Web services ,without rewriting the implementation code of these applications. The objective of this work is to propose an approach based on ontological reengineering for the description and classification of SWS WSMO applications. The approach consists of two phases: first, extracting useful information from source code of web service of WSMO SWS, then the analysis of data derived by the ontology domain by using a similarity criterion.

ETL-XDesign : outil d'aide à la modélisation de processus ETL

Mahfoud Bala, Zaia Alimazighi
LSI, Université des Sciences et de la technologie Houari Boumediene, Alger, Algérie
Mahfoud.bala@gmail.com
alimazighi@wissal.dz

Résumé. Un processus ETL est très complexe en termes de flux de données et des tâches chargées de nettoyer, filtrer, normaliser et charger les données dans l'entrepôt de données. Ces processus sont pris en charge par des moulinettes logicielles classées en 03 catégories (1) L'extraction des données à partir des sources, (2) transformation permettant de livrer des données de qualité ayant une valeur pour l'analyse (3) chargement des données préparées dans l'entrepôt. Nous proposons dans ce papier un outil pour la modélisation des processus ETL aux niveaux conceptuel et logique, les modèles obtenus sont stockés sous forme de documents XML. Nous nous sommes basés sur l'approche de Panos Vassiliadis et al. (Dolap 2002) tout en adaptant le métamodèle conceptuel et proposant un métamodèle au niveau logique.

1 Introduction

Pour faire face à un marché très concurrentiel, les entreprises doivent disposer de grandes capacités d'analyse pour s'adapter aux changements et à une évolution du marché sans cesse. La grande difficulté est au niveau des quantités de données très volumineuses constituées au fur et à mesure et stockées sous forme de fichiers classiques, bases de données dans différents modèles, documents Excel, etc. En plus de cette hétérogénéité, ces données sources n'ont pas été constituées avec des perspectives d'analyse, ce qui rend leur exploitation très difficile.

Un projet décisionnel consiste à exploiter et mettre en valeur des quantités de données créées et stockées durant des années dans des formats et modèles éventuellement très divers. Ces données, malgré qu'elles soient hétérogènes avec une partie non structurée, sont d'une valeur inestimable pour des applications d'analyse et d'aide à la décision.

L'objectif du projet décisionnel est de mettre en place un entrepôt de données qui regroupera l'ensemble des données pertinentes ayant la qualité et la valeur requises par les applications d'analyse et de datamining.

Avant d'être chargées dans l'entrepôt, les données sources passent par un processus d'extraction, de transformation et de chargement, connu sous le nom ETL, pour les préparer et leur donner les propriétés nécessaires en termes de format, de fiabilité, de précision et de pertinence. Des outils dédiés ETL existent sur le marché et sont destinés à prendre en charge ces phases d'extraction, de transformation et de chargement.

Un grand intérêt a été réservé au domaine de l'ETL par la communauté des chercheurs. Il y a ceux qui ont travaillé sur l'aspect modélisation comme 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). D'autres ont travaillé sur la sémantique de l'ETL comme Simitsis (2005), Simitsis & Vassiliadis (DSS 2008), Skoutas et Simitsis (DOLAP 2006, IJSWIS 2007), Skoutas et Simitsis (NLDB 2007), etc.

Un panorama complet sur les différentes contributions dans le domaine de l'ETL a été présenté par Vassiliadis (2009).

Nous nous intéressons dans ce papier à l'analyse et la modélisation d'un processus ETL indépendamment des outils. Il s'agit de formaliser les besoins des utilisateurs en termes d'analyse (mesures et dimensions), de sélection des sources disponibles pouvant répondre à ces besoins, des tâches de transformations à faire subir aux données afin de les préparer et enfin du mappage permettant de préciser les correspondances entre les données sources et les données cibles.

Plusieurs contributions ont été faites dans ce domaine, nous citons particulièrement celles de Panos Vassiliadis et al. (Dolap2002) et Juan Trujilio et Sergio Lujan Mora (ER2003). Le premier a proposé un formalisme graphique ad hoc permettant de modéliser les données (sources, DSA, entrepôt) sous forme de concepts de données caractérisés par un ensemble d'attributs qui serviront à décrire, à travers un certain nombre de transformations, le cheminement des données jusqu'au mappage avec les données de l'entrepôt (tables de dimension et tables de faits). La puissance du formalisme de Vassiliadis et *al.* est dans la description détaillée de la nature des tâches de transformation avec leurs inputs/outputs, leurs paramètres ainsi que leur sémantique d'exécution. L'approche est décrite par un métamodèle qui assure une extensibilité et une ouverture afin de prendre en charge tout type de processus ETL.

La deuxième contribution est basée sur une extension UML avec des profils pour la modélisation multidimensionnelle. Plusieurs schémas sont prévus sous forme de diagramme de classes avec de nouveaux stéréotypes multidimensionnels pour le schéma de l'entrepôt de données en trois niveaux (schéma global en constellation, différents schémas en étoile et hiérarchies des dimensions) ainsi que le schéma du processus ETL et de mappage.

Le présent article est organisé comme suit : la section 1 présentera le processus ETL de manière générale ainsi que les problèmes qu'il pose dans le cadre d'un projet décisionnel. On montrera pourquoi modéliser le processus ETL et que faut-il modéliser. La section 2 sera consacrée à l'approche de Vassiliadis pour la modélisation ETL en présentant ses concepts de base, ses principes et son métamodèle. Nous résumons, dans la section 3, notre contribution sous forme de propositions par rapport à l'approche de Vassiliadis. Nous présenterons l'outil « ETL-XDesign » dans la section 4 et nous terminons cet article par une conclusion et des perspectives dans la section 5.

2 Processus ETL

Un processus ETL sert à extraire des données à partir de diverses sources existantes et hétérogènes, à nettoyer, à filtrer, à normaliser, à agréger et enfin à charger dans l'entrepôt de données. En d'autres termes, le système ETL donne de la valeur et de la pertinence aux données avant de les charger dans l'entrepôt. Le rôle de l'entrepôt de données est de publier ces données « prêtes à l'emploi » pour l'analyse et l'aide à la décision.

La difficulté du processus ETL est dans la diversité des données et leur hétérogénéité. Chaque projet décisionnel comporte ses tâches spécifiques liées aux données propres à l'organisation en question. Ceci fait de l'ETL la partie la plus coûteuse (70%) et la plus longue dans un projet décisionnel.

La figure 1 décrit l'environnement d'un processus ETL. Elle montre les différentes zones de stockage des données ainsi que les tâches ETL qui opèrent sur ces données pour les préparer et les charger dans l'entrepôt.

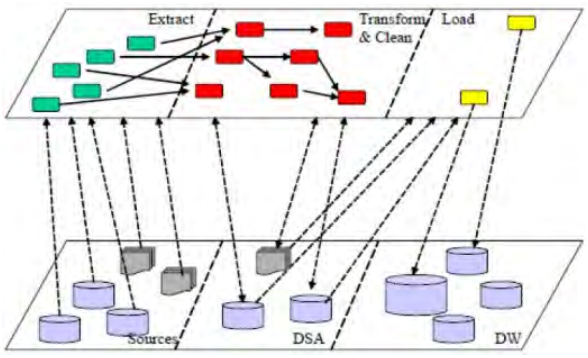


FIG. 1 – *Environnement d'un processus ETL. Panos Vassiliadis et al. (Dolap2002)*

La couche inférieure de la figure 1 présente la partie statique qui montre les sources à partir desquelles les données sont extraites, le data staging area (DSA) dans laquelle sont copiées les données sources et où s'opèrent les tâches de nettoyage, filtrage, conversion, etc. Dans la partie droite, on trouve l'entrepôt de données (DW) représentant le stockage final des données après transformations. Dans la couche supérieure, sont représentées les 03 phases ETL à savoir extraction, transformation et chargement.

2.1 Pourquoi modéliser un processus ETL ?

Selon C. Shilakes et J. Tylman (Merrill Lynch, 1998), les processus de gestion de données (ETL et qualité des données) sont complexes et laborieux.

La plupart des sociétés estiment à 1/3 du coût et 1/3 du temps consacrés à l'ETL pendant le processus de développement. Cependant, la grande majorité des entreprises développent leurs propres outils ETL (non réutilisables) en interne plutôt que d'acheter une solution ETL commercialisée.

	1998	1999	2000	2001	2000	TCAM [1]
Solutions ETL	$101,00	$125,00	$150,00	$180,00	$210,00	20,1%
Taux de croissance		24%	20%	20%	17%	

TAB. 1 – *Estimation du taux de croissance du marché des solutions ETL (en M $)*

Afin de maîtriser cette complexité et de réduire les risques et les coûts de cette étape stratégique dans un projet décisionnel, les concepteurs préfèrent mettre toute la lumière sur ce processus avant d'aborder l'implémentation physique. La modélisation du processus ETL aux niveaux conceptuel et logique en est une des solutions qui contribue à la maîtrise de la complexité et des aléas du processus ETL.

2.2 Modélisation du processus ETL ?

Comme le montre la figure 1, un processus ETL regroupe les éléments suivants :

- Sources de données candidates au processus d'entreposage
- Le DSA pour le stockage intermédiaire des données impliquées dans l'entreposage
- L'entrepôt de données, destination finale des données à préparer
- Les tâches d'extraction, de nettoyage, conversion, filtrage, d'agrégation et de chargement

La modélisation devra formaliser tous ces éléments de manière à comprendre le circuit des données depuis les systèmes sources jusqu'à l'entrepôt en passant par les différentes tâches de transformation. Le mappage des données à différents niveaux est un aspect très important pour comprendre le cheminement des données.

[1] Taux de Croissance Annuel Moyen (TCAM) ou CAGR en anglais.

3 L'approche de Vassiliadis

L'approche de P. Vassiliadis et *al.* est considérée comme l'une des meilleures contributions dans ce domaine. Pour comprendre l'architecture générale d'un processus ETL, la figure 1 montre la partie statique à trois niveaux (sources, DSA, entrepôt) ainsi que la partie dynamique (tâches d'extraction, de transformation et de chargement). Il est intéressant

de visualiser cet environnement statique et dynamique aux niveaux conceptuel, logique et physique.

Le niveau conceptuel consiste à donner une idée très générale sur l'environnement du projet décisionnel à savoir : les besoins des utilisateurs en termes d'analyses (mesures et dimensions), les sources disponibles qui répondent à ces besoins, principales transformations à faire subir aux données sources avant leur chargement dans l'entrepôt. Ce niveau permet aux concepteurs d'aborder les premières réunions pour discuter de la pertinence des données disponibles, leur qualité et sous quels formats existent-elles. Il s'agit aussi de vérifier si tous les besoins peuvent être satisfaits à partir des sources disponibles et quelles sont les transformations nécessaires pour répondre aux exigences de qualité pour l'entrepôt de données.

Le passage du niveau conceptuel vers le niveau logique consiste à affiner le modèle en intégrant plus de détails : séquence d'exécution des activités, planification d'exécution du processus, plan de reprise et de restauration en cas de panne, plan d'administration consistant à faire de l'audit, gestion des accès et sécurité, etc.

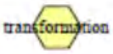
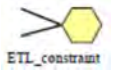
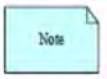
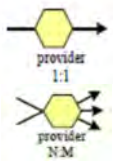
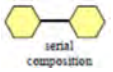
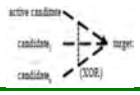
Le niveau physique qui décrit de manière complète le processus ETL doit préciser l'environnement sous lequel s'exécutera le processus : OS, nature des systèmes sources (SGBDs, fichiers plats, Excel, XML, ...), ressources matérielles en termes de serveurs et stations nécessaires ainsi que les profils utilisateurs nécessaires pour faire aboutir le processus de bout en bout jusqu'au rafraichissement de l'entrepôt de données sans échecs.

L'approche de Vassiliadis se caractérise par rapport aux autres au niveau de la modélisation des données et leur cheminement le long du processus ETL. Pour faire toute la lumière sur les tâches de transformations, l'approche représente l'activité à un niveau de granularité très fin : les attributs. Ces derniers sont valorisés et modélisés comme des entités à part entière et participent dans la représentation des détails des tâches de transformation et ce pour comprendre celles-ci à un niveau attribut jusqu'au chargement de celui-ci dans l'entrepôt. Dans ce qui suit, nous résumons de manière succincte le formalisme de modélisation au niveau conceptuel et logique.

3.1 Modélisation conceptuelle

Le formalisme proposé par Vassiliadis permet de représenter les différents objets manipulés dans le processus ETL (sources de données, transformations, contraintes, concepts, attributs, . . .) ainsi que les associations reliant ces objets (mappage) vers ceux de l'entrepôt de données. Il prend en charge les activités fréquemment utilisées telles que l'assignation d'une clé de substitution (surrogate key), contrôle des valeurs NULL, contrôle de type clé primaire, les agrégations, etc. Ainsi, le formalisme proposé permet une personnalisation et une extension du processus ETL. Voici les notations graphiques retenues pour la représentation du processus au niveau conceptuel :

Symbol	Désignation
	Représente une entité dans la source de données, dans le DSA ou dans l'ED

	Comme dans une modélisation E/R, ils permettent de définir les concepts
	Cette association permet de relier le concept à ses attributs
	Abstraction d'un bout ou d'un module complet de code exécutant une tâche ETL : (1) nettoyage/ filtrage de données (comme violation de la contrainte PK/FK, (2) transformations de données (comme une agrégation).
	Permet d'exprimer certaines contraintes sur le contenu de l'ED à travers les attributs de celui-ci (PK, FK, NOT NULL, …)
	Permet d'expliquer des choix de conception, de préciser une sémantique ou une contrainte à vérifier en temps réel (temps d'exécution, événements, erreurs, …)
	Représentent le mappage entre les données sources (input) et les données de l'ED (output) via une transformation. Le cas simple (1 :1) représente le cas où une donnée source (input) donne en sortie, à travers une transformation, à une donnée de l'ED (output). Le cas général (N:M) représente le cas où un ensemble de données sources (inputs) transformées donneront naissance à plusieurs données de l'ED (outputs)
	Permet de modéliser le cas où dans une association « Provider » les données passent par plusieurs transformations avant de donner naissance aux données de l'ED.
	Permet de préciser les sources de données candidates à l'alimentation de l'ED en mettant en relief la source active. Une des sources pourra être utilisée en même temps (XOR)

TAB. 2 – *Formalisme de Vassiliadis pour la modélisation au niveau conceptuel*

Pour que le formalisme soit générique et prendra en charge n'importe quel type de processus sur le niveau conceptuel, Vassiliadis et al. ont proposé un métamodèle (FIG. 4) regroupant l'ensemble des éléments pouvant intervenir dans un processus : métaclasse concepts, métaclasse attributs, métaclasse transformations, ...

Pour mieux comprendre l'utilisation de ce formalisme, voici un exemple de processus ETL sur la gestion des étudiants d'un établissement universitaire :

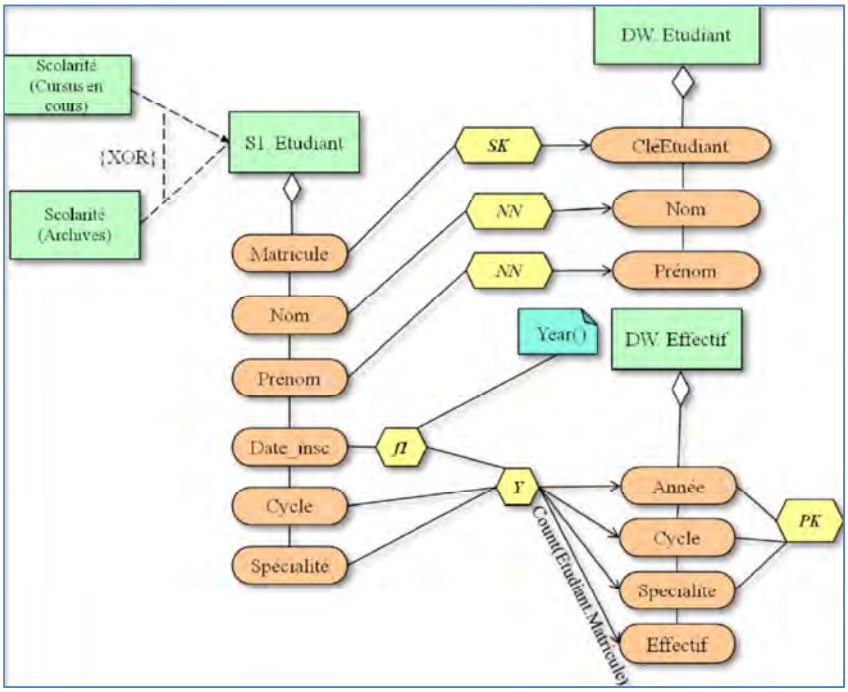


FIG. 3 – *Exemple d'un processus ETL relatif à un établissement universitaire*

Les sources *scolarité (en cours)* et *scolarité (archives)* contiennent des données intéressantes pour alimenter l'entrepôt des données. Le schéma nous renseigne que la source *scolarité (cursus en cours)* est celle qui a été retenue pour l'entreposage (*active candidate*).

Les données extraites à partir de cette source sont copiées dans le concept *S1.Etudiant* pour opérer les transformations nécessaires avant leur chargement dans l'entrepôt. Comme expliqué précédemment, le schéma montre les attributs du concept *S1.Etudiant* comment sont traités au niveau des transformations *SK* (*serrogate key*), *NN* (*not null*), *f1* (*fonction year()*), *agrégation γ* (*count()*). *PK* étant une contrainte (*Primary key*) sur les données *Année*, *Cycle* et *Spécialité*.

Les classes manipulées dans le processus ETL (*attributs*, *concepts*, *transformations*, …) sont des instances des métaclasses du métamodèle présenté sur la figure 4: « *Concept* », « *Attribut* », « *Candidate* », « *Active candidate* », « *transformation* », « *ETL_contraint* »
et les associations manipulées dans le processus ETL (*part-of*, *provider*, …) sont des instances des métaclasses: *Part-of*, *Serial-composition*, *Provider*. C'est pour cela que ces dernières héritent de la métaclasse générale « *Relationship* »

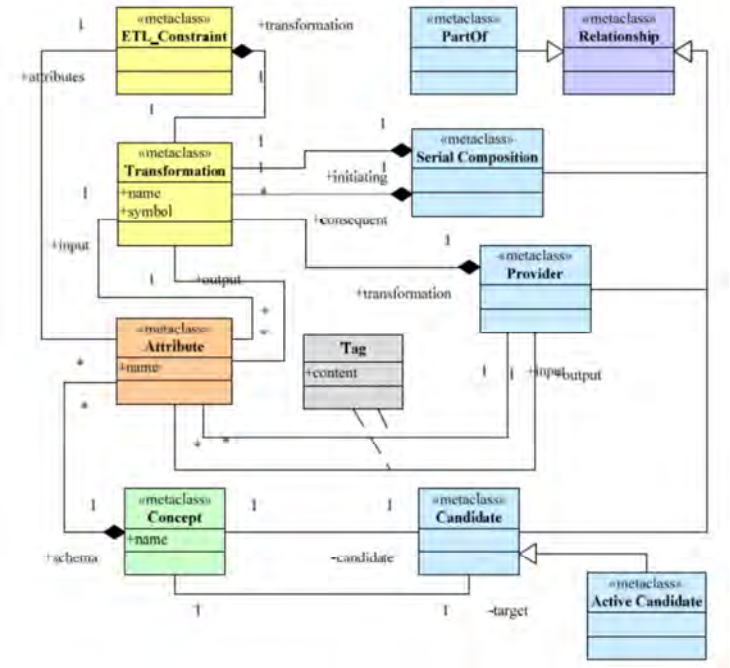


FIG. 4 – *Métamodèle proposé pour la modélisation au niveau conceptuel*

Afin de faire le lien entre le métamodèle de la FIG. 4 et le processus à modéliser, la FIG.5 présente les trois couches correspondantes au métamodèle, modèle et le schéma conceptuel du processus.

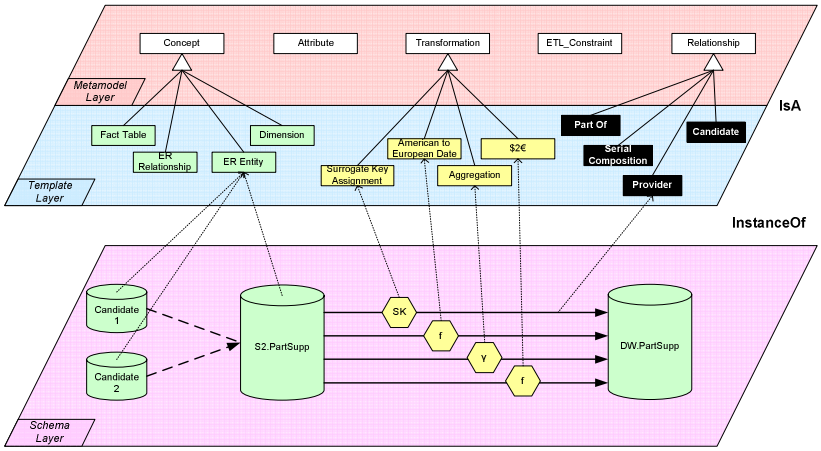


FIG. 5 – *Couches métamodèle, modèle et schéma conceptuel d'un processus*

La couche métamodèle étant le moyen permettant l'extensibilité de l'approche pour intégrer des spécificités de certains processus. La couche modèle étant la palette des éléments de modélisation utilisée par le concepteur pour modéliser des processus ETL.

En plus du formalisme décrit précédemment, une démarche a été proposée pour la modélisation du processus :

- Étape 1 : Identification des sources appropriées de données
- Étape 2 : identification des sources candidates et candidates actives
- Étape 3 : mappage entre les données sources et celles de l'ED
- Étape 4 : Annotation du diagramme avec des contraintes d'exécution

3.2 Modélisation logique

Le modèle logique, appelé aussi graphe d'architecture, est une représentation du système tel qu'il sera implémenté dans la machine mais sans faire référence à un langage de programmation. Il capture les flux de données à partir des sources vers l'entrepôt de données grâce à une suite d'activités synchronisées chargées de préparer les enregistrements de données. Il s'agit de préciser le type des données utilisées lors des traitements, les inputs/outputs et paramètres de chaque activité, la séquence des activités qui montre comment les outputs d'une activité A1 sont utilisés comme inputs dans une activité A2. D'autres activités, par contre, peuvent s'exécuter en parallèle.

Le graphe d'architecture comporte des sommets et des arêtes. Les sommets représentent les types de données utilisés, les types de fonctions, les constantes, les attributs, les activités, les enregistrements, les paramètres et les fonctions. Les arêtes représentent les différents types de relations entre ces entités. Par conséquent, un graphe d'architecture couvre :

- Les enregistrements de données (RecordSet) du scénario avec leurs composants.
- Les activités (Activity)
- Les flux des données (Dataflow)
- Le typage des entités (DataTypes)
- La séquence (Execution sequence) d'un scénario par des paramètres spécifiques.

On distingue plusieurs types de sommets dans le graphe d'architecture :

a) Les entités élémentaires
   - DataTypes: Chaque type de données T est caractérisé par un nom et un domaine,
   - Attributes: Les attributs sont caractérisés par un nom et un type de données.
   - Schema: est une liste finie d'attributs. Toute entité caractérisée par un ou plusieurs schémas est appelée entité structurée.

b) RecordSet: Un enregistrement est défini comme l'instanciation d'un schéma à une liste de valeurs appartenant aux domaines des attributs respectifs au schéma.

c) Function: Il est supposé l'existence d'un ensemble fini de types de fonctions. Un type de fonction comporte un nom, une liste finie de types de données des paramètres et un seul type de retour de données.

d) Activity: Les activités sont considérées comme des abstractions logiques représentant des parties ou des modules de codes complets. Elles sont représentées par un langage LDL qui, d'une part, définit le code source d'une activité et d'autre part évite le traitement des spécificités d'un langage particulier.

Une activité est formellement décrite par un nom (*Name*), schéma d'entrée (*Input Schemata*), schéma de sortie (*Output Schema*), schéma des rejets (*Rejections Schema*), liste des paramètres (*Parameter List*), sémantique opérationnelle des sorties (*Output Operational Semantics*), sémantique opérationnelle des rejets (*Rejection Operational Semantics*)

Les différents types de relations constituant un graphe d'architecture sont :

- *PartOf*: Met en relation les attributs et les paramètres avec les activités, enregistrements ou fonctions auxquels ils appartiennent.
- *Instance-Of*: Permet de capturer les informations sur le typage des attributs et des fonctions.
- *Regulator*: Cette relation est définie entre les paramètres d'une activité et les termes (attributs ou constantes) qui alimentent cette activité.
- *Provider*: Cette relation capture le passage des données entre fournisseurs (*Providers*) et consommateurs (*Consumers*) par la relation Provider entre les attributs des schémas concernés.
- *Derived provider*: Cas particulier de la relation Provider. Cette relation est utilisée lorsque les attributs de sortie sont générés par la composition des attributs d'entrée et des paramètres de l'activité.

La figure 6 présente un modèle au niveau logique de l'exemple universitaire :

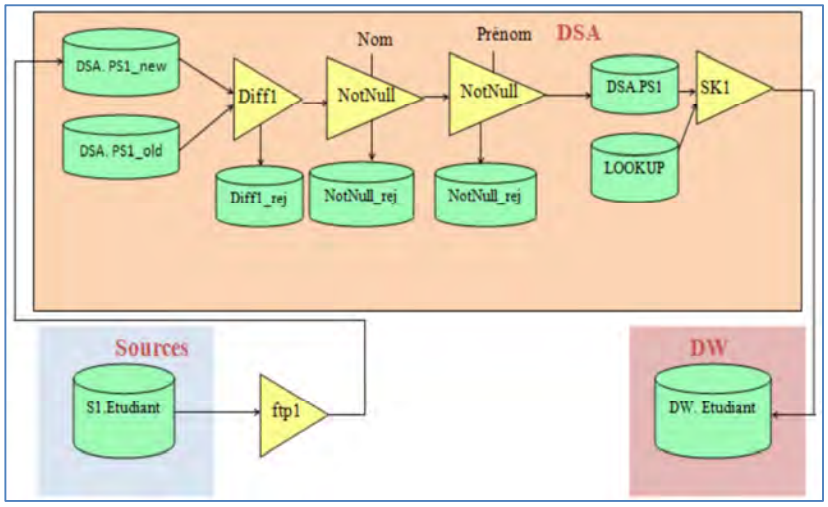


FIG. 6 – *Exemple du processus ETL au niveau logique relatif à un établissement universitaire*

4 Contribution

Une bonne analyse du métamodèle, des notations graphiques de modélisation et de l'environnement d'un processus ETL nous a permis de comprendre le lien entre les différentes parties de l'approche de Vassiliadis, de mettre en valeur certains aspects implicites et d'apporter quelques améliorations. Au niveau conceptuel, nous avons proposé quelques modifications au niveau du métamodèle dans le but de le rendre plus fin et plus exhaustif. Pour le niveau logique, nous avons proposé un métamodèle qui décrit, de manière similaire à celui du niveau conceptuel, les éléments manipulés au niveau d'un modèle logique. Voici un résumé de notre contribution :

1) En analysant l'environnement d'un processus ETL (figure 1) et la notation proposée pour la modélisation d'un processus ETL (figure 3), nous avons découvert le lien non explicité dans les papiers de Vassiliadis qui consiste à délimiter les différentes phases (sources de données, extraction, DSA, transformation, entrepôt de données) dans le schéma d'un processus comme le montre l'exemple de la figure 7

2) Dans le métamodèle de Vassiliadis, une partie seulement des schémas sources, DSA et entrepôts de données est représentée, i.e les concepts de données nécessaires aux besoins d'analyse dépourvues des relations entre eux. En rajoutant une association réflexive au niveau de la métaclasse « *Concept* » avec « *Attribut* » comme metaclasse-associative pour exprimer l'attribut assurant le lien (*Foreign Key*), on aura représenté tout le contenu des schémas des données. Pour l'entrepôt en particulier, on aura le schéma en étoile (voir figure 8)

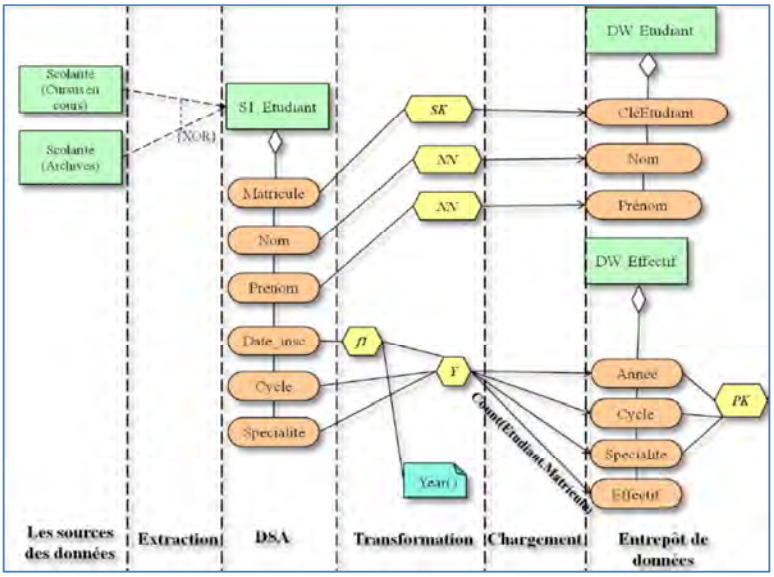


FIG. 7 – *Délimitation des différentes étapes dans le schéma du processus*

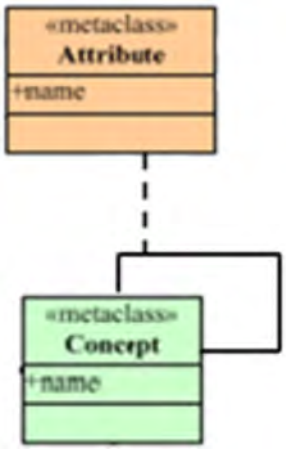


FIG. 8 – *Représentation des relations entre concepts dans le métamodèle*

3) Dans le métamodèle, la Cardinalité (*1:1)* de la métaclasse « *candidate* » avec « concept » ne représente pas de manière naturelle la réalité. Une relation « *candidate* » possède plusieurs candidats sources et un seul candidat cible. Pour exprimer ceci, nous proposons alors de mettre une cardinalité (*1..**) dans l'association *candidate* au lieu de '1'. Si on note *R* la relation candidate, *C1, C2, C3* les concepts candidats et *C4* comme concept cible. Les instances des deux associations *Candidate* et *Target* se présentent comme suit :

Relation	*Concept candidat*
R	*C1*
R	*C2*
R	*C3*

Relation	*Concept cible*
R	*C4*

TAB. 5 – Instances des relations *Candidate* et *Target* avec une cardinalité *'1..*'*

Si on se réfère au métamodèle, les instances se présentent comme suit :

Relation	*Concept candidat*
R1	*C1*
R2	*C2*
R3	*C3*

Relation	*Concept cible*
R1	*C4*
R2	*C4*
R3	*C4*

TAB. 6 – Instances des relations *Candidate* et *Target* avec une cardinalité '1'

Il est clair que la représentation proposée (TAB. 5) est plus naturelle et optimisée.

4) Pour lui donner plus de sens dans le métamodèle, la relation "*Part-Of*" représentée comme métaclasse doit être associée avec la métaclasse "*Concept*" et la métaclasse "*Attribut*" pour représenter que tel attribut est un constituant de tel concept.

5) La métaclasse "*Tag*" telle que conçue ne permet pas d'utiliser une note partout dans le modèle conceptuel. Il serait plus intéressant de la relier à la métaclasse "*Relation*" (afin d'impliquer toutes les relations), les attributs, les concepts et les transformations.

6) Pour le niveau logique, nous avons proposé un métamodèle (voir figure 6) absent sur les papiers de P.Vassiliadis et *al.*

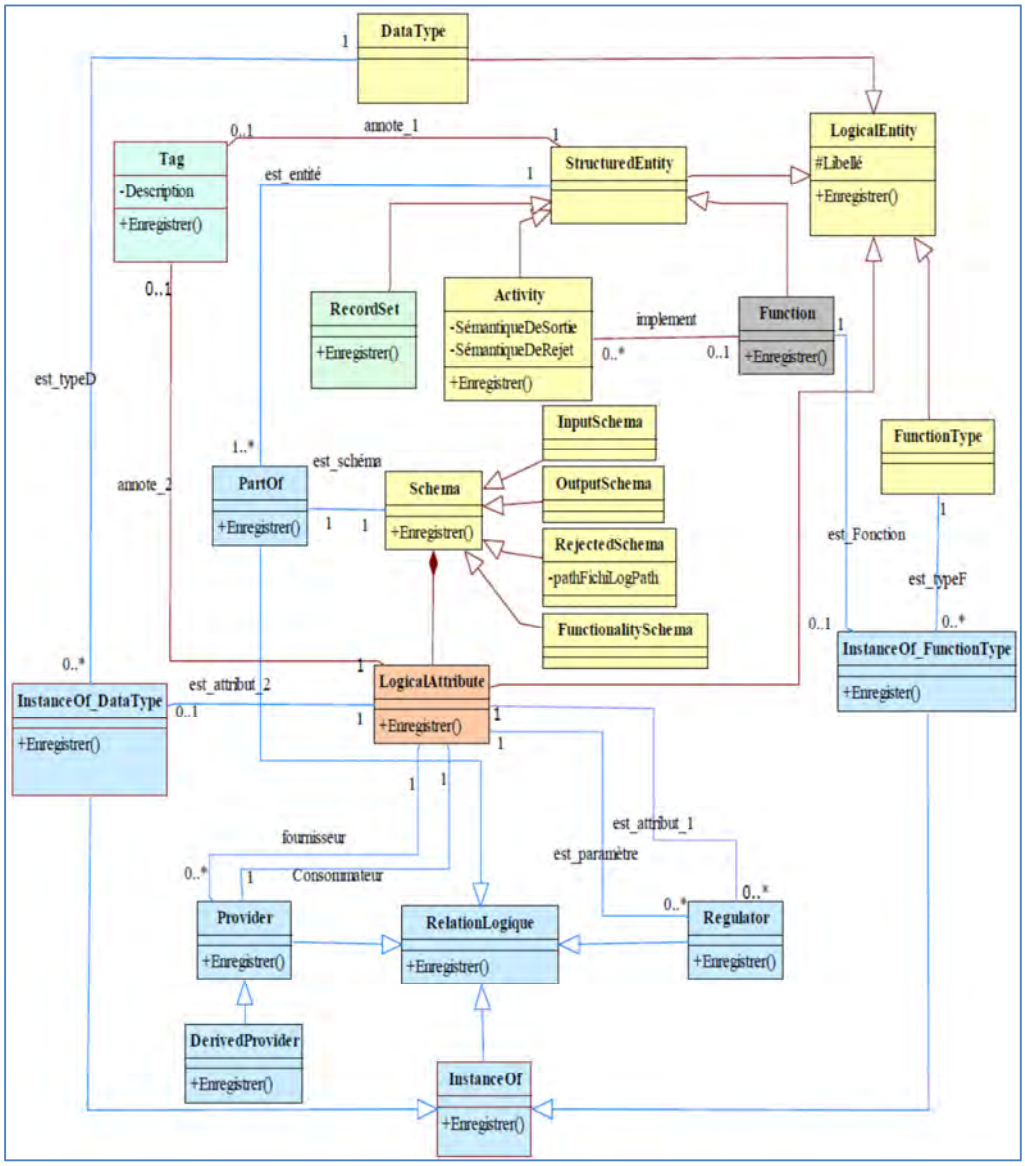


FIG. 9 – *Métamodèle proposé pour le niveau logique*

5 Présentation de l'outil ETL-XDesign

En se basant sur le métamodèle et formalisme de P.Vassiliadis, nous avons mis en œuvre un outil d'aide à la modélisation d'un processus ETL sur les deux niveaux conceptuel et logique. Dans la figure qui suit, nous présentons les fonctionnalités offertes par l'outil :

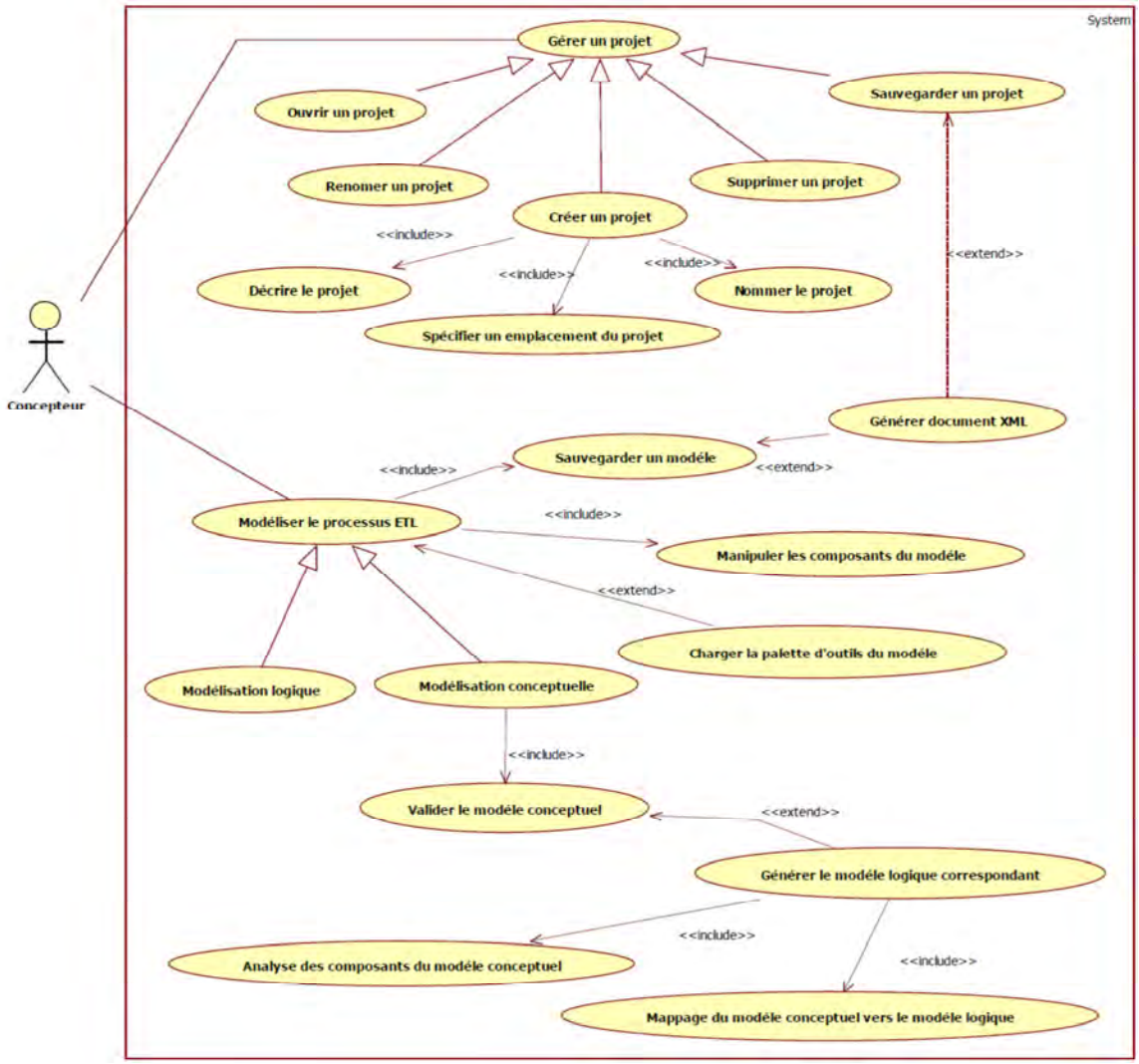


FIG. 10 – *fonctionnalités de l'outil ETL-XDesign*

ETL-XDesign dispose d'une première interface de paramétrage pour la création d'un projet (Nom du projet, description, date création, emplacement du projet et de tous les modèles

conceptuels et logiques associés). Une fois le projet paramétré et créé, le concepteur pourra aborder le travail de modélisation en utilisant une interface organisée en trois compartiments: Gestionnaire d'objets (à gauche), Espace de modélisation (au milieu) et la Palette d'outils (à gauche).

