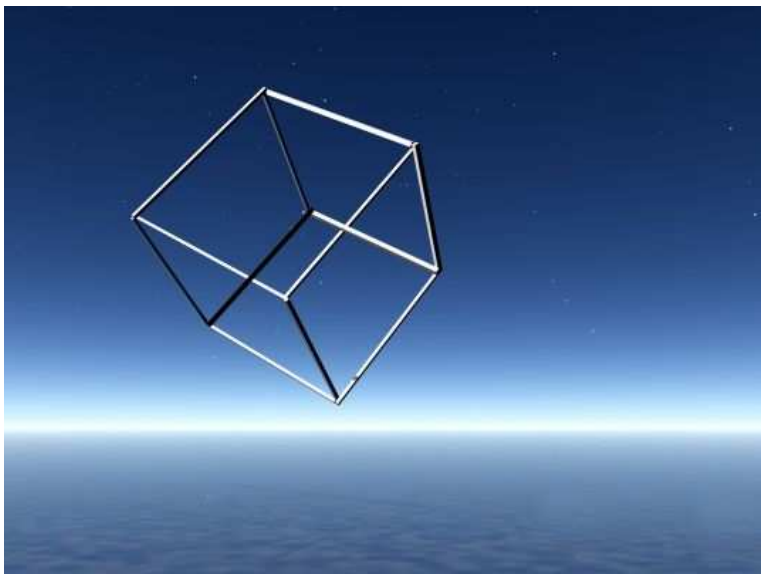
LES SYSTEMES DECISIONNELS

APPLICATION ET PERSPECTIVES

ASD 2009

Atelier des **S**ystèmes **D**écisionnels



Editeurs

Azedine BOULMAKOUL
Omar BOUSSAÏD
Jamel FEKI
Faiez GARGOURI
Ali MELIT

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 et Mohammedia-Maroc 2008), ASD 2009 (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 quatrième édition de l'atelier, en particulier après le succès des trois 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 qutrième édition ASD. ASD 2009 a reçu 59 soumissions d'articles de nombreux pays (Algérie, Canada, France, Maroc, Suisse, Tunisie,...). Après évaluation par les membres du comité scientifique, composé par des experts internationaux du domaine, 21 articles ont été retenus. Ces derniers couvrent différents thèmes de recherche et d'application sur les systèmes décisionnels.

ASD 2009 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, le Ministère Algérien de l'enseignement supérieur et de la recherche scientifique, Monsieur le Recteur de l'université de Jijel, Monsieur le Doyen de la faculté des sciences et de la technologie de l'université de Jijel (Algérie). Nous sommes reconnaissants de leur soutien.

Le succès de cette quatrième 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

A. BOULMAKOUL, O. BOUSSAID, J. FEKI, F. GARGOURI, A. MELIT

Comité de pilotage

Azedine BOULMAKOUL (FST, Université Hassan II, Mohammedia, Maroc)
Omar BOUSSAID (ERIC, Université Lumière Lyon 2, Lyon, France)
Jamel FEKI (MIRACL, Université de Sfax, Sfax, Tunisie)