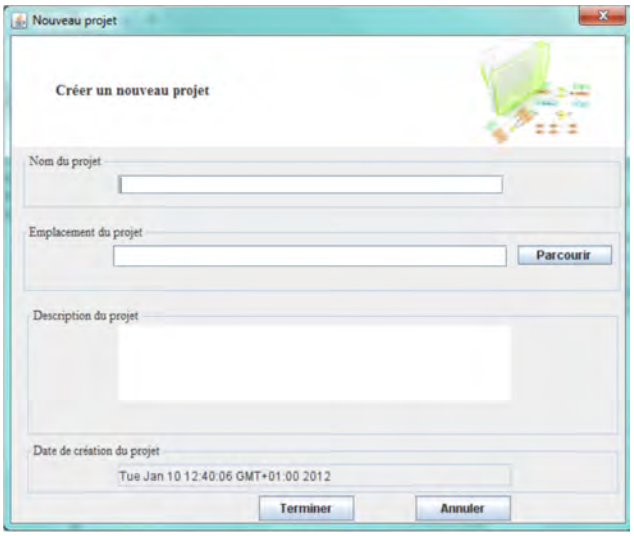


FIG. 11 – *Interface de création et de paramétrage d'un projet*

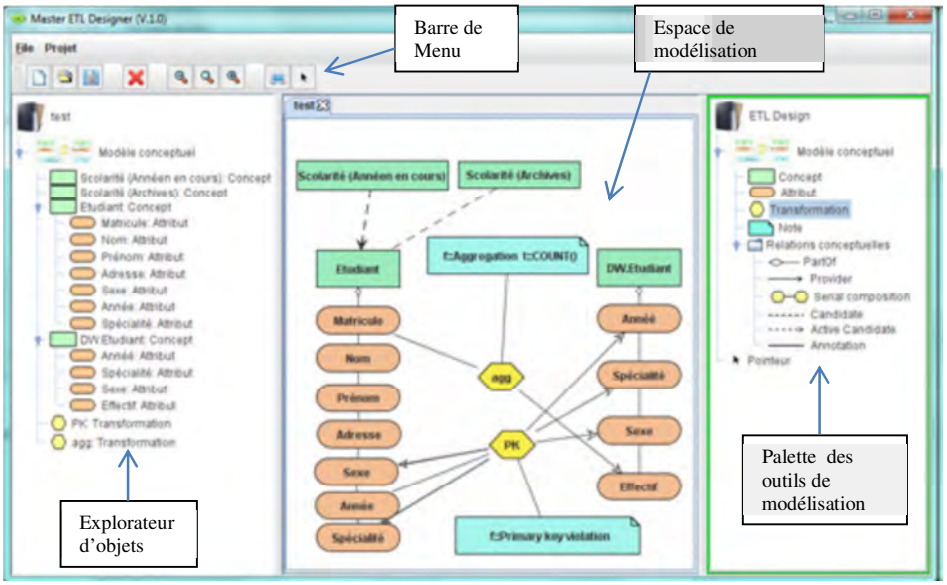


FIG. 12 – *Interface de modélisation*

Le modèle conceptuel représenté ci-dessus est stocké sous forme de document XML. Ce dernier étant une technologie standard avec des qualités en termes de légèreté et de portabilité, ce qui permettra d'exporter les modèles construits sous ETL-XDesign vers d'autres outils. L'utilisation de ETL-XDesign ne nécessite pas l'installation d'un SGBD pour stocker les données et les modèles ETL.

```
<?xml version="1.0" encoding="UTF-8"?>
- <test>
  - <Concept>
      <ID>0</ID>
      <nom>Scolarité (Annéen en cours)</nom>
    - <point>
        <ptX>18</ptX>
        <ptY>16</ptY>
      </point>
    - <dimension>
        <height>32</height>
        <width>163</width>
      </dimension>
    </Concept>
  - <Concept>
      <ID>1</ID>
      <nom>Scolarité (Archives)</nom>
    - <point>
        <ptX>31</ptX>
        <ptY>89</ptY>
      </point>
    - <dimension>
        <height>29</height>
        <width>131</width>
      </dimension>
    </Concept>
  - <Concept>
      <ID>2</ID>
      <nom>Etudiant</nom>
    - <point>
        <ptX>219</ptX>
        <ptY>57</ptY>
      </point>
    - <dimension>
        <height>40</height>
        <width>90</width>
      </dimension>
    </Concept>
```

FIG. 13 – *document XML contenant le modèle du processus de la figure 12*

6 Conclusion et perspectives

L'article a mis en valeur l'importance du processus ETL dans le cadre d'un projet décisionnel. La modélisation de ce processus 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 d'aborder l'implémentation ou le paramétrage de l'outil ETL. L'approche de Vassiliadis étant une des plus intéressantes dans ce domaine. L'étude de cette approche nous a permis de découvrir des aspects très intéressants dans les processus ETL. Notre contribution au niveau de l'approche se résume en un ensemble de remarques et de propositions pour affiner davantage le métamodèle et refléter au mieux la réalité d'un processus ETL.

La mise en œuvre de l'outil ETL-XDesign pour la modélisation des processus a permis d'approfondir davantage l'approche de Vassiliadis et en même temps nous a fait découvrir l'importance et la complexité de l'ETL. Le niveau physique, qui modélise le processus ETL dans un environnement caractérisé par les moyens techniques (équipements et environnement logiciel) et les profils utilisateurs nécessaires pour le bon fonctionnement du processus, mérite aussi une bonne analyse, proposition d'un formalisme, d'un métamodèle ainsi qu'une implémentation.

Pour en faire un outil de modélisation complet, ETL-XDesign devra être étendu en intégrant la modélisation au niveau physique. Le mappage Conceptuel-Logique et Logique-physique est un aspect très intéressant qui mérite du travail en profondeur afin d'assurer la génération des modèles logiques et physique de manière semi-automatique mais avec un moindre effort pour le concepteur. Aussi, Un benchmark sera nécessaire pour évaluer la robustesse de l'outil et sa fiabilité face à des modèles complexes et volumineux.

Références

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

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

Lujàn-Mora S, 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. (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 process is very complex in terms of data flows and tasks entrusted to clean, filter, normalize and load the data into the data warehouse. These processes are supported by pieces of software classified in 03 categories (1) Data extraction from sources, (2) processing to deliver quality data with value to the analysis (3) loading data prepared in the warehouse. We propose in this paper a tool for modeling ETL processes in the conceptual and logical models whose products are stored as XML documents. We based on the approach and metamodel of Panos Vassiliadis and *al.* (Dolap 2002) while adapting the conceptual metamodel and propose a logical metamodel.

Gestion des connaissances médicales par l'intégration des données hétérogènes

Zerf Boudjettou Nadjet, Oukid Khouas Saliha
LRDSI, département informatique, Université Saad Dahleb Blida
BP270 Route de Soumaa, Blida
+213 (0) 662 416 736
zerf_na@yahoo.fr, zerf.na@gmail.com, osalyha@yahoo.com, oukhouas@univ-blida.dz

Abstract. La gestion des connaissances consiste à acquérir et représenter les connaissances utiles à un domaine, une tâche ou une organisation particulière dans le but d'en favoriser l'accès, la réutilisation et l'évolution. Cela revient généralement à construire, maintenir et faire évoluer une représentation explicite de ces connaissances. Il s'agit ensuite de fournir un accès à ces connaissances, c'est-à-dire de les diffuser dans le but d'en permettre une utilisation efficace. La gestion des connaissances dans le domaine médical vise à améliorer les performances de l'organisation médicale en permettant aux individus de l'établissement de soins (médecins, infirmières, paramédicaux, etc.) de capturer, partager et appliquer des connaissances collectives pour prendre des décisions optimales en temps réel. Dans cet article nous proposons une approche de gestion des connaissances basée sur les techniques d'intégration des données hétérogènes dans le domaine médical en réalisant un entrepôt de données, d'extraction des connaissances à partir des données médicales en choisissant une technique du data mining, et enfin d'exploitation de ces connaissances dans un système de raisonnement à base de cas.

Mots-clés. Entrepôt des données, Data mining, Extraction des connaissances à partir de données ECD, Gestion des connaissances médicales, règle d'association.

1 Introduction

L'orientation des patients vers les différents services et les différentes spécialités est un problème majeur dans les grands établissements de santé tels que les CHU (Centre Hospitalo-universitaire), (Voros S et al. 2006). Si le patient ne peut pas reconnaître les premiers signes d'un problème médical ou d'une lésion, il peut prendre un rendez-vous de plus de deux ou trois mois chez un médecin spécialiste dans un service sans qu'il soit le bon médecin.

La prise de décision médicale pertinente tant au niveau diagnostic que thérapeutique nécessite une bonne connaissance du cas à traiter. Le médecin a besoin de connaître ses données cliniques, ce qui suppose une forte collaboration entre les différents professionnels de santé et une interopérabilité entre les systèmes utilisés. Etant donné la complexité du domaine médical, nous rencontrons plusieurs problèmes tels que:

- La diversité des sources d'informations distribuées et leurs hétérogénéités,
- Les problèmes d'accès aux informations pertinentes pour les soins sont liés à la dispersion des informations médicales sur différentes structures de santé dont les systèmes d'information sont souvent autonomes et hétérogènes.

– La formalisation de l'information ne permet pas l'extraction, le partage, la diffusion et l'exploitation de ces connaissances médicales.
– La difficulté de comprendre les mécanismes d'interprétation et de raisonnement médical.

2 Objectif du travail

La prise en compte des problèmes cités est une des clés de la mise en place d'un système qui permet à la gestion des connaissances d'intervenir dans le domaine médical, nous proposons notre système qui peut :
– Donner une solution d'intégration des données hétérogènes,
– Proposer une démarche d'extraction des connaissances pertinentes à partir des données,
Proposer un système de raisonnement à base de cas d'orientation des patients vers des services médicaux.

Notre objectif est de concevoir un système de gestion de connaissances dans le domaine médical. Nous allons montrer dans ce travail comment nous avons pu répondre à cet objectif tout en essayant de prendre en compte la synergie entre les différentes approches telles que les entrepôts de données, l'extraction de connaissances à partir de données et le raisonnement à base de cas.

3 Etat de l'art

La gestion des connaissances médicales s'attache à développer et à évaluer des méthodes et des systèmes pour l'acquisition (Voros S et al. 2006), le traitement et l'interprétation des connaissances extraites à partir des données « patient » avec l'aide des connaissances issues de la recherche scientifique. Pour réaliser ces objectifs, le domaine médical utilise des méthodes scientifiques qui héritent de l'informatique, l'intelligence artificielle, des mathématiques et des sciences de gestion (Soualmia L. et Darmoni 2005).

Les premiers systèmes intelligents ou d'aide au diagnostic en médecine étaient des systèmes de raisonnement à base de cas (Stefanelli M 2002). Puis, au fur et à mesure du temps, et suivant les besoins, des méthodes de data mining ont été intégrées à ces systèmes. Ces méthodes sont venues combler les lacunes des systèmes de Raisonnement à Base de Cas (Naiditch M 2005).

4 Approche Proposée

Dans cette partie nous allons décrire l'approche proposée, nous présentons le processus de construction de l'entrepôt de données médical pour résoudre le problème d'hétérogénéité des sources de données, ensuite nous choisissons une des techniques du data mining pour l'extraction des connaissances à partir de données, enfin le résultat de l'approche proposée est une base de connaissances utilisée pour le système de raisonnement à base de cas.

4.1 Intégration des données hétérogènes

Cette partie du travail consiste à extraire, à l'avance, les données médicales pertinentes pour l'usage des utilisateurs (infirmier, médecin, radiologue, biologiste…), à les filtrer, les transformer et les stocker. Pour répondre à nos besoins nous allons construire un entrepôt des données médical qui stocke physiquement les données des sources réparties, ces dernières correspondent généralement aux sources de données opérationnelles de l'établissement de santé.

4.1.1 La conception de l'entrepôt de données médicales

Dans notre travail, nous avons choisi l'architecture réelle, elle est généralement retenue pour les systèmes décisionnels, c'est une phase de préparation des données qui vont être exploitées par la suite, le stockage de ces données est réalisé dans un SGBD séparé du système opérationnel, ces systèmes sont conçus d'une façon autonome et distribuée. Le système intégral regroupe l'ensemble des systèmes de production tels que: la gestion des consultations, la gestion des RDV médicaux, la gestion des admissions, la gestion d'hébergement hospitalier, la gestion de laboratoires et analyses médicales, la gestion de radiologie et imagerie médicale, la gestion gynécologie et bloc obstétrical, la gestion financière et comptabilité, la gestion du patrimoine, la gestion de la pharmacie et des stocks de médicaments, la gestion d'hémodialyse, la gestion des ressources humaines (GRH).

4.1.2 La construction de l'entrepôt de données médicales

Pour la construction de l'entrepôt de données médicales, nous avons suivi le processus d'ETL (Extraction, Transformation, Loading). (Benitez.E 2002), (Aberer.K et Hemm.K 2005):

1) L'extraction des données

Cette phase collecte les données utiles des sources de données opérationnelles à partir des différentes sources hétérogènes, (Jérôme.D 2006).

2) La transformation des données

Les sources peuvent également présenter des problèmes d'hétérogénéité sémantique, ces problèmes sont traités en spécifiant un ensemble de règles de transformation pour arriver à une représentation uniforme, par exemple, remplacer le mot « genre » par « sexe » ou bien «Féminin» par «F» et «Masculin» par «M».

3) Le chargement des données intégrées dans le système cible

Le processus de rafraîchissement dans notre système est incrémental, la mise à jour utilise les changements dans les sources chaque fois qu'une source change ou bien de manière périodique avec une période qui dépend des besoins des utilisateurs et de la charge d'accès à l'entrepôt.

4.1.3 La construction du schéma de l'entrepôt

Le résultat du traitement de construction de l'entrepôt de données consiste à définir un schéma global fournissant une vue intégrée des sources qui vont être exploitées par la suite dans le processus d'extraction des connaissances à partir des données (María Trinidad et Serna Encinas 2005).

Nous avons définis dans le schéma global une représentation des différents paquetages et classes qui composent l'entrepôt de données médicales.

4.1.4 Data Mart Consultation

Le data mart « Consultation » est ciblé et piloté par les besoins de notre système. Il a la même vocation que l'entrepôt de données médical (fournir une architecture décisionnelle), mais il vise à résoudre notre problématique avec un nombre d'utilisateurs plus restreint.

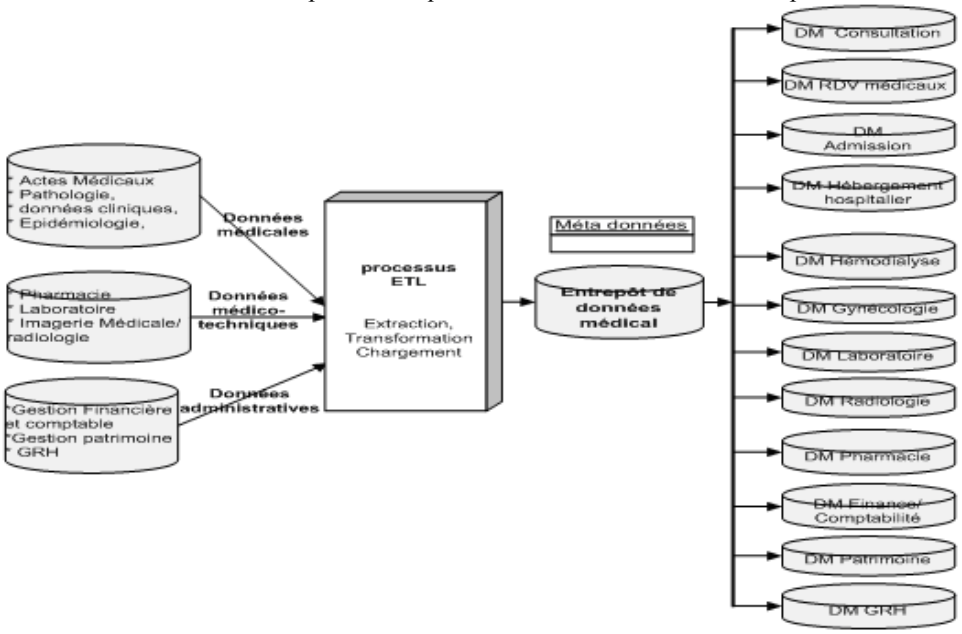


Fig. 1- Décomposition de l'entrepôt de données en plusieurs Data Marts.

Notre Travail est basé sur le data mart « Consultation », il vise à résoudre notre besoin d'orientation médicale vers les services. Il regroupe trois paquetages :

1. Paquetage consultation : contient les classes qui détaillent les données de la consultation telles que code consultation, date consultation, les signes fonctionnels, les douleurs, les symptômes, le diagnostic et le traitement.
2. Paquetage patient : contient les classes qui détaillent les données du patient telles que nom, prénom, sexe, age, poids, date de naissance, adresse….
3. Paquetage affectation : contient les services de la consultation.

4.1.5 Modélisation multidimensionnelle du Data Mart « consultation »

La modélisation que nous avons adoptée pour notre Data Mart est le schéma en étoile. Ce schéma est composé d'une relation de fait « Consult » et de dix relations de dimensions (Dossier_Patient, Service, Age, Type_piods, Fievre , Douleur, Signe1, Signe2, Signe3, Signe4).

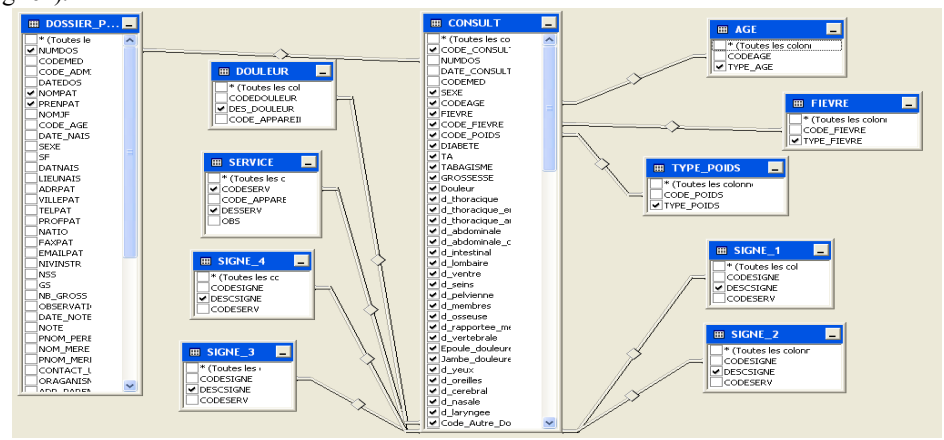


Fig.2 -. Schéma en étoile du Data Mart « Consultation ».

4.2 Extraction des Connaissances à partir des Données

Dans cette partie du système, nous allons réaliser un processus d'extraction des connaissances médicales à partir de l'entrepôt de données (data mart « consultation ») construit dans la partie précédente.

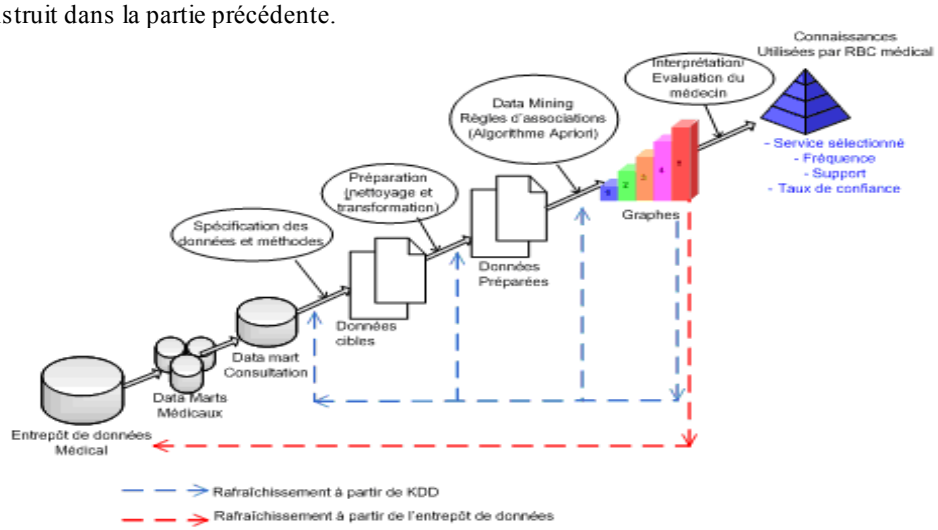


Fig. 3- Processus d'extraction des connaissances à partir de données médicales.

4.2.1 La sélection des données

Cette phase se résume en la sélection des données dont l'exploitation permet de répondre à notre problématique, elle implique : spécification des données, spécification de méthodes de data mining, spécification de la mesure, représentation des résultats du data mining, représentation de la connaissance extraite.

4.2.2 La préparation des données

Dans cette étape, plusieurs procédures sont nécessaires (Azuaje.F et al. 2004): la procédure de nettoyage des données (Nous avons procédé à plusieurs méthodes qui permettent de compléter les données manquantes dans l'entrepôt de données médicales), la procédure de transformation (Cette procédure est déjà réalisée dans la construction de l'entrepôt de données).

4.2.3 Le data mining

C'est le coeur du processus d'ECD (Brachman.J and T. Anand 1996). Il s'agit à ce niveau de trouver des connaissances à partir des données. Le travail consiste à appliquer des méthodes intelligentes dans le but d'extraire cette connaissance.

1. Choix de la méthode

Tout le problème du data mining réside dans le choix de la méthode adéquate à un problème donné. La méthode que nous avons optée dans notre étude et qui répond à nos besoins est celle des règles d'association.

Dans notre travail nous avons adopté cette technique pour la détection des règles d'association de type « ensemble de signes fonctionnels => service ». La recherche de règles d'association est une méthode non supervisée car on ne dispose en entrée que de la description des signes fonctionnels. Une règle d'association est une règle de la forme : *Si <condition> Alors <résultats>*. (Agrawal.R 1993)

2. Choix de l'algorithme : Algorithme Aprior, (Agrawal.R 1993)

Notre data mart contient une masse très importante de données, et notre objectif est de découverte la connaissance pertinente et utile pour pouvoir orienter les patients vers les différents services. Chaque enregistrement stocké dans le data mart « Consultation » contient les données d'un patient: Identification (nom, prénom, age, sexe, poids, adresse), Signes fonctionnels du patient (fièvre, diabétique, tension artérielle, la liste des douleur, la liste des autres signes), Données administratives (code consultation, numéro de dossier, le médecin traitant, date de consultation, service de consultation).

Le calcul des critères globaux (fréquence, support et le taux de confiance) de tous les ensembles des enregistrements qui existent dans le data mart de données est effectué. Ainsi, un enregistrement ayant un taux de confiance supérieur à un seuil est qualifié de fréquent.

(a) La fréquence

La fréquence d'une règle correspond au nombre d'apparitions simultanées des signes fonctionnels d'un enregistrement sans tenir compte du service de consultation :

$$Fréquence = freq\ (signe\ fonctionnel_1, signe\ fonctionnel_2, \ldots\ signe\ fonctionnel_n) \quad (1)$$

(b) Le Support

Le support d'une règle correspond à la fréquence simultanée d'apparition des signes fonctionnels d'un enregistrement qui figurent dans la condition et qui donnent le même service de consultation qui correspond au résultat:

$$Support = freq\ (signe\ fonctionnel_1, signe\ fonctionnel_2, \ldots\ signe\ fonctionnel_n\ et\ Service_k) \quad (2)$$

(c) Le taux de confiance

La confiance est le rapport entre le support et la fréquence, Soit:

$$Confiance = freq((signe\ fonctionnel_1, signe\ fonctionnel_2, \ldots\ signe\ fonctionnel_n)\ et\ Service_k)/freq(signe\ fonctionnel_1, signe\ fonctionnel_2, \ldots\ signe\ fonctionnel_n) \quad (3)$$

$$Confiance\ k = Support\ k\ /\ fréquence. \quad (4)$$

Pour notre application, nous avons considéré une centaine d'attributs avec plus de 10 000 enregistrements, l'application de l'algorithme **apriori** a obtenu **327 règles d'association** qui a donnée lieu à **161 cas**.

- La liste des attributs qui constituent la partie condition de la règle d'association:

Total Cas 161

>	Sexe	Type d'a...	Fièvre	Type fiè...	Type poi...	Diabète	TA	Tabagis...	Grossesse	douleur	d.thoraci...	d.thoraci...
	M	Enfant ...	1	progress...	poids no...	0	0	0	0	1	0	0
	M	Jeune ...	0	ND	poids no...	1	1	1	0	1	0	0
	M	Jeune ...	1	aigu ...	obésité ...	0	0	0	0	1	0	1
	M	Enfant ...	1	progress...	poids no...	0	0	0	0	1	0	0

Fig. 4- Sélection d'une règle d'association.

- Partie résultat de la règle d'association du cas sélectionné:

Nombre de service est: 3

SERVICE	Frequence	Support	Confiance
Service Cardio-vasculaire	46	10	21,7391304347826
Pneumologie	46	10	21,7391304347826
Chirurgie vasculaire et thoracique	46	26	56,5217391304348

Fig. 5- La fréquence, le support et le taux de confiance de la règle sélectionnée.

4.2.4 Evaluation et présentation des résultats

– l'interaction avec l'expert du domaine est privilégiée dans notre système ; la validation par expertise est réalisée par un expert du domaine (médecin) qui jugera la pertinence des résultats produits (la pertinence des règles d'association).
– La phase de présentation consiste à interpréter les résultats à l'utilisateur grâce à des techniques de visualisation, (Jiawei Han et al. 2000). Les résultats obtenus du data mining sont représentés sous la forme d'histogramme.

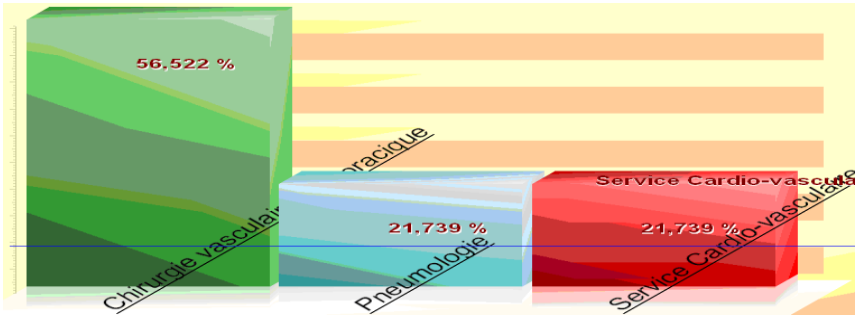


Fig. 6- Présentation graphique de la règle d'association sélectionnée.

4.3 Raisonnement à Base de Cas

Dans cette section, nous présentons notre système de raisonnement à base de cas par l'orientation médicale qui aide l'utilisateur à orienter le patient vers le service adéquat. La base de connaissance est construite par le processus décrit dans la partie précédente.

4.3.1 Construction de la Base de cas

La structure des enregistrements, de forme (attribut, valeur) dans un SGBD s'apparente bien avec la forme des caractéristiques dans une base de cas. Notre solution consiste à stocker les cas obtenus à partir du processus de data mining dans une base de données relationnelle. Les cas sont alors stockés dans une table de la base de données : chaque attribut correspond à un champ, chaque cas correspond à un enregistrement.

4.3.2 Représentation de cas

La description du cas dans le cadre de la prise en charge de l'orientation médicale est celle des cas cliniques déjà stockés dans notre base de cas. Trois éléments principaux apparaissent dans le contenu du cas : (1) le but (représente l'objectif que tentera d'accomplir la solution). (2) les caractéristiques (quelque soit la représentation des connaissances au niveau « symbole » choisie, la caractérisation est décrite par un ensemble fini de couples <attribut, valeur>. Il s'agit de donner le numéro d'ordre de la valeur). (3) la solution au problème (dans notre travail la solution est donnée le « service médical », « le taux confiance» du service par rapport au cas, ainsi que « la fréquence et le support » du cas).

4.3.3 Processus du raisonnement à base de cas

Le fonctionnement de cette partie de notre travail repose sur les quatre parties qui composent les systèmes RBC, (Kolodner et al. 1996):

1. Partie recherche: lors de la présentation d'un nouveau cas, cette phase permet de déterminer les cas de la base qui sont les plus similaires au problème donné du patient, après l'extraction des indices connus qui vont servir à effectuer la recherche de cas analogues.

Il existe deux types d'attributs :

Les attributs à valeurs discrètes (sexe, type d'age, type de poids, type de fièvre, signe1, signe2, signe3, signe4…)

Les attributs à valeurs booliennes (diabète, tabagisme, tension artérielle, liste des douleurs, liste des signes fonctionnels…).

(a) Calcul de similarité locale : la similitude entre deux attributs est calculée en utilisant l'équation:

$$\text{Sim}(ai, bi) = \begin{cases} 1 & \text{si } ai = bi \\ 0 & \text{si } ai \neq bi \end{cases} \quad (5)$$

(b) Calcul Similarité globale : la similarité globale de deux cas C1 et C2 est calculée par la formule suivante:

$$d_2(x,y) = \left(\sum_{i=1}^{k} w_i \times |x_i - y_i|^2\right)^{1/2} \quad (6)$$

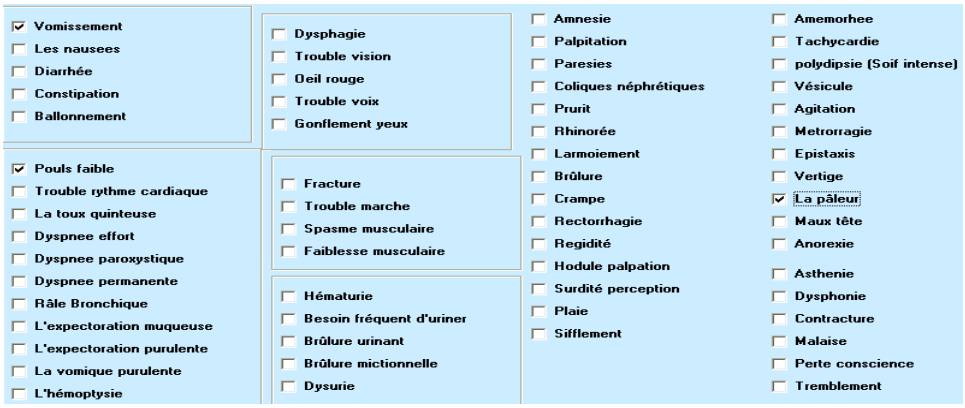


Fig. 7- Saisir d'un nouveau cas.

(c) Sélection des cas similaires : la procédure de recherche est implantée par une sélection des plus proches voisins ("k-proches voisins ") (Gierl and Schmidt 1998).

Les cas similaires sont classés selon les mesures de similarité (poids) et le cas le mieux classé et qui à un taux de confiance supérieure au seuil est proposé comme solution. Dans notre travail la partie du raisonnement à base de cas est basé sur la partie précédente du data mining et le problème de définition des poids de similarité des cas est résolu, on considère que le degré de similarité de chaque cas dans le raisonnement à base de cas est le taux de

confiance extrait par la partie précédente. Dans le cas saisi, nous avons quatre ces similaires avec des degrés de similarité différents, le cas proposé est le service « gastro-enterologie », avec un degré de similarité de 62.5% (on considère seuil=50%).

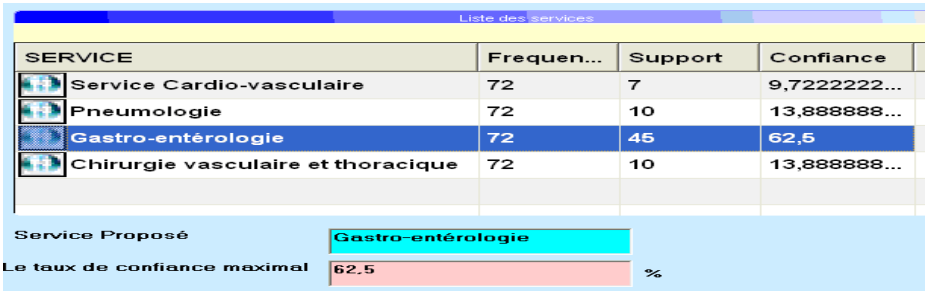


Fig. 8- Cas Similaires et la solution du nouveau cas.

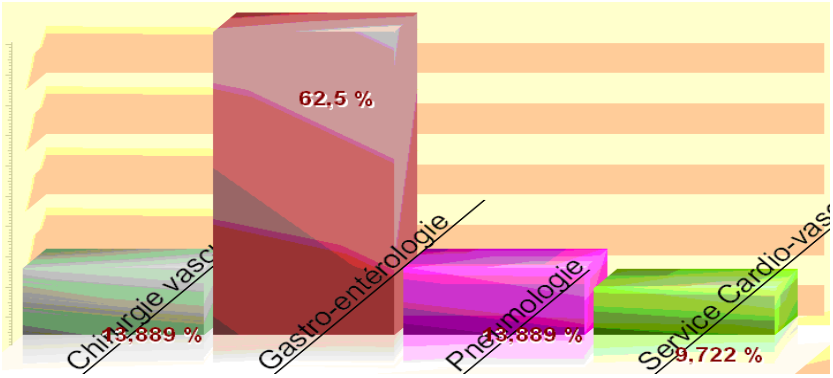


Fig. 9- Représentation graphique des cas similaires.

2. Partie adaptation : cette partie n'est pas prise en charge dans notre système, pour cela nous allons la proposer comme un travail futur dans les perspectives.

3. Partie révision : la révision consiste à évaluer la solution proposée en la testant dans un environnement réel (Nilsson.M et al. 2003). Dans notre système, la révision est faite par une intervention humaine.

4. Partie mémorisation : si après le calcul des mesures de similarité entre le nouveau cas et ceux de la base, le système génère un message qu'il n'existe pas de cas similaire ou assez proche de ce nouveau cas, l'expert a accès au système pour insérer le nouveau cas.

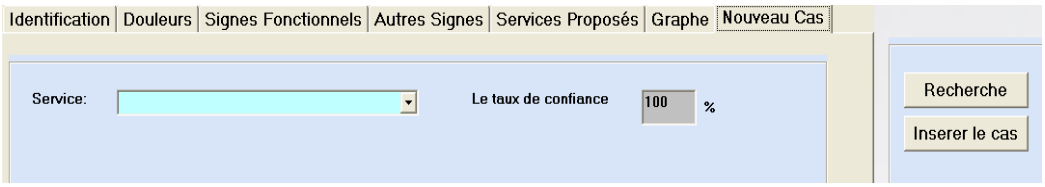


Fig. 10- Mémorisation d'un nouveau cas.

5 Conclusion

Dans cet article, nous avons montré l'importance de combiner de la gestion de connaissances dans le domaine médical, par la combinaison des techniques des entrepôts de données, data mining et le raisonnement à base de cas.

Notre approche est décomposée en trois phases fondamentales:

- En premier lieu, nous avons construit un entrepôt de données médical pour résoudre le problème de l'hétérogénéité des sources de données médicales. Pour cela nous avons suivi le processus ETL (Extraction, transformation, Loading).

- En second lieu, nous avons réalisé un processus d'extraction des connaissances médicales à partir de l'entrepôt de données déjà construit, le processus que nous avons retenu se décompose en plusieurs étapes : la sélection des données, la préparation des données sélectionnées, l'utilisation de la technique des règles d'association du data mining appliquée sur les données traitées, et enfin évaluation et présentation des résultats.

- En dernier lieu, nous avons présenté notre système de raisonnement à base de cas d'orientation médicale qui aide l'utilisateur à orienter le patient vers le service adéquat, nous avons exploité les connaissances extraites du data mining.

Références

Aberer.K and Hemm.K, "A Methodology for building Data Warehouse in a Scientific Environment", 2005.

Agrawal R, Imielinski T, Swami A, "Mining Association rules between sets of items in large database", Proceedings of the ACM SIGMOD International Conference on Management of Data, Washington, DC, pp 207-216, May 26-28, 1993.

Azuaje.F, C. Gertosio and A. Dussauchoy. "Knowledge discovery from industrial databases". Journal of Intelligent Manufacturing, 15: 29–37, 2004.

Benitez.E « Infrastructure adaptable pour l'évolution des entrepôts de données ». Phd thèse, Université Joseph Fourier, Grenoble, France, Septembre 2002.

Brachman.J and T. Anand. "The process of knowledge discovery in databases. Advances in knowledge discovery and data mining", pages 37–57, 1996.

Gierl and Schmidt. "Cbr in medicine, case-based reasoning technology. From Foundations to Applications", pages 273–297, 1998.

Kolodner, Janet L. et Leake, David B. «A tutorial introduction to case-based reasoning ». In Leake, David B., editor, Case-Based Reasoning Experiences, Lessons, & Future Directions, pages 31–66. American Association for Artificial Intelligence, Menlo Park, 1996.

Jérôme.D, « Optimisation et évaluation de performance pour l'aide à la conception et à l'administration des entrepôts de données complexes ». Université Lumière Lyon 2 Ecole Doctorale de Sciences Cognitives. Novembre 2006.

Jiawei Han, Jian Pei, and Yiwen Yin. "Mining frequent patterns without candidate generation". In Weidong Chen, Jeffrey Naughton, and Philip A. Bernstein, editors, 2000 ACM SIGMOD Intl. Conference on Management of Data, pages 1–12. ACM Press, 05, 2000.

María Trinidad Serna Encinas. « Entrepôts de données pour l'aide à la décision médicale : conception et expérimentation ». Thèse phd. Université Joseph Fourier, Grenoble, France, Juin 2005.

Naiditch M. Modélisation des trajectoires : « Problèmes méthodologiques. Innovation et technologie en biologie et médecine », 21(5), 307.12. 2005.

Nilsson.M, P.Funk, and M.Sollenborn. "Complex measurement classification in médical applications using a case-based approach" . In Workshop on CBR in the Health Sciences, pages 63–72. ICCBR'03, 2003.

Soualmia L. & Darmoni S. "Combining different standards and different approaches for health information retrieval in a quality-controlled gateway". International Journal of Medical Informatics (IJMI), p. 141–150. 2005.

Stefanelli M. "Knowledge management to support performance-based medicine. Methods of Information in Medicine", 1, 36.43. 2002.

Voros.S, Orvain e., Long j. & Cinquin p. "Automatic detection of instruments in laparoscopic images : a first step towards high level command of robotized endoscopic holders" . In International Conference on Biomedical Robotics and Biomechatronics, Pisa, Italy, 2006.

An extension of K-mode algorithm to cluster OLAP-CUBE Schemas

Nouha Arfaoui*, Jalel Akaichi**

BIRT- Institut Supérieur de Gestion
41, Avenue de la liberté,Cité Bouchoucha,
Le Bardo, 2000
Tunisie
*Arfaoui.nouha@yahoo.fr
**Jalel.akaichi@isg.rnu.tn

Abstract. The Data Mining (DM) offers different methods and techniques in order to extract information and knowledge from different kind of data sets. The k-mode is one of many algorithms used to cluster categorical data into groups by increasing the similarity within the same group and decrease between these groups. We propose, in this work, the extension of this algorithm and its application to OLAP-CUBE schemas which are specific type of categorical data and more precisely "complex categorical data" since a pattern is composed of dimensions, measures and level names.

1 Introduction

The Data Mining (DM) is "the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner" (Hand et al, 2001). It deals with different types of data such as: continuous numerical variables (Huang,1998),(Huang,1997a), binary variables (Balcan and Gupta, 2010), categorical variables (Huang,1997a), etc.
Many techniques and algorithms are used: classification (classification by decision tree induction, Bayesian Classification, etc) clustering (partitioning methods, hierarchical agglomerative methods, etc), regression, artificial intelligence, neural networks (back propagation), association rules (multilevel association rule, quantitative association rule, etc), decision tree, genetic algorithm, nearest neighbor method, etc (Ramageri, 2006),and according to (San et al, 2004), Clustering is one of fundamental operations in DM.

Concerning clustering, it is the unsupervised classification of patterns into groups called Clusters (Jain et al, 1999), it involves dividing a set of data points into non-overlapping groups, or cluster of points (Faber,1994). The objects in one cluster must be more similar to one another to objects in other clusters. To determine the notion of similarity we use some measure of proximity.
The purpose of the cluster is to maximize the homogeneity of a partition of a set of variables into K disjoint clusters (Chavent et al, 2010).

Different clustering algorithms exist; such iterative relocation algorithm and hierarchical algorithm (Chavent et al, 2010). In the first one, the variable can move in and out groups at different stages. Concerning the hierarchical algorithm, it can be divided into two types: ascendant and divisive
In the ascendant hierarchical clustering algorithm, one recursively merges two clusters, starting from the stage in which each variable is considered to form a cluster by itself to the stage where is a single cluster containing all variables. The divisive hierarchical approach consists on reversing the process of agglomerative hierarchical cluster, by starting with all variables in one cluster and successively dividing each cluster into two sub-clusters.

In this work, we propose to extend the k-mode algorithm to deal with new type of data which is OLAP-CUBE schema.
The OLAP CUBE corresponds to the users queries which are made against multidimensional CUBE (Niemi et al, 2001), and it is defined in (Datta et Thomas, 1999) as "the fundamental underlying construct of the multidimensional database and serves as the basic unit of input and output for all operators defined on a multidimensional database".
Concerning the k-mode, it was proposed by (Huang,1997a), (Huang,1998) to cluster categorical data. It is an extension of k-means using: the simple matching dissimilarity measure, the mode instead of measure, and the frequency based method to update modes. It serves to overcome the inconvenient of k-means since this latter is efficient to cluster large data sets but it is destined to deal only with numeric values.

The choice of k-mode is done because in our case we deals with OCSs which are considered as a complex category data since they are composed by a set of dimensions, a set of level names and a set of measures. Their number can be different from one schema to another.

This paper is organized as follow:
In section 2, we present some algorithms used to cluster categorical data, the in section 3, we define the OLAP-CUBE Schema and we give an example of a cube. Section 4 contains the description of k-mode algorithm as well as the extension of this algorithm to deal with OLAP-CUBE schema, and we finish with the conclusion in section 5.

2 State of art

In this part we propose the different methods used to cluster categorical data.

K-Mode: the k-means has the capacity to deal with large databases (San et al, 2004) and it is efficient with numerical data (Huang,1998), (Ng et al, 2007), but it doesn't work with categorical data because of two main problems: the formation of cluster center and the calculation of dissimilarity between objects and cluster centers (San et al, 2004). The idea is to extend this algorithm to deal with real world data (including categorical data). The authors in (Huang,1998), (Ng et al, 2007), (San et al, 2004) propose then k-mode that uses the 'simple matching dissimilarity measure'; it replaces the means of clusters with modes, and to update modes in the clustering process, it uses frequency-based method.
The following set of modifications is applied to the k-means algorithm to be able to cluster the categorical data:

- Using a simple matching dissimilarity measure for categorical objects.
- Replacing means of clusters by modes.
- Using a frequency-based method to find the modes.

ROCK: the authors in (Guha et al, 2000) propose the ROCK (**RO**bust hierarchical **C**lustering with lin**K**s) as a way to solve the problem of the use of distance which is not appropriate to deal with categorical data. The authors propose as solution a novel concept of links to measure the similarity/proximity between a pair of data points. The solution is: hierarchical clustering algorithm ROCK that uses links and not distance when merging clusters.

Concerning the measure, the best pair of clusters is this having a maximum goodness measure.
Computation of Links, they propose the use of **n*n** adjacency matrix A. A [**i**, **j**] takes 0 or 1 depending on whether or not points **i** and **j** respectively are neighbors. The numbers of links between a pair of points **i** and **j** can be obtained by multiplying row **i** with column **j**.

QROCK: an improvement has been proposed in (Dutta et al, 2005) through QROCK (Quick ROCK) that computes the clusters by determining the connected components of the graph which ensures having a drastic reduction of the computing time compared to ROCK.

CACTUS: the authors propose in (Aranganayagi and Thangavel, 2009) a novel formalization of a cluster for categorical attributes. They describe a very fast summarization based algorithm called CACTUS (CAtegorical ClusTering Using Summarie) (it discovers exactly such clusters in the data). The proposed algorithm requires only two scans of the datasets which makes it a fast and scalable algorithm.
This algorithm is composed by three steps: summarization, clustering and validation:

- Summarization: they compute the summary information from the dataset.
- Clustering: they use the summary information to discover a set of candidate clusters.
- Validation: they determine the actual set of clusters from the set of candidate clusters.

COOLCAT: the proposed algorithm in (Barbara et al, 2002) uses the notion of entropy measure to group the records. The choice of the entropy is because it is a more natural and intuitive way of relating records and it does not rely in arbitrary distance metrics. This algorithm is composed by two steps: initialization and incremental step.
In the first step, the algorithm groups the objects in 'k' most dissimilar records maximizing the minimum pairwise entropy of the chosen points. In the second step, the remaining records are placed into clusters which are calculated through computing the expected entropy. The choice of the cluster is done in function of the expected entropy (the minimum one).

3 The OLAP-CUBE Schema

In this section we start by defining the OLAP-CUBE schema (OCS) and we give, then, an example of cube extracted from a star schema.

3.1 Presentation

The traditional relational data models are not powerful enough for the DW applications (Vassiliadis and Sellis, 1999). The solution is the use of data cubes that provide the functionality needed for summarizing, viewing and consolidating the data existing in DW. In addition, they have twofold benefits. Indeed, they are close to the way of thinking of data analyzers; therefore, they help users to understand data, also, they support performance improvement as their simple structures allow designers to predict the user intentions (Rizzi, 2008).

So the cube represents the data in a multidimensional space. It is composed by a set of dimensions (so called hypercube or just cube):, each one has an associated hierarchy of levels of consolidated data. The measures correspond to columns in a relational database table whose values functionally depend on the values of other columns. A value in a single cell may represent an aggregation measure computed from more specific data at some lower level of the same dimension that comprises a set of aggregation level (Franconi and Sattler, 1999).

3.2 Example

In this part we present an example of OLAP-CUBE schemas (figure 1) extracted from the star schema (figure 2), which is composed by three dimensions “Patient”, “Doctor” and “TimeStamp”, and one fact table “UniversityHospital”. The “Patient” table contains one primary key “Patient_ID”, and three attributes “Name”, “Total bill”, and “Length of stay”. The “Doctor” table is composed by one primary key “Doctor_ID” and four attributes “Specialization”, “Department”, “Position” and “Years of Experience”. The “TimeStamp” table contains “TimeStamp_ID” as primary key and “Year”, “Month”, “Day” and “Hour” as attributes.

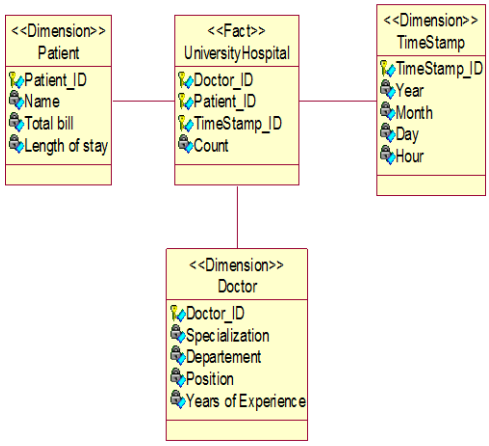


FIG. 1- Example of star schema

Through the Cube, we aim to analysis the evolution of the number of patients for a specific doctor over the time. Our example contains three dimensions: Doctor, Patient, and TimeStamp, but the concept of Cube is not limited to three areas, it can be spreading in hyper-cube with a number of axes more than several dozen.

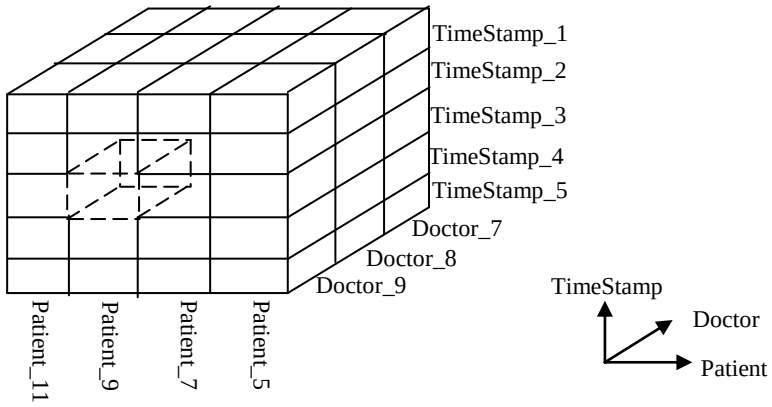


FIG. 2- The OLAP-Cube corresponding to the Star schema of the FIG. 1

4 The OLAP-CUBE schema Cluster

In this part, we start with presenting the principle of OCS cluster, then we describe the k-mode algorithm, and we finish by defining its extension to deal with OCSs.

4.1 The principle

For a set of OLAP-CUBE Schemas (OCSs), each one is composed by "d" Dimensions, "m" Measures and "l" Level names.
$OCS_j = D_{1j} \cup D_{2j} \cup \ldots \cup D_{dj} \cup M_{1j} \cup M_{2j} \cup \ldots \cup M_{mj} \cup L_{1j} \cup L_{2j} \cup \ldots \cup L_{lj}$, with D > 1, M > 1 and L > 1, and the name of different components are disjoints; i.e. {Dimension name} $\cap$ {Measure name} $\cap$ {Level name} = ϕ.

We have 'n' OCSs and we want classify them into clusters according to their degree of similarity. To realize this purpose, we propose to present each OCS as xml file, so each one describes the structure of a specific schema.
In order to apply the k-mode in our case, we propose the extraction of useful information i.e. extract from the file the different dimensions, measures and level names composing the current schema, and we compare them to the elements of the other files in order to find the closest ones.

The figure 3 presents the steps to cluster an OCS. In fact, we start by extracting the useful information from the cube by presenting it as an XML file that separates the different elements (dimensions, measures and level names). Than we use the extension of the k-mode to cluster the different files by making comparisons between the new file and the modes of the clusters and we choose the most similar one. The algorithm will be detailed later.

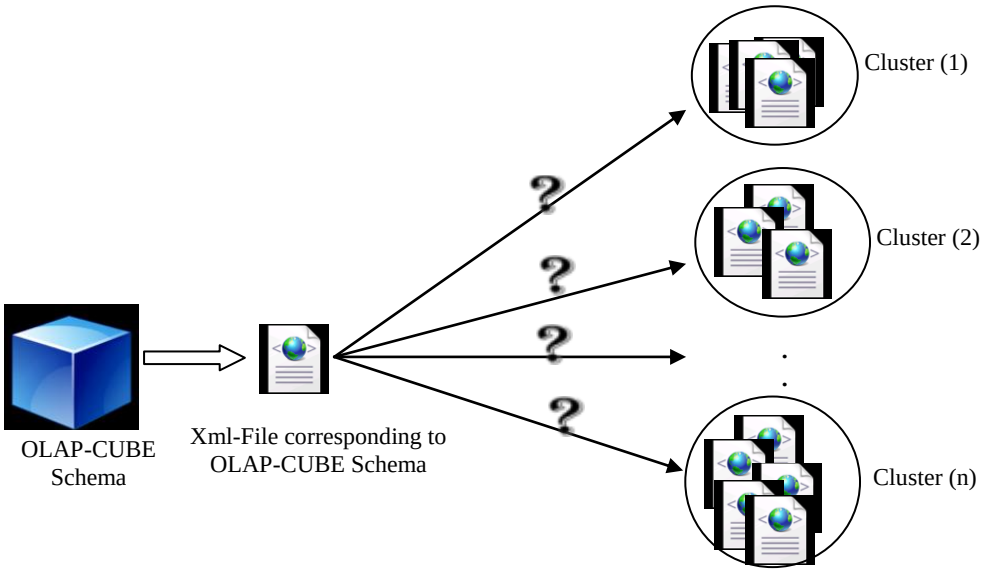


FIG. 3- Clustering xml files presenting the schemas of the OLAP-CUBEs

4.2. The K-mode algorithm

In this section, we present the algorithm of k-mode as presented in (Huang,1997b). In fact this algorithm produces locally optimal solutions depending on the initial modes, and it is described as follow:

a) Select k initial modes, one for each cluster.
b) Allocate an object to the cluster whose mode is the nearest to it using the simple matching dissimilarity measure (**1**). Update the mode of the cluster after each allocation according to the **Theorem**.
c) After all objects have been allocated to clusters, retest the dissimilarity of objects against the current modes. If an object is found such that its nearest mode belongs to another cluster rather than its current one, reallocate the object to that cluster and update the modes of both clusters.
d) Repeat 'c' until no object has changed

$$d(X, Y) = \sum_{j=1}^{m} \delta(Xj, Yj) \qquad \text{where } \delta(Xj, Yj) = \begin{cases} 0 & (Xj = Yj) \\ 1 & (Xj \neq Yj) \end{cases} \qquad (1)$$

Theorem: The function D(Q, X) is minimized iff $f_r(A_j = q_j \mid X) \geq f_r(A_j = c_{k,j} \mid X)$ for $q_j \neq c_{k,j}$ for all j=1…m.

With:

- $D(X, Q) = \sum_{i=1}^{n} d_1(X_i, Q)$

- Mode of X= { X_1,X_2,X_3,…,X_n} is a vector Q= [q_1, q_2, q_3,…,q_m]
- $n_{ck,j}$: is the number of objects having category $c_{k,j}$ in the attribute A_j
- f_r ($A_j = q_j \mid X$) = $n_{ck,j}$ / n: is the relative frequency of category $c_{k,i}$ in X.

Concerning the distance, many measures are proposed such as (He et al, 2005) and (San et al, 2004). In our case, it is calculated using the simple matching similarity measure which take the two values '0' or '1'(Huang,1998), (Ng et al, 2007). The main idea is to use the relative attribute frequencies of the cluster modes in the similarity measure in the k-modes objective function. This modification allows the algorithm to recognize a cluster with weak intra-similarity, and therefore assign less similar objects to such cluster, so that the generated clusters have strong intra-similarities.

4.3 The extension of k-mode

To cluster OCSs we propose the use of k-mode since it is efficient to cluster categorical data and in our case the OCSs are complex categorical data since they are composed by elements (including dimensions, levels and measures), also, the different schemas do not have the same size (number of the existing elements). So, we should modify the calculation of the dissimilarity coefficient as well as the determination of the mode of the cluster. Since, the center of the k-mode algorithm is a virtual object that contains the most frequent attribute values in the cluster (San et al, 2004), we propose in our case to select a specific number of the most frequent elements.

As hypothesis (**hyp**), we propose that
- "a" takes as value the smallest number of dimensions existing in one schema belongs to the current cluster,
- "b" takes as value the smallest number of measures existing in one schema belongs to the current cluster,
- "c" takes as value the smallest number of level names existing in one schema belongs to the current cluster.

Input : XML files
Output: Set of clusters

k: the number of clusters

Begin
- Select "k" initial xml-file, one for each cluster.
- Allocate the xml-file:
  - Extract the dimensions, measures and levels names from each file as "Vectors".
  - Compare the current file to those that exist in the "k" clusters.
  - Allocate the current file to the cluster whose mode is the nearest to it according to **Coef** (figure 5)
- Update the mode of the cluster:
  - Calculate the frequencies of dimensions, measures, and levels names.
  - Allocate "a" dimensions, "b" measures and "c" level names to the mode (**hyp**)

- Update the mode of the cluster after each allocation (figure 6)

- Retest the dissimilarity of xml-files against the current modes.
- Repeat the last step until no objects has changed.

End

FIG.4- Algorithm “OCSs Clustering”

```
Input:    XMLfile1, XMLfile2
Output:  Coef_IS

Begin

V_1D : Dimension Vector from XML file1
V_2D: Dimension Vector from XML file2.
V_1M: Measure Vector from XML file1.
V_2M: Measure Vector from XML file2.
V_1L: Level name Vector from XML file1.
V_2L: Level name Vector from XML file2.

Coef_D=0; Coef_M=0; Coef_L=0; Max_D =0 ;
 Max_M =0 ; Max_L=0 ;

Max_D = Max (V_1D.Size, V_2D.Size)
Max_M = Max (V_1M.Size, V_2M.Size)
Max_L = Max (V_1L.Size, V_2L.Size)

     For i=1 ; i≤ V_1D.Size ; i++
          For j=1; j≤ V_2D.Size ; j++
               If  V_1D[i]== V_2D[j] then
                   Coef_D++
                End if
           End for
      End for

     For i=1 ; i≤ V_1M.Size ; i++
          For j=1; j≤ V_2M.Size ; j++
               If  V_1M[i]== V_2M[j] then
                   Coef_M++
                End if
          End for
     End for

     For i=1 ; i≤ V_1L.Size ; i++
          For j=1; j≤ V_2L.Size ; j++
               If  V_1L[i]== V_2L[j] then
                    Coef_L++
```

End if
End for
End for

$$\text{Coef} = \frac{Max_D - Coef_D}{Max_D} + \frac{Max_M - Coef_M}{Max_M} + \frac{Max_L - Coef_L}{Max_L}$$

End

FIG.5- Algorithm "**Coef**"

Input: Cluster
Output: Mode

Begin

f_D: corresponds to "a" most frequent dimensions in a cluster.
f_M: corresponds to "b" most frequent measures in a cluster.
f_L: corresponds to "c" most frequent level names in a cluster.

[ ]f_D = $Freq_D$ (Cluster)
[ ]f_M = $Freq_M$ (Cluster)
[ ]f_L = $Freq_F$ (Cluster

Mode = $f_D + f_M + f_L$

End

FIG.6- Algorithm "**ModeMAJ**"

5 Conclusion

In this work, we proposed the extension of k-mode which is one of the algorithm used in DM to cluster the data into groups, to deal with new kind of data which is "OLAP-CUBE schema". This latter is considered as a complex categorical data, since it is composed by sub-elements: dimensions, measures, and level names. The numbers of elements can be different from one schema to another. So, we have to take all those details into consideration when we apply the k-mode algorithm.
In order to improve our algorithm, we propose, as future work, the integration of the ontology to group the words having the same meaning. We assume that this can improve the result of comparison.

References

Aranganayagi, S., and Thangavel, K (2009). Clustering Categorical Data using Bayesian Concept, *International Journal of Computer Theory and Engineering*. 119-125.

Balcan, M. F., Gupta, P (2010). *Robust Hierarchical Clustering*, In Proceedings of the Conference on Learning Theory (COLT).

Barbara, D., Couto, J., and Li, Y (2002). COOLCAT: An entropy-based algorithm for categorical clustering, *In Proceedings of the eleventh international conference on Information and knowledge management*, 582-589.

Chavent, M., Kuentz, V., and Saracco, J (2010). Clustering of categorical variables around latent variables, *Cahiers du GREThA 2010-02, Groupe de Recherche en Economie Theorique et Appliquee*.

Dutta, M., Mahanta, A. k., and Pujari, A. K (2005). QROCK: A Quick Version of the ROCK Algorithm for Clustering of Categorical Data, *Pattern Recognition Letters; with M. Dutta, T.U. and A.K. Puzari, University of Hyderabad*, 2364-2373.

Datta, A. and Thomas. H (1999). The Cube Data Model: A Conceptual Model and Algebr for On-Line Analytical Processing in Data Warehouses, *Decision Support Systems*, 289-301.

Faber, V (1994). Clustering and the Continuous k-means Algorithm, *Los Alamos Science*.138-144.

Franconi, E., and Sattler, U (1999) A Data Warehouse Conceptual Data Model for Multidimensional Aggregation: a preliminary report, *Journal of the Italian Association for Artificial Intelligence AI*IA Notizie, 1*.

Guha,S., Rastogi, R., and Shim, K (2000). ROCK: A Robust Clustering Algorithm for Categorical Attributes, *In: Inf. Syst., UK: Elsevier Science Ltd*. 345-366.

Hand, D., Mannila, H., and Smyth, P (2001). *Principles of Data Mining*, MIT Press, Cambridge, MA.

He, Z., Deng, S. and Xu, X (2005). Improving k-modes algorithm considering frequencies of attribute values in mode, International Conference on Computational Intelligence and Security, LNAI 3801,157-162.

Huang, Z (1997). Clustering large data sets with mixed numeric and categorical values, *In: KDD: Techniques and Applications (H. Lu, H. Motoda and H. Luu, Eds.). — Singapore: World Scientific*. 21-34.

Huang, Z (1997). A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining, *In: SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*.1-8.

Huang, Z (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values, *Data Mining and Knowledge Discovery*. 2:283–304.

Jain, A. K.., Murty, M. N., and Flynn, P. J (1999). Data Clustering: A Review, *ACM Comput. Surv.*, 264-323.

Ng, M. K., Li, M. J., Huang, J. Z., and He, Z (2007). On the Impact of Dissimilarity Measure in k-modes Clustering Algorithm, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 503-507.

Niemi, T., Nummenmaa, J., and Thanisch, P (2001). Constructing OLAP cubes Based on Queries, *In Journal Hammer, editor, DOLAP 2001, ACM Fourth International Workshop on Data Warehousing and OLAP, ACM*. 9-11.

Ramageri, B. M (2006). *Data Mining Techniques And Applications*, Indian Journal of Computer Science and Engineering, Vol. 1 No. 4, Modern Institute of Information Technology and Research, Department of Computer Application, Yamunanagar, Nigdi Pune, Maharashtra, India-411044. 301-305

Rizzi, S (2008). Conceptual Modeling Solutions for the Data Warehouse, *In Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications, J. Wang (Ed.), Information Science Reference*. 208-227.

San, O. M., Huynh, V. N., and Nakamori, Y (2004). An Alternative Extension Of The K-Means Algorithm For Clustering Categorical Data, *Journal of Applied Mathematics and Computer Science*, 241-247.

Vassiliadis, P., and Sellis, T (1999). A Survey on Logical Models for OLAP Databases, *In: SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, 64-69.

Résumé

Le Data Mining (DM) offre plusieurs méthodes et techniques pour extraire les informations et les connaissances à partir de plusieurs types d'ensemble de données. Le K-mode est l'un de plusieurs algorithmes utilisés pour grouper des données catégoriques tout en augmentant la similitude dans le même groupe et la diminuant entre ces groupes. Nous proposons, dans ce travail une extension de cet algorithme ainsi que son application sur les schémas OLAP-CUBE qui sont considérés comme un type spécifique des données catégoriques et plus précisément des données catégoriques complexes vu qu'un motif est composé de dimensions, de mesures et des noms de niveaux.

Quel Critère Discriminant À Choisir Pour Grouper Les Analystes Des Entrepôts De Données?

Eya Ben Ahmed*, Ahlem Nabli**
Faïez Gargouri***

*Institut Supérieur de Gestion de Tunis
eya.benahmed@gmail.com
**Faculté des Sciences de Sfax
ahlem.nabli@fsegs.rnu.tn
***Institut Supérieur d'Informatique de Multimédia de Sfax
faiez.gargouri@isimsf.rnu.tn

Résumé. Un défi de la personnalisation consiste à enrichir le profil individuel en utilisant des informations relatives aux préférences communes entre un groupe d'utilisateurs. Cette méthode est connue sous le terme '*groupization*'. Elle permet d'adapter efficacement les résultats des requêtes aux intentions des membres du groupe. Dans cet article, nous visions l'identification optimale des groupes d'analystes d'entrepôts de données. Pour ce faire, nous étudions la similarité entre les requêtes sélectionnées dans les historiques de navigation. Quatre axes fondamentaux d'identification de groupe d'analyste sont cernés. Un algorithme de clustering hiérarchique semi-supervisé est employé pour identifier les groupes d'analystes basés sur ces critères. Les résultats des expérimentations menées sur un entrepôt de données mis en place dans le domaine boursier, démontrent que la groupization améliore le résultat de la personnalisation, particulièrement pour des groupes basés sur la fonction de l'analyste et ceux explicitement identifiés.

1 Introduction

Les progrès des technologies de l'information et l'amélioration des capacités de stockage ont induit à une surabondance des quantités des données. La prise de la meilleure décision dans un tel contexte est devenue un défi majeur expliquant le recours aux entrepôts de données où les informations stratégiques sont stockées et sont exploitables à travers la technologie OLAP.

Afin d'améliorer la qualité des informations et perfectionner les performances des analyses OLAP, plusieurs contributions ont été proposées pour placer l'analyste au cœur du processus d'analyse. Deux courants de ces approches peuvent être distingués (Ben Ahmed et *al.*, 2011) : (i) *la personnalisation* où l'intervention du système pour rapprocher les résultats des requêtes aux préférences de l'utilisateur est implicite (Bellatreche et *al.*, Favre et *al.*, Ravat et *al.*, Rizzi) ; (ii) *la recommandation* où l'intervention du système est explicite se

manifestant à travers la suggestion d'un ensemble de propositions facilitant ainsi l'opération d'analyse (Giacometti et *al.*, Jerbi et *al.*).

Généralement, le recensement des préférences des analystes est explicitement accompli dans la littérature. A l'encontre des contributions de (Giacometti et *al.*, 2008), (Aligon et *al.*, 2011) et (Ben Ahmed et *al.*, 2012) où un apprentissage des préférences d'analyste à partir de l'historique de navigation des analyses OLAP est assuré.

Dans un travail antérieur (Ben Ahmed et *al.*, 2012), nous avons proposé une approche de construction de profil d'analyste multi-vues en se basant sur les historiques des navigations dans l'entrepôt de données. En effet, nous avons stocké les préférences fonctionnelles dans une vue comportementale et nous les avons classés selon le degré de conflit par rapport aux analystes. Ces préférences sont enrichies par une vue professionnelle et une vue personnelle. Toutefois, cet apprentissage des préférences d'analyste est étroitement lié aux nombres d'analystes. En effet, l'accroissement du nombre d'analyste d'entrepôt peut empêcher cet apprentissage automatique des préférences.

Pour pallier à cet inconvénient, nous proposons de combiner les données des analystes similaires afin de faciliter le processus d'apprentissage de préférences. Cette extension du processus de personnalisation s'appelle la « groupization» qui prend en considération les données de tous les membres du groupe au lieu des données d'un seul utilisateur. En fait, elle vise le rapprochement entre le résultat fourni d'une requête et les attentes du groupe.

Un défi capital dans l'utilisation des données du groupe pour la personnalisation réside dans l'identification des groupes de personnes ayant des préférences similaires. Certes, certains attributs sont plus utiles que d'autres pour rassembler les membres selon leurs historiques d'analyse sous forme de requêtes lancées et leurs jugements des résultats pertinents.

Le constant essentiel suite au survol de l'état de l'art de la groupization est la restriction du nombre des travaux qui se sont focalisés sur l'identification des groupes dans les communautés (Morris et *al.*, 2008), encore moins ceux qui ont abordé le domaine des entrepôts de données.

Dans cet article, nous nous intéressons à l'identification des facteurs qui permettent de définir les groupes d'analystes dans les entrepôts de données. La principale originalité de notre contribution est la construction des groupes dans le cadre des bases de donnés multidimensionnelles. La principale caractéristique est la justification de la méthode optimale de groupization des analystes. Finalement, une étude expérimentale est menée sur un entrepôt de données dans le domaine boursier pour valider l'identification optimale des groupes.

L'article est organisé comme suit. La section 2 motive notre contribution à travers un exemple motivant. La section 3 passe en revue les travaux de recherche dédiés à la groupization. La section 4 est dédiée à la présentation de notre contribution visant à cerner les facteurs d'identification des groupes d'analystes d'entrepôt de données. Les résultats des expérimentations montrant l'utilité de notre approche dans le cadre de l'entrepôt de données boursier sont présentés dans la section 5. La conclusion et les travaux futurs font l'objet de la section 6.

2 Exemple motivant

Dans cet article, nous utilisons les données stockées dans un entrepôt de données que nous avons mis en place dans le domaine de la bourse Tunisienne comme exemple motivant à notre proposition.

La bourse est un marché financier regroupant deux partenaires : d'une part, les entreprises et l'Etat qui nécessitent de l'argent sous forme de capitaux « *actions* » ou d'emprunts « *obligations* »; et d'autre part, les particuliers qui désirent épargner leurs argents.

Chaque entreprise introduite en bourse aura sa propre *cotation* qui en fonction de l'offre et de la demande peut varier. Cette valeur se traduit par *le cours de l'indice boursier* de cette entreprise. En outre, le *portefeuille boursier* est la représentation de l'ensemble des *titres* sur lesquels un agent économique a investi sur le marché financier.

Un *indice boursier* est un outil statistique permettant de mesurer l'évolution du cours des *titres* qui le composent. En effet, il est représentatif soit d'un marché ; soit d'un secteur d'activité particulier.

Tel entrepôt de données permet l'évaluation de la performance du marché des *actions* et des *indices obligataires* à travers l'analyse des *indices boursiers* (*cours boursiers*) de chaque *société cotée* en bourse dans un secteur particulier, soit à travers la mesure du *cours d'indice boursier sectoriel* ou à travers la mesure du *cours d'indice boursier total* dans tous les secteurs. Une partie du schéma de la base de données multidimensionnelle associée est illustrée dans la figure 1. Pour représenter son schéma, nous adoptons des notations graphiques proches de (Golfarelli et *al.*, 2008). Cet entrepôt de données permet d'analyser les fluctuations des *indices boursiers* afin d'évaluer la performance du marché boursier. En effet, notre schéma comporte un fait nommé *Mouvement Boursier* mesuré à travers le cours de l'indice boursier selon plusieurs axes d'analyse (*Société cotée en bourse*, *Temps*, *Titre*).

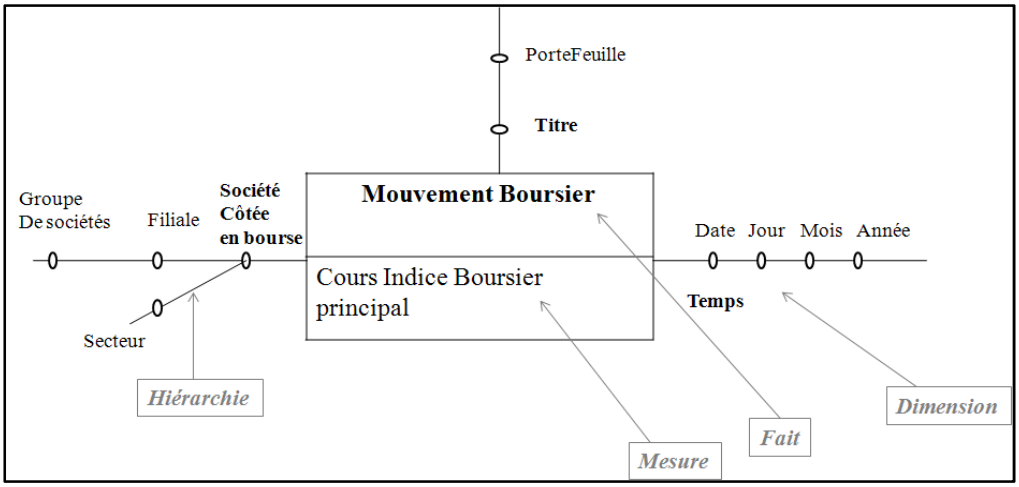


FIG. 1 – *Schéma de l'entrepôt de données dans le domaine de la bourse.*

La vocation de cet entrepôt de données est l'analyse des données boursières pour l'aide à la décision des agents économiques. L'exploration de la masse d'informations entreposées permettra d'orienter plusieurs décideurs tels que les *gestionnaires de portefeuille*, les *investisseurs*, les *gestionnaires privés* (sous mandat) et les *investisseurs privés*.

Face au nombre abondant d'analystes qui peuvent se servir de l'entrepôt boursier pour les orienter dans leurs prises de décision, le processus d'adaptation du résultat d'une requête d'analyse aux préférences individuelles de chaque analyste parait une tâche pénible surtout si nous considérons un historique de navigation assez daté.

Bien que ces acteurs puissent partager les mêmes préférences, par exemple, les *gestionnaires de portefeuille* ont tous tendance à comparer *le cours d'indice boursier du portefeuille* par rapport au *cours de l'indice boursier du secteur* et celui de tous les secteurs.
Pour pallier à ces inconvénients, nous nous proposons de regrouper les analystes ayant des préférences similaires. Certes, cette solution permettra de réduire l'écart entre le résultat obtenu et le résultat attendu suite à une requête d'analyse lancée en utilisant les informations relatives au groupe auquel appartient le décideur.

Néanmoins, l'identification des groupes d'analystes reste un problème complexe. En effet, les analystes peuvent être regroupés selon plusieurs dimensions telles que (i) *la fonction exercée* : nous admettons que les analystes occupant la même position auront des préférences communes, (ii) *les responsabilités engagées pour la réalisation des objectifs* : en fait, deux gestionnaires de portefeuille peuvent ne pas assurer les mêmes responsabilités, (iii) *la source de l'identification des groupes* étant explicite quand un analyste choisit d'appartenir à un tel groupe ou cette opération est effectuée implicitement, (iv) *l'aspect dynamique dans l'identification des groupes* : reflétant si les groupes identifiés seront mis à jour ou resteront figés.

Pour choisir le meilleur facteur d'identification des analystes, nous abordons après la revue de l'état de l'art dans ce qui suit, la présentation de notre approche.

3 Etat de l'art

Dans cette section, nous présentons une revue de l'état de l'art des travaux relatifs dans un premier volet au concept de groupization et dans un second volet au concept de clustering hiérarchique semi-supervisé.

3.1 Groupization

Initialement, la groupization est appliquée dans le domaine du web. En effet, la collaboration sur le web se prolifère dans de nombreux secteurs comme l'éducation, l'enseignement et la recherche scientifique. Des exemples typiques de cette collaboration concernent le regroupement des étudiants lors de l'attribution de devoirs collectifs ou le rassemblement des universitaires qui collaborent dans le cadre de projets de recherche.

Selon une étude menée par Microsoft (Morris et *al.*, 2009), l'analyse des expériences de la recherche collaborative sur le web a permis de dégager trois techniques, à savoir :

(i) *Groupization* : en se basant sur l'historique de navigation de tous les utilisateurs, l'idée de base de la groupization est d'attribuer plus de poids aux articles préférés par plus d'utilisateurs dans le même groupe.

(ii) *Fractionnement intelligent* : des outils de recherche collaborative s'appuient sur la division du travail. En fait, les résultats de la recherche peuvent être évalués et classés par les différents membres du même groupe.

(iii) *Groupe Hit-Highlighting* est une technique permettant à l'utilisateur d'assimiler la pertinence des résultats à travers l'accentuation des mots-clés comme le soulignement du titre ou la description ou l'URL. En supposant que nous avons accès à un historique de requêtes de groupe, un système peut exécuter du groupe *Hit-Highlighting* où les mots-clés apparaissent dans les résultats de recherche soulignés pour tous les membres du même groupe.

Dans le cadre de la groupization, Morris et *al.* (2008) et Teevan et *al*. (2009) ont analysé la similitude des choix de requêtes et des jugements de pertinence pour les recherches lancées sur le web entre les groupes d'utilisateurs. Ces derniers ont été regroupés selon deux axes d'analyse : (i) le premier axe porte sur *la longévité de l'appartenance aux groupes,* en effet, les groupes peuvent être formés selon leurs tâches à court terme ou selon leurs traitements qui durent plus dans le temps. (ii) Le deuxième axe concerne *la manière dont l'appartenance au groupe* est déterminée soit par des informations explicites d'appartenance aux groupes par les membres du groupe ou par une déduction implicite de cette appartenance. Les résultats fournis ont montré que les groupes explicitement définis sont caractérisés par une similitude des préférences par contre ceux identifiés implicitement manquent remarquablement de cohésion. En outre, l'identification des groupes à travers leurs traitements communs reflétant des intérêts communs à long terme, étant donc plus fiable que l'identification sur la base de tâches temporellement délimitées.

3.2 Clustering hiérarchique semi-supervisé

Le clustering permet d'organiser une collection d'objets en clusters, tels que les objets les plus similaires sont regroupés au sein du même cluster. Les versions semi-supervisé du clustering ont essayé d'améliorer le résultat en incorporant des connaissances externes dans le processus de clustering. Ces connaissances externes sont introduites sous forme de contraintes. Ces contraintes peuvent être directement dérivées des données d'origine à travers l'utilisation des données partiellement étiquetées ou fournies par un expert, en essayant d'adapter les résultats de clustering à ses attentes. Il existe deux principaux types de contraintes : (i) *Lien-Obligatoire* indiquant que deux objets de l'ensemble des données doivent appartenir au même cluster ; et (ii) *Lien-Interdit* qui inversement impose que deux objets appartiennent à deux clusters différents.

Peu de travaux ont étudié le clustering hiérarchique semi-supervisé dans la littérature. Parmi les principaux travaux, Klein et *al.* (2002) intègrent les contraintes au niveau instance par paires (*Lien-Obligatoire* et *Lien-Interdit*) lors de l'exécution du clustering semi-supervisé basé sur l'algorithme de lien complet proposé par Jain et Dubes (1988). L'intégration des contraintes comporte deux phases :

- *l'imposition* : les contraintes sont ajoutées aux paires d'objets. Si deux objets x_i et x_j ont une contrainte de *Lien- Interdit*, leur distance est mise à zéro. Sinon, si elles ont une contrainte *Lien-Obligatoire*, la distance est fixée à la distance maximale repérée dans la matrice de similarité.
- *la propagation* : l'algorithme estime que si x_k est un exemple à proximité de x_i et x_i a une contrainte de *Lien-Obligatoire* ou une contrainte de *Lien-Interdit* avec x_j, donc x_k est également à proximité ou loin de x_j. La distance entre les nouvelles x_k et x_j est calculée en utilisant le triangle d'inégalité.

Kestler et *al.* (2006) utilisent les contraintes par paire au premier niveau de l'algorithme de clustering hiérarchique lors de la génération des clusters initiaux. Ces contraintes ne sont pas propagées aux niveaux postérieurs. Les objets étiquetés sont utilisés par Bade et *al.* (2007) dans une étape de post-traitement. La méthode utilise les instances étiquetées pour générer les contraintes *Lien-Obligatoire* et *Lien-Interdit* entre les paires d'objets. Ainsi, ces contraintes sont utilisées pour déterminer après s'il faut fusionner ou fractionner les clusters résultants. Dans Böhm et des plantes (2008), un algorithme de clustering semi-

supervisé basé sur la densité est proposé. Les données étiquetées sont utilisées pour générer une première hiérarchie, qui sera par la suite élargi. Ainsi, les données non étiquetées sont affectées aux clusters les plus cohérents, en fonction de la structure du cluster prédéfini.

Exclusivement, Klein et *al.* (2002) est le seul qui a proposé une approche d'apprentissage active où les contraintes sont intégrées par paire. Le redémarrage du clustering est effectué d'une manière non supervisée jusqu'à l'étape de fusion. Ensuite, l'utilisateur spécifie si les racines du prochain fusionnement doivent être fusionnées. Selon le résultat, les contraintes sont propagées. Ainsi, nous optons pour l'utilisation de cette méthode pour regrouper les fichiers logs des analystes.

4 L'approche IGED pour à l'Identification des Groupes d'analystes dans les Entrepôts de Données

L'étude des analystes et leurs relations nous a permis de dégager des axes d'analyse pour l'identification des groupes à partir de l'historique de l'analyse OLAP. Ces facteurs peuvent être résumés comme suit :

1. La *fonction* : nous admettons que les analystes occupant la même position auront des préférences similaires, deux alternatives sont plausibles soit la restriction à la fonction actuelle soit la prise en compte de l'expérience et de l'expertise acquise ;
2. Les *responsabilités engagées pour la réalisation des objectifs* : étant des décideurs, les analystes de l'entrepôt auront des objectifs qui peuvent être opérationnels à court terme, tactiques à moyen terme ou stratégiques à long terme. L'exemple de deux gestionnaires de portefeuille qui peuvent ne pas assumer les mêmes responsabilités, et par moment n'auront pas les mêmes objectifs tout au long du temps illustre cette idée ;
3. La *source de l'identification des groupes* étant explicite quand un analyste choisit d'appartenir à un tel groupe sinon cette identification est apprise automatiquement ;
4. La *dynamicité* des groupes identifiés : autrement dit les groupes identifiés seront mis à jour dynamiquement suite aux changements survenus dans l'environnement ou resteront statiques.

Afin de déterminer le facteur le plus discriminant dans la liste de ces axes, nous détaillons notre approche IGED dédiée à l'**I**dentification des **G**roupes d'analystes dans les **E**ntrepôts de **D**onnées. Nous étudions la similarité entre les préférences des décideurs à base de leurs historiques de navigation. Pour ce faire, on fait recours au clustering hiérarchique semi-supervisé des requêtes sauvegardées dans les fichiers logs.

Ce choix s'explique d'une part par la capacité du clustering hiérarchique de s'adapter aux objets volumineux, particulièrement les documents. D'autre part, la supervision du processus d'apprentissage permettra de guider les éventuelles fusions entre les objets selon les contraintes dictées par chaque critère de groupization.

La figure 2 présente l'architecture paramétrique de notre approche IGED. Partant de l'ensemble des fichiers logs des analystes de l'entrepôt de données et d'un des critères de groupization fixé par l'expert, on applique le processus illustré par la figure 2 quatre fois en variant à chaque fois le critère de regroupement. En effet, le clustering hiérarchique semi-supervisé selon un lien complet introduit par Klein et *al.* est utilisé pour regrouper les fichiers logs de l'entrepôt selon chaque critère spécifié, à savoir, la fonction, les responsabilités, la source et la dynamicité. En fait, cet algorithme fusionne à chaque étape les deux plus

proches clusters ayant la plus grande similitude. Cette similarité est calculée en utilisant une distance entre ces fichiers logs. Cette distance peut être modifiée selon les types de contraintes : (i) Contrainte de *Lien-Obligatoire* : la distance entre les clusters est remplacée par la plus grande distance existante dans la matrice de similarité, (ii) Contrainte de *Lien-Interdit* : la distance entre les clusters qui ne devraient pas être fusionnées est substituée par zéro.

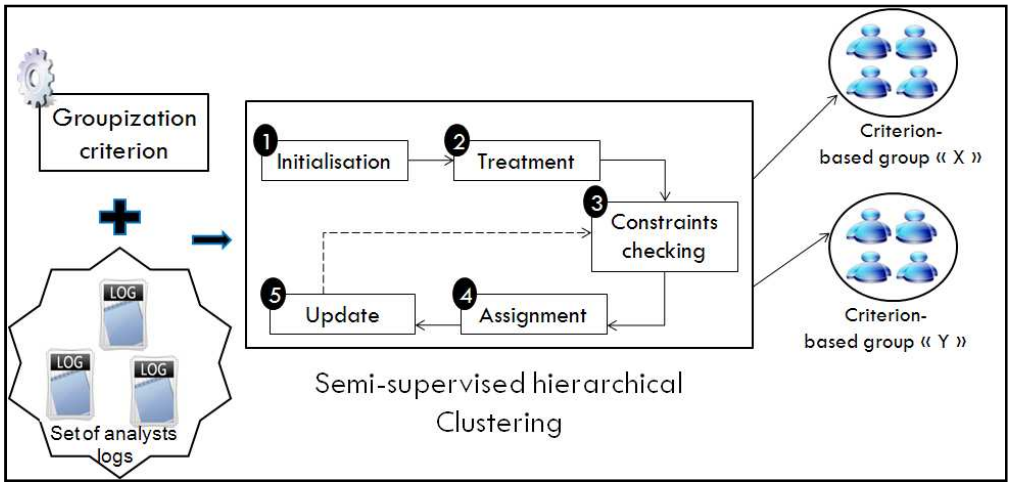


FIG. 2 – *Architecture paramétrique de notre approche IGED*

Nous présentons dans ce qui suit, la mesure de similarité entre les historiques des analystes puis nous détaillons le principe de l'algorithme.

4.1 Mesure de similarité entre les fichiers d'historiques des analystes

Une panoplie de mesures de similarité est utilisée dans le clustering hiérarchique, particulièrement pour découvrir les documents les plus similaires à fusionner.
Parmi ces mesures, la mesure du cosinus est majoritairement la plus utilisée dans le clustering de documents en particulier lorsque le nombre de concepts fréquents pour chaque document est radicalement différent. En outre, cette mesure est fondée sur les composants du document et n'est pas sensible à la longueur du document. Nous avons choisi la distance de Jaccard, car elle convient les documents volumineux. Ainsi, nous étendons cette mesure dans le contexte multidimensionnel.

Un extrait du fichier log est illustré par la figure 3. En effet, la distance de Jaccard adaptée au contexte des historiques multidimensionnelles s'appuie sur la structure de la requête MDX (i.e., *MultiDimensional eXpressions*), en particulier sur la similarité entre les faits, les mesures, les attributs de dimension, ainsi que les membres de spécification (utilisés dans la clause WHERE pour restreindre les résultats) dans les requêtes.

```
<?xml version="1.0" encoding="UTF-8"?>
    <Log session = "Gestionnaire Portefeuille X">
        <Query id = "1">
            SELECT [Measures].[Cours Indice Boursier] ON COLUMNS,
            [Titre].[All] ON ROWS
```

```
            FROM Mouvement Boursier
            WHERE ( [Société Cotée en bourse]. [Secteur].[Médecine] )
        </Query>
        <Query id = "2">
            SELECT [Measures].[ Cours Indice Boursier] ON COLUMNS,
             [Entreprise Cotée en bourse].[All] ON ROWS
            FROM Mouvement Boursier
            WHERE ([Temps].[Année].[2011] )
        </Query>
        <!–...-
    </Log session>
</xml>
```

FIG. 3 – *Extrait du fichier log du gestionnaire de portefeuille X*

La mesure de similarité est donnée par le nombre de requêtes communes ayant des faits, des mesures, des dimensions et des membres de spécification communs dans les deux fichiers d'historiques, divisé par le nombre total de requêtes dans les deux fichiers, elle est calculée selon la formule suivante :

$$J(Hi, Hj) = \frac{C_{Requêtes(Hi,Hj)}}{\sum Requêtes\ (Hi) + \sum Requêtes\ (Hj) - C_{Requêtes(Hi,Hj)}}$$

Avec $C_{Requêtes(Hi,Hj)}$: Nombre de requêtes communes dans les deux fichiers d'historiques Hi et H j,

$\sum Requêtes\ (Hi)$: Somme de toutes les requêtes existantes dans le fichier d'historique Hi.

Par exemple, nous considérons deux fichiers d'historique de deux gestionnaires de portefeuille, le premier contient 2124 requêtes MDX et le deuxième 2438 requêtes. Le nombre des requêtes communes dans les deux fichiers est 781, la distance entre les deux gestionnaires est donnée par la distance de Jaccard entre leurs deux fichiers d'historiques :

$$J(H1, H2) = \frac{C_{Requêtes(H1,H2)}}{\sum Requêtes\ (H1) + \sum Requêtes\ (H2) - C_{Requêtes(H1,H2)}} = \frac{781}{2124 + 2438 - 781} = 0.206.$$

4.2 Algorithme de clustering hiérarchique semi-supervisé

Partant de l'ensemble des historiques de tous les analystes et du nombre de groupes fixé selon le critère à étudier, (*i.e.* si le critère est la fonction, ayant quatre postes aptes à manipuler les données de l'entrepôt boursier, nous fixons le nombre de clusters à quatre) ; l'algorithme de clustering hiérarchique semi-supervisé peut être exécuté. Son fonctionnement est résumé dans les étapes suivantes :

1. *Initialisation* : Considérer chaque fichier comme un cluster ;
2. *Traitement* : Calculer la matrice de similarité selon la mesure Jaccard adaptée au cadre de l'historique OLAP multidimensionnel ;
3. *Vérification des conditions* : Selon les types de contraintes à intégrer : (i) Contrainte de *Lien-Obligatoire* : la distance entre les clusters qui doivent être fusionnés est remplacée par la plus grande distance ; (ii) Contraintes de *Lien-Interdit* : la distance entre les clusters qui ne doivent pas être fusionnés est remplacée par un zéro ;

4. *Affectation* : Fusionner les deux clusters ayant le maximum de distance de similarité;
5. *Mise à jour* : mettre à jour la matrice de similarité ;
6. *Itération* : Répéter les étapes 3, 4 et 5 jusqu'à ce que le nombre de clusters trouvés atteigne le nombre de clusters passé en paramètre.

Nous notons que les contraintes du *Lien-Obligatoire* et du *Lien-Interdit* dépendent étroitement du critère de groupization étudié. Par exemple, dans le cas où la groupization est articulée sur la fonction, la contrainte du *Lien-Obligatoire* sera les deux clusters pratiquent la même fonction et la contrainte du *Lien-Interdit* sera l'occupation de postes différents.

Afin d'illustrer le fonctionnement de notre approche IGED, nous présentons dans ce qui suit l'illustration de son processus à travers un exemple. Nous considérons le critère fonction de groupization ainsi le nombre de clusters à passer en paramètre est quatre, et l'ensemble des historiques illustré par le tableau 1.

	Gestionnaire de porte feuille *X*	Gestionnaire de porte feuille *Y*	Gestionnaire privé *Z*	Investisseur *U*	Investisseur privé *V*
Nombre de requêtes dans le fichier log	2742	2518	1313	847	563

TAB. 1 – *Nombre de requêtes dans le fichier log de chaque analyste*

D'abord l'initialisation est exécutée et cinq clusters sont dégagés. Puis, le calcul des distances entre les cinq clusters se produit pour fournir la matrice de similarité. Supposons le nombre de requêtes communes entre X et Y est 876, la distance de Jaccard est calculée comme suit :

$$J(HX, HY) = \frac{C_{Requêtes(HX,HY)}}{\sum Requêtes\ (HX) + \sum Requêtes\ (HY) - C_{Requêtes(HX,HY)}} = \frac{876}{2742 + 2518 - 876} = 0.2.$$

Nous supposons que la matrice de similarité obtenue dans notre cas est donnée par le tableau 2.

Distance	C_1	C_2	C_3	C_4	C_5
C_1	0	0.2	0.5	0.1	0.8
C_2		0	0.4	0.2	0.6
C_3			0	0.3	0.2
C_4				0	0.5
C_5					0

TAB. 2 –*Matrice de similarité obtenue suite à la première itération*

L'intégration des contraintes suivantes est accomplie : *Lien-Obligatoire* où les deux clusters pratiquent la même fonction et *Lien-Interdit* où les deux clusters ne pratiquent pas la même fonction. Suite à la vérification des contraintes, la distance entre C_1 et C_2 sera modifiée à 0.8

en respectant la contrainte du *Lien-Obligatoire* et la distance entre tout le reste des clusters sera égale à zéro en s'alignant à la contrainte du *Lien-Interdit*. L'affectation fusionnera les deux clusters C_1 et C_2 ayant la plus grande distance de similarité. Atteignant les quatre clusters, l'algorithme s'arrête.

5 Etude expérimentale

L'objectif de cette section est d'étudier les facteurs d'identification des groupes d'analystes et d'évaluer les performances de chaque groupization selon les divers critères. Durant les expérimentations effectuées, nous avons utilisé les fichiers OLAP d'un entrepôt de données mis en place dans le domaine de la bourse. Chaque fichier log contient moyennement 3000 requêtes OLAP interrogeant l'entrepôt pour effectuer des analyses.

Nous avons choisi de comparer notre algorithme à l'algorithme d'arbre de décision ID3 et à l'algorithme du réseau bayésien naïf disponibles sur la plateforme Weka[1] édition 3.6.5. Basé sur des résultats obtenus par notre approche, l'expert dénote les fichiers logs. La classe de chaque fichier OLAP est attribuée en fonction du cluster assigné par notre approche. Un fichier sous format CSV (fichier d'entrée de Weka) est généré pour chaque fichier journal et utilisé comme entrée pour les algorithmes de classification choisis.

Dans notre contexte d'identification de groupes d'analystes d'entrepôt de données, trois mesures importantes peuvent être mises en exergue pour évaluer le meilleur critère de détection de groupes :

- Le Taux des vrais positifs (**TP**) mesure la proportion d'objets qui ont été classés comme classe X, parmi tous les objets qui appartiennent vraiment à la classe X. Il est équivalent au rappel ;
- Le Taux des Faux positifs (**TF**) mesure la proportion des objets qui ont été classés comme classe X alors qu'ils appartiennent à une autre classe. Le TF est relatif au taux de la précision.
- Le *Receiver operating characteristic* (**ROC**) est la relation entre les taux TP et FP.

À travers les expérimentations, nous visons un double objectif : d'abord, nous analysons l'impact de chaque critère de groupization. Ensuite, nous mettons l'accent sur la comparaison de l'ensemble des critères d'identification du groupe.

5.1 Analyse de la performance de notre approche pour chaque critère

L'ensemble d'apprentissage formé des fichiers logs des divers analystes est annoté par un expert. Dépendamment du critère d'identification de groupes à analyser, l'expert identifie la classe de chaque fichier log.

5.1.1 Analyse du critère fonction

D'abord, nous nous concentrons sur le critère fonction. Ainsi, les classes apprises sont *: (i) les gestionnaires des portefeuilles, (ii) les investisseurs, (iii) les gestionnaires privés* et *(iv) les investisseurs privés*.

[1] http://www.cs.waikato.ac.nz/ ml/weka/.

L'ensemble d'apprentissage	ID3		Réseau bayésien naïf	
	TP	TF	TP	TP
Logs des *gestionnaires des portefeuilles*	0.52	0	0.51	0
Logs des *investisseurs*	1	0.2	1	0.01
Logs des *gestionnaires privés*	1	0	0.99	0
Logs des *investisseurs privés*	1	0	0.99	0.157

TAB. 3 –*Taux de vrai positif vs. Taux de faux positif générés par la groupization basée sur la fonction*

Comme l'illustre le tableau 3, l'algorithme ID3 engendre un TP égal à 52% pour les gestionnaires de portefeuille, à 100% pour les investisseurs, à 100% pour les gestionnaires privés et à 100% pour les investisseurs privés. Toutefois, la technique du réseau bayésien naïf génère un TP égal à 51% pour les gestionnaires de portefeuille, à 100% pour les investisseurs, à 99% pour les gestionnaires privés et à 99% pour les investisseurs privés. Ainsi, le TP généré pour les gestionnaires de portefeuille est le plus grand, car une telle fonction exercée se concentre sur une mission spécifique. Par conséquent, les requêtes généralement lancées sont fréquemment répétées et radicalement dissemblables des autres fichiers logs. Néanmoins, chaque gestionnaire privé ou investisseur analyse d'une manière différente l'entrepôt de données selon sa mission. Ainsi, les fichiers logs générés contiennent des requêtes divergentes et leurs contenus sont hétérogènes. En conséquence, le TF est relativement élevé.

5.1.2 Analyse du critère responsabilités engagées pour la réalisation des objectifs

En se basant sur le critère responsabilités engagées pour la réalisation des objectifs, nous dégageons trois classes : (i) court terme concerne des *objectifs opérationnels* ; (ii) moyen terme imputant des *objectifs tactiques* et (iii) long terme correspondant à des *objectifs stratégiques.*

L'ensemble d'apprentissage	ID3		Réseau bayésien naïf	
	TP	TF	TP	TP
Logs rassemblés selon les *objectifs opérationnels* des analystes	0.882	0.342	0.883	0.345
Logs rassemblés selon les *objectifs tactiques* des analystes	0.589	0.066	0.587	0.066
Logs rassemblés selon les *objectifs stratégiques* des analystes	0.517	0.098	0.515	0.098

TAB. 4 –*Taux de vrai positif vs. taux de faux positif générés par la groupization basée sur les responsabilités*

Le tableau 4 montre la précision de la groupization basée sur les responsabilités accordées. Pour les fichiers logs regroupés conformément aux objectifs opérationnels, l'algorithme ID3 engendre un TP égal à 88,2% alors que le réseau bayésien naïf produit un TP égal à 88,3%. Toutefois, pour les fichiers rassemblés selon les objectifs tactiques, l'algorithme ID3 induit à un TP égal à 58,9% et le réseau bayésien naïf génère un TP égal à 58,7%. En ce qui concerne les fichiers collectés conformément aux objectifs stratégiques, l'algorithme ID3 fournit un taux de TP égal à 51,7% et le réseau bayésien naïf engendre

un TP égal à 51,5%. La comparaison de ces taux nous permet de déduire que le plus grand taux de vrai positif est assuré par les analystes ayant des objectifs opérationnels similaires. En effet, les décideurs partagent des préférences similaires pour aboutir à la réalisation d'objectifs opérationnels. Travaillant sur des projets en cours, ils mèneront des analyses précises et axées sur l'entrepôt de données. Ainsi, les fichiers logs générés seront plus semblables dans le contexte de la réalisation des objectifs opérationnels que les objectifs stratégiques. Ce fait peut expliquer la baisse du taux de TP en fonction du passage du temps.

5.1.3 Analyse du critère source d'identification des groupes

La source d'identification des groupes peut être soit (i) *implicite* sans aucune intervention de l'analyse en question, soit (ii) *explicite* selon le choix de l'analyse à quel groupe il veut appartenir.

L'ensemble d'apprentissage	**ID3**		**Réseau bayésien naïf**	
	TP	**TF**	**TP**	**TP**
Logs *implicitement* collectés	0.796	0.082	0.797	0.087
Logs *explicitement* collectés	0.918	0.204	0.913	0.204

TAB. 5 –*Taux de vrai positif vs. taux de faux positif générés par la groupization basée sur la source d'identification*

Comme l'illustre le tableau 5, les fichiers journaux explicitement collectés, à la fois les deux algorithmes ID3 et réseau bayésien naïf génèrent des TP respectivement égaux à 91,8% et 91,3%. Cependant, les journaux implicitement rassemblés apportent un TP égal à 79,6% pour les ID3 et 79,7% pour le réseau bayésien naïf. Il est à noter que c'est préférable de choisir le groupe de l'analyste explicitement que de l'apprendre automatiquement, car l'analyste connaît ses collègues et leurs préférences de sorte qu'il est plus qualifié pour sélectionner à quel groupe appartenir. La capacité à identifier les groupes est importante pour garantir les meilleurs résultats des techniques de groupization.

5.1.4 Analyse du critère dynamicité

L'identification des groupes peut se faire soit (i) *statiquement* d'une façon figée, ou (ii) *dynamiquement* selon les événements qui surgiront.

L'ensemble d'apprentissage	**ID3**		**Réseau bayésien naïf**	
	TP	**TF**	**TP**	**TP**
Logs rassemblés *statiquement*	1	0.25	1	0.251
Logs rassemblés *dynamiquement*	0.75	0	0.749	0

TAB. 6 –*Taux du vrai positif vs. taux de faux positif générés par la groupization basée sur la dynamicité*

Comme l'illustre le tableau 6, l'algorithme ID3 et le réseau bayésien naïf génèrent respectivement des TP égaux à 75% et 74,9% pour les fichiers collectés de façon dynamique. Alors que les deux algorithmes produisent des TP égaux à 100% pour des journaux statiquement collectés. Ceci s'explique par le fait que dans la groupization basée sur la dynamicité est sensible aux changements constants et aux fluctuations variantes de la

bourse. Ainsi, les requêtes vont radicalement évoluées et la similitude entre les fichiers logs se réduira progressivement.

5.2 Comparaison des capacités d'identification des groupes d'analystes d'entrepôts de données

Pour cerner quel critère est le plus discriminant pour la détection des groupes d'analystes d'entrepôts de données, nous présentons une comparaison des mesures ROC générés pour chaque critère, comme l'illustre le tableau 7.

L'ensemble d'apprentissage	**ROC de l'ID3**	**ROC du réseau bayésien naïf**
Logs regroupés selon la *fonction*	0.963	0.967
Logs regroupés selon les *responsabilités*	0.841	0.842
Logs regroupés selon la *source*	0.932	0.934
Logs regroupés selon la *dynamicité*	0.939	0.94

TAB. 7 –*Comparaison des ROC pour chaque critère*

En comparant les critères de groupization étudiés, nous remarquons que la fonction est le critère le plus discriminant. En effet, le ROC de la groupization basé sur la fonction est égal à 96,3% pour l'ID3 et 96,7% pour l'algorithme du réseau bayésien naïf. Le second critère est la source de l'identification du groupe. En fait, le ROC de la groupization basée sur la source d'identification est de 93,2% pour les ID3 et 93,4% pour le réseau bayésien naïf. Le troisième critère classé est la dynamicité et finalement les responsabilités.

En résumé, nous soulignons l'importance de la fonction exercée par l'analyste face à la source de l'identification. Ce critère peut être considéré comme un critère assez discriminant car tout analyste connait ses collègues et il est le plus habile de choisir à quel groupe il préfère appartenir. Quant au critère de la dynamcité, il est bien reconnu que l'identification peut prendre en considération les changements tels que les variations probables sur le marché boursier. Enfin, les responsabilités accordées visant à atteindre des objectifs précis au fil du temps est le critère le moins tranchant car les objectifs sont en constante évolution, d'où les responsabilités accordées varient constamment ainsi le clustering basé sur ce critère n'aboutit pas à des résultats très satisfaisants.

6 Conclusion

Dans cet article, nous avons traité le problème de la groupization dans le cadre des entrepôts de données. Nous avons analysé la similarité entre les requêtes à partir des historiques d'activités des analystes pour mesurer la distance entre ces décideurs. L'approche proposée se base sur le clustering hiérarchique semi-supervisé des historiques des analystes de l'entrepôt afin de cerner le meilleur critère d'identification des groupes. L'analyse des résultats des expérimentations de notre approche a permis de constater que la fonction est le meilleur critère discriminant pour identifier des groupes d'analystes d'entrepôt.

Cette contribution suggère des perspectives de travaux futurs concernant : (i) le couplage de plusieurs critères d'identification de groupes pour améliorer la qualité des groupes détectés ; (ii) la proposition de notre propre approche d'algorithme hiérarchique semi-supervisé répondant adéquatement aux exigences de notre contexte ; (iii) l'exploration de méthodes alternatives afin de guider la supervision du clustering, telles que la sélection supervisée d'attributs.

Références

Aligon, J., Golfarelli, M., Marcel, P., Rizzi, S. et Turricchia E. (2011) Mining Preferences from OLAP Query Logs for Proactive Personalization. *Proceedings of the 15th Advances in Databases and Information Systems (ADBIS'11)*, 84-97, Springer-Verlag.

Bade, K., M. Hermkes, et A. Nürnberger. (2007). User oriented hierarchical information organization and retrieval. *Proceedings of the 18th European conference on Machine Learning*, (ECML '07), 518–526, Berlin, Heidelberg. Springer-Verlag.

Bellatreche, L., A. Giacometti, P., Marcel, H., Mouloudi, D., Laurent. (2005). A personalization framework for OLAP queries, International Workshop on Data Warehousing and OLAP (DOLAP'05) 9-18

Ben Ahmed, E., A. Nabli et F. Gargouri (2011). A Survey of User-Centric Data Warehouses : From Personalization to Recommendation, *The International Journal of Database Management Systems (IJDMS)*, May 2011, Volume 3, Number 2, 59-71.

Ben Ahmed, E., A. Nabli et F. Gargouri (2012). Building MultiView Analyst Profile From Multidimensional Query Logs : From Consensual to Conflicting Preferences, *The International Journal of computer science issues (IJCSI)*, To appear.

Böhm, C. et C. Plant. (2008). Hissclu : a hierarchical density-based method for semi-supervised clustering. Proceedings of the 11th international conference on Extending database technology, (EDBT '08), 440–451, New York, NY, USA.

Favre, C., F. Bentayed, O., Boussaid. (2007). Evolution et personnalisation des analyses dans les entrepôts de données : une approche orientée utilisateur, *Congrès Informatique des organisations et systèmes d'information et de décision*, Perros-Guirec 308-323.

Giacometti A., P. Marcel, and E. Negre (2008). A Framework for Recommending OLAP Queries, *In International Workshop on Data Warehousing and OLAP*, US 307-314.

Golfarelli, M. (2008) From User Requirements to Conceptual Design in Data Warehouse Design – a Survey, In Data Warehousing Design and Advanced Engineering Applications: Methods for Complex Construction.

Jain, A.K., R.C. Dubes. (1988). *Algorithms for clustering data*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Jerbi, H., F. Ravat, O. Teste, G., Zurfluh. (2009) . Applying RecommendationTechnology in OLAP Systems, *International Conference on Enterprise Information Systems* (ICEIS'2009), Milan, Italie 220-233

Kestler, H. A., J.M. Kraus, G. Palm, et F. Schwenker. (2006). On the effects of constraints in semi-supervised hierarchical clustering. *Artificial Neural Networks in Pattern Recognition*, 57–66. Springer-Verlag.

Klein, D., S. D. Kamvar, et C. D. manning (2002). From instance-level constraints to space-level constraints : making the most of prior knowledge in data clustering. *Proceedings of the 19th international conference on machine learning, (ICML '02)*, 307–314, san francisco, CA, USA.

Morris, M.R., et J. Teevan (2008). Understanding Groups' Properties as a Means of Improving Collaborative Search Systems.*8th Workshop on Collaborative Information Retrieval, JCDL 2008* .Pittsburgh, USA, Juin 2008.

Morris, M.R., J. Teevan, et S. Bush (2008). Enhancing Collaborative Web Search with Personalization : Groupization, Smart Splitting, and Group Hit-Highlighting, *Proceedings of the 2008 ACM conference on Computer supported cooperative work.*

Ravat, F., O. Teste. (2008). Personalization and OLAP Databases, Annals of Information Systems, *New Trends in Data Warehousing and Data Analysis*, Vol. 3 7192

Rizzi, S. (2010). New Frontiers in Business Intelligence : Distribution and Personalization, *Advances in Databases and Information Systems (ADBIS'10)*. Springer-Verlag 23-30

Teevan. J, R. M. Morris, et S. Bush (2009). Discovering and Using Groups to Improve Personalized Search, *Proceedings of Web Search and Data Mining (WSDM).*Février 2009.

Summary

A challenge of personalization is to enrich the individual preferences using similar individual's data. This method is known as 'groupization'. It may efficiently adapt the query results to the user expectations. In this paper, we aim to optimally identify the analyst'groups in data warehouse. For that reason, we study the similarity between the selected queries in the analytical history. Four axis for group identification are assessed. A semi-supervised hierarchical algorithm is used to discover the most discriminating criterion. Carried out experiments on real data warehouse confirm the soundness of our approach. And our findings demonstrate that groupization improves upon personalization for several group types, mainly for function-based groupization and explicitly identified groups.

Une approche parallèle optimisée de distribution intégrale de la fouille de données basée sur une infrastructure CORBA

Abdelfettah Idri, Azedine Boulmakoul

Département Informatique, Laboratoire Informatique de Mohammedia
Faculté des Sciences et Techniques de Mohammedia, Maroc
abdelfattah_id@yahoo.com
azedine.boulmakoul@yahoo.fr

Résumé. L'objectif principal de la fouille de données est l'extraction de l'information utile depuis des entrepôts de données souvent volumineux et sa transformation en un outil de prise de décision. Ce processus 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. Dans nos travaux antérieurs, l'accent a été mis sur la distribution du processus de la fouille de données relativement au traitement. La distribution de la mémoire laisse espérer une amélioration substantielle de la performance. Du moment que notre démarche repose sur le treillis de Gallois pour la prospection des données, on propose dans ce papier une approche parallèle optimisée de distribution intégrale de la construction du treillis de Galois basée sur une infrastructure CORBA. Le treillis ainsi généré sera réutilisé pour la génération des règles d'association.

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, puisque la résolution du problème dans l'analyse formelle des concepts peut directement servir dans la prospection de données. L'analyse des concepts formels (FCA) représente en fait un socle pour la prospection de données vu que ces deux domaines sont étroitement liés. 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.

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.

Généralement dans le domaine d'analyse de concepts formels, les données sont formulées sous forme de contexte. Un contexte est constitué d'un triplet (O, M, I) où O représente l'ensemble des objets, M l'ensemble des attributs et I une relation binaire entre O et M. Sur la base de ce contexte, un ensemble de concepts peut être construit. Lorsque ce dernier satisfait une relation d'ordre partiel, on parle alors de treillis de Galois ou de treillis de concepts Barbut et Montjardet (1970).

En général le processus de la fouille de données est décliné sur plusieurs étapes qui collaborent toutes pour aboutir au résultat final qui est la découverte de la connaissance représentée par les règles d'association. L'optimisation de chacune de ces étapes implique directement l'optimisation du processus global. Suivant ce raisonnement, des interventions au niveau de la préparation des données et de l'implémentation des structures de données ont eu lieu afin de réduire le temps d'exécution et de permettre le stockage du treillis de Galois en mémoire pour des bases de données consistantes. Ainsi, dans la phase de préparation des données, ces dernières ont subi des transformations permettant leur gestion a travers les indexes de tableaux assurant un accès mémoire rapide et minimisant l'espace mémoire occupé. En plus, on a adopté pour l'implémentation des ensembles et leur gestion la notion de *bitsets* qui a énormément optimisé l'utilisation mémoire et qui a permis la manipulation de Trie pour des bases de données denses sans faire recours à la sérialisation. Dans ce papier, on présente un algorithme parallèle distribué pour la construction de treillis de Galois en se basant sur les mêmes techniques des algorithmes séquentiels tels que celui de Bordat (1986) et Choi (2006). Nous présentons également l'architecture du système supportant cet algorithme ainsi que son implémentation. L'accent sera mis sur la distribution intégrale du processus de la fouille de données et donc aussi bien du traitement que de l'utilisation mémoire. Ce document est organisé comme suit. Le paragraphe 2 rappelle la théorie et la terminologie des treillis de Galois. Le paragraphe 3 présente l'architecture du système. L'algorithme parallèle distribué est abordé dans le paragraphe 4. Le paragraphe 5 traite l'implémentation de l'algorithme et expose les résultats de cette approche. Dans le paragraphe 6 on présente les règles d'association. On conclut dans le paragraphe 7 avec nos suggestions et recommandations.

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.

Généralement dans le domaine d'analyse de concepts formels, les données sont formulées sous forme de contexte. Un contexte est constitué d'un triplet (O, M, I) où O représente l'ensemble des objets, M l'ensemble des attributs et I une relation binaire entre O et M. Sur la base de ce contexte, un ensemble de concepts peut être construit. Lorsque ce dernier satisfait une relation d'ordre partiel, on parle alors de treillis de Galois ou de treillis de concepts Barbut et Montjardet (1970).

Par ailleurs, il est important de considérer la relation entre les treillis de Galois et la prospection de données. En fait 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, puisque la résolution du problème dans l'analyse formelle des concepts peut directement servir dans la prospection de données.

Dans ce papier, on présente un algorithme parallèle pour la construction de treillis de Galois en se basant sur les mêmes techniques des algorithmes séquentiels tels que celui de Bordat (1986) et Choi (2006). Nous présentons également l'architecture du système supportant cet algorithme ainsi que son implémentation. Ce document est organisé comme suit. Le paragraphe 2 rappelle la théorie et la terminologie des treillis de Galois. Ensuite nous verrons dans le paragraphe 3 l'architecture du système. Le paragraphe 4 aborde l'algorithme parallèle. Le paragraphe 5 aborde l'implémentation de l'algorithme et expose les résultats de cette approche. Dans le paragraphe 6 on traite les règles d'association. On conclut dans le paragraphe 7 avec nos suggestions et recommandations.

2 Analyse formelle des concepts, théorie et terminologie

L'analyse formelle des concepts (ou FCA) est un domaine de recherche vaste et elle est dérivée de la théorie des treillis basée sur la notion de concepts. FCA s'intéresse à la construction des treillis de concepts fournissant ainsi un outil efficace pour la fouille de données et la génération des règles d'associations. Dans la suite du paragraphe on aborde les notions de base de FCA.

2.1 Définitions

Définition 1 Contexte : Dans FCA, on nomme un contexte le triplet (O, M, I), où $O = \{g_1, g_2, ..., g_n\}$ désigne un ensemble de n éléments appelés objets ; $M = \{1,2,...,m\}$ désigne un ensemble de m éléments appelés attributs et $I \subseteq O \times M$ la relation binaire entre les objets et les attributs.

Le contexte est représenté souvent sous forme d'un tableau dont les objets sont en ligne et les attributs sont en colonne comme le montre le Tableau 1. On appelle ensemble d'objets un sous-ensemble $X \subseteq O$. De même, on appelle un ensemble d'attributs un sous-ensemble $J \subseteq M$. Par convention, on écrit un ensemble d'objet $\{b, d, e\}$ sous la forme bde, et un ensemble d'attribut $\{3,4,6\}$ sous la forme 346.

Définition 2 Listes adjacentes : L'ensemble des objets communs d'un élément $i \in M$ est défini par $nbr(i) = \{g \in O : (g, i) \in I\}$ et appelé liste adjacente de i. L'ensemble des attributs communs d'un élément $g \in O$ est défini par $nbr(g) = \{i \in M : (g, i) \in I\}$ et appelé liste adjacente de g. Dans le Tableau 1, on peut lire $nbr(a) = \{1,6\}$ et $nbr(1) = \{a, b, c\}$.

	1	2	3	4	5	6	7
a	1					1	
b	1		1	1	1	1	
c	1			1		1	
d		1	1		1		
e		1					1

Tableau 1 Exemple de contexte (O,M,I) *avec* $O=\{a,b,c,d,e\}$ *et* $M=\{1,2,3,4,5,6,7\}$ *. Le tableau représente la relation binaire* I

Définition 3 Les fonctions attr et obj : La fonction $attr:2^O \to 2^M$ fait correspondre à un ensemble d'objets donnés leurs attributs communs : $attr(X)=\bigcap_{g\in X} nbr(g)$ avec $X \subseteq O$. De la même façon, la fonction $obj:2^M \to 2^O$ fait correspondre à un ensemble d'attributs donnés leurs objets communs : $obj(J)=\bigcap_{j\in J} nbr(j)$ avec $J \subseteq M$.

Définition 4 La fermeture des ensembles : Un ensemble d'objets $X \subseteq O$ est fermé si $X=obj(attr(X))$. Un ensemble d'attributs $J \subseteq M$ est fermé si $J=attr(obj(J))$.

Dans Tableau 1, l'ensemble $X=abc$ est fermé puisque $obj(attr(X))=abc$. $attr(abc)=16$ et $obj(16)=abc$.

Définition 5 Concepts : Un concept est un couple de la forme $C=(X,J)$ avec $X \subseteq O$ et $J \subseteq M$ dans lequel $X=obj(J)$ et $J=attr(X)$.

L'ensemble X est nommé l'extension (extent) de C et noté $X=ext(C)$. L'ensemble J est nommé l'intention (intent) de C et noté $J=\text{int}(C)$. Par définition, X et J sont tous les deux fermés. L'ensemble de tous les concepts du contexte (O,M,I) est noté par $B(O,M,I)$ ou B. La relation d'ordre définie sur B de la manière suivante :

$$(A_1,B_1) \prec (A_2,B_2) \Leftrightarrow A_1 \subseteq A_2 (B_2 \subseteq B_1)$$

Où (A_1,B_1) et (A_2,B_2) sont deux concepts de B, est une relation d'ordre partielle sur B.

Définition 6 treillis de Galois : $L=<B,\prec>$ est un treillis de concepts (Galois) complet.

2.2 L'algorithme séquentiel

Construire un treillis de Galois revient à générer tous les concepts en identifiant les successeurs de chacun d'entre eux. L'idée principale d'un algorithme séquentiel est de commencer par le concept parent $(O, attr(O))$ et de générer ensuite tous ses successeurs. D'une manière récursive, on génère chacun de ces successeurs selon l'algorithme de parcours en largeur (BFS : Breadth First Search). La figure ci-dessous montre l'architecture de l'algorithme séquentiel. Il s'agit de trois étapes principales :

- La préparation de données
- La génération du treillis de Galois en utilisant un trie global et un trie local
  - La génération des concepts enfants
  - Le test de la fermeture d'un concept candidat
  - Le test d'existence d'un concept
- La visualisation du treillis

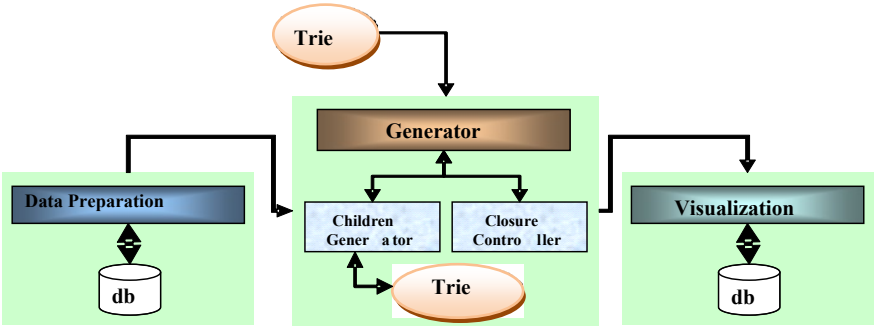


Figure 1 Architecture de l'algorithme séquentiel

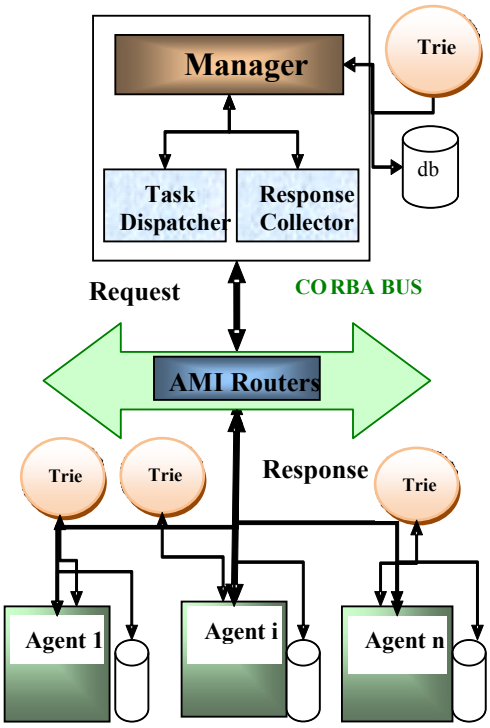
3 Architecture du système cible

En général, le nombre de concepts issu d'un contexte donné est exponentiel par rapport à la taille des données initiales. Par conséquent, la génération des concepts (treillis de Galois) peut devenir très coûteuse en termes de complexité temporelle et spatiale. De ce fait, on s'est penché sur l'étude de possibilités pour améliorer les performances du processus de construction du concept Galois en s'intéressant à l'aspect distribution et parallélisme d'exécution de l'algorithme.

Une approche parallèle optimisée de distribution intégrale de la fouille de données basée sur une infrastructure CORBA

3.1 Aspect traitement

3.1.1 Architecture



Algorithme parallèle Manager

1: Initialise concept queue Q= {*C*}; *C* = (*O*, attr(*O*))
2: Initialise Context, Agenda, Trie
5: Initialise and Synchronise Agents
6: **while** *Q* is not empty or Agenda is not empty
or waiting for reponses **do**
7: **while** Q is not empty **do**
8: *C* = dequeue(*Q*)
9: Enqueue C to Agenda
10: **end while**
11: **for each** available Agenti **do**
12: **if** Agenda is not empty **then**
13: C = dequeue(Agenda)
14: Send_request(generateChildren(C)) to Agenti
15: Mark Agenti as busy
16: **end if**
17: **end for**
18: **for each** buzy agenti **do**
19: **if** response of agenti is available **then**
20: get response
21: **for each** child in response **do**
22: **if** child not exists in Trie **then**
23: insert child into Trie
24: Identify child as successor of C
25: Enqueue child to Q
26: **else**
27: Identify child as successor of C
28: **end if**
29: **end for**
30: mark agenti as free
31: **end if**
32: **end for**
33: **end while**

Figure 2 Architecture du système et Algorithme

L'architecture proposée dans le schéma ci-dessus est constituée de trois composantes principales dans l'objectif d'améliorer la performance de l'algorithme séquentiel (voir Figure 1) comme illustré dans le paragraphe 3.1.2 :

- 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. Une tâche comprend en fait la génération des concepts enfants d'un concept donné.
  - 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*).

3.1.2 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 alourdir la communication entre elles. 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 :

- 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 et la distribution des services et ceci coïncide bien avec notre objectif.

La génération des enfants d'un concept est un processus complexe et utilise un algorithme spécial ainsi qu'un arbre local. Cette tâche peut être déléguée aux Agents puisqu'elle peut s'exécuter d'une manière totalement indépendante. La multiplication du nombre d'agents implique directement la réduction du temps d'exécution et permet d'éviter les piques de mémoires pendant la génération des enfants. Par ailleurs ceci exige une implémentation efficace pour le transport des concepts enfants entre le Manager et les Agents.

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 de codification (Trie). La clé se compose des éléments de l'intention du concept. Cette tâche ne peut pas être totalement déléguée aux Agents puisque l'arbre contient au fur et à mesure tous les concepts générés par tous les Agents et donc il doit être partagé par eux pour pouvoir tester l'existence d'un concept donné. C'est donc le Manager qui prend en charge la gestion de l'arbre.

Le Manager utilise un dispatcher pour distribuer les tâches aux Agents et un collecteur pour collecter les résultats envoyés par ces derniers. 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 Distribution de la 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 résultant à l'assignation des tâches de génération des concepts enfants et de test de fermeture aux Agents. 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 puisque ces informations ne sont disponibles qu'au moment de l'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. Notre approche est la suivante : on s'est inspiré du monde de la biologie (spécifiquement la division cellulaire qui se déclenche une fois certaines conditions sont satisfaites) et des réseaux de neurones. Dans notre cas, notre souci est la taille du Trie et donc on a définie la notion de poids d'un nœud qui est le nombre de nœuds se trouvant au dessous du nœud concerné. Une fois que ce poids atteigne une certaine valeur prédéfinie, le processus de subdivision se déclenche. Comme on a statué que les branches du Trie sont uniques puisqu'elles représentent des clés 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 implémentée par une table de hachage ou même un Trie. Ce processus est décrit dans la Figure 4. On a pris le cas d'un Trie ordinaire mais un Trie binaire est aussi valable. Du moment que l'Agent utilise lui aussi un Trie local pour la génération des concepts enfants, la distribution mémoire peut être bien appliquée à ce niveau. Encore plus, comme notre architecture est récursive, la distribution mémoire peut accompagner cette logique et s'effectuer récursivement. Pour optimiser la gestion du Trie distribué, on peut prévoir une référence dans le nœud de subdivision vers la machine contenant le sous-Trie correspondant, autrement dit vers la ligne de la table de répartition du Trie. Dans l'exemple de la figure, une fois le poids du nœud « 3 » ait atteint le seuil de distribution, le processus de subdivision se déclenche et l'hiérarchie au dessous de ce nœud (3) est migrée vers une autre machine « host i ». Pour assurer la transparence du Trie, 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 distribution intégrale du processus de génération du treillis de Galois comme illustré dans la Figure 3.

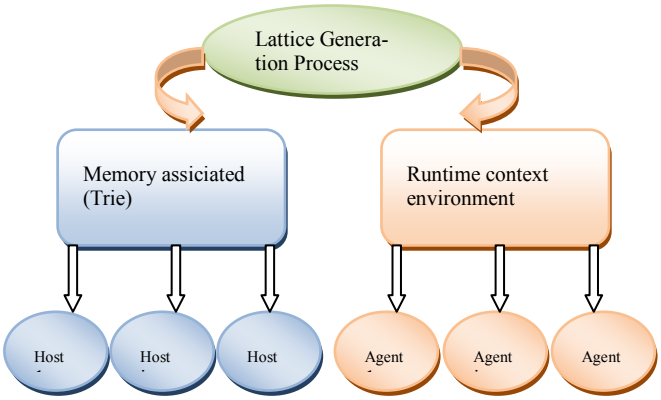


Figure 3 Vue abstratie du processus de génération du treillis de Galois

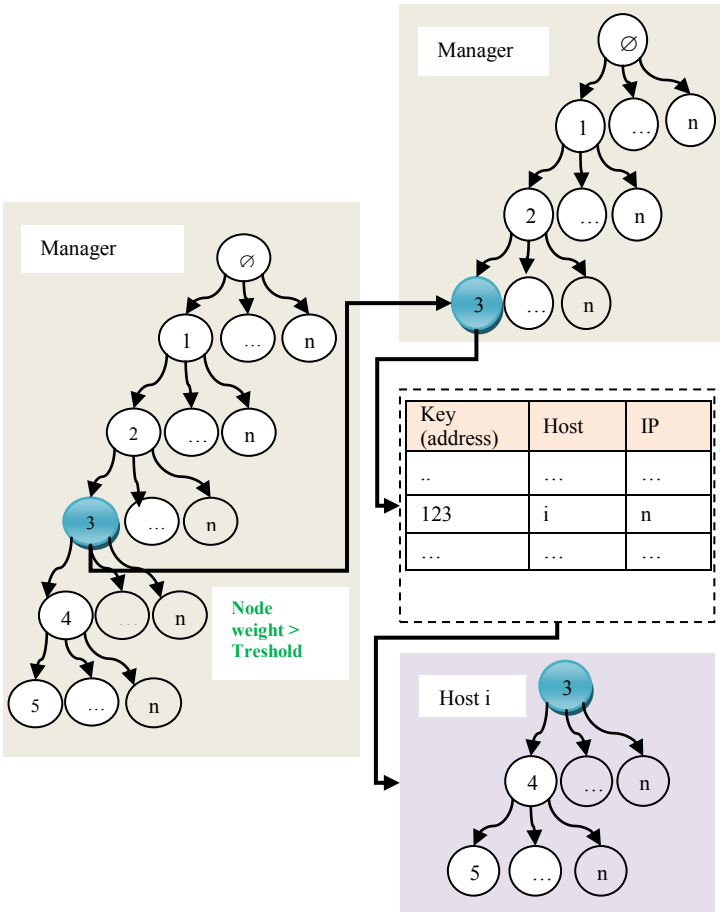


Figure 4 Distribution du Trie (mémoire)

4 L'algorithme parallèle de construction de treillis de Galois

La Figure 2 expose l'algorithme du système. On explique dans ce qui suit les principales étapes de l'approche adoptée.

Selon notre schéma proposé, la construction du treillis de Galois est réalisée dans deux phases principales réparties sur le Manager et les Agents.

Première phase :

Tout d'abord, l'Agent s'occupe de la génération des concepts enfants candidats d'un concept donné. Ensuite, l'Agent applique simultanément la fermeture au résultat obtenu de façon à n'envoyer au Manager qu'un ensemble de concepts déjà traité.

Deuxième phase :
Le Manager envoie progressivement les concepts disponibles dans l'Agenda (une file de concepts) aux Agents sélectionnés par le Dispatcher. En retour, le collecteur reçoit les réponses des Agents sous forme d'ensembles de concepts représentant les concepts enfants des concepts envoyés. Le Manager procède alors à la mise à jour de l'arbre des concepts : soit par insertion du concept enfant et connexion avec son concept parent ; soit par connexion seulement de ces deux concepts en cas d'existence préalable du concept enfant dans l'arbre. Ce processus est répété jusqu'au traitement de tous les concepts dans la file des concepts.

5 Implémentation et résultats

Dans ce paragraphe, nous traitons les aspects d'implémentation de l'algorithme dans la première section. Notre implémentation est applicable aussi bien dans le domaine de l'analyse formelle des concepts que dans le domaine de la fouille de données. On expose nos résultats dans la deuxième section.

Environnement de travail
Pour l'environnement de travail et de test on a utilisé la configuration suivante :

- Plateforme : Windows XP, C++ de Visual Studio
- Communication : CORBA de Orbacus 4.3
- Performance machine : centrino avec un processeur de 2,26 GH et une mémoire de 1 GB
- Comme outil de visualisation du treillis nous avons utilisé Galicia 3.2

5.1 Implémentation

La première section aborde le diagramme de collaboration du modèle. La deuxième section spécifie les données d'entrée et de sortie. La troisième section traite les structures de données principales utilisées dans l'algorithme. Dans la quatrième section on discutera la communication entre le Manager et les Agents ainsi que l'outil CORBA. La cinquième section se focalise sur l'aspect optimisation au niveau de la préparation des données.

5.1.1 Diagramme de collaboration

La figure ci-dessous expose le diagramme de collaboration (UML) reflétant les relations entre les différentes classes. On a choisi celui de la classe *cchildren_impl* (l'implémentation de l'interface CORBA côté agent) du fait qu'il est le plus significatif et il couvre la majorité des classes qu'on désire aborder. La signification de ces dernières est la suivante :

- *cchildren_impl et POA_lattice::cchildren* : implémentent les fonctionnalités de l'Agent intégrant l'interface CORBA
- *attributeSet* : implémente l'ensemble des attributs (l'intention d'un concept)
- *objectSet* : implémente l'ensemble des objets (l'extension d'un concept)
- *binaryRelation* : implémente la relation binaire entre l'ensemble des attributs et celui des objets (voir les liens vers les deux ensembles)

- *Concept* : implémente le concept et se compose entre autres de l'ensemble des attributs et des objets.
- *Context* : implémente le contexte du treillis de Galois. Il est évident qu'il comporte l'ensemble des attributs, l'ensemble des objets et la relation binaire comme indiqué dans le schéma.
- *Triespr* : implémente le Trie local servant à la génération des concepts enfants. Dans le schéma, cette classe pointe sur *triespr_node* pour exprimer qu'il s'agit d'une structure de donnée récursive (arbre).
- *Lattice* : implémente les fonctionnalités de génération des concepts enfants.

Dans ce qui suit on fera référence aux classes de ce diagramme pour exprimer le lien avec les concepts correspondants.

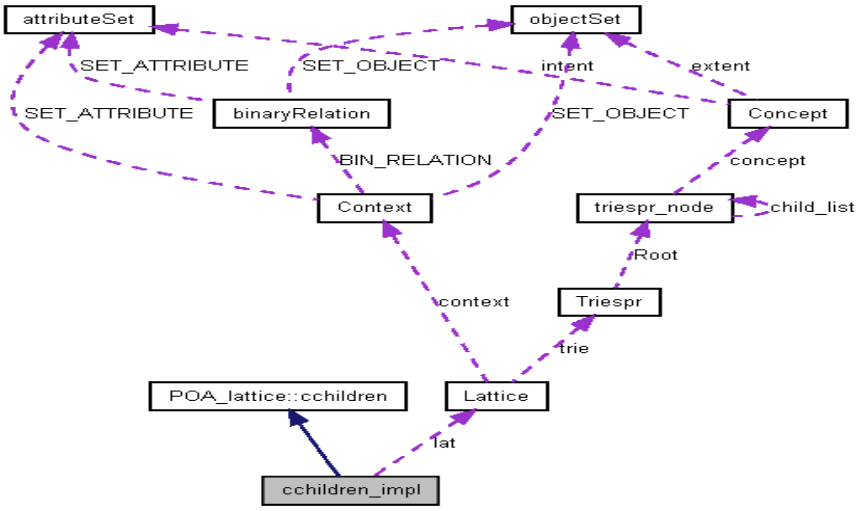


Figure 5 Digramme de collaboration de cchildren_impl

5.1.2 Les entrées et sorties

On a adopté deux formats pour les données de notre algorithme : le format SLF de Galicia et le format transactionnel matriciel. De même pour les sorties, on génère deux formats : le format GSH-XML de Galicia qui est un fichier XLM et un format interne spécifique pour des fins d'analyse et de recherche. La classe *Context* se charge de la lecture du contexte alors que la classe *Lattice* supporte la génération du format GSH-XML.

5.1.3 Structures de données

En général, pour les structures de données standards telles que les ensembles, les files, les listes et les vecteurs, on utilise la librairie standard de C++ : STL. Dans cette section on aborde spécialement les structures de données les plus spécifiques et relevantes.

Le contexte

Celui-ci est constitué d'un ensemble d'objets, un ensemble d'attributs et une relation binaire, conformément à sa définition originale. La relation binaire comporte sa table de données sous forme d'un ensemble de structure comportant un objet, un attribut et une booléenne indiquant leur relation. Dans le domaine de la fouille de données, les données sont sous forme d'une matrice où la première colonne indique les transactions (objets) et les autres colonnes représentent les items (attributs). Par conséquent, la relation binaire est définie implicitement. Ce sont les classes *Context, binaryRelation, attributeSet et ObjectSet* du diagramme de collaboration qui implémentent ces structures de données.

L'arbre des concepts (Trie)

Quant à l'arbre des concepts, on a adopté une codification lexicographique pour mémoriser les concepts candidats. Un concept candidat est identifié par une clé se composant des éléments de son intention (ou extension). La classe *Triespr* (agent) ou *Trie* (Manager) implémente le Trie (dans ce cas *Triespr* relativement à l'agent). . La Figure 5 montre l'état du trie après l'insertion des deux concepts (abd,16) et (de,2) correspondant au contexte du Tableau 1

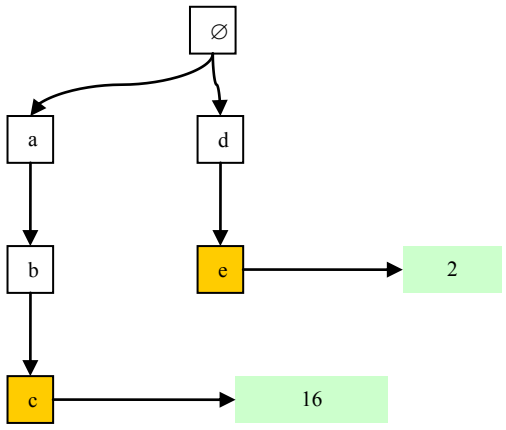


Figure 6 Le Trie après insertion des concepts (abc,16) et (de,2)

Le concept

Un concept est constitué d'un ensemble d'attributs (extension), un ensemble d'objets (intention), une identité unique, une liste des identités des parents et une liste des identités des enfants. Voir la classe *Concept* dans le diagramme.

5.1.4 CORBA et communication Manager/Agent

Parmi les alternatives présentées par CORBA, on a choisi la technique ***AMI Poller*** pour son avantage qu'elle n'impacte pas l'agent en plus qu'elle offre le mode asynchrone.

Une approche parallèle optimisée de distribution intégrale de la fouille de données basée sur une infrastructure CORBA

Par ailleurs, on peut se contenter d'une communication CORBA simple tout en adoptant le multithreading pour le Manager. Une explication plus élaborée sur cet aspect est exposée dans Idri et al.(2008).

5.1.5 Optimisation de la préparation des données

Parmi les étapes qu'on peut identifier dans le processus de la fouille de données Han et Kamber (2006), Ye (2003) et Kantardzic (2003), on trouve celle de la préparation des données. La projection de ce principe sur notre contexte s'avère nécessaire puisqu'on doit adapter les données brutes afin qu'elles puissent être exploitées d'une manière optimale. Les ensembles Attributs et Objets sont de type alphanumérique alors qu'on aimerait assigner à ces ensembles une structure de données abstraite qui est indépendante des données initiales de telle manière que les étapes postérieures du processus de la fouille de données restent transparentes. Trois transformations ont eu lieu pour assurer le passage de l'alphanumérique aux entiers optimisé comme le montre la Figure 7.

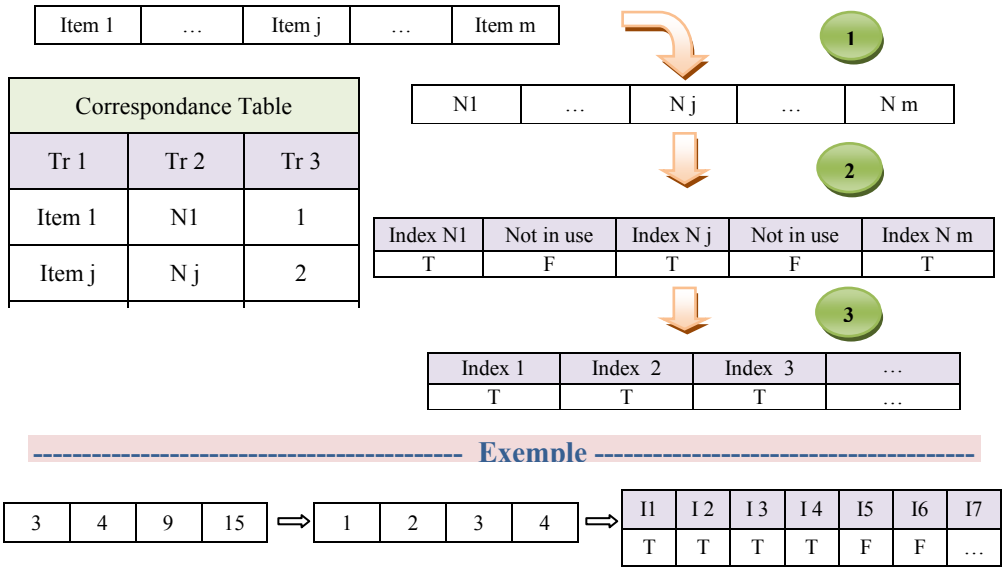


Figure 7 Préparation des données

Ces transformations sont aussi bien valables pour les objets que pour les attributs. Pour retrouver les valeurs initiales lors de l'affichage des résultats, les transformations sont tracées dans des tables de correspondances qui à bas niveau peuvent être traduites par des ensembles d'objets et d'attributs temporaires comme mentionné ci-dessus. La première transformation convertit les items alphanumériques en entiers sous forme d'un tableau (ou ensemble) par exemple. Une fois on a obtenu des nombres entiers à manipuler, la deuxième transformation sert dans cette modélisation à substituer chaque entier par son indexe correspondant dans le tableau de l'étape précédente et ainsi manipuler les index d'un tableau au lieu de manipuler et stocker les entiers à l'état brut et par conséquent le temps de recherche d'un nombre (attri-

but ou objet) devient linéaire et puisqu'il s'agit d'un tableau indexé il équivaut $\theta(1)$ en plus que l'accès à un tableau est nativement plus rapide qu'une liste dynamique. Pour indiquer qu'un élément du tableau est actif et qu'il appartient à l'ensemble représenté par ce tableau, on a associé une booléenne au contenu du tableau de telle façon que la valeur vraie signifie que l'index appartient à l'ensemble représenté et vis-versa. La taille du tableau équivalente à l'image d'un ensemble peut croitre inutilement du moment que la distribution des index qui est aléatoire et liée à la nature des données et ceci peut causer la dilatation du tableau suite à quelques index de grandes valeurs et laisser derrière des ruptures d'index inutilisés. C'est pour cela que l'objectif de la troisième transformation est de garantir la continuité des index et ainsi réduire la taille du tableau d'une manière optimale : cette correspondance projette chaque index de l'étape deux vers une plage d'index commençant par 1 et consommant la valeur suivante des index séquentiellement jusqu'à épuisement de l'ensemble des items à stocker.

5.2 Résultats et expérimentations

5.2.1 Exemple

Visualisation

Pour nos expérimentations et tests, on a intégré notre algorithme dans Galicia (voir adresse site web de Galicia).On a adopté le format de sortie GSH-XML pour s'interfacer avec Galicia. L'utilisateur peut soit lancer L'algorithme directement depuis Galicia, soit le lancer séparément et utiliser le résultat ainsi généré sous le format GSH-XML dans Galicia pour le visualiser.

Afin d'illustrer le fonctionnement de notre architecture, on présente ci-dessous le résultat de l'exemple de données transactionnelles listé dans Figure 8 à l'aide de Galicia. Les données de cet exemple représentent en fait un sous-ensemble du fichier de test « mushroom » utilisé dans la fouille de données. Ces données sont sous la forme **SLF** et **transactionnelle** ((a) respectivement (b)). Le treillis de Galois relatif au contexte de la Figure 8-b est présenté par le graphe de la Figure 8-c. Les nœuds représentent les concepts. Le concept racine se trouve tout en haut de l'arborescence. Les concepts enfants sont liés directement au concept parent du niveau hiérarchique supérieur. On retrouve pour chaque concept (nœud) les références vers son intention et son extension.

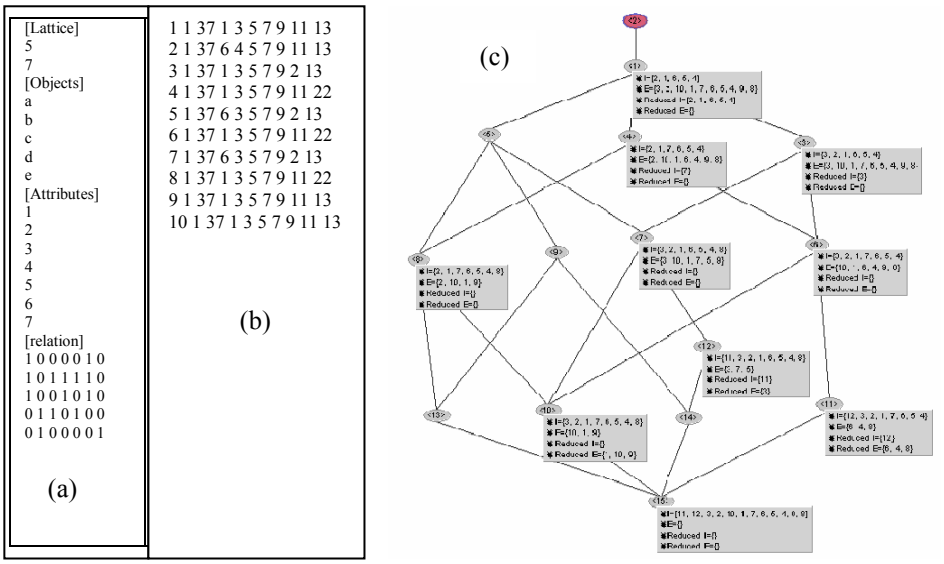


Figure 8 Exemple : (a) SLF; (b) Transactionnel;(c) Treillis de Galois sous Galicia

5.2.2 Expérimentations

Pour nos expérimentations, nous avons réalisé les tests avec la configuration suivante :

Machine	Composant
PC1	2 Agents
PC2	1 Agent
PC3	Manager + 1 routeur AMI

Tableau 2 Configuration de la plateforme d'exécution

Sur la machine PC1, les deux agents partagent le même processeur physique alors que l'agent sur la machine PC2 bénéficie de la capacité du processeur dans sa totalité. Les tests ont été réalisés sur deux bases de données : dense (mushroom) et éparse (T40I10D100K) ayant les caractéristiques suivantes :

Base de données Benchmark	Type	# attributs	# transactions
Mushroom	Dense	119	8124
T40I10D100K	éparse	870	100000

Tableau 3 Caractéristiques des bases de données de test

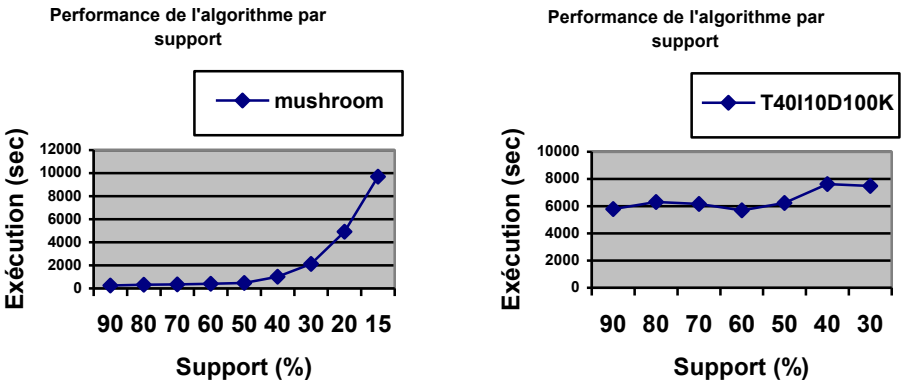


Figure 9 Temps d'exécution par rapport au support

La figure ci-dessous montre la performance de l'algorithme vis-à-vis d'un fichier dense « mushroom » et un fichier éparse «T40I10D100K ». On rappelle qu'il s'agit ici de la génération du treillis de Galois qui englobe la génération exhaustive de tous les concepts ainsi que leur relation hiérarchique. Il est insensé de se comparer aux algorithmes qui ne génèrent que les motifs fermés du moment que ceux-ci ignorent la notion du treillis et même du concept qui est constitué de l'extension et l'intension (le motif fermé). De même on souligne le fait que ces expérimentations ne représentent qu'un cas de figure de l'architecture (dans cet exemple 3 machines et 3 agents) qui est ouverte et dynamique offrant la possibilité d'améliorer en continu la qualité du résultat grâce à sa scalabilité. Initialement, lors de la distribution logique des agents sur une même machine physique, il était difficile de traiter le fichier « mushroom » en mémoire et en générer son treillis. Alors que la distribution des agents sur différentes machines a permis aisément son traitement et même avec un support allant jusqu'à 15% ce qui n'est pas évident pour plusieurs algorithmes n'étant spécialisés que dans la génération des motifs fermés fréquents. Avec cette configuration, le résultat par exemple pour le support de 30% est de 2129s dans le cas de « mushroom » (voir Figure 9) et allant jusqu'à 9688s pour un support de 15%. Il est à remarquer que le temps d'exécution croît exponentiellement au fur et à mesure que le support décroit. Du fait que le choix a été fait sur la mise en mémoire du treillis dans sa globalité (donc pas de sérialisation) pour favoriser une exécution rapide, le facteur mémoire a influencé négativement le déroulement de ce processus et par conséquent, pour le fichier « T40I10D100K » l'exécution a été alourdie par la manipulation de la grande quantité de données. Dans l'optique d'une amélioration de l'algorithme, on a pensé à mieux optimiser l'utilisation de la mémoire en étudiant la possibilité d'une mémoire et donc d'un Trie distribué.

Selon nos constatations et nos mesures ces résultats peuvent être améliorés en multipliant le nombre d'agents et en les exécutant sur des machines physiques séparées.. D'autres techniques peuvent être appliquées pour améliorer la performance d'avantage. Celles-ci seront mentionnées dans la partie conclusion et perspective. Il faut noter que l'objectif principal de

ce travail est dans un premier temps la conception et l'implémentation de l'architecture parallèle distribuée supportant la génération du treillis de Galois et par conséquent les règles d'association. La performance de l'algorithme viendra dans une seconde place surtout qu'il est à considérer, pour se comparer à des algorithmes déjà existants, d'utiliser les techniques améliorant la performance telle que *diffset* de Zaki et ceci nécessite bien évidement leur adaptation à notre architecture.

Par ailleurs, une étude comparative a été menée pour monter l'avantage de la distribution physique des agents sur des machines séparées. Pour cela, le fichier « mushroom » a été traité encore une fois avec des agents tournant en tant que composants logique sur une même machine. Le résultat est présenté dans la Figure 11. Il est clair que la différence est consistante quand on compare par exemple le temps d'exécution pour le support de 30% qui est 15000s contre 2000s dans le cas de la configuration distribuée ci-dessus. Le cas du support 15% et 20%, a fait planter le processus d'exécution en architecture monoposte.

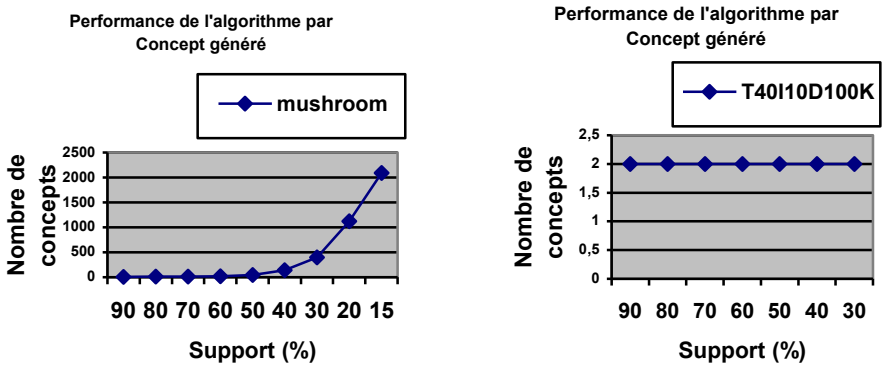


Figure 10 Nombre de concepts générés par rapport au support

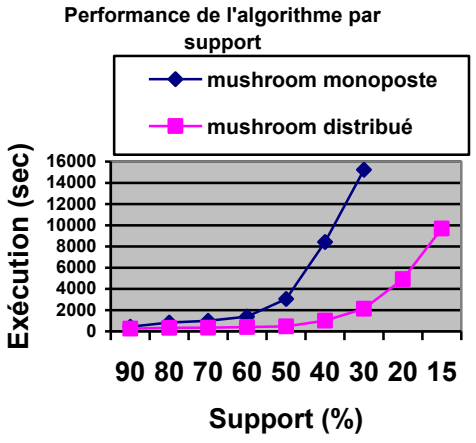


Figure 11 Performance de la distribution physique des agents

6 Règles d'association

Notre algorithme génère le Treillis de Galois et par conséquent, tous les motifs fermés. Le support étant donné (cardinalité de l'extension d'un concept), la génération des motifs fermés fréquents devient évidente. Seulement les règles d'association restent la connaissance la plus précieuse à explorer dans la base de données. Pour se faire, on est amené à exploiter le même treillis déjà généré.

On a examiné les techniques d'extraction des règles d'association basées sur les Treillis de Galois. Comme modèle, on s'est inspiré du framework Mirage de Zaki et Phoophakdee pour les générer vue que c'est algorithme repose sur le treillis de Galois aboutir au résultat. La figure ci-dessous expose l'architecture adoptée pour générer les règles d'association.

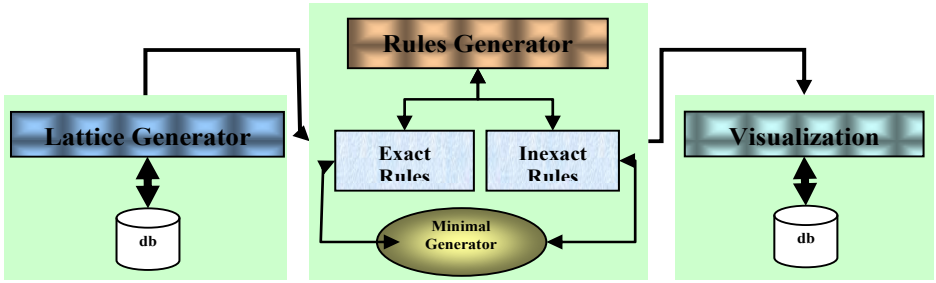


Figure 12 Génération des règles d'association

La génération du Treillis de Galois est en fait la tâche la plus coûteuse, celle-ci est traitée dans les paragraphes précédents en adoptant l'approche parallèle distribuée. On exploite le treillis déjà obtenu pour générer les règles d'association. L'élément clé dans ce processus est celui du générateur minimal d'un motif fermé donné qui est en fait un de ses sous-ensembles mais qui ne doit être contenu dans n'aucun de ses enfants directs dans le Treillis. La génération des règles exactes et inexactes entre deux motifs fermés est basée directement sur ces générateurs minimaux. On travaille actuellement sur la partie visualisation des règles d'association. L'implémentation du générateur est globalement achevée.

7 Conclusions et perspectives

La distribution de l'algorithme de construction de treillis de Galois nous a permis :

- De générer la totalité des concepts et par conséquent tous les motifs fermés fréquents et les règles d'association sur la base d'un treillis de Galois qui est construit dans une phase préliminaire.
- d'offrir un moyen pour dépasser la limite naturelle de l'algorithme séquentiel au moyen, d'une part de la scalabilité qui est une suite logique de la distributivité de l'architecture adoptée et qui donne la possibilité de multiplexer les agents selon le besoin, d'autre part, du parallélisme de l'architecture qui est le levier de performance derrière l'accélération du processus de génération du treillis de Galois.

- Une bonne maîtrise du processus de génération du Treillis et ceci en dissociant ses tâches principales : la génération des concepts enfants (Agents) et la gestion de l'arbre des concepts (Manager). La première est gourmande en capacité du processeur et la deuxième en mémoire. Ceci nous a permis de tester et d'optimiser chacun de ces processus séparément et d'atteindre des résultats encourageants qui n'étaient pas possible avec l'algorithme séquentiel. Par ailleurs, le parallélisme nous a permis d'améliorer la performance et la scalabilité de l'algorithme en multiplexant les agents selon le besoin. Cependant comme cette approche construit le treillis en mémoire (Trie global) et effectue des calculs intenses (inclusions et intersections) pour tester la fermeture, deux points sont candidats d'optimisation si on veut la préconiser pour le datamining notamment : mémoire au niveau du manager et contrôle de fermeture au niveau des agents. On propose ci-dessous des alternatives pour surmonter ces difficultés.

Perspectives :

- 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.
- Cette architecture peut s'étendre à une utilisation web similairement aux réseaux « peer to peer » et « cloud computing » qu'on pourrait qualifier de « peer to peer Data mining ».

Références

Barbut M. et Montjardet B. (1970), Ordre et Classification : Algèbre et Combinatoire. Hachette.

Ben Yahia S. et Nguifo E.M. (2004), Approches d'extraction de règles d'association basées sur la correspondance de Galois, RSTI-ISI, pages 23-55

Berry A. et Sigayret A. (2004), Discrete Applied Mathematics, volume 144, Issue 1-2, Discrete Mathematics & Data mining (DM & DM), pages 27-42.

Bordat J. P. (1986), Calcul pratique du treillis de Galois d'une correspondance, Math. Sci. Hum. 96 31-47.

Bouchahda A., Ben Yahia S. et Slimani Y., Une approche pour l'extraction des itemsets (fermés) fréquents, Université de Tunis El Manar

Chein M. (1969), Algorithme de recherche de sous-matrice première d'une matrice, Bull. Math. R. S. Roumanie 13.

Choi V. (2006), Faster Algorithms for Constructing a Concept (Galois) Lattice, Department of Computer Science, Virginia Tech, USA.

Ganter B. et Reuter K. (1991), Finding all closed sets : a general approach. Order, 8:283-290.

Une approche parallèle optimisée de distribution intégrale de la fouille de données basée sur une infrastructure CORBA

Ganter B. et Wille R. (1999), Formal Concept Analysis : Mathematical Foundations. Springer Verlag.

Han, J. & Kamber, M. (2006): Data mining: Concepts and techniques. Morgan Kaufmann

Idri A.F., Boulmakoul A. et Marghoubi R (2008), Une approche parallèle pour la construction des Treillis de Galois, SITA 2008.

Kantardzic M. (2003): Data Mining: Concepts, Models, Methods, and Algorithms, Wiley J. & Sons.

Lakhal L. et Stumme G., Efficient Mining of Association Rules Based on Formal Concept Analysis

Levy G. et Baklouti F., A distributed version of the Ganter algorithm for general Galois Lattices

Njiwoua P. et Nguifo E. M., A Parallel Algorithm to build Concept Lattice, In proceedings of 4 Groningen Intl. Information Tech. Conf. for Students, pp. 103-107, 1997.

Pasquier N., Bastide Y., Taouil R. et Lakhal L (1999). Efficient mining of association rules using closed itemset lattices. Information systems. 24(1), p25-46.

Pasquier N., Bastide Y., Taouil R. et Lakhal L (1999), Closed set based discovery of small covers for association rules. In Actes des 15èmes journées Bases de Données Avancées (BDA'99), pages 361- 381.

Stumme G. (1999), Conceptual knowledge discovery with frequent concept lattices. FB4-Preprint 2043, TU Darmstadt.

www.orbacus.com

www.iro.umontreal.ca/~galicia/publication.html

Ye N., (2003): The Handbook of Data Mining, Arizona State University, IEA.

Zaki M. et Hsiao Ching-Jui, CHARM : An Efficient Algorithm for Closed Itemset Mining

Zaki M. J. et Ogihara M. (1998), Theoretical foundations of association rules. Proc. 3rd ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, p1-7.

Zaki M. et Phoophakdee B. MIRAGE: A Framework for Mining, Exploring and Visualizing Minimal Association Rules Rensselaer Polytechnic Institute.

Summary

The usefulness of the concept Lattice is proven in Data mining since the two fields are extremely dependent of each other in terms of closed item sets. The generation process of concept lattices have often an exponential time and space complexity, especially when dealing with very large databases in the domain of Data Mining. 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, an integral parallel distributed approach to improve the performance of the sequential

algorithm is proposed. In addition to the distribution of the data mining process, also the memory distribution is taken in account. The algorithm and the system architecture are exposed as well as the aspects of its implementation. FCA (Concept Lattice) will be also applied for the generation of the association rules. This process is based on concept of minimal generators as proposed by Zaki.

Nouvelle approche coopérative de classification dynamique de données évolutives et multimodales

Abdelouahab Moussaoui*
Mohamed Amir Abbas**

*Département d'informatique, Université Ferhat Abbas de Sétif, ALGERIE
moussaoui.abdel@gmail.com
**Département d'informatique, Université Saad Dahleb de Blida, ALGERIE
m.amir.abbas@gmail.com

Résumé. Dans ce papier nous présentons une nouvelle approche pour la classification dynamique de données non-stationnaires et multimodales, basée sur une architecture multi-agents. Cette approche est adaptative à l'évolution des classes et de données, dans le but d'optimiser l'affectation des entités provenant de sources hétérogènes aux classes, et à renforcer le mécanisme de la classification incrémentale ainsi la détermination des cas de création de nouvelles classes. L'approche proposée va permettre aux agents classificateurs de collaborés dans la prise de décisions finales. Elle a été implémentée sur la plate forme JADE, où chaque démarche sera assurée par des agents spécialisés collaborant et communicant entre eux.

Mots-clés. Système Multi-Agents, Classification dynamique, Coopération, Interaction, Communication.

1 Introduction

Pour La classification dynamique est une discipline difficile à mettre en œuvre avec une maximisation des performances, où le système doit se doté d'une capacité d'auto-adaptation capable de faire face aux changements brusques de l'environnement d'exécution (évolution des classes), ainsi la non-stationnarité de données manipulées. Nous avons orienté notre solution vers un système collaboratif fondé sur les systèmes multi-agents, où plusieurs agents classificateurs coopèrent entre eux.

A travers ce travail nous souhaitons adapter les méthodes de classifications dynamiques aux données évolutives et multimodales.

La complexité de la méthode proposée ne se situe pas au niveau du modèle de la classification lui-même, mais plutôt dans le suivi et la coordination des tâches entre les agents classificateurs, ainsi le suivi du nombre et dimension des classes.

Cet article est organisé comme suit : la deuxième section présente les principales méthodes de classification dynamique avec un rappel sur les principaux concepts des systèmes multi-agents. Dans la section trois, nous développons l'architecture globale de notre approche avec une définition des entités Agents et leurs niveaux d'interaction. La quatrième

section, quant à elle, décrit brièvement l'implémentation avec les données (Météo-Insolation) et discute les performances de la solution proposée, suivie par une conclusion et perspectives.

2 Etat de l'art

Concevoir un algorithme de classification dynamique est un travail assez délicat vu la difficulté de manipuler des données non-stationnaires, avec des classes évolutives.

2.1 Méthodes de classification dynamique

L'objectif est d'associer chaque individu non-stationnaire X={X1,..., Xn}, à l'une des k classes évolutives définie dans C={C1,...,Ck}, selon un critère de ressemblance. La non stationnarité de données implique l'évolution des classes (apparition, scission et élimination). Nous avons deux types de méthodes : supervisée et non-supervisée (A-B.Habiboulaye, 2006).

2.1.1 Méthodes supervisées

Chaque méthode commence par un apprentissage sur des échantillons représentatifs pour définir les règles de classification et les propriétés de classes, suivi par l'affectation de l'objet à sa classe en fonction des règles découvertes (A.BLANSCHE, 2006). Parmi ces méthodes nous citons :

- **Méthodes de Classification Bayésiennes** (Réseau naïf de Bayes, champs de Markov, k plus proche voisins) : L'objectif est de calculer pour l'individu X toutes les probabilités d'appartenance conditionnelles aux classes afin de choisir sa classe la plus probable selon la fonction φ tq:

$$\varphi(X) = y_j \ / \forall y \in C, P(y_j/X) \geq P(y/X) \qquad 1$$

$P(y_j/X)$ est donnée par la loi de Bayes (V.GUNES, 2001) suivante :

$$P(y_j/X) = \frac{P(X/y_j) \bullet P(y_j)}{P(X)} \qquad 2$$

- **Arbres de Décision**: C'est une représentation graphique d'une procédure de classification et d'aide à la décision très efficace. Parmi les algorithmes les plus utilisés: ID3, C4.5 et enfin le modèle CART.
- **Réseaux de Neurones**: C'est un réseau d'unités élémentaires (nœuds) interconnectées entre eux, regroupées en plusieurs groupes (couches). La classification consiste à soumettre un individu à classer dans la couche d'entrée du réseau, afin de récupérer sa classe d'appartenance la plus probable à partir de la couche de sortie.

2.1.2 Méthodes non supervisées

Nous avons seulement des individus ou des objets non étiquetés, nous ne connaissons pas leurs classes d'appartenances a priori. Le but est de les classer selon un critère de similarité, en différents groupes nommés classes (clusters) (A.BLANSCHE, 2006). Parmi ces méthodes [3]:

- **C-moyennes** ("Hard C-Means" HCM) : L'objectif est de définir un partitionnement de l'ensemble de données $X=(X_1,...,X_N)$ en C classes. X_j est représentée par un vecteur d'attributs $\{x_{j1}, x_{j2}, ..., x_{jn}\}$. Une classe Wi est caractérisée par un centre P_i, qui va permettre de calculer le degré d'appartenance u_{ij} de l'objet X_j à cette classe. En finalité, chaque donnée X_j sera attribuée à une et une seule classe W_i parmi les C classes proposées. Pour cela, à partir des centres de classes, l'algorithme va minimiser la fonction nommée WGSS (Within Group Sum of Squared errors) (E.ANQUETIL, 1997). suivante :

$$J(P,U,X) = \sum_{i=1}^{C}\sum_{j=1}^{N} u_{ij}\, d^2(X_j, P_i) \quad (3)$$

  Où, $P = (P_1, ..., P_C)$ est l'ensemble des centres de classes, avec :

$$P_i = \frac{\sum_{j=1}^{N} u_{ij} X_j}{\sum_{j=1}^{N} u_{ij}} \quad (4)$$

  $d^2(X_j, P_i)=\sum_{k=1}^{n}(x_{jk} - P_{ik})^2$: distance euclidienne entre la donnée X_j et le centre de la classe W_i.
  $U = [u_{ij}]$: la partition recherchée de l'ensemble X, c'est la matrice des degrés d'appartenance du modèle de classification. Nous avons :

$$u_{ij} = \begin{cases} 1 \text{ si } d^2(X_j, P_i) < d^2(X_j, P_k) \forall k \neq i \\ 0 \text{ sinon} \end{cases} \quad (5)$$

- **C-moyennes floues** ("Fuzzy C-Means" FCM) : Permet d'associer une donnée à une ou plusieurs classes en même temps. L'appartenance d'une donnée à une classe est définie par un degré d'appartenance qui représente l'élément de base d'une partition U. L'idée de FCM est de construire une matrice de C-partition flou U : $C \times N$, ayant des degrés d'appartenance $u_{ij} \in [0,1]$ des données X_j aux classes Wi. L'objectif de FCM et défini par la minimisation de la somme pondérée des carrés des distances entre les données à regrouper et les centres de classes, décrite comme suit :

$$J(P,U,X) = \sum_{i=1}^{C}\sum_{j=1}^{N} (u_{ij})^m d^2(X_j, P_i) \quad (6)$$

  Où $m \in [1, \infty[$: indice de flou « fuzzy index» qui détermine le degré de flou de la partition obtenu.

- **C-moyennes possibilistes** ("Possibilistic C-means" PCM) : C'est une nouvelle modélisation possibiliste de l'algorithme HCM basée sur la théorie de possibilité. Il est

basé sur la théorie de possibilité pour définir la partition des données X sur l'ensemble des classes W, représentées par leurs centres P. En effet, les degrés d'appartenances u utilisés en FCM sont traduits en PCM autant que degrés de vérité relatifs, décrivant l'appartenance d'une donnée à chacune des classes possibles. La nouvelle formulation de la fonction objective à minimiser est décrite par:

$$J(P,U,X) = \sum_{i=1}^{C}\sum_{j=1}^{N}(u_{ij})^{m}d^{2}(X_j,P_i) + \sum_{i=1}^{C}\eta_i\sum_{j=1}^{N}(1-u_{ij})^{m} \quad (7)$$

Où η_i : est le carré de la distance entre P_i (centre de la classe W_i) et l'ensemble de données X_j ayant leur degrés d'appartenance $u_{ij} = 0.5$.

2.2 Systèmes Multi-Agents

Un SMA est un ensemble d'agents intelligents et autonomes capables de communiquer, collaborer et d'agir dans un environnement commun afin d'effectuer plusieurs tâches communes ou individuelles, dans un but de résoudre un problème complexe (J.Ferber,1995).

L'application des systèmes multi-agents a connue une grande expansion, touchant une large sélection de domaines techniques liés à l'Intelligence artificielle, système distribués et le génie logiciel. Dans le contexte d'une classification dynamique, les systèmes multi-agents sont très efficaces pour appliquer plusieurs méthodes en même temps d'une façon autonome, avec une possibilité d'adaptation aux évolutions des classes et de données. L'utilisation d'une approche coopérative pour classifier des données non stationnaires et multimodales dans des classes évolutives nécessite un SMA (A.SAIDANE et al, 2005).

Notre conception été basée sur la méthode GAIA (F.ZAMBONELLI et al, 2003). Nous avons commencé par définir les entités d'agents avec leurs types (rôles, responsabilités, connaissances, protocoles), ensuite nous avons modélisé l'interaction entre ces agents en utilisant les diagrammes d'organisation.

3 Architecture de l'approche proposée

Notre méthode est le résultat d'une fusion entre cinq phases de traitement de données. Elle est la conjonction de deux processus initiales : la collecte, la modélisation et l'échantillonnage de données, suivi par les trois processus celui de la classification, l'interprétation et la fusion des interprétations et enfin la présentation de la décision finale.

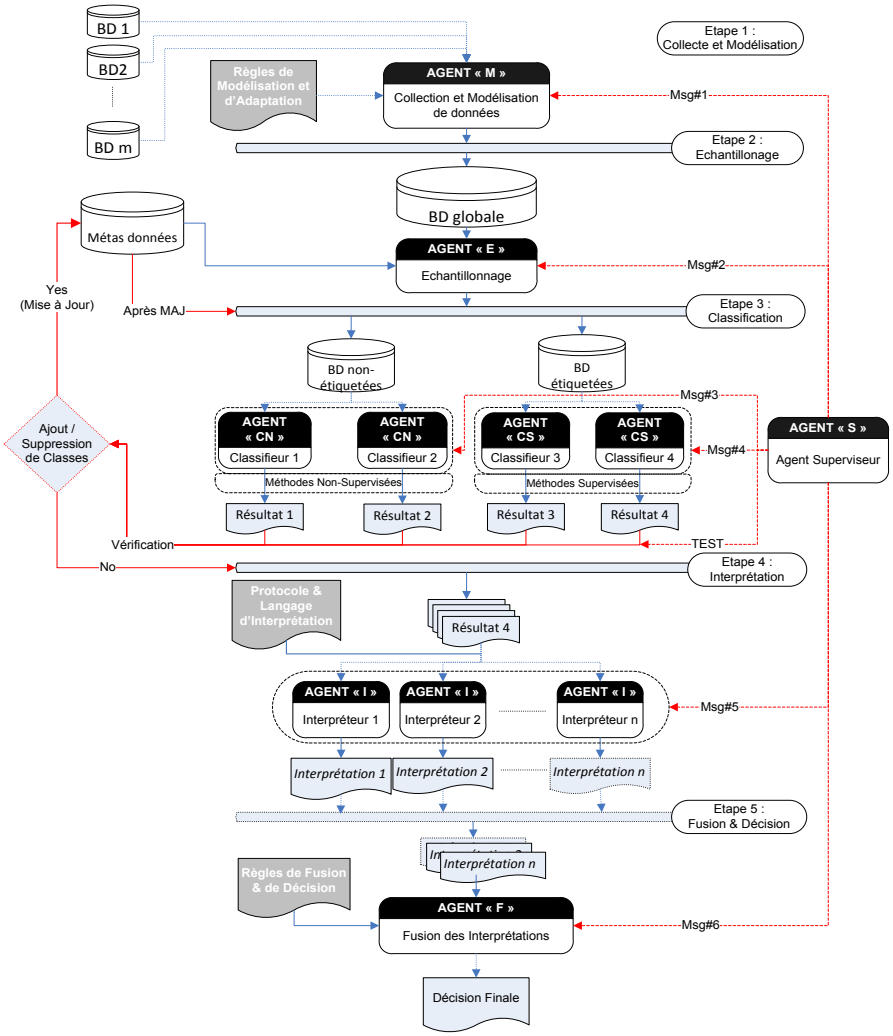


FIG. 1 – *Architecture d'un SMA pour la classification dynamique de données évolutives et multimodales*

La Fig.1 présente l'architecture de notre approche avec les différents agents hybrides (M.OCCELLO, 2003) suivant :

AgS : l'agent Superviseur, joue le rôle du leader et coordinateur entre les agents,
AgM : l'agent de modélisation, prépare les données au processus de classification,
AgE : l'agent d'échantillonnage, divise la base de données globale en deux sous bases.
AgCS & **AgCN** : les agents de classification supervisée et non-supervisée.
AgI : les agents d'interprétation, valorisent les résultats de classification,
AgF : l'agent de fusion, fourni le résultat ou la décision finale à l'utilisateur.

3.1 Algorithme de la phase de Collecte et Modélisation de données

1: Détection des nouvelles données par AgS ;
2: Envoi msg#1 de réveil par AgS à AgM ;
3: Accès aux sources de données par AgM;
4: Adaptation et agrégation des nouvelles données ;
5: Regroupement des nouvelles données modélisées dans une BD temporaire;
6: Jointure Externe entre l'ancienne BD globale et la nouvelle BD → *Insérer* les nouvelles données seulement & *Mettre* à jour les données existantes évoluées ;
7: Envoi msg d'achèvement de AgM à AgS → AgM s'endort.

3.2 Algorithme de la phase d'Echantillonnage

1: Détection d'une mise à jour de la BD globale par AgS;
2: Envoi msg#2 de réveil de AgS à AgE;
3: Répartition de la nouvelle BD globale sur deux sous bases, étiquetées et non-étiquetées par AgE;
4: Envoi msg d'achèvement par AgE à AgS → AgE s'endort.

3.3 Algorithme de la phase de Classification

1: Possibilité d'une mise à jour de la base des classes ;
2: Possibilité d'une mise à jour de la BD étiquetée;
2': Possibilité d'une mise à jour de la BD non-étiquetée;
3: Si (2 ou 1) est vrais → Envoi msg#4 de réveil de AgS aux AgCS ;
3': Si (2'ou 1) est vrais → Envoi msg#3 de réveil de AgS aux AgCN ;
4: Classification des nouvelles données étiquetées par les AgCS;
4': Classification des nouvelles données non-étiquetées par les AgCN;
5: Envoi msg d'achèvement par les agents classificateurs à AgS;
6: Si (Mise à Jour de la base des classes) est vrais → Réitération d'une nouvelle phase de classification (*aller à 1*);
7: Fin de classification → les agents classificateurs s'endorment.

3.4 Algorithme de la phase d'Interprétation

1: Détection de nouveaux résultats de classification par AgS;
2: Envoi msg#5 de réveil de AgS aux agents AgI;
3: Chaque AgI délibère une interprétation selon ses connaissances et sa spécialité ;
4: Envoi msg d'achèvement par les AgI à AgS → AgI s'endorment.

3.5 Algorithme de la phase de Fusion et de décision

1: Détection de nouvelles interprétations par AgS;
2: Envoi msg#6 de réveil de AgS à AgF ;
3: Application des règles de fusion et de décision sur les nouvelles interprétations par AgF → Délibérer une décision final à l'utilisateur du système ;
4: Envoi msg d'achèvement par AgF à AgS → AgF s'endort.

4 Expérimentations

Nous avons appliqué notre système de classification sur une base de données réel contenant des données évolutives représentant les durées d'insolation journalières sur plusieurs sites nationales. La base de données mise à notre disposition nous provient du Laboratoire de Biologie de l'Université des sciences et technologies HOUARI BOUMEDIENNE (Alger, Algérie), elle est constituée de 52 stations météorologiques du réseau ONM (l'Office National de la Météorologie), couvrant pratiquement le territoire national. Chaque station contient l'enregistrement journalier du paramètre météorologique et climatique « INSOLATION » (en dixième heure, sur une période d'observation de 11 ans [1992-2002]).

Pour notre implémentation, nous avons utilisé la plate-forme Jade (Java Agent DEvelopment Framework) développée en Java conformément aux spécifications FIPA au sein du laboratoire TILAB.

Le but du processus de classification est d'affecter chaque station météorologique vers l'une des classes représentant une zone climatique selon le paramètre de classification ***: fraction d'insolation*** (SS/SS0). Il est définit comme étant la fraction de la durée d'insolation mesurée brute SS (durées d'insolation journalière enregistrés par la station) sur la durée d'insolation théorique SS0. Cette dernière définie la durée du jour astronomique, elle représente la durée comprise entre le lever et le coucher du soleil pour un lieu donné. Elle évolue naturellement au cours de l'année avec le cycle des saisons, elle dépend uniquement de la latitude du lieu Lat et du numéro du jour dans l'année j. Elle est donnée par la formule suivante :

$$SS0(j) = (2/15)\ Arc\ (cos)\ [-Tan\ (\ Lat\ )\ Tan\ [23{,}45\ Sin\ (\ 0{,}98\ (\ j+284\ )\ )]\ ] \qquad (8)$$

La fraction d'insolation élimine l'effet déterministe de l'insolation brute (*variation selon la latitude et le temps*), de même elle permet de soustraire l'effet de la tendance saisonnière. Prenant comme exemple la station de Constantine [693 m; 36°17 N; 06°37 E], où nous pouvons remarquer l'aspect dynamique de l'insolation qui réside dans ses changements brusques.

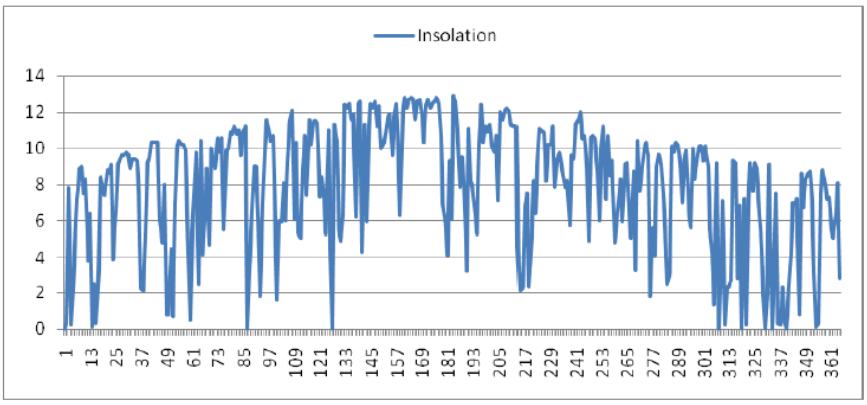


FIG. 2 – *Evolution de l'insolation journalière brute au cours de l'année 2002 dans la station de Constantine.*

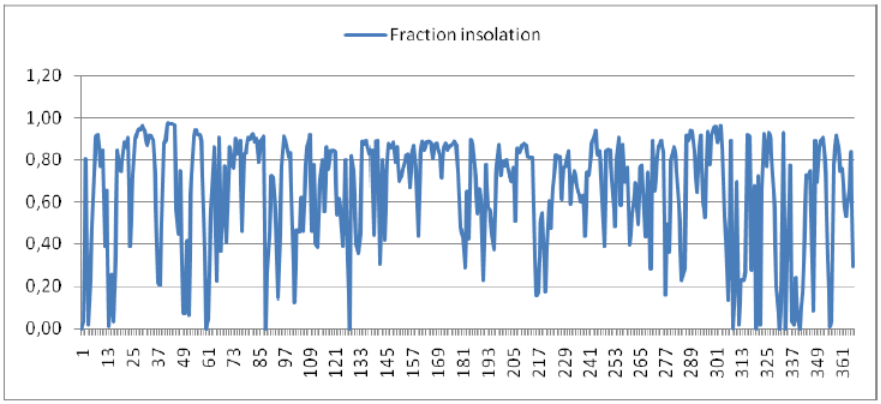


FIG. 3 – *Evolution de la fraction d'insolation journalière au cours de l'année 2002 dans la station de Constantine.*

La bonne répartition des tâches et l'échange des messages de communications entre les agents nous a permis de suivre le bon déroulement du processus décisionnel.

IdClasse	Désignation	MIN(Fraction d'Insolation)	MAX(Fraction d'Insolation)
C1	Classe n°1	0,62	0,65
C2	Classe n°2	0,65	0,69
C3	Classe n°3	0,69	0,73
C4	Classe n°4	0,73	0,77
C5	Classe n°5	0,77	0,81
C6	Classe n°6	0,81	0,84

TAB. 1 – *Classes des zones climatiques résultantes après la classification*.

Pour une meilleure discussion des résultats d'application de notre approche, nous avons recouru à des agents d'évaluation spécialisés, où chaque agent est dédié à utiliser l'une des métriques d'évaluation suivantes : *Temps, Précision, Charge de Communication*.

4.1 Evaluation suivant la métrique Temps

Le calcul des taux *Tagent* , représentant la fraction du temps d'exécution d'un agent par rapport au temps d'exécution global du système *Ttotal* , nous a permis de déterminer les agents qui ont une charge de travail plus importante. Or cette charge n'a pas une grande influence sur la qualité du service offerte par le système, suite à l'exécution des cycles des agents en parallèle.

Agent	Durée Max (*Tagent* / *Ttotal*)	Durée Min (*Tagent* / *Ttotal*)
Agent_M	85,47 %	10,21 %
Agent_E	70,45 %	10,21 %
Agent_CS	85,47 %	10,21 %
Agent_CN	73,01 %	10,21 %
Agent_I	59,51 %	9,21 %
Agent_F	54,24 %	8,45 %

TAB. 2 – *Les durées d'exécution par agent*.

Certaines *Tagent* se rapprochent au temps global de l'exécution du système *Ttotal*. Cela explique bien le gain de temps que nous avons obtenu suite à l'utilisation d'un système distribué pour l'analyse de données, car dans un système non distribué nous devons avoir : $Ttotal = \sum Tagent$.

4.2 Evaluation suivant la métrique Précision

L'objectif est de vérifié l'adéquation de notre système de classification avec le domaine d'application. A cet effet, nous avons associé à chaque classe résultante de notre classification sa métrique de précision relative. La précision représente la probabilité qu'un objet prédit de la classe C_i soit effectivement de la classe C_i. Soit :

- **Pc** est un objet correctement classifié dans la classe C_i ;
- **Pf** est un objet mal classifié dans la classe C_i ;

La métrique de précision Pr est calculée pour chaque classe selon la formule suivante :

$$Pr = Pc / (Pc + Pf) \qquad (9)$$

Classe	Taux de Précision
Classe n°1	75,00 %
Classe n°2	50,00 %
Classe n°3	37,50 %
Classe n°5	33,33 %
Classe n°	66,67 %

TAB. 3 – *Taux des précisions estimées sur l'ensemble des classes.*

Avec une moyenne du taux de précision dépassant les 50%, notre méthode de classification nécessitent plus d'amélioration pour augmenter cette métrique. L'utilisation d'autres résultats de classification comme référence dans l'analyse de la métrique de précision seraient plus favorable pour l'évaluation de notre système, ainsi l'implémentation des méthodes de classification adéquates au domaine d'application augmente sa fiabilité.

5 Conclusion

A travers ce travail nous avons voulu introduire une nouvelle architecture collaborative fondée sur une solution multi-agents qui consiste à utiliser plusieurs méthodes supervisée et non-supervisée en même temps pour l'amélioration des résultats de classification de données multimodales et évolutives. Notre approche peut être déployé et utiliser sur plusieurs domaines d'application en classification dynamique, elle est conceptualisée d'une façon adaptatif.

Le travail à venir consiste en une amélioration des processus d'interprétation sur les résultats de classification.

Les références sont données en fin d'article avant le « Summary ». Elles doivent être listées par ordre alphabétique. Elles sont en Times New Roman 10 Points. Utiliser le style « RNTI référence à un article ». Merci de suivre les exemples donnés à la fin de cet article.

Dans le corps du texte, on utilise Sauwens (2000), Hölldobler et Wilson (1990) pour faire référence à un article avec un ou deux auteurs, et Lioni et al. (2001) lorsque trois auteurs ou plus sont présents. Selon les cas, on utilisera (Sauwens, 2000) ou (voir Breiman et al., 1984, chapitre 4), ou encore pour des citations multiples d'un même auteur, Quinlan (1986, 1993).

Références

A-B.Habiboulaye. (2006).*Classification Dynamique De Données Non-Stationnaires Apprentissage Et Suivi De Classes Evolutives*. Doctoral Thesis, Sciences and Technologies Lille University, France.

A.BLANSCHE. (2006). *Classification non supervisée avec pondération d'attributs par des méthodes évolutionnaires*. Doctoral Thesis, Louis Pasteur – Strasbourg I University, France.

A.MOUSSAOUI, M.SEMCHEDINE, L.TOUMI. (20-21 Novembre 2007).*Un Système Multi-agents pour La Classification Coopérative D'images IRM Cérébrales*. The first national seminar on natural language and artificial intelligence (LANIA'2007), Chlef, Algérie.

A.SAIDANE, H.AKDAG, I.TRUCK. (March, 2005).*Une Approche SMA de l'Agrégation et de la Coopération des Classifieurs*. 3rd International Conference: Sciences of Electronic, Technologies of Information and Telecommunications (SETIT 2005), Tunisia.

E.ANQUETIL. (1997). *Modélisation et reconnaissance par la logique floue : application à la lecture automatique en-ligne de l'écriture manuscrite omni-scripteur*. Doctoral Thesis, RENNES I University, France.

F.ZAMBONELLI, N-R.JENNINGSY and M.WOOLDRIDGE. (Juillet 2003).*Developing Multiagent Systems: The Gaia Methodology*. Journal ACM Transactions on Software Engineering and Methodology (TOSEM) Volume 12 Issue 3.

J.Ferber.(1995). *Les systèmes multi-agents : vers une intelligence collective*. InterEditions.

M.OCCELLO.(Juillet 2003). *Méthodologie et architectures pour la conception de systèmes multi-agents*. Doctoral Thesis, Joseph Fourier University, Grenoble, France.

V.GUNES. (2001). *Reconnaissance des formes évolutives par combinaison, coopération et sélection de classifieurs*. Doctoral Thesis, Rochelle University, France.

La prédiction d'ordre pour le filtrage collaboratif

Abderrahmane Kouadria *,
Omar Nouali, **
* Département d'Informatique
Université Ibn Khaldoun
Tiaret – Algérie
ab_kouadria@esi.dz
** Division théorie & ingénierie des systèmes informatiques, CE.R.I.S.T.,
Rue des 3 frères Aïssiou, Ben Aknoun, Alger, Algérie
**onouali@cerist.dz

Résumé. L'objectif des systèmes de recommandation est d'estimer la préférence d'un utilisateur et d'offrir une liste d'articles qui pourraient être privilégiées par un utilisateur donné. Le Filtrage Collaboratif (FC) représente l'une des techniques de recommandation les plus populaires, dont la plupart de leur méthodes fondent leur approche sur la prédiction de notes pour générer les recommandations. Dans cet article, nous nous basons sur l'approche de prédiction d'ordre, consistant à ordonner correctement les articles selon les goûts des utilisateurs, à cet effet nous proposons une adaptation d'une méthode d'ordonnancement de recherche d'information ListMLE pour le filtrage collaboratif qui consiste a combine cette dernière avec la méthode de factorisation matricielle FM.

1 Introduction

Les systèmes de recommandations ont été introduits comme une technique informatique intelligente pour traiter le problème de la surcharge de l'information. Ils peuvent être utilisés pour fournir efficacement des services personnalisés dans plusieurs domaines tels que la recherche documentaire, le commerce électronique, les loisirs, etc. Dans le e-business, par exemple, on fait bénéficier à la fois le client par des suggestions sur les produits ou les articles les plus susceptibles de l'intéresser et le commerçant par l'augmentation des ventes (Emmanouil et al 2003).

Une fonction très importante de la plupart des systèmes de recommandation est la génération de la liste des N meilleurs articles pour chaque utilisateur, afin de faire des recommandations personnalisées, qui consiste essentiellement à résoudre un problème d'ordonnancement. Pour ordonner les articles, la plupart des algorithmes de filtrage collaboratif formule cela comme un problème de prédiction de notes qui permet de prédire tout d'abord les notes potentielles d'un utilisateur sur les articles, puis ordonner les articles selon les évaluations prévues (Liu et al 2008, 2009). Cependant, une plus grande précision dans la prédiction de notes ne conduit pas nécessairement à l'efficacité meilleure d'ordonnancement comme l'illustre l'exemple simple suivant. Soient [2, 3] les notes de deux

articles A et B, r_1 = [2.5, 3.6] et r_2 = [2.5, 2.4] deux vecteurs de prédictions obtenus par deux méthodes différentes. Bien que r_1 et r_2 soient équivalents en terme d'erreur carrée (les deux sont égales à 0.52 + 0.62), seule r_1 prédit l'ordre correctement, puisque le score qu'elle attribue à B est supérieur à celui de A. par contre le r_2 n'assure pas le bon ordonnancement (Jean-François 2007).

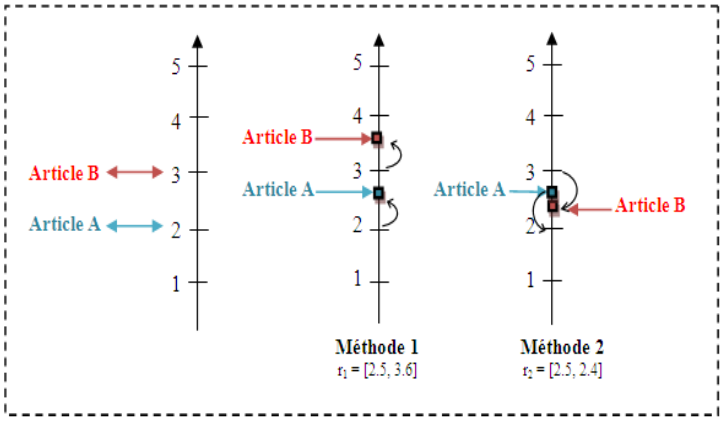


Figure 1 : Exemple d'ordonnancement

Ces dernières années, et comme de plus en plus les bases standards de filtrage collaboratif avec les jugements de pertinence sont disponibles, les méthodes à base d'apprentissage d'ordonnancement supervisé (Learn To Rank methods) ont donné lieu à différentes études et développements (Amini et al 2005, Clémençon et al 2005, D.Cossock et al 2008), par exemple l'utilisation d'une fonction d'ordonnancement efficace à partir des données d'apprentissage pour l'apprentissage automatique (Burges et al 2005, Cao et al 2007, Crammer et al 2002, Crammer et al 2002, Freund et al 2003, Herbrich et al 2000).

L'apprentissage d'une fonction d'ordonnancement peut être vu comme l'apprentissage d'une fonction score : une fonction à valeurs réelles, qui prend en entrée un élément d'un ensemble à ordonner. L'ordre est ensuite prédit en triant les éléments selon les scores croissants ou décroissants (Vinh 2009).

Dans cet article, nous nous intéressons à l'application des méthodes d'apprentissage d'ordonnancement pour le filtrage collaboratif. L'objectif est donc de proposer une adaptation d'une méthode d'ordonnancement de recherche d'information RI pour la tâche de prédiction d'ordre pour le filtrage collaboratif, qui consiste à ordonner correctement les articles plutôt que de prédire correctement leurs notes. Cette approche combine la méthode de factorisation matricielle (MF) avec une méthode d'ordonnancement de l'approche listewise. Une liste ordonnée d'articles est obtenue en réduisant au minimum une fonction d'erreur qui représente l'incertitude entre la liste d'apprentissage en entrée et la liste résultante en sortie selon la méthode proposée.

Dans les prochaines sections, nous résumons l'état de l'art et les méthodes clés liés à notre proposition, à savoir : factorisation matricielle (FM) et méthode d'ordonnancement List-MLE. Ensuite, nous présentons notre proposition ainsi que la validation expérimentale, et enfin, nous terminons par une conclusion et une présentation des perspectives de ce travail.

2 Etat de l'art

2.1 Filtrage collaboratif

Actuellement, le filtrage collaboratif (FC) a été l'une des technologies les plus efficaces pour la recommandation personnalisée (Sarwar et al 2000, Matthew et al 2004). Leur principe est de suggérer de nouveaux articles ou de prédire l'utilité des articles inconnus pour un utilisateur donné, en se basant sur les évaluations déjà exprimées par cet utilisateur à propos d'autres articles.

Selon Breese et al (1998), les algorithmes de filtrage collaboratif peuvent être regroupés en deux catégories : Les algorithmes basées mémoire, et les algorithmes basées modèle.

Les algorithmes basées mémoire consistent à utiliser la totalité des informations disponibles pour générer les prédictions, les exemples les plus analysés dans ses algorithmes comprennent des approches basé-utilisateur (Herlocker et al 1999, Ding et al 2008) et basées articles (Deshpande et al 2004, Sarwar et al 2001).

Les algorithmes basés modèle utilisent la base de données des évaluations des utilisateurs pour créer ou apprendre un modèle de prédiction via un processus d'apprentissage, telles que : le clustering, les réseaux bayésiens, les arbres de décision, le modèle Latent semantic (Hofmann 2004) et factorisation matricielle (Koren et al 2009).

2.2 L'apprentissage d'une fonction d'ordonnancement

L'apprentissage d'une fonction d'ordonnancement (Learn To Rank) est une approche d'apprentissage automatique, qui construit automatiquement une fonction d'ordonnancement à partir de données d'apprentissage.

Beaucoup de méthodes d'apprentissage d'une fonction d'ordonnancement ont été proposées dans la littérature, avec des motivations et des formulations différentes. Ces méthodes se divisent en général en trois catégories (Cao et al 2007) : L'approche pointwise, comme la régression (D. Cossock et al 2008) et McRank (P. Li et al 2008), qui prend un seul objet comme instance d'apprentissage.

L'approche pairwise, telles que RankingSVM (Herbrich et al 2000), RankBoost (Freund et al 2003), et RankNet (Burgeset al 2005), qui prend une paire d'objets comme instance d'apprentissage. L'approche listwise, tels que ListNet (Cao et al 2007) et ListMLE (F. Xia et al 2008) prend une liste ordonnée d'objets comme instance d'apprentissage. Toutes ces méthodes apprennent en général leurs fonctions d'ordonnancement en minimisant certaines fonctions d'erreurs pour ces trois approches.

Actuellement et afin d'améliorer la qualité de la recommandation, l'attention de la recherche dans le domaine du filtrage collaboratif a déplacée le problème de prédiction de note en un problème de prédiction d'ordre. Il existe des méthodes qui utilisent la préférence par paires (pair-wise) des articles des utilisateurs. Parmi elles l'EigenRank (Liu et al 2008), ainsi que probabilistic latent preference analysis (Liu et al 2009). D'autres méthodes ont été proposées telle que CoFiRank (Weimer 2007) qui optimise directement la mesure d'ordonnancement NDCG (Normalized Discounted Cumulative Gain) et ListRank-MF (Y. Shi et al 2010) qui consiste a combiné l'approche liste-wise et la méthode factorisation matricielle FM.

2.3 Formalisme

Considérons *R* la matrice de notes avec des données manquantes qui contient n utilisateurs et *p* articles, chaque utilisateur ayant noté au moins un article. Un exemple d'apprentissage est un triplet *(u, a, r)* où u est un indice d'utilisateur dans $U = \{1, \ldots, n\}$, *a* est un indice d'article $A = \{1, \ldots, p\}$ et *r* est une note dans $V = \{1, \ldots, v\}$. Nous proposons d'apprendre pour chaque utilisateur une fonction score permet de représenter les préférences de manière simple et intuitive. Une fonction d'utilité $f : U \times A \rightarrow R$ modélise la préférence d'un utilisateur pour un article par un score réel. Si un utilisateur $u \in U$ préfère un article *A* à un article *B*, cette préférence est simplement représentée par l'inégalité $f(u,A) > f(u,B)$. Dans le contexte des systèmes de recommandation, les articles sont présentés à chaque utilisateur par ordre décroissant des scores $f(u,.)$.

2.4 Factorisation Matricielle

La méthode de factorisation matricielle FM (Koren et al 2009) permet d'apprendre la représentation vectorielle des données en utilisant une factorisation matricielle régularisée. Le principe est de trouver deux matrices $U \in R^{nxd}$, $A \in R^{pxd}$ telle que le produit matriciel UA^T est la matrice de scores $(n \times p)$ associée à la fonction d'utilité *f*, où $(UA^T)_{ua} = f(u, a)$. Remarquons que le paramètre de taille *d*, commun aux matrices *U* et *A*, définit la dimension des espaces de représentation des utilisateurs et des articles. Chaque ligne de *A* peut être vue comme une représentation vectorielle des articles. Nous noterons A_j le vecteur représentant le j^{ème} article. De même, chaque ligne de *U* caractérise le classifieur linéaire pour un utilisateur donné. Nous noterons u_i le vecteur poids pour l'^{ième} utilisateur.

2.5 ListMLE

C'est l'une des méthodes de l'approche listwise, qui emploie le modèle de Plackett-Luce paramétré et l'estimateur de maximum de vraisemblance et formalise l'apprentissage d'ordonnancement comme problème de minimisation de la fonction d'erreur (F. Xia et al 2008).

La fonction d'erreur de vraisemblance utilisé dans ListMLE est définit de la manière suivante :

$$R(f(X), Y) = -\log P(Y|X; f)$$

Où $$P(Y|X; f) = \prod_{i=1}^{n} \frac{\exp(f(X_{Y(i)})}{\sum_{k=i}^{n} \exp(f(X_{Y(k)})}$$

Il est à noter la définition réelle de la distribution de probabilité exponentielle paramétrée sur toutes les permutations obtenues par la fonction d'ordonnancement, ainsi que la fonction d'erreur comme étant le négatif log-vraisemblance de la liste obtenu par les scores réels (ordre souhaité). La distribution de probabilité s'avère être un modèle de Plackett-Luce (Marden 1995).

3 SOLUTION PROPOSÉE

Notre objectif est d'obtenir un ordonnancement correct de tous les articles en préservant les préférences de chaque utilisateur et de recommander les k-top articles. Notre méthode proposée consiste à combiner les deux méthodes citées auparavant (FM et ListMLE). Pour cela, nous utilisons une fonction d'erreur basée sur le principe de ces dernières. Cette fonction est définie comme suit:

$$L(U,A) = \sum_{i=1}^{n} \left\{ -\sum_{j=1}^{p} I_{iZ(j)} \log \frac{\exp(U_i A_{Z(j)}^T)}{\sum_{k=j}^{p} I_{iZ(j)} \exp(U_i A_{Z(j)}^T)} \right\} + \frac{\lambda}{2}(\|U\|_F^2 + \|A\|_F^2)$$

$$L(U,A) = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{p} I_{iZ(j)} \left[\log\left(\sum_{k=j}^{p} I_{iZ(j)} \exp(U_i A_{Z(j)}^T) \right) - U_i A_{Z(j)}^T \right] \right\} + \frac{\lambda}{2}(\|U\|_F^2 + \|A\|_F^2)$$

Où :

Z : représente l'ordonnancement réel des articles donné par l'utilisateur u;
Z(j) : représente la position des articles selon la préférence de l'utilisateur ;
$I_{iZ(j)}$: égale 1 si $Y_{iZ(j)} > 0$ sinon 0

Articles / Utilisateurs	A_1	A_2	A_3	A_4	A_5	A_6
u=1	3	?	5	?	2	1
u=2	4	3	2	5	1	2
u=3	?	2	5	?	5	3

u=2	$A_4 > A_1 > A_2 > A_6 > A_3 > A_5$						
Z	J	1	2	3	4	5	6
	Indice de l'article Selon les préférences de l'utilisateur u=2	4	1	2	6	3	5

La fonction L n'est pas convexe en U et A simultanément, en revanche elle l'est en chacune des matrices séparément. Le problème d'optimisation associé s'écrit simplement:

$$(U^*, A^*) = \underset{U,A}{\operatorname{argmin}} L(U, A)$$

Pour minimiser la fonction d'erreur *L*, nous optons pour une approche itérative en 2 étapes, où chaque étape consiste à fixer une des deux matrices et minimiser *L* par rapport à l'autre grâce à la méthode du gradient conjugué.

Nous donnons les formules permettant de calculer les gradients de *L* par rapport à *U* et à *A*, nécessaires à l'optimisation :

$$\frac{\partial L}{\partial U_{iq}} = \sum_{j=1}^{p} I_{iZ(j)} \left[\frac{\left(\sum_{k=j}^{p} I_{iZ(k)} \exp(U_i A_{Z(k)}^T) A_{Z(k)q} \right)}{\sum_{k'=j}^{p} I_{iZ(k')} \exp(U_i A_{Z(k')}^T)} - A_{Z(j)q} \right] + \lambda U_{iq}$$

$$\frac{\partial L(U,A)}{\partial A_{jv}} = \sum_{i=1}^{n} \left\{ \sum_{x=1}^{p} I_{iZ(x)} \left[\frac{\sum_{k=j}^{p} I_{iZ(k)} \exp(U_i A_{Z(k)}^T) \delta_{Z(k)x} U_{iv}}{\sum_{k'=j}^{p} I_{iZ(k')} \exp(U_i A_{Z(k')}^T)} - \delta_{Z(x)j} U_{iv} \right] \right\} + \lambda A_{jv}$$

$$\text{avec} \quad \begin{array}{llll} \delta_{Z(k)x} = 1 & si & Z(k) = x & \qquad \delta_{Z(x)j} = 1 \quad si \quad Z(x) = j \\ \delta_{Z(k)x} = 0 & si & Z(k) \neq x & \qquad \delta_{Z(x)j} = 0 \quad si \quad Z(x) \neq j \end{array}$$

Algorithme

Entrée:
- L'ensemble de préférences de chaque utilisateur

Initialiser:
- Initialiser $U^{(1)}$ et $A^{(1)}$ aléatoirement
- $I \longleftarrow 1$

Repeat
- $U^{(I+1)} \longleftarrow \underset{U^{(I)}}{\arg\min} L(U^{(I)}, A^{(I)})$
- $A^{(I+1)} \longleftarrow \underset{A^{(I)}}{\arg\min} L(U^{(I+1)}, A^{(I)})$
- $I \longleftarrow I+1$

Until convergence de *L (U, A);*

Sortie: U et A

4 Expérimentation et discussion des résultats

Nous utilisons pour l'expérimentation, le jeu de données réelles du système de recommandation de films MovieLens[1]. Ce jeu contient 100 000 évaluations[2] (la note donnée

à un film est sur une échelle de 1 à 5), fournies par 943 utilisateurs sur 1682 films. Trois cas se présentent, tel que pour chaque utilisateur, nous choisissons aléatoirement 10, 20 et 50 films évalués dans le but de l'apprentissage et on utilise le reste de ces films notés dans le profil de l'utilisateur pour le test. Pour chaque cas, nous supprimons les utilisateurs qui ont noté moins de 20, 30 et 60 films respectivement afin de nous assurer que nous pouvons évaluer au moins 10 films notés par utilisateur.

L'exécution de notre algorithme a été refaite une dizaine de fois, et une moyenne a été prise pour chaque cas. Les résultats expérimentaux sont consignés dans le tableau suivant :

	Cas d'apprentissage		
	10	**20**	**50**
CoFiRank-NDCG	**0.6400 ±0.0061**	0.6307 ±0.0062	0.6076 ±0.0077
Proposition	0.6381 ±0.0073	**0.6340 ±0.0055**	**0.6210 ±0.0030**

Nous avons adopté le protocole d'évaluation appelé «généralisation faible» dans l'estimation de CoFiRank (Weimer 2007) et nous avons comparés notre proposition avec celle de CoFiRank-NDCG. Le NDCG (Normalized Discounted Cumulative Gain) (Kalervo et al 2000) est choisi comme métrique d'évaluation, du fait de sa grande sensibilité à la pertinence des k-top films ordonnés. Dans les trois cas de l'expérimentation nous avons pris *k* égal à 10 (NDCG@10), le coefficient de régularisation $\lambda=1$ et d =5 qui représente la taille du paramètre commun aux matrices *U* et *A*. D'après le tableau ci-dessus, nos résultats se rapprochent de ceux de CoFiRank-NDCG où l'on remarque pour les deux premiers cas une intersection entre les résultats dans les intervalles respectifs [0.6453, 0.6339], [0.6369, 0.6245] et un léger dépassement pour le troisième cas.

5 Conclusion

Ces dernières années les systèmes de recommandation ont connus des progrès significatifs et de nombreuses techniques ont été proposées pour améliorer la qualité de recommandation. Parmi ces techniques la prédiction d'ordre qui consiste à ordonner correctement les articles selon les goûts des utilisateurs. Dans cet article, nous avons utilisé une approche qui exploite l'apprentissage d'une fonction d'ordonnancement dans le filtrage collaboratif consistant à combiner la méthode ListMLE avec la technique de factorisation matricielle. Nous souhaitons pour les futurs travaux l'amélioration de cette proposition afin de la rendre plus performante. Nous signalons que la présente proposition, comme la plupart des algorithmes de recommandation de filtrage collaboratif, pourrait être considérée comme une approche de recommandation variationelle, où les mesures d'évaluation, telles que, la précision moyenne MAP et NDCG, ne sont pas directement liées à notre modèle. Par contre les récentes recherches dans le domaine d'apprentissage d'une fonction d'ordonnancement pourraient être plus exploitables pour améliorer la performance de recommandation par optimisation directe des mesures d'évaluation.

1 http://movielens.umn.edu/
2 http://www.grouplens.org/data/

Références

Amini M.-R., Usunier N., Gallinari P. (2005) Automatic Text Summarization Based on Word-Clusters and Ranking Algorithms. p. 142-156.

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., and Hullender, G. (2005) Learning to rank using gradient descent. In Proceedings of the 22nd International Conference on Machine Learning.

Breese, D. Heckerman, and C. Kadie. (1998) Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI-98), pages 43–52.

B.Sarwar, G.Karypis, J.Konstan, J.Riedl, (2000) Analysis of recommendation algorithms for E-commerce , In: ACM Conference on Electronic Commerce, pp.158-167,.

Cao, Z., Qin, T., Liu, T. Y., Tsai, M. F., and Li, H. (2007) Learning to rank: From pairwise approach to listwise approach. In Proceedings of the 24th International Conference on Machine Learning.

Crammer, K., and Singer, Y.(2002) PRanking with ranking. Advances in Neural Information Processing Systems, 14: 641-647.

Clémençon S., Lugosi G., Vayatis N., (2005) Ranking and Scoring Using Empirical Risk Minimization. , AUER P., MEIR R., Eds., COLT, vol. 3559 de Lecture Notes in Computer Science, Springer, , p. 1-15.

Craswell N., Robertson S., Zaragoza H., Taylor M., (2005) Relevance weighting for query independent evidence, SIGIR ’05 : Proceedings of the 28th annual international ACM SIGIR conference,.

D.Cossock and T. Zhang. (2008) Statistical analysis of bayes optimal subset ranking. Information Theory, 54:5140–5154,.

Ding, S., Zhao, S., Yuan, Q., Zhang, X., Fu, R. and Bergman, L., (2008). Boosting collaborative filtering based on statistical prediction errors. In RecSys ′08, 3-10.

Deshpande, M., and Karypis, G., (2004). Item-based top-N rec-ommendation algorithms. ACM TOIS, 22, 1, 143-177.

Emmanouil Vozalis, Konstantinos G. (2003). Margaritis Analysis of Recommender Systems’ Algorithms

Freund, Y., Iyer, R., Schapire, R. E., and Singer, Y. (2003) An efficient boosting algorithm for combining preferences. Journal of Machine Learning Research, 4: 933-969..

F. Xia, T. Liu, J. Wang, W. Zhang, and H. Li, . (2008) “Listwise approach to learning to rank: theory and algorithm,” Proceedings of the 25th international conference, Helsinki, Finland, pp. 1192-1199.

Herbrich, R., Graepel, T., and Obermayer, K. (2000) Large margin rank boundaries for ordinal regression. Advances in Large Margin Classifiers, 115-132.

Herlocker, J., Konstan, J., Borchers, A., and Riedl, J., (1999). An algorithmic framework for performing collaborative filtering. In SIGIR ′99, 230-237

Hofmann, T., (2004). Latent semantic models for collaborative filtering. ACM TOIS, 22, 1, 89-115.

Jean-François Pessiot, Vinh Truong, Nicolas Usunier, Massih-Reza Amini, Patrick Gallinari, (2007) Filtrage Collaboratif avec un Algorithme d'Ordonnancement, CORIA, pp. 165–180.

Koren, Y., Bell, R., and Volinsky, C., (2009). Matrix factorization techniques for recommender systems. IEEE Computer, 42, 8, 30-37.

Kalervo Järvelin and Jaana Kekäläinen (2000) . In evaluation methods for retrieving highly relevant documents. In Proceedings of the 23rd annual international ACMSIGIR conference on Research and development in information retrieval, SIGIR '00, pages 41–48,New York,NY, USA,.

Liu, N. N., and Yang, Q., (2008) . EigenRank: a ranking-oriented approach to collaborative filtering. In SIGIR ′08, 83-90.

Liu, N. N., Zhao, M., and Yang, Q., (2009). Probabilistic latent preference analysis for collaborative filtering. In CIKM ′09, 759-766.

Matthew R. McLaughlin, Jonathan L. Herlocker, (2004) Content-based filtering & collaborative filtering: A collaborative filtering algorithm and valuation metric that accurately model the user experience", In Proc. Of the 27th ACM SIGIR Conf, pp. 329-336.

Marden, J. (1995) Analyzing and Modeling Rank Data.

P. Li, C. Burges, and Q. Wu. (2008) Mcrank: Learning to rank using multiple classification and gradient boosting. In NIPS '07: Advances in Neural Information Processing Systems 20, pages 897–904, Cambridge.

Sarwar, B., Karypis, G., Konstan, J., and Reidl, J., (2001). Item-based collaborative filtering recommendation algorithms. In WWW ′01, 285-295.

Vinh Truong, (2009), Apprentissage de Fonctions d'Ordonnancement avec peu d'Exemples Étiquetés: une Application au Routage d'Information, au Résumé de Textes et au Filtrage Collaboratif, Thèse de doctorat

Weimer, M, Karatzoglou, A., Le, Q., and Smola, A., (2007). CoFi rank-maximum margin matrix factorization for collabo-rative ranking. In NIPS ′07, 20, 1593-1600.

Y. Shi, M. Larson, and A. Hanjalic, (2010), List-wise learning to rank with matrix factorization for collaborative filtering, In Proceedings of the fourth ACM conference on Recommender systems, pp. 269–272.

MMGA : une approche hybride bio-inspirée pour l'alignement multiple de séquences

Rached YAGOUBI*, Abdelouahab MOUSSAOUI**

* Département d'informatique. Université Amar Telidji - Laghouat,
Route de Ghardaia - B.P. 37G laghouat 03000 Algérie
rached.yagoubi@gmail.com
**Département d'informatique. Université Ferhat Abbas SETIF,
Cité Elmaabouda. BP 19000. SETIF 19000 Algérie.
moussaoui.abdel@gmail.com

Résumé. Dans cet article, nous présentons MMGA une nouvelle approche hybride bio-inspirée pour l'alignement multiple de séquences. MMGA combine trois algorithmes et qui sont MUSCLE, MAFFT et un algorithme génétique. La population initiale de ce dernier est générée par MUSCLE et MAFFT, après cela nous appliquons différents opérateurs génétiques afin d'accroître la précision des alignements. L'évaluation effectuée en utilisant le Benchmark populaire BAliBASE (version 3.0) prouve que MMGA réalise une amélioration importante de la précision par rapport à d'autres algorithmes performants d'alignements multiple y compris MUSCLE, MAFFT, ClustalW et ProbCons tout en maintenant un temps de calculs réduit.

1 Introduction

Afin de comprendre les mécanismes de fonctionnement du vivant, les biologistes ont besoin d'extraire des informations et des connaissances à partir de données biologiques, afin de les analyser et de les interpréter. La Bioinformatique se propose comme une science capable de fournir des moyens et des outils pour satisfaire les besoins des biologistes.

L'alignement de séquences est un des problèmes les plus importants de la bioinformatique. Il permet de découvrir des similitudes biologiques entre les séquences (nucléiques ou protéiques) et de déterminer les correspondances entre résidus.

Nous pouvons distinguer deux types de problèmes dans le domaine d'alignement de séquences selon le nombre de séquences traitées :

- *L'alignement par paires :* aussi appelé l'alignement de deux séquences. Ce problème peut être résolu de manière exacte à l'aide de la programmation dynamique ;
- *L'alignement multiple :* qui est tout alignement de plus de deux séquences. Il a été démontré NP-complet et il ne peut être résolu par une méthode exacte que pour des séquences de petites tailles et dont le nombre est très réduit.

Différents algorithmes existent pour l'alignement multiple de séquences. Nous pouvons les classer selon trois catégories :

1. *Les algorithmes exacts:* sont en général basés sur la programmation dynamique. Il existe un nombre très petit d'algorithmes exacts et ils ne peuvent être utilisés que pour des alignements ne comportant que peu de séquences car ils sont très gourmands en ressources CPU et mémoire. Nous citons comme exemples d'algorithmes exacts l'algorithme MSA Carrilo et Lipman (1998) et l'algorithme DCA Stoye et al. (1997).
2. *Les algorithmes progressifs:* consistent à aligner des sous-groupes de séquences. En général, ces algorithmes utilisent la notion de profil pour réaliser les alignements ainsi que le calcul d'un arbre appelé guide-tree pour indiquer l'ordre dans lequel il convient d'aligner les séquences. Ces algorithmes commencent par réaliser des alignements de deux séquences, puis ces alignements sont à leur tour alignés entre eux. L'algorithme s'arrête lorsque toutes les séquences sont alignées. Ces méthodes sont rapides et donnent généralement des alignements de bonnes qualités. Cependant, leur inconvénient majeur est la perte d'informations au cours du processus d'alignement, car traiter des séquences deux à deux et moins précis que de les traiter toutes ensemble. Dans cette catégorie d'algorithmes progressifs nous citons MAFFT Katoh et al. (2002), MUSCLE Edgar (2004), Clustal W Thompson et al. (1994), ProbCons Do et al. (2005).
3. *Les algorithmes itératifs:* sont basés sur des méthodes plus variées que les algorithmes progressifs. Ils ont comme point commun de réaliser l'alignement de séquences en prenant en compte toutes les séquences simultanément. Le problème de ces algorithmes est qu'ils nécessitent en générale des temps de calculs très importants. Nous citons comme algorithmes itératifs SAGA Notredame et Higgins (1996), Multiple alignment using hidden markov models Eddy (1995), Multiple Sequence Alignment Based on ABC_SA Xu et Lei (2010).

Malgré l'existence d'un nombre très important d'algorithmes d'alignement multiple de séquences, aucun d'entre eux ne permet d'obtenir la solution optimale dans tous les cas Notredame (2007).

Dans ce travail, nous proposons une nouvelle approche hybride bio-inspirée pour résoudre le problème d'alignement multiple de séquences baptisée MMGA (Muscle Mafft Genetic Algorithm). Cette approche tire profit de la vitesse des algorithmes progressifs en combinant les résultats des deux algorithmes réputés par leur rapidité (MAFFT Katoh et al. (2002) et MUSCLE Edgar (2004)) à l'aide d'un algorithme génétique, ce qui permet d'accroître la précision des résultats tout en préservant un temps d'exécution réduit.

Le reste du papier est organisé comme suit : D'abord nous décrivons les méthodes d'alignements utilisées. Dans la section 3, nous détaillons l'approche proposée. Nous présentons les données de test et les critères d'évaluations dans la section 4. La section 5 contient les résultats des différents tests effectués suivi par une conclusion dans la section 6.

2 Méthodes d'alignements

Dans notre approche, nous avons utilisé deux algorithmes d'alignement multiple de séquences très populaires qui sont MUSCLE Edgar (2004) et MAFFT Katoh et al. (2002).

Leurs rôle est de produire deux alignements qui vont être utilisés comme population initiale pour l'algorithme génétique.

2.1 MAFFT

MAFFT Katoh et al. (2002) est un algorithme progressif d'alignement multiple. Il exploite les caractéristiques physico-chimiques des acides aminés qui composent les protéines pour établir le degré de similitude ou de divergence entre elles. Il utilise une transformée de Fourier rapide pour construire le guide-tree qui lui sert pour calculer l'alignement. Cette opération permet de réduire le temps nécessaire pour générer le guide-tree.

De plus, MAFFT apporte quelques modifications pour l'évaluation des alignements. La matrice de substitution utilisée est normalisée pour ne contenir que des valeurs positives. Le coût des brèches est lui aussi adapté pour correspondre aux valeurs de la matrice de substitution qui a été calculée.

L'algorithme MAFFT est actuellement un des algorithmes les plus rapides.

2.2 MUSCLE

L'algorithme MUSCLE Edgar (2004) respecte le principe général des algorithmes d'alignement progressif, mais en apportant plusieurs modifications majeures. Pour obtenir le résultat, deux alignements sont réalisés, avec des méthodes différentes pour obtenir les deux guide-trees. Une phase de raffinement est ensuite proposée, pour vérifier la validité du second guide-tree.

L'originalité de MUSCLE est qu'il utilise une nouvelle mesure de distance qui n'exige pas un alignement pour déterminer la distance entre deux séquences, elle donne un avantage significatif de vitesse.

Tout comme MAFFT, Muscle nécessite un temps de calculs très réduit.

3 Nouvelle méthode d'alignement multiple MMGA

Notre approche consiste à combiner trois algorithmes. Les deux premiers sont MUSCLE Edgar (2004) et MAFFT Katoh et al. (2002), Le troisième algorithme est un algorithme de type génétique, il a le rôle d'améliorer les alignements obtenus par les deux premiers algorithmes. Cet algorithme doit évaluer la qualité de chaque alignement, les combinés à l'aide des différents opérateurs génétiques, pour produire de nouveaux alignements et donner le meilleur résultat. Dans (Fig. 1) nous donnons le schéma général de l'approche proposée.

3.1 Génération de la population initiale

Contrairement à SAGA Notredame et Higgins (1996) (Sequence Alignment by Genetic Algorithm), un algorithme itératif d'alignement multiple réputé d'être gourmand en ressources et très lent, qui génère une population initiale de 100 individus par insertion aléatoire de brèches. Notre algorithme utilise les alignements obtenus par MAFFT et MUSCLE ce qui permet une stabilisation beaucoup plus rapide de la population.

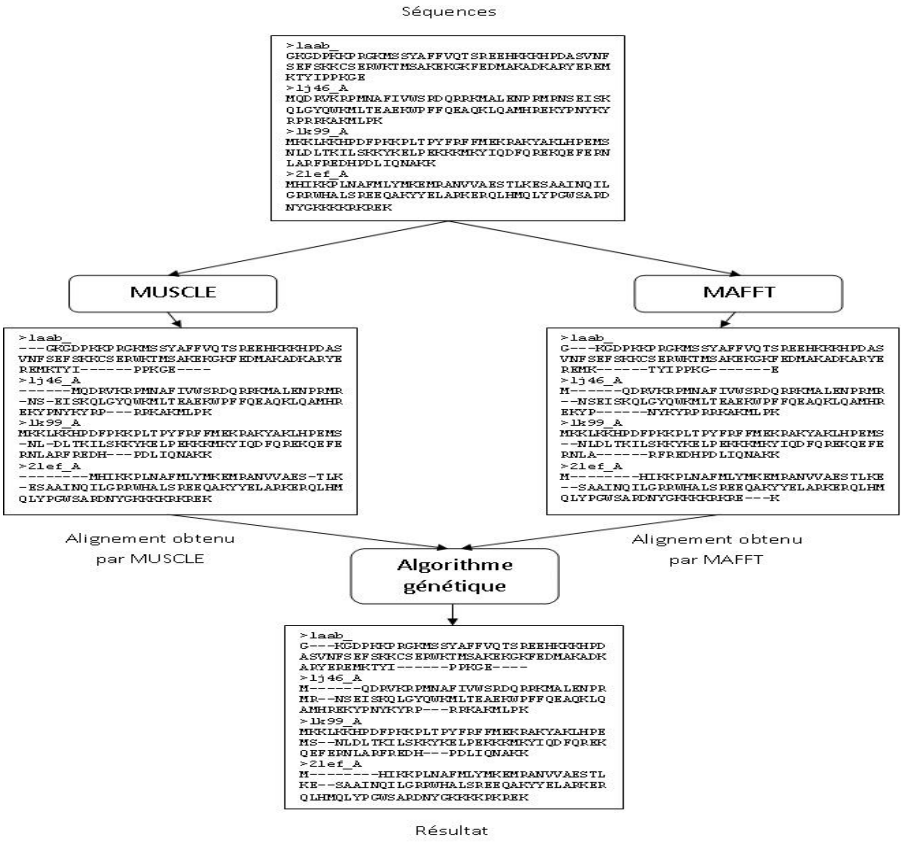


FIG. 1 – *Schéma générale de l'approche proposé.*

3.2 La fonction d'évaluation

La fonction d'évaluation que nous utilisons pour déterminer la qualité d'un alignement est la fonction de *somme des paires* (SP). Elle peut être définie comme suit :

Soit S un ensemble de k séquences, et soit A un alignement multiple de S de longueur l. Soit f une fonction permettant d'évaluer un couple de résidus, on définit la fonction de somme des paires par :

$$SP(A) = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \sum_{p=1}^{l} f(S_i(p), S_j(p))$$

Cette définition de SP dépend de la fonction f permettant d'évaluer une paire de résidus. Cette fonction a besoin de deux éléments :

- Une matrice de substitution permettant d'associer une valeur aux opérations d'appariement et de substitution,

- Un modèle d'évaluation pour les brèches (gaps) associant une valeur aux opérations d'insertion et de deletion.

La matrice de score que nous avons utilisé est BLOSUM 62 Henikioff et Henikoff (1992) (Fig. 2). C'est une des matrices de substitution (pour les protéines) les plus connue.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

FIG. 2 – *La matrice BLOSUM* 62.

Le calcul du coût des brèches est effectué suivant le modèle de coût affine car d'un point de vue biologique, la création de la brèche est beaucoup plus pénalisante que son élongation. Pour cela, il faut associer un coût plus important pour une ouverture de brèche par rapport à une extension de cette dernière. Nous avons utilisé comme paramètres par défaut le coût d'ouverture d'une brèche est $gop = -10$, et le coût de l'extension de la brèche $gep = -1$. Ces paramètres peuvent être changés selon le type des séquences en entrées. Par exemple si les séquences initiales sont de tailles très différentes, l'alignement résultat va contenir de très grandes brèches donc il serait avantageux de ne pas trop pénaliser les extensions des brèches, en leur affectant un coût $gep = -0.1$.

Avec cette matrice de substitution et ce modèle de calcul de coût des brèches, la fonction *Somme des paires* donne en générale une bonne évaluation des alignements multiples dans un temps très petit.

3.3 Le Croisement

Le croisement a pour but d'enrichir la diversité de la population en manipulant la structure des chromosomes. Il consiste à combiner deux individus (parents) pour générer deux individus (enfants ou descendants) héritant des caractéristiques de leurs parents (Fig. 3).

Dans notre cas, nous prenons deux alignements (parents) qu'on coupe en deux parties. La position du point de découpage est un paramètre important dans la convergence vers l'alignement optimal. Nous changeons cette position plusieurs fois afin d'obtenir différents descendants.

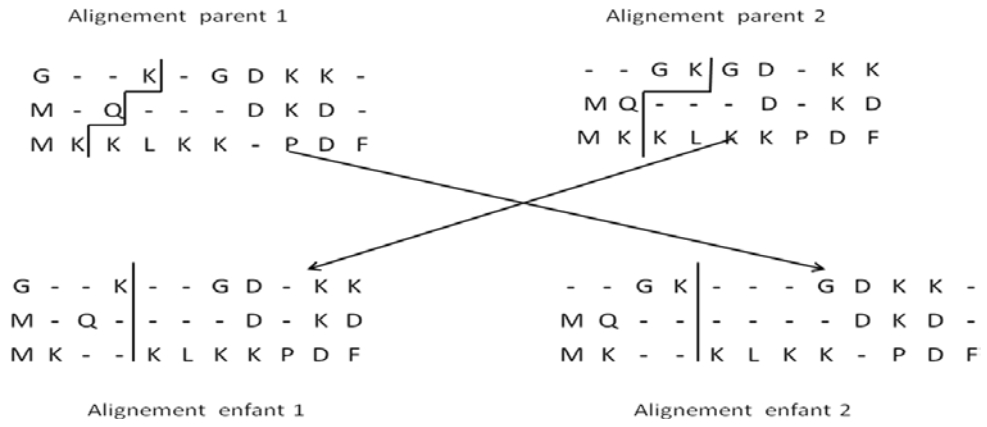


FIG. 3 – *Exemple d'un croisement.*

3.4 La sélection

La sélection permet d'identifier statistiquement les meilleurs individus d'une population et d'éliminer les mauvais. Dans notre approche, nous évaluons les alignements descendants obtenus à l'aide de la fonction Somme des paires, puis nous choisissons les trois meilleurs alignements afin d'être utilisés pour produire d'autres générations. Cette méthode de sélection directe des meilleurs individus a l'avantage de permettre une convergence rapide des solutions.

3.5 L'algorithme général

Après avoir défini les différents paramètres de notre algorithme génétique, nous présentons les étapes de l'algorithme MMGA

Algorithme MMGA

1. Générer la population initiale en utilisons MUSCLE et MAFFT
2. Evaluer les deux alignements parents
3. Faire le croisement et générer 10 descendants
4. Evaluer les descendants
5. Sélectionner les trois meilleurs descendants
6. Sauvegarder les trois meilleurs descendants (*d1, d2, d3*) et abandonner le reste
7. Si (la population est stabilisée) Ou (condition d'arrêt) alors Aller a 9
8. Aller à 3 et faire le croisement des meilleurs descendants (*d1 d2, d1 d3, d2 d3*)
9. Afficher le meilleur alignement
10. Fin

4 Données de test

Afin de déterminer la qualité d'un algorithme d'alignement multiple, nous avons besoin de comparer l'alignement produit par cet algorithme avec un alignement référence qui prend en compte le point de vue biologique. Pour cela, des benchmarks ont été créés à partir de séquences réelles, et ils ont été alignés par des biologistes. Dans notre cas, nous avons utilisé *BAliBASE* dans sa version 3.0 Thompson et al. (2005).

BAliBASE (Benchmark Alignment dataBASE) Thompson et al. (2005) est considérée comme la base référence pour les problèmes d'alignement. Elle contient différents ensembles de tests répartis en catégories, appelées *références*. Chacune des références correspond à une classe différente de problèmes. Chaque ensemble de tests de *BAliBASE* est proposé avec la meilleure solution.

Afin de pouvoir réaliser des comparaisons entre le résultat d'un algorithme et la solution optimale, deux fonctions de comparaison ont également été ajoutées, chacune permettant de montrer un critère de qualité pour les alignements. La première SP score correspond à un critère de comparaison local, la deuxième TC score réalise une comparaison plus globale de l'alignement.

La fonction de comparaison *SP score* (Sum-of-Pairs score) est basée sur le principe de la fonction de somme des paires. Toutes les paires de séquences sont parcourues, aussi bien dans l'alignement de référence que dans l'alignement résultat. Pour chacun des deux alignements, les paires de résidus identiques ont la valeur 1, et les autres ont la valeur 0. Cette méthode consiste donc à déterminer le nombre de paires de résidus identiques entre la référence et le résultat (Fig. 4). En divisant cette somme par le nombre total de paires de résidus de la référence, nous obtenons un pourcentage de similarité entre les paires de résidus des deux alignements.

FIG. 4 – *Principe du critère SP score.*

La fonction de comparaison TC score (Total Column score) est quant à elle basée sur un point de vue différent du concept d'alignement. La qualité que l'on attribue dans ce cas à un alignement multiple dépend d'une colonne complète bien alignée. Réussir à obtenir une paire de résidus identique entre la référence et le résultat ne suffit plus, il faut obtenir l'identité entre tous les résidus d'une même colonne (Fig. 5). Le critère de comparaison TC score se calcule en faisant la somme de toutes les colonnes identiques entre l'alignement de référence et l'alignement résultat. Pour obtenir un pourcentage, ce nombre est divisé par le nombre de colonnes de l'alignement de référence.

FIG. 5 – *Principe du critère TC score.*

5 Résultats et évaluation

Afin de valider notre approche, nous avons effectué des tests sur 60 ensembles de séquences protéiques à partir de différentes références de *BAliBASE 3.0*. Le nombre de séquences contenues dans chaque ensemble et les tailles des séquences sont très variées ce qui permet de bien tester notre approche.

5.1 La qualité des résultats

Nous avons comparé MMGA avec quatre algorithmes d'alignements multiples de séquences très populaires tel que : MAFFT, MUSCLE, Clustal W et ProbCons.

Pour les tests pratiques nous avons utilisé MUSCLE 3.70, MAFFT 6.717-1, Clustal W 2.0.10-1 et ProbCons 1.12-4.

Tous ces programmes ont été utilisés avec les paramètres par défaut (automatique). Dans les quatre programmes le paramètre par défaut favorise la précision des résultats et non la vitesse d'exécution.

Le tableau (TAB. 1) présente les moyennes des scores *SP score* et *TC score* dans chaque référence de *BAliBASE*. Nous noterons par Globale la moyenne du score dans tous les tests effectués. Les meilleurs résultats sont mis en gras.

	MUSCLE		MAFFT		Clutal W		ProbCons		MMGA	
	SP	TC	SP	TC	SP	TC	SP	TC	SP	TC
Ref 1	0.474	0.242	0.536	0.300	0.412	0.161	0.546	0.298	**0.554**	**0.331**
Ref 2	0.624	0.211	**0.744**	0.213	0.584	0.060	0.730	**0.218**	**0.744**	0.213
Ref 3	0.747	0.310	0.804	0.431	0.696	0.266	0.795	0.400	**0.809**	**0.437**
Ref 4	0.817	0.441	0.839	0.552	0.796	0.432	**0.844**	0.557	**0.844**	**0.559**
Ref 5	0.566	0.375	**0.752**	0.433	0.600	0.302	0.734	0.439	0.747	**0.441**
Globale	0.566	0.282	0.629	0.348	0.511	0.209	0.632	0.346	**0.641**	**0.369**

TAB. 1– *Tableau récapitulatif des résultats obtenus avec BAliBASE 3.0.*

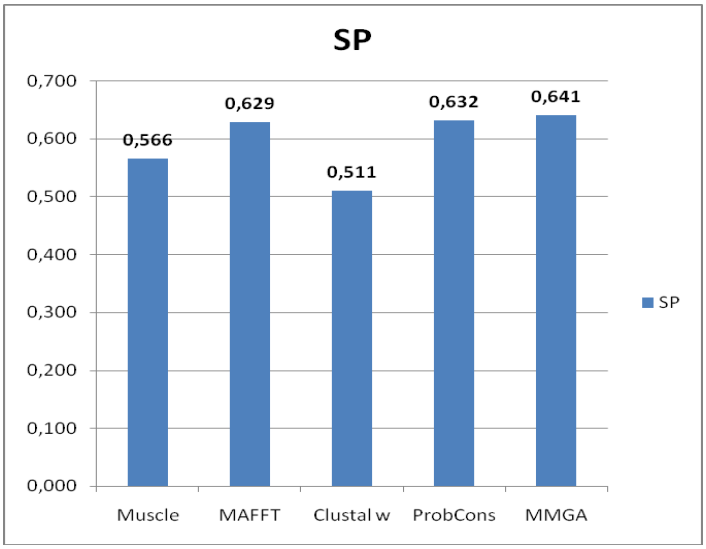


FIG. 6 – *Histogramme des moyennes des scores SP de chaque algorithme.*

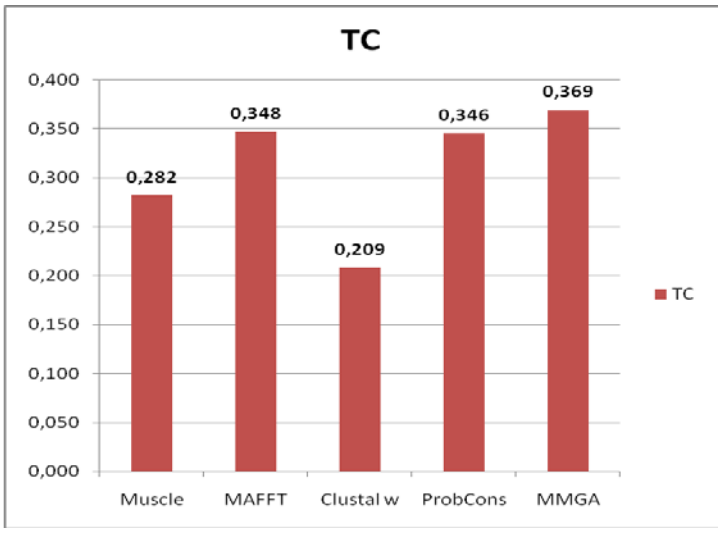


FIG. 7 – *Histogramme des moyennes des scores TC de chaque algorithme.*

D'après le tableau (TAB. 1) et les histogrammes (Fig. 6 et Fig. 7) nous constatons que MMGA donne les meilleurs scores avec *SP=0.641* et *TC=0.369* et donc elle permet de produire des alignements multiples de meilleurs qualité.

5.2 Les temps d'exécutions

L'objectif de notre travail est de développer un algorithme permettant d'améliorer la qualité des résultats d'alignements. Cet objectif est atteint comme nous l'avons déjà présenté dans la section précédente. Pour les temps d'exécutions nous avons calculé le temps pris par chaque algorithme dans plusieurs cas. Nous présentons ici les tests effectués sur BB20002 qui contient la plus longue séquence parmi tous les tests effectués (avec 1520 acides aminés) et BB30003 qui contient le plus grand nombre de séquences (142 séquences). Ces deux tests donnent une idée générale sur les temps d'exécutions pris par chaque algorithme.

Les tests sont effectués sur une machine dotée d'un processeur Pentium(R) Dual-Core E5500 @ 2.80GHZ 2.80GHZ, 2,00 Go de RAM. Système linux (Ubuntu version 10.04 (lucid), Noyau Linux 2.6.32-25-generic, GNOME 2.30.2).

TAB. 2 illustre les temps d'exécution (en seconde) pris par MUSCLE, MAFFT et l'algorithme génétique proposé (GA). MMGA contient la somme des temps d'exécutions des trois algorithmes précédents. MMGAp (pour MMGA parallèle) c'est le temps pris par notre méthode si MUSCLE et MAFFT sont lancés en parallèle. Le temps d'exécution de MMGAp est donné par la formule suivante :

$$T(MMGAp) = MAX(T(MUSCLE), T(MAFFT)) + T(GA)$$

Tel que : T(X) est le temps d'exécution de l'algorithme X.

TAB. 2 contient aussi les temps d'exécution de Clustal W et de ProbCons.

	BB20002	BB30003
MUSCLE	22	36
MAFFT	17	12
GA	3	50
MMGA	42	98
MMGAp	25	86
Clustal W	16	69
ProbCons	54	1226

TAB. 2– *Temps d'exécutions de chaque algorithmes dans les cas BB20002 et BB30003.*

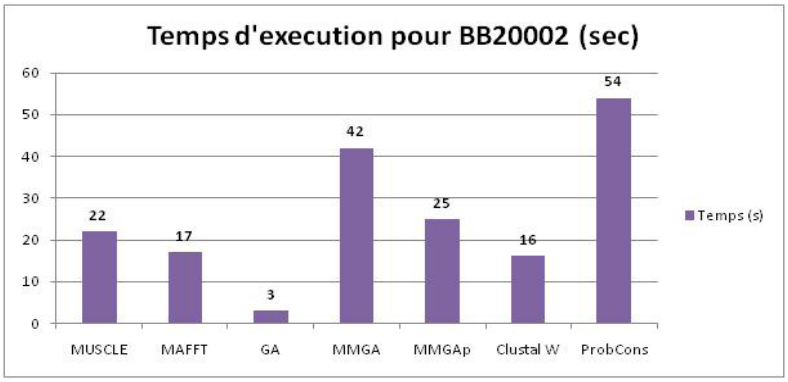


FIG. 8 – *Histogramme des temps d'exécutions de chaque algorithme dans le cas BB20002 (sec).*

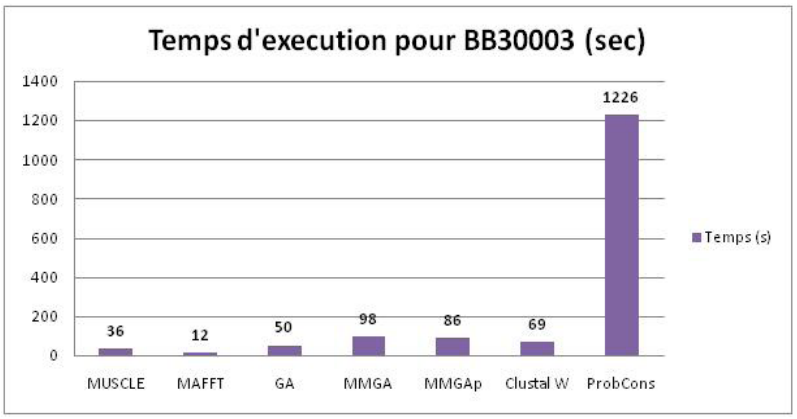


FIG. 9 – *Histogramme des temps d'exécutions de chaque algorithme dans le cas BB30003 (sec).*

Les résultats prouvent que l'algorithme ProbCons est l'algorithme le plus lent, il prend un temps d'exécution de 54 secondes dans le test BB20002 et il reste plus de 20 minutes et 26 secondes dans le cas BB30003. Même si MMGA est une hybridation de trois algorithmes lancés séquentiellement il reste largement plus rapide que ProbCons.

6 Conclusion

Dans ce travail, nous avons proposé une nouvelle approche hybride bio-inspirée pour la résolution du problème d'alignement multiple de séquences que nous avons appelé MMGA.

MMGA combine trois algorithmes, deux algorithmes progressifs d'alignement multiple (MAFFT et MUSCLE) et un algorithme évolutionnaire de type génétique.

Nous avons testé notre approche sur différents ensembles de tests de BAliBASE (version 3.0). Cette dernière fournie avec chaque ensemble de test un alignement référence qui prend en compte le point de vue biologique. Ensuite, nous avons comparé les résultats obtenus par notre approche avec ceux de différents algorithmes d'alignement multiple réputés d'être parmi les meilleurs qui sont : MUSCLE, MAFFT, Clustal W et ProbCons. Les résultats ont prouvé que MMGA donne des alignements meilleurs tout en préservant un temps de calcul très petit.

Références

Carillo, H. et D.J. Lipman (1998). The multiple sequence alignment problem in biology. SIAM J. APPL. MATH, Vol. 48 No. 5.

Do, C.B., M.S.P. Mahabhashyam, M. Brudno, et S. Batzoglou (2005). ProbCons : Probabilistic consistency-based multiple sequence alignment. Genome Research, Vol. 15.

Eddy, S.R. (1995). "Multiple alignment using hidden markov models". In CA AAAI Press, Menlo Park, editor, Third International Conference on Intelligent Systems for Molecular Biology (ISBM).

Edgar, R.C. (2004). MUSCLE : multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, Vol. 32 No.5.

Henikoff, S. et J.G. Henikoff (1992). Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Science.

Katoh, K., K. Misawa, K. Kuma, et T. Miyata (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research, Vol. 30 No. 14.

Notredame, C (2007). Recent Evolutions of Multiple Sequence Alignment Algorithms. PLoS Computational Biology, Vol. 3 No. 8.

Notredame, C. et D.G. Higgins (1996). SAGA : Sequence alignment by genetic algorithm. Nucleic Acid Research, Vol. 24 No. 8.

Stoye, J., V. Moulton, et A.W. Dress (1997). DCA : an efficient implementation of the divide-and-conquer approach to simultaneous multiple sequence alignment. Comput. Apll. Biosci. , Vol. 13 No. 6.

Thompson, J.D., D.G. Higgins, et T.J. Gibson (1994). Clustal w : improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Research, Vol. 22 No.22.

Thompson, J.D., P. Koehl, R. Ripp, et O. Poch (2005). BAliBASE 3.0 : Latest developments of the multiple sequence alignment benchmark. PROTEINS : Structure, Function, and Bioinformatics, Vol. 61.

Xu, X. et X. Lei (2010). Multiple Sequence Alignment Based on ABC_SA. Proceedings of Artificial intelligence and computational intelligence : Part II.

Summary

In this paper, we describe MMGA a new hybrid bio-inspired approach for multiple sequence alignment. The design of MMGA is based on a combination of three algorithms: MUSCLE, MAFFT and a genetic algorithm. The initial population of the genetic algorithm is generated by MUSCLE and MAFFT, followed by the application of different genetic operators in order to improve the accuracy of alignments. Assessed using the popular benchmark BAliBASE (version 3.0), MMGA achieves statistically significant accuracy improvements over the existing top performing aligners, including MUSCLE, MAFFT, ClustalW and ProbCons while keeping a reduced computation time.

Towards Generic Moving Object Trajectories' Framework

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

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

*Université Joseph Fourier, Grenoble
Laboratoire LIG BP 72- 38402 St Martin d'Hères, Grenoble, France
ahmed.Lbath@ujf-grenoble.fr

Abstract. A rapid growth and fierce competition in the market of localisation technologies (e.g. mobile phones, vehicles with navigational equipments...) have motivated studies and development of various services based on time-geographical records generated by mobile devices. Based on object pattern, we integrate, in our proposed framework, existing trajectories models' as raw, structured and semantic trajectories data models. Also, we add space–time path trajectories to describe activities in geographical space-time and we express recursive region of interest. This framework integrates new patterns of moving objects trajectories' for efficiently answering a wide range of complex trajectory queries. These specifications could also be used for trajectories warehousing. To deploy proposed specifications, minimum components of a generic architecture and trajectories' queries are also described in this paper.

1 Introduction

In recent years, Global Positioning System (GPS) devices are available and easily installed in most moving objects: persons, animals, buses, aircraft, trucks, boats... Also, with evolution and availability of GPS-enabled devices, clients' needs are no longer know the spatial coordinates (x, y) of a moving object but rather find similar trajectories, understand human being behaviours and complex phenomenon, mining , tracking, storing, visualizing and exploring trajectories: e.g. system for destination and future route prediction based on trajectory mining (Chen and al., 2010), real-time monitoring of water quality using temporal trajectory of live fish (Heng and al., 2010), analyzing bird migrations trajectory (Spaccapietra and al., 2008).

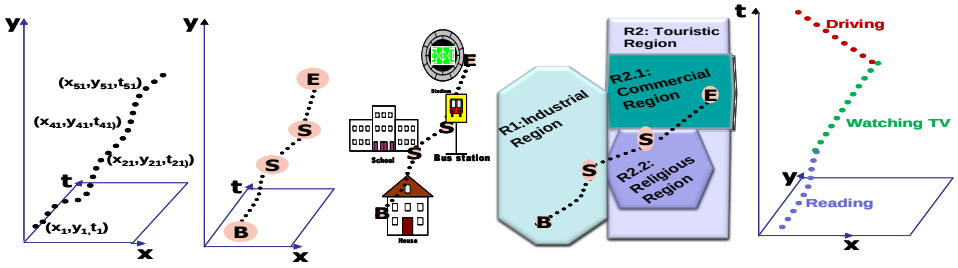
The first essential step in these applications studies is to define how to present and model the trajectory. The first response that comes to mind is to present it as a sequence of spatio-temporal events (x, y, t), but is-it expensive? Can this presentation answer any trajectory's request? Which system architecture must be used?

This paper is organized as follows. In Section 2 we present basic concepts relating to moving objects' trajectories whereas in Section 3 we discuss the issues related to the representation of trajectories. In Section 4 we discuss objective of our research. Then in section 5

we present our moving objects trajectories meta-model. Section 6 focuses on presenting the proposed trajectories' system architecture and analyzing it by providing spatio-temporal, semantic, region of interest and space time path. Finally, Section 7 presents some expressiveness requests algebra for spatial database trajectories'.

2 Basic concepts of trajectories

A trajectory is a description of physical movement of moving objects (points, lines, areas or volumes) changing over time, in the following basic presentation of trajectories: (i) Raw trajectory (figure 1-a) is the recording of positions of an object at specific space-time domain, for a given moving object and a given time interval, it's presented as a sequence of geometric location in 2D spatial system (x_i, y_i, t_i). (ii) Structured trajectory (Spaccapietra and al., 2008) (figure 1-b) defined as a raw trajectories structured into segments corresponding to meaningful steps in the trajectory trace (e.g. travel). (iii) Semantic trajectory (Spaccapietra and al., 2008) express the application oriented meaning using four component (stop, move, begin and end). Stop, move, begin and end are no more spatio-temporal position, but semantic objects linked to general geographic knowledge and application geographic data (figure 1-c). (iv) Other recent approach describes movement patterns in both spatial and temporal contexts based on Region of Interest (Giannotti and al., 2007) by defining spatial neighbourhood and temporal tolerance (figure 1-d).



(a) Raw trajectory (b) Structured trajectory (c) Semantic trajectory (d) Region of Interest (e) Space Time Path

FIG. 1 – *Basic presentations of trajectory.*

3 Related work

Meng and Ding (2003) proposed a Discrete Spatio-Temporal Trajectory Moving Object Database system destined to model moving objects' trajectory by a set of straight line within a constant speed. Patterns presented, by Spaccapietra and al., (2008) provide a trajectory structure as a sequence of moves and stops in between, with begin and end events to represent a trajectory in a relational model. However, there are many inconveniences when using a relational trajectory model, such as complex maintenance when upgrading application. Moreover, the functional approach is not suited for developing applications managing complex phenomena that are constantly changing. Also, in view of enriching, the modelling of network-constrained trajectories, design pattern presented in by Spaccapietra and al., (2008)

needs to further explore the interaction between trajectory modelling strategies and the multiple network models that have been proposed in the literature. Furthermore, the current trajectory design pattern models trajectory just from the semantic point of view.

Works proposed by Yan and Spaccapietra (2009) introduced a conceptual and computational approach for semantically trajectory data analyzing which redefines trajectory as the trace of a moving object that has geometric spatio-temporal and semantic features (the meaning of a movement). Yan and al. (2009) describes a framework for semantic trajectory relying on the definition of trajectory related ontology. Yan and al. (2010) proposed a hybrid spatio-semantic model and a computing platform for trajectories of moving objects.

Giannotti and al. (2007) describes movement patterns in both spatial and temporal contexts, based on static and dynamic RoI (Region of Interest) by allowing approximation in both spatial neighbourhood and temporal tolerance.

However, it is still very hard to explain and understand movement behaviours based on these patterns, because most of these research methods define trajectory and neglect having observation and description of trajectory/moving object (e.g. when taking taxi or bus, it's very interesting to have information of passengers number, full tank of gas in litters…). Also, working with geometrical facet is interesting for some users, e.g. doctors need to have spatial coordinate and time reference of his material when doing critical chirurgical operation (lithotripsy, radiotherapy), also when tracking a terrorist we need to know in which region he is located, but what can interest a marketing study is what was activities of the moving object: his physical activities (e.g. drive to work) and virtual activities (e.g. receive a call) in a specific space time.

With advances and multitude of mobile devices and location-based services, we can't find, with current models, which mechanism of detection was provided in each zone? Camera, NFC/RFID device, cash machine, e-mail (ip location), GSM call, etc...

4 Research objectives

The aim of this research is to provide a system architecture for a Unified trajectories' meta-model which several application domains could benefit. Hence, our proposed Oriented Object Trajectory Patterns use object approach and integrates previous models of geometric, structured and semantic of trajectory. Our model includes also the hybrid spatio-semantic model given in Yan and al. (2010) and should models the following, by using the space-time event ontology: (i) Spatial Model according to OGC Spatial Data Model, (ii) Observation domain of trajectory according to OGC Sensor Meta Model and OGC Feature Type, (iii) All activities between the begin and the end of Space Time Path Shaw (2011), (iv) Mechanism of detection used to collect generated positioning data, (v) Movement patterns using composite Region of Interest. Moreover, our generic moving objects framework aims to provide most useful trajectories' services based on our unified trajectory data-model.

5 The proposed Unified Trajectories Patterns'

In proposed trajectories patterns', we use object approach to increase developer productivity and higher quality applications. In the following we present package and class diagram for our trajectory patterns.

5.1 Class diagram for our trajectory patterns

In the following, we present needed classes to be instantiate for modelling and producing the different models integrated and modelled in our proposed meta-model**.**

5.1.1 Producing Structured Trajectory Model

For a given moving object and time interval, GPS allows collecting a huge of spatio-temporal coordinates (x_i, y_i, t_i). Figure 1(b) illustrates the structured presentation of trajectory. Class diagram, shown on figure 2, presents classes to instantiate for modelling a structured trajectory model. Below is a description of this instance:

SpaceTimeEvent class presents an event as an occurrence that happens in a small space and lasts a short time, e.g. changing position. From temporal point of view, *SpaceTimeEvent* class is a composition of *TemporalObject* class according to specification of *TimeReference* class. From spatial point of view, it is a composition of *SpatialObject*. Each *SpatialObject* class is a specialisation of OGC Spatial Data *Domain*::*Geometry* class. In our model, we present *SpatialObject* as polymorphic class (point, line or polygon class). Using UML notation, we models reflexive composition relation where the parent and child are basically from the same *SpaceTimeEvent* class. For example, workshop could be considered as an event but inside this event we find other events, like invited talks, Coffee Break... *AbstractSpaceTimePath* class trace the path taken by a moving object, it's a composition of ordered *SpaceTimeEvent* class, have begin *SpaceTimeEvent* and end *SpaceTimeEvent*. *TimeRankingFunctor* class added to order times and *SpatioTemporalFonctor* class use *SpaceTimeEvent* class to order *SpaceTimePaths***.**

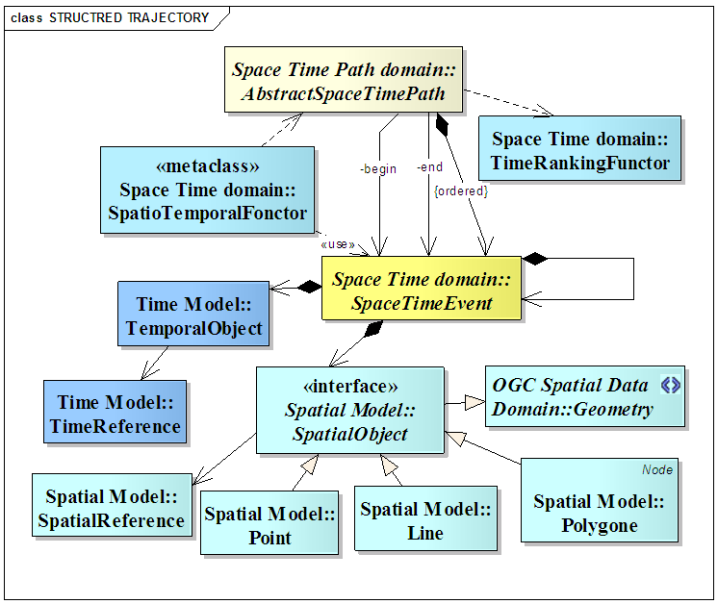


FIG. 2 – *Structured Trajectory Model class diagram.*

5.1.2 Producing Semantic Trajectory Model

In preceding paragraph, we found that using structured presentation to describe trajectory of a person, is better in terms of performance and speed of trajectory querying: saving just spatio-temporal position where trajectory begin, end, stopped and moved. But, this presentation is poor of meaning, and it's hard to analyze this trajectory.

By definition, a semantic trajectory is a structured trajectory with added semantic information, figure 3 shows semantic trajectory class diagram, in addition to structured trajectory classes (*AbstractSpaceTimePath*, *SpaceTimeEvent*, *SpatialObject*, *TimeRankingFunctor* …), user needs to add *SemanticSpaceTimeEvent* class, which is a specification of *SpaceTimeEvent*, each *SemanticSpaceTimeEvent* is *SpaceTimeEvent* with semantic information that have begin and end *SpaceTimeEvent*, Begin, End, Stop and Move are no more spatio-temporal position, but *SemanticSpaceTimeEvent* were moving object begin, end, stop end move.

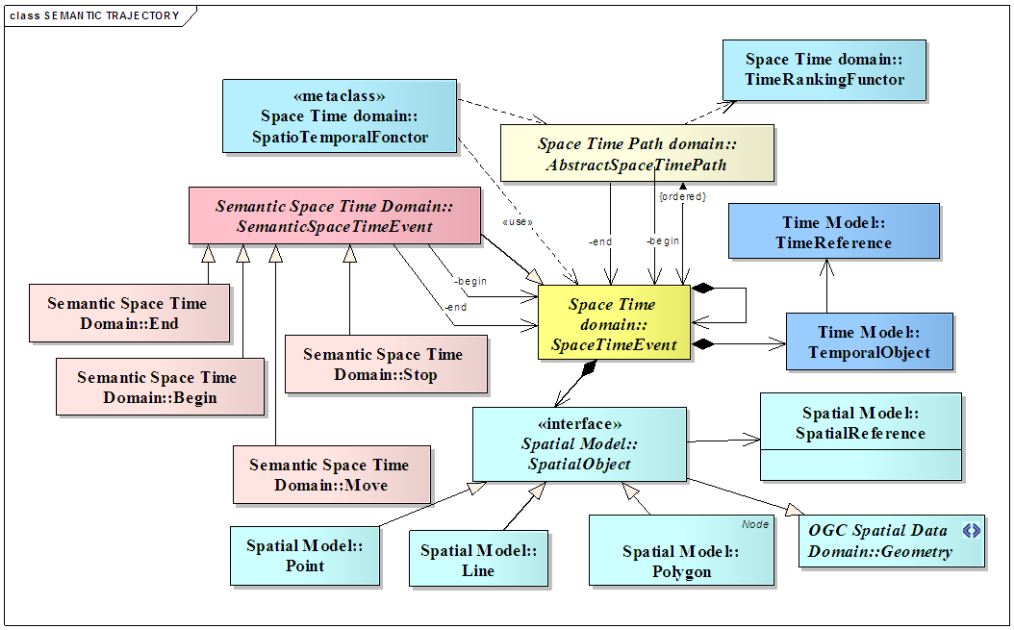


FIG. 3 – *Semantic Trajectory Model class diagram.*

5.1.3 Producing Trajectory with Region Of Interest model

As shown in the second paragraph, Trajectory with Region of Interest model describes movement patterns in both spatial and temporal contexts based on Region of Interest. For example, a person trajectory's could be presented as follows: Person X was at 9 a.m. in neighbourhood "R1" then moved to neighbourhood "R2" at 10 a.m. and finally the end of his trajectory was in neighbourhood "R3". Our meta-model allow creating this trajectory's presentation, also we modelled that region of interest could be composite of other region of interest. For example, supermarket is region of interest when tracking a moving object, but if we want more details and precisions, we need to compose it into other regions of interest: food region

and non food region. Class diagram, in figure 4, presents instantiated Trajectory with Region of Interest class diagram.

A set of *SemanticSpaceTimeEvent* class are located in *AbstractObjectOfInterest* class. Using the inheritance relationship, *AbstractObjectOfInterest* could describe a *PointOfInterest* (supermarket, bank…), area of interest (industrial region, tourism region), modal network, and Voronoi diagram. We added reflexive composition Relationship on the *AbstractObjectOfInterest* class to model that an *ObjectOfInterest* is composed of other *ObjectOfInterest*.

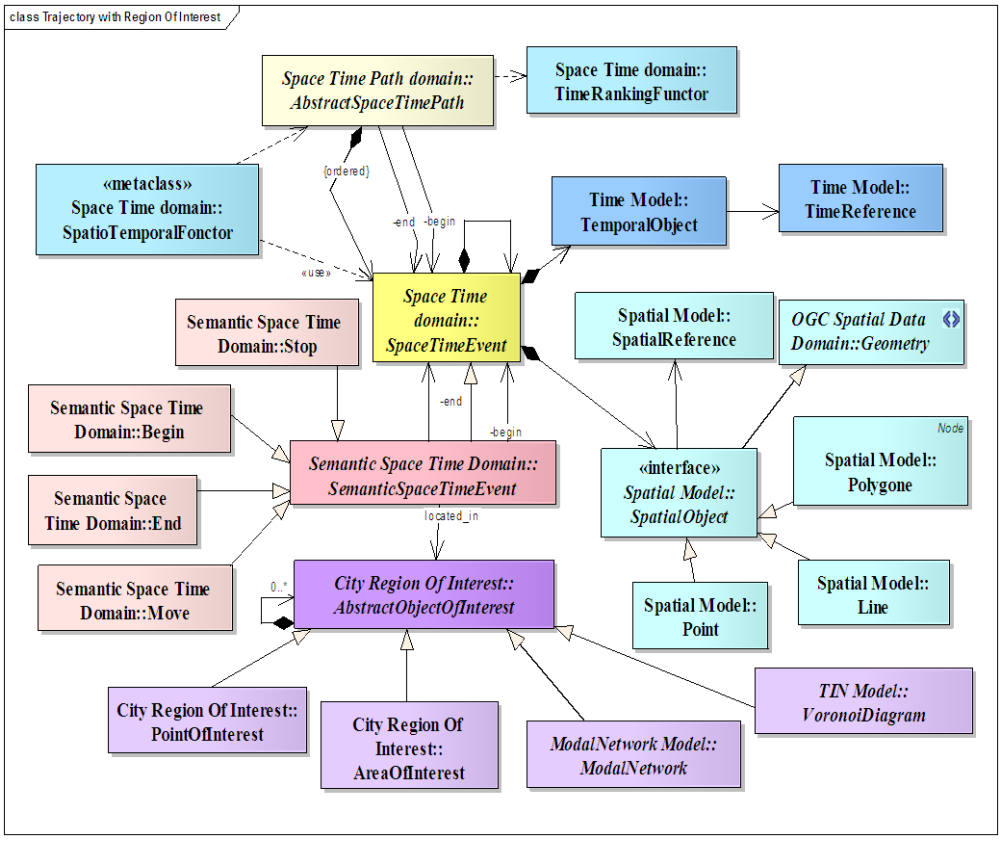


FIG. 4 –*Trajectory with composed Region of Interest class diagram*.

5.1.4 Producing Space Time Path Model

Space Time Path is a trajectory as sequence of activities describing a moving object by representing both virtual and physical activities in an integrated space-time. In our proposed meta-model, we extend raw, structured, semantic trajectory's model and trajectory based on region of interest to deals activities and process of a moving object. Figure 1(e) gives an example of space time path of a moving object.

Space Time Path class diagram, shown in figure 5, allows creating Space Time Path from previous models of trajectories (raw trajectory, structure, trajectory, semantic trajectory or trajectory with region of interest).

To get visibility of a moving object's activities, at each space time, we added for each SpaceTimeEvent class a begin activity and end activity associations with an AbstractActivity class. Thanks to UML Generalization relationship, we have modelled both virtual activity, like sending email and receiving a call, and physical activity like drive to work, walk to school or shopping. Also, in order to support spatio-temporal analysis, we added composition relationship between process and activities to capture activities as a process.

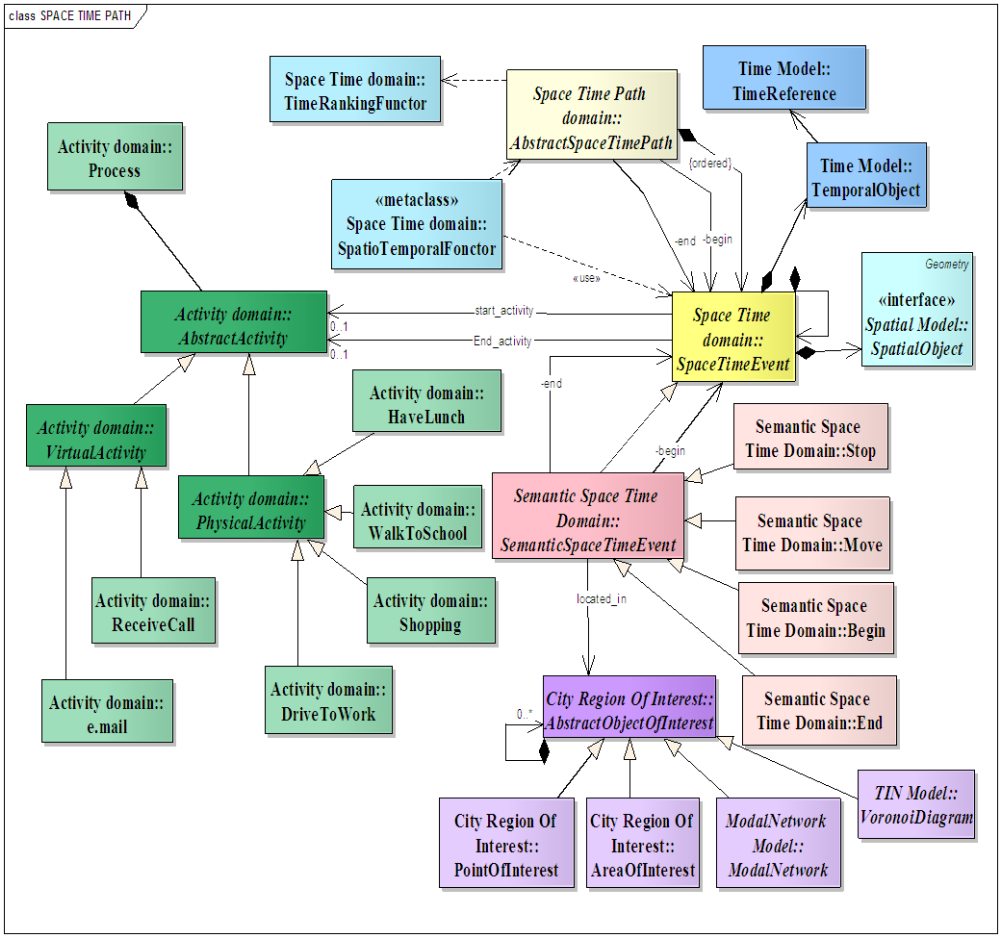


FIG. 5 – *Space Time Path Model class diagram.*

5.1.5 Trajectory Observation Measure Domain Model

In figure 7, we present our trajectory Observation Model to store pertinent and useful observations. To get observations of a moving object at each space time, we created Observation model; it consists of association between SpaceTimeEvent class and Observation class according to OGC feature Observation reference class.

Trajectory Measure Domain Model created not only to get information about mechanism of detection, used to collect generated positioning data, at a specific space time, but also to make it possible to analyze the reliability of data collected.

ProxyPositionMeasure class, presented in figure 6, functioning as interface to several measure devices: *ProxyDeviceCamera*, *ProxyDeviceCallCellularLocation*, *ProxyDeviceGPS* and *ProxyEpayment*. Thanks to the association between a *SpaceTimeEvent* class and *Proxy-PositionMeasure* class, we can know, at any moment, device's characteristics used to capture data, whatever the model used to present the trajectory.

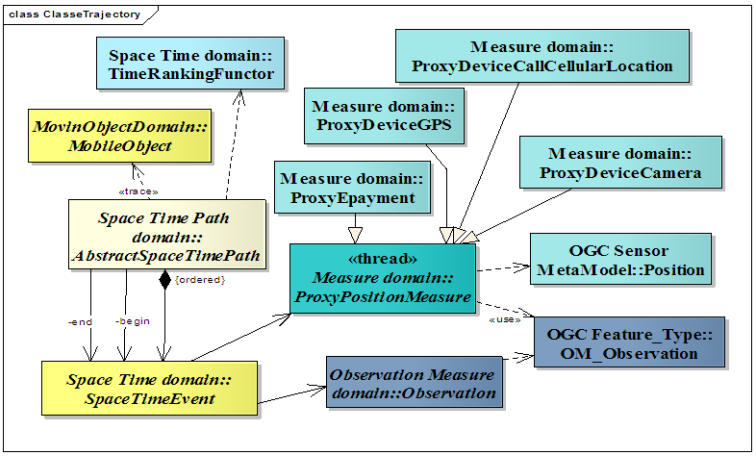


FIG. 6 – *Trajectory's observation and measure domain class diagram.*

5.2 Package diagram for our trajectory patterns

The most important packages used in our trajectory patterns are: Space Time Path Domain package, Activity Domain package, Observation Domain package, Measure Domain package and City Region of Interest packages. A description these packages will be reported in future works.

6 Unified Moving object System Architecture

This paragraph describes the proposed system architecture for our unified moving object data model. Target system architecture, shown in figure 7, consists of the three main layers: Trajectory pre-processing layer, trajectory database layer and Trajectory applications layer.

1. Trajectory pre-processing layer components are described as follows:

(a) Data collector: all depends on constraints imposed by trajectories services' user and the criticality of data to collect, data collector component is used to collect online or offline data.

(b) Data reducer: mobile devices with location positioning capability generate a huge of redundant spatio-temporal locations of (x_i, y_i, t_i), the aim of data reduce component is to reduce number of discrete spatio-temporal points to record in order to save storage

space and increase quality and trajectories' query speed. in literature, there are two reduction techniques (Wang and Krumm, 2011):

- Batched compression techniques: is an off-line compression uploading strategies where the full positioning data of tracked moving object is taken into consideration by the compression algorithms, the results aim to approaching the global optimal better than the techniques in the other category. There are various algorithms used to replace the original trajectory by an approximate line segment (Hershberger and Snoeyink (1992)).
- On-line data reduction techniques: batched compression techniques cannot deal with the second category of application that requires an instantaneous update of position, e.g. traffic monitoring service must display traffic situation constantly. Thus, depending on precision requirement and criticality of application, a selective update is applied (Vitter, (1985)).

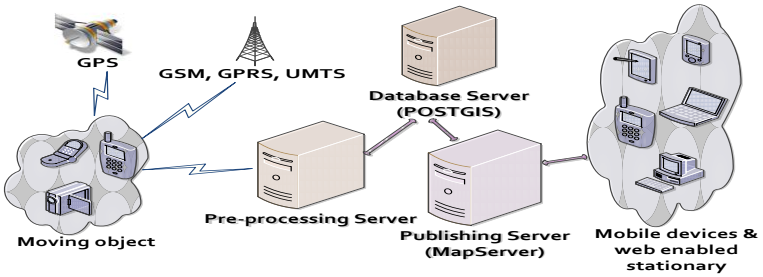


FIG. 7 – *Unified moving objects system architecture.*

(c) Error measures: data reduction, especially using online technology, can cause errors. Hence, we added error measures component to control the efficiency and performance of the technique of reduction used in a previous step. Evaluation criteria are: (i) the execution time spent to run a trajectory data reduction algorithm, (ii) size of an approximate trajectory vs. the size of its original trajectories, (iii) the deviation of reduced trajectory from its original trajectory (Potamias and al., 2006).

(d) Activity recognition: activities could be captured using sensors, e.g. GPS, WIFI, cameras and Self Terminal SSL. Reddy and al. (2010) deal the problem of activities recognition using mobile phones.

(e) Transform process: Through this component, a set of spatio-temporal points are transformed from a cleaned raw trajectory presentation to structured, semantic, region of interest or space time path.

(f) Reverse geo-coding process: convert a coordinate recorded to a readable street address which is easier to understand by the end user, this component is needed when transform process transform trajectory from raw to semantic trajectory or Region of Interest.

2. In trajectory database layer, we used POSTGIS database that support geographic objects of open source object-relational database PostgreSQL. In effect, PostGIS follows the OpenGIS Simple Features Specification for SQL and has been certified as compliant with the "Types and Functions" profile.

3. Trajectory applications layer: our framework use MapServer as an open source platform for publishing spatial data and interactive mapping applications to the web. Application

server layer contains specialised Web services that are self-contained, self-describing, modular applications of trajectory mining, tracking that can be published, located and invoked across the web using wide spectrum of Web-enabled stationary (desktops, workstations, Web TV) and mobile devices (PDAs, mobile phones, laptops, handheld computers, etc.)

7 Proposed Unified Moving Object trajectories' queries

Exploitation of trajectories data is built on various queries in a trajectory database. This paragraph is interested in evaluating the efficiency, performance and utility of models instantiated through a set of spatio-temporal queries. Using our proposed unified meta-model, trajectory queries can be classified into six types according to their instantiated spatiotemporal data-model:

1. Raw trajectory queries, asks for spatio-temporal coordinates of a specified moving object (mo) at a given time t to specified trajectory segment(s), but as raw trajectory store no semantic information, just sequence of (x,y,t), asking for semantic information need using of spatial join, e.g. find all places (restaurant, supermarket and administrations) visited by a moving object MO:

   select r.name
   from rawtrajectory t , restaurant r
   where t.id='MO' and intersects (t.spatialpoint.geometry , r.geometry)
   Union
   select s.name
   from rawtrajectory t , supermarket s
   where t.id='MO' and intersects (t.spatialpoint.geometry , s.geometry)
   Union
   select a.name
   from rawtrajectory t , administrations a
   where t.id='MO' and intersects (t.spatialpoint.geometry , a.geometry)

2. Structured trajectory queries, asks for spatio-temporal coordinates where the moving object begin, end, stops, and moves in specified trajectory segment(s), e.g. find all roads where a moving object MO crossed and stayed more than 10min.

   select r.name
   from StructuredStop t , road r
   where t.id='MO' and intersects (t.spatialpoint.geometry , r.geometry) and (t.timeEndStop - t.timeBeginStop)>10

3. Semantic trajectory queries, asks for trajectories where moving object MO stayed in a given semantic place (restaurant, cinema, stadium…) for a while (e.g., 1 hour). e.g. find all trajectories crossed a road and stayed more than 10min.

   select st.name
   from SemanticStop t, SemanticTrajectory st
   where t.cat='road' and (t.timeEndStop - t.timeBeginStop)>10
   and t.id=st.id

4. Trajectories and regions of interests queries, ask for trajectories crossed a point of interest, area of interest, modal network or voronoi diagram in a given time interval, e.g. How many trajectories visited each commercial region.

select t.nameRegion, count(*) nb_visits
from RoITrajectory t
where t.cat='commercial'
group by t. nameRegion

5. Space time path queries, asks for activities or process of a moving object in a spatio temporal location, e.g find the space time path of a specific person.
   select t.name, t.time, t.physicalActivity, t.virtualActivity
   from SpaceTimePath t,
   where t.id='MO'
6. Trajectories and mechanism of detection queries, asks for devices and their reliability degree used to capture information in a specific spatio-temporal location, e.g. which mechanism of detection used to capture information when the moving object was at the airport.
   select d.name, d.reliability
   from RoITrajectory t , Devices d
   where t.name like='%airport%' and t.id= d.tid and t.time=d.time

Conclusion

In this paper, we proposed our unified moving object-oriented trajectories framework to increase trajectories applications development and studies. Also, we presented generic trajectories system architecture and types of queries that are usually used in a trajectory database, to easily query their raw, structured and semantic trajectories as well as space time path. In future work, we project target system on a light application.

References

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

Eagle, N., and A. Pentland (2009). Eigenbehaviors: Identifying structure in routine. Behavioral Ecology and Sociobiology 63(7), 1057–1066.

Frentzos, E. K. (2008). Trajectory Data Management in Moving Object Databases. PhD Thesis, University Of Piraeus, Department of Informatics.

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

Guting, R. H., V.T. Almeida, Z. Ding (2006). Modeling and Querying Moving Objects in network, in Very Large DataBase (VLDB), 165-190.

Heng, M., T. F. Tsai, and C. Cheng (2010). Real-time monitoring of water quality using temporal trajectory of live fish. Expert Systems with Applications. 5158–5171.

Hershberger, J., and J. Snoeyink (1992). Speeding up the Douglas-Peucker Line simplification Algorithm. International Symposium on Spatial Data Handling, 134–143.

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

Pfoser, D., and C. S. Jensen. Indexing of Network Constrained Moving Object (2003). ACM International Symposium on Advances in Geographical Information Systems, Louisiana, USA, 25-32.

Potamias, M., K., Patroumpas, and T. Sellis (2006). Sampling Trajectory Streams with Spatio-Temporal Criteria. International Conference on Scientific and Statistical Database Management (SSDBM), 275–284.

Reddy, S., M. Mun , J. Burke, D. Estrin, M.H. Hansen, M.B. Srivastava (2010). Using mobile phones to determine transportation modes. ACM Transactions on Sensor Networks (TOSN) 6(2).

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

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, 26–146.

Vitter, J. (1985). Random sampling with a reservoir. ACM Transactions on Mathematical Software (TOMS) 11(1).

Wang. L. and J. Krumm (2011). Trajectory Preprocessing, Springer, chapter 1.

Yan, Z. , C. Parent, J. Macedo and S. Spaccapietra (2009). Trajectory Ontologies and Queries. Blackwell Publishing Ltd Transactions in GIS, 75–91.

Yan, Z., and S. Spaccapietra (2009). Towards Semantic Trajectory Data Analysis: A Conceptual and Computational Approach. 35th Very Large Data Base (VLDB) PhD Workshop, Lyon, France.

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

Résumé

La croissance rapide et la concurrence accrue sur le marché des technologies de localisation, tels que les téléphones mobiles et les véhicules avec des équipements de navigation, ont motivé les études et le développement de différents services basés sur les données spatio-temporelles générées par les appareils mobiles. Le présent papier propose un système permettant d'intégrer les modèles existants de trajectoires tels que les modèles de trajectoires des données brutes, structurées et sémantiques. En outre, nous ajoutons le chemin d'espace-temps pour décrire les activités dans l'espace-temps. Le système proposé intègre de nouveaux modèles de trajectoires des objets mobiles pour répondre efficacement à un large éventail de requêtes sur des trajectoires complexes. Ces spécifications pourraient également être utilisées pour l'entreposage des trajectoires. Pour déployer les spécifications proposées, les composants de l'architecture générique et les requêtes sur les trajectoires sont également décrits dans ce document.

Indexation sémantique des sources hétérogènes et distribuées en vue de médiation

Imène Saidi*, Sid Ahmed Djallal Midouni**
Lotfi Sofiane Settouti***

*Département d'informatique , Faculté des sciences
Université de Tlemcen - Abou bekr Belkaid , Tlemcen
B.P.230 - Tlemcen 13000, Algérie
s.imenedz@hotmail.fr,
**Département d'informatique , Faculté des sciences
Université de Tlemcen - Abou bekr Belkaid , Tlemcen
B.P.230 - Tlemcen 13000, Algérie
djmidouni@hotmail.com
***LIRIS - Bât Nautibus - UFR Informatique
Université Claude Bernard Lyon 1 / F-69622 Villeurbanne Cedex
lotfi.settouti@gmail.com

Résumé. Ce travail s'inscrit dans le cadre des travaux de recherches concernant la composition des services web pour l'interrogation des sources de données hétérogènes et distribuées de nature médicale (rapports médicaux, imagerie médicale annotée, ..).
L'objectif de cet article est de proposer des techniques d'indexation sémantique automatiques (indexation textuelle liée à une ontologie du domaine) et manuelles (annotations faites par des médecins à base d'une ontologie du domaine). Il s'agit également de spécifier sous forme de services Web une interface permettant l'exploitation des index sémantiques proposés.

Mots-clés : Recherche d'informations, Indexation sémantique, Services web.

1 Introduction

La diversité des sources d'information distribuées et leur hétérogénéité est l'une des principales difficultés rencontrées par les utilisateurs du Web aujourd'hui. Cette hétérogénéité peut provenir du format ou de la structure des sources (sources structurées : bases de données relationnelles, sources semi-structurées : documents XML, ou non structurées : textes) Laublet et al. (1999). Le futur web, dénommé *web sémantique*, compte parmi ses objectifs la résolution de cette problématique : fournir des mécanismes d'accès à des sources de données distribuées et hétérogènes de manière normalisée Tim Berners-Lee (2001). Des systèmes connus sous le nom de systèmes de médiation sont alors très utiles, en présence de données hétérogènes, car ils donnent l'impression d'utiliser un système homogène. Afin d'utiliser les systèmes de médiation, des services web sont spécifiés.

L'objectif de cet article est de spécifier des services web qui peuvent être utilisés par un système de médiation dans une architecture d'interrogation de sources de données hétérogènes et distribuées de nature médicale (rapports médicaux, imagerie médicale,...)(voir Figure 1),le principal objectif de ces services web est de permettre l'exploitation des index issus d'une indexation sémantique des sources de données hétérogène (semi ou non structurées) déjà faite.

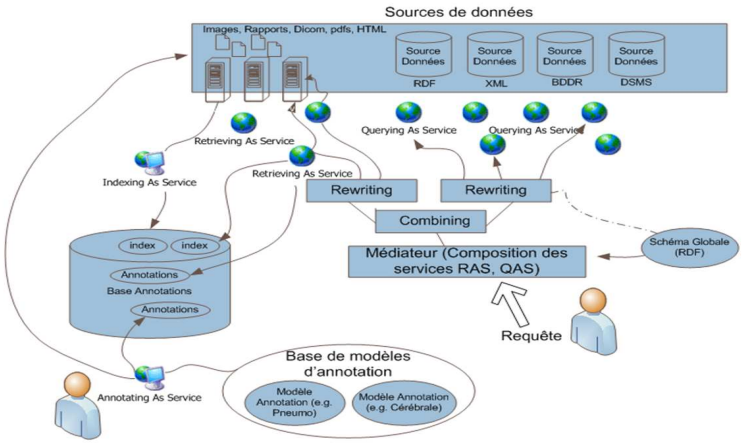


FIG. 1 – *Architecture globale du système*

Dans ce article, nous avons proposé des services web d'indexation ainsi qu'une approche d'indexation. Notre proposition est basée dans un premier temps sur la proposition de l'approche d'indexation des sources de données semi ou non structurées et les techniques utilisées pour la réalisation des index exploitables par nos services web et dans un deuxième temps sur la spécification des services web d'indexation qui peuvent être intégrés et combinés avec d'autres services web d'interrogation (concernant les sources de données structurées, ex : BDD) de l'architecture proposée. Cette spécification des services web d'indexation est faite en RDF (Ressource description framework) qui est un modèle de description des données. Nous donnerons aussi un exemple d'utilisation des services web d'indexation.
Le présent article est organisé comme suit : Dans la section suivante, nous présentons notre approche d'indexation afin de montrer les différentes étapes à suivre pour la création des index. La section 3, est réservée à la spécification des services web d'indexation proposés. A la fin de cet article, une conclusion est présentée pour synthétiser ce travail et un ensemble de perspectives sont proposées.

2 Approche d'indexation proposée

Dans ce que suit nous présentons l'approche de l'indexation suivie dans ce travail, c'est à dire les étapes de la construction des index.

Nous avons au départ un ensemble de ressources (fichiers) hétérogènes. Selon le type de la ressource, deux traitements sont possibles (Figure 2) :

1. Indexation automatique : liée aux documents textuels, ex : fichiers textes.
2. Indexation manuelle : ça concerne les annotations faites par des humains (e.g. médecins, etc.) sur les ressources non textuelles, e.g. une vidéo, une image.

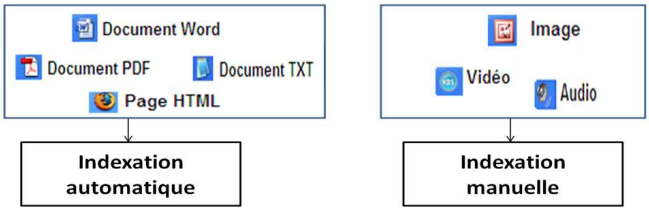


FIG. 2 – *Types d'indexation*

2.1 Indexation automatique

Dans ce type d'indexation, nous nous inspirons des travaux de Baziz et al. Baziz et al. (2005). La construction des index dans l'indexation automatique, se fait en deux phases principales : l'indexation syntaxique et l'indexation sémantique, comme il est montré dans la Figure 3.

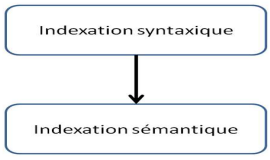


FIG. 3 – *Phases de l'indexation automatique*

a- Indexation syntaxique

L'objectif de l'indexation syntaxique est d'extraire les termes des documents et de les stocker avec leur nombre d'occurrences dans un index physique (Figure **??**).

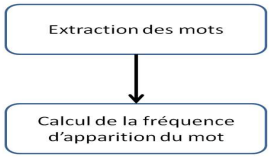


FIG. 4 – *Phases de l'indexation syntaxique*

- Extraction des mots
Dans cette première phase de construction des index, il s'agit d'extraire les termes significatifs du document, c'est-à-dire de voir pour chaque terme rencontré, s'il ne fait pas partie d'une liste de mots vides (non utiles), exemple : les articles, les pronoms, ... etc, nommés aussi les " stop-words ". Si le terme est un stop-word il sera alors ignoré et ne sera pas pris pour indexer le document, Sinon il sera pris dans l'index.

- Calcul de la fréquence d'apparition du mot
Pour les mots significatifs, la fréquence d'occurrences doit être calculée. A chaque fois qu'un mot est rencontré dans le document, on incrémente sa fréquence (nombre d'apparition). Cette valeur obtenue est nommée *tf* (Term Frequency).

Le résultat de cette phase est le fichier " Index_Syn " : Index Syntaxique.
Il est constitué comme suit :
Pour chaque document, on trouve tous les mots qui appartiennent à ce document avec leurs fréquences d'occurrence.
Num document => {(terme$_1$, i fois),..., ($terme_n$, j fois)},
i,j : nombre d'occurrence du mot.

b- Indexation sémantique
Cette phase accepte en entrée les informations issues de l'indexation syntaxique. Elle est constituée des étapes suivantes (voir la Figure cyan5) :

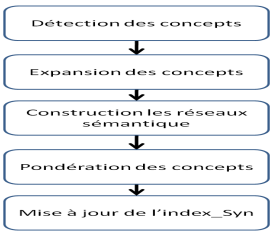


FIG. 5 – *Phases de l'indexation sémantique*

Dans l'Index_Syn (issu de l'indexation syntaxique) et pour chaque document, on fait :

– **Détection des concepts**
Un terme est dit concept s'il appartient à au moins une entrée de l'ontologie.
La détection des concepts se fait en prenant les termes un à un et en les projetant sur l'ontologie pour détecter ce qui est concept de ce qui ne l'est pas.
Exemple : "Diabète" est un concept car il correspond à une entrée de l'ontologie qui est de type maladie alors que le terme "agir" par exemple ne sera pas pris comme concept.
Dans cette étape, on fait la projection du document sur l'ontologie.

– **Expansion des concepts en utilisant l'ontologie**
Pour chaque concept détecté, un traitement spécifique est fait, il consiste à :

- Détecter les liens entre les différents concepts et de les relier ensemble en se basant sur l'ontologie.
- Etendre les concepts par leurs synonymes, dérivés et concepts de la même famille.
Une liste de synonymes et de sens sera attribuée à chaque concept, des liens et des relations entre concepts seront créés, ce qui forme une sorte de réseau pour chaque concept. Dans cette étape on fait la projection de l'ontologie sur le document.

– **Construction du réseau sémantique**
Dans les deux étapes précédentes, nous avons utilisé notre ontologie pour représenter le contenu des documents sous forme de :
- Concepts.
- Relations entre concepts.
Nous avons eu comme résultat et pour chaque concept un nombre de termes qui lui sont associés. On appellera cette association le réseau sémantique du terme.

Nous définissons un réseau sémantique comme suit :

Définition
L'ensemble de termes R_k= $\{t_1, ..., t_p\}$ forme un réseau avec un terme t du document *doc*, c'est-à-dire :
- R_k est formé par l'union des éléments n_i de type terme qui sont en relation avec t, où :
relation ={synonymie, dérivé, même_famille, ...}.
- **ET** $\forall\ t(n_i) \in R_k,\ t(n_i) \in doc$.

– **Pondération des concepts**
Il existe plusieurs approches pour pondérer les termes significatifs d'un document ou d'une requête. Nombre d'entre elles se basent sur les facteurs *tf* et *idf* qui permettent de considérer les pondérations locales et globales d'un terme.
Dans ce travail, on distingue la fréquence d'apparition d'un terme dans un document (term frequency, *tf*) qui a été déjà calculée dans l'indexation syntaxique et la fréquence d'apparition de ce même terme dans toute la collection considérée (inverse document frequency, *idf*).
La mesure *tf*idf* permet d'approximer la représentativité d'un terme dans un document. Dans notre approche, nous allons utiliser une variante de cette mesure qui le *Cf*idf*.

*** Calcul du Cf**

Le Cf est alors non pas la fréquence du terme mais celle du concept. Pour être calculé, on doit ajouter au *tf* du terme initial tous les *tf* des termes qui lui ont été associés dans le même document lors des phases précédentes, i.e les *tf* des termes du réseau sémantique.

Cf = *tf(t)* + $\sum$ *tf*(réseau). tel que t est le terme en cours et *tf*(réseau) sont tous les termes du réseau.

*** Calcul du *idf***

Le *idf* est le nombre d'occurrences d'un terme dans tous les documents.
idf sera calculé comme suit :
$idf(t) = \sum_{j=1}^{nbr_Doc} tf(t)$.

– **Mise à jour de l'index_Syn**
Pour faire la mise à jour de l'index_syn on ne prend en compte que les concepts ayant une entrée dans l'ontologie (c'est à dire les concepts ayant été détectés). Les termes qui appartiennent à leur réseau seront pris comme étant des concepts aussi avec la même pondération (qui est la somme) (enrichir l'index_Syn avec les termes du réseau créés ex : "synonymes").

On aura en résultat, un index sémantique (Index_Sem) qui est constitué comme suit :

Num `document -> {{(concept_`1_k`,i),`
`(`*concept*`_`1_{k+1}`,i)},...,{(concept_`n_k`,j), (concept_`n_{k+1}`, j)}}.`
`i,j : pondération du concept selon` *`Cf.idf`* `déjà calculé, k :1..m`
`et m : entier.`

Création d'un index inversé

Cette étape n'est pas nécessaire pour l'indexation sémantique, mais elle est utile pour faire la création et l'exploitation de l'index.
La création de l'index inversé consiste à : faire correspondre à tous les concepts, les documents qui les contiennent avec leurs pondérations, pour cela on doit regrouper tous les documents dans lesquels un concept apparait.

```
Concept 1 -> <document et pondération>
...
Concept n -> <document et pondération>
```

2.2 Indexation manuelle

Comme nous avons dit précédemment, ce type d'indexation est fait manuellement, c'est-à-dire que les documents seront annotés manuellement par des médecins (des utilisateurs) sur les ressources non textuelles, ex : une vidéo, une image (Figure 6).

Création de l'index Sémantique manuel

Après l'annotation manuelle d'un fichier non textuel, faite par un spécialiste (médecin par exemple) en se basant sur les concepts d'une ontologie, ces derniers (les concepts) seront stockés dans un index et serviront pour indexer le fichier. On considère que pour l'indexation manuelle, un fichier est pertinent pour un concept, s'il est annoté par ce concept et qu'on n'a

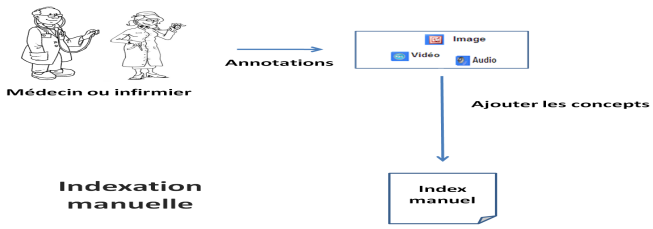


FIG. 6 – *Indexation manuelle*

pas la notion de tf et $Cf.IDF$. Le fichier sera alors retourné en résultat s'il contient le concept recherché.
On aura alors pour un fichier f, un ensemble de couple (concept, instance) :

Num `fichier -> {{concept_1, instance }, ..., { concept_`n`, instance}}.`

Création de l'index Sémantique manuel inversé

Pour chaque instance de concept, on recherche les documents qui ont été annotés par cette même instance, et on les regroupe, ce qui permet d'avoir :

```
Instance 1 -> <documents contenant l'instance>
...
Instance n -> <documents contenant l'instance>
```

3 Conception des services web

Comme nous l'avons déjà mentionné, notre indexation s'inscrit dans le cadre d'une approche d'intégration de données par médiation Gio (1992). Dans une telle approche, il est courant de définir, conceptuellement et de manière centralisée, un schéma global ou une ontologie regroupant l'ensemble des prédicats modélisant le domaine d'application du système médiateur. Dans notre cas qui est le domaine médical et afin de soutenir l'intégration des données des différentes sources, l'utilisateur posera ses requêtes dans les termes du vocabulaire structuré du domaine médical fourni par l'ontologie représentant l'ensemble des termes modélisés et utilisés par les différentes sources intégrées.

Le rôle de cette ontologie est d'établir la connexion entre les différentes sources accessibles en se fondant sur la définition de vues abstraites décrivant de façon homogène et uniforme le contenu des sources d'informations en termes des concepts de l'ontologie. Les sources d'informations pertinentes, pour l'évaluation d'une requête, sont calculées par réécriture de la requête en termes de ces vues (partie interrogation). Parmi ces vues, les services web d'indexation que nous allons proposer peuvent être utilisés.

Un exemple d'ontologie médicale de médiation

Afin de montrer un scénario global dans lequel nous allons exemplifier notre approche d'indexation à base de services web, nous présenteront dans un premier temps le schéma global décrivant l'ontologie de notre système (Figure 7).

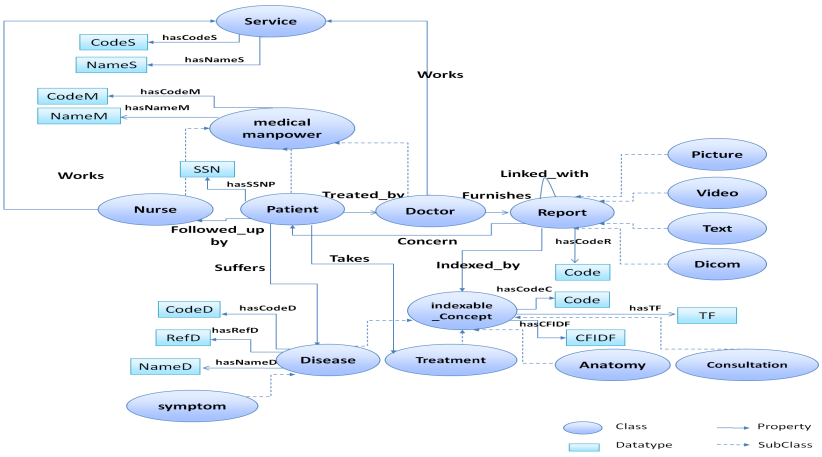


FIG. 7 – *Ontologie de médiation proposée*

L'usage d'une ontologie lors de la phase d'indexation permet de rendre un certain nombre de services dont le plus important est la levée des ambiguïtés des sens des termes utilisés pour l'indexation. L'usage d'une ontologie permet aussi une meilleure représentation des connaissances contenues dans les documents. En termes d'indexation sémantique, les concepts de l'ontologie sont associés à chaque document selon les sémantiques qui y sont véhiculées. Ainsi, en plus de lier les documents à des termes pondérés comme dans les approches classiques Faraj et al. (1996), ces documents sont liés à des termes inter-connectés faisant partie d'une ontologie où les relations disposent d'une sémantique claire et non ambiguë (synonymie, équivalence, relation hiérarchiques, etc.). Dans cet exemple (Figure 7), notre ontologie du domaine d'application a été définie comme un ensemble de classes, et chaque classe dispose de :

– *propriétés*, e.g. la classe `Service` a un code de service (propriété : `hasCodeS`).
– *sous-classes*, e.g. le lien (`rdfs:subClass`) entre la classe `Doctor` et la classe `Medical_Manpower` signifiant que la classe `Doctor` est sous-classe de la classe `Medical_Manpower`.

Une classe peut être également liée à une ou plusieurs classes, e.g. `Report` est fourni par un médecin `Doctor` et concerne un `patient`.

Afin de présenter les services Web pour l'indexation, nous allons présenter dans un premier temps les services web permettant l'interrogation des données dans le système de médiation. Ceci est nécessaire pour garder une cohérence notamment lorsque la réponse à une requête doit être fondée sur des services d'interrogation. Ces services d'interrogation doivent prendre

en compte les caractéristiques de l'indexation pour permettre une future réécriture des requêtes et une combinaison des résultats.
Les différents services d'interrogation et d'indexation, seront décris par des vues RDF à partir de l'ontologie de médiation.

L'interrogation effective des sources se fait via un médiateur, qui traduit ou réécrit les requêtes en termes de vues. La partie interrogation des sources hétérogènes qui se base sur la réécriture des requêtes et de l'intégration des différents services web n'est pas notre objectif immédiat (c'est la seconde phase), néanmoins, il est nécessaire pour notre propos d'expliciter quelques exemples de services web d'interrogation afin de montrer comment ces derniers vont être combinés à nos services web d'indexation.

Le tableau suivant (Tableau 1) contient quelques exemples des services web de la partie interrogation du système (l'autre partie avec laquelle les services web qu'on va proposer devraient être intégrés).

Service	Fonctionnalités
S_3($a, ?b)	Donne les médecins(b) travaillant dans un service(a)
S_8($a, ?b)	Donne les rapports (b) fournis par un médecin (a)

TAB. 1 – *Services web d'interrogation*

Pour notre système, nous avons proposé deux services web d'indexation décrits dans le tableau suivant (Tableau 2).

Service	Fonctionnalités	Contraintes
S1i($a, ?b)	Donne tous les concepts indexables (b) cités dans le rapport (a) ordonnés selon n	n > Seuil1 / Seuil1 : entier
S2i($a, ?b)	Donne les m premiers rapports (b) concernant un concept (a) ordonné par t, Rapport =texte et/ou vidéo et/ou dicom et/ou image	m > Seuil2 / Seuil2 : entier

TAB. 2 – *Services web d'indexation*

Syntaxe utilisée

La syntaxe des services est celle de RDF.

Exemple de description du service en RDF/RDFS

Prenonc comme exemple la définition du service S1i en RDF est la suivante :

```
S1i($a,$seuil, ?b) :-
    (?C1 rdf:type O:Concept)  .
    (?C1 O:CodeC ?b) .
    (?C1 O:hasCFIDF ?CFIDF) .
    (?R1   rdf:type   O:Rapport) .
    (?R1   O:CodeR  ?a) .
    (?R1   O:Indexé_par  ?C1)
    FILTER (?CFIDF > $seuil$)
    ORDER BY  ?CFIDF
```

Explication

a (inputs) de type Rapport et b(outputs) de type index_Conceptable.

- '$C1$' est une variable de type `Concept` qui aura un `Codec` comme code correspondant à chaque valeur de sortie qui est $?b$. Chaque `C1` a une valeur `CFIDF`.
- Les `CFIDF` sont filtrés et seulement les concepts ayant le `CFIDF > seuil` seront pris dans les résultats.
- `R1` est une variable de type `Rapport` qui a comme code `CodeR` ayant comme valeur a (l'entrée).
- Le rapport `R1` est indexé par `C1`
- Les résultats seront ordonnés par `CFIDF`.

On aura alors comme résultat : les concepts (b) cités dans le rapport (a).

Afin de schématiser le service 'S1i' représenté en RDFS, on a la figure suivante (Figure 8) :

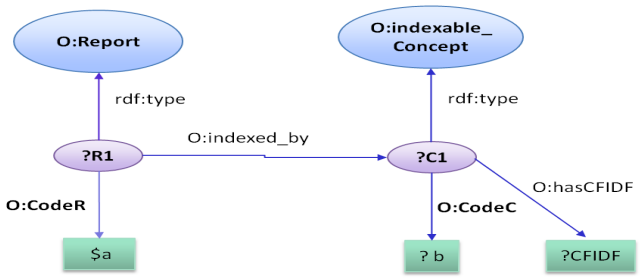


FIG. 8 – *RDFS du service S1i d'indexation*

4 Exemple d'intégration des services web

Les services web d'indexation que nous avons proposé ne sont pas capables de résoudre tous les problèmes, la composition de services web permet de répondre aux besoins complexes des utilisateurs, par la combinaison de plusieurs services web. Dans l'exemple suivant, nous

illustrons l'utilisation et l'intégration des différents services d'interrogation et d'indexation.

*** Soit la requête de l'utilisateur suivante :**

Donner les rapports contenant le concept 'y1' fournis par un médecin travaillant dans un service 'ser0'./ y1 est une maladie par exemple.

Solution de la requête
La solution de cette requête est la suivante (Figure 9) :

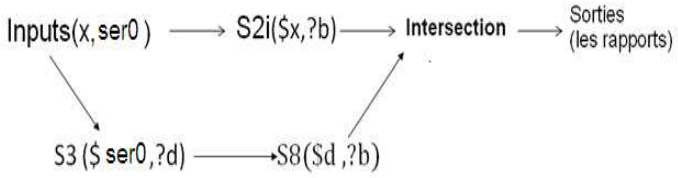


FIG. 9 – *Solution de la requête*

Pour répondre à la requête, des services d'interrogation et d'indexation ont été intégrés. Nous avons en entrée la maladie x et le service $ser0$:

1. Le service web d'interrogation S3 est lancé avec $ser0$ en entrée afin de trouver les médecins qui travaillent dans le service $ser0$.
2. Le service d'interrogation S8 est lancé après, avec comme entrée les médecins obtenus de (1) pour avoir les rapports fournis par ces médecins.
3. En parallèle à (1) et (2), le service web d'indexation S2i est lancé avec x comme entrée qui est une maladie (c'est un concept car c'est une sous classe de 'indexable_Concept') afin de trouver les rapports contenant le concept x.
4. Une intersection est faite entre le résultat de (2) et le résultat de (3) pour avoir les rapports parlant de x et fournis par les médecins travaillant dans $ser0$.

5 Conclusions et perspectives

Dans cet article, nous avons présenté une approche d'indexation sémantique des sources de données hétérogènes et distribuées qui se base sur des services web. Ces services web exploitent les index générés de l'indexation.
Le but des services web d'indexation proposés est non seulement d'exploiter les index, et alors de rendre possible la recherche dans des sources de données (semi ou non structurées), mais aussi d'être intégrés et réutilisés dans un cadre global d'intégration, et par d'autres services web d'interrogation qui concernent les sources de données structurés.
Le but de cette intégration est de gérer l'hétérogénéité, de considérer toutes les sources d'information, et les rendre en réponse aux utilisateurs. Cela est dû grâce à une réécriture des requêtes. Cette réécriture est faite selon des vues spécifiques aux différents services web définis en RDF

dans le système d'intégration global.

Notre approche n'est donc qu'une première phase dans le cadre général d'un travail de recherche concernant la composition des services web pour l'interrogation des sources de données hétérogènes et distribuées.

Pour une continuation de notre travail, plusieurs perspectives peuvent être envisagées :
- Nos travaux futurs consisteront à implémenter notre approche d'indexation et l'intégrer avec la partie interrogation du système global.
- Tester les performances de notre approche d'indexation en la comparant à d'autres approches selon des métriques qu'on définira.
- Proposer et implémenter une approche d'indexation manuelle pour tous les fichiers non textuels, ou intégrer d'autres travaux d'indexation manuelle existants.

Références

Baziz, M., M. Boughanem, N. Aussenac-Gilles, et C. Chrisment (2005). Semantic cores for representing documents in ir. In *SAC*, pp. 1011–1017.

Faraj, N., R. Godin, R. Missaoui, S. David, et P. Plante (1996). Analyse d'une méthode d'indexation automatique basée sur une analyse syntaxique de texte. In *Canadian Journal of Information and Library Science*, Canada, pp. 1–21.

Gio, W. (1992). Mediators in the architecture of future information systems. In *Computer*, pp. 38–49.

Laublet, P., C. Reynaud, et J. Charlet (1999). Sur quelques aspects du web sémantique. In *IPPS/SPDP '99/JSSPP '99 : Proceedings of the Job Scheduling Strategies for Parallel Processing*, London, UK, pp. 162–178.

Tim Berners-Lee, James Hendler, O. L. (2001). The semantic web. *Scientific American 22*, 3.

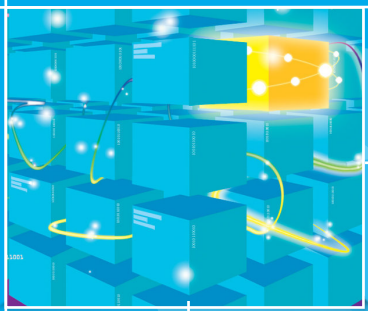
Summary

This work is a part of research regarding the composition of web services for querying the heterogeneous and distributed data sources of a medical nature (medical reports, medical imaging annotated, ...).
The purpose is to propose techniques for automatic semantic indexation (indexing texts based on an ontology) and manual indexation (annotations made by doctors based on an ontology). An interface is also specified in a form of Web services that allows the exploitation of the semantic index.
We will propose in a first time a scenario of use of the information sources and then specify (formally) the different elements allowing the implementation (e.g. index extracted from the semantic indexation, the search in this index, etc.).

Key words : Information retrieval, Semantic indexing, web services.

Atelier des Systèmes Décisionnels

ASD 2012



ASD'2012

1-3 avril 2012
Université Saad Dahlab, Blida, Algérie
 Prix : .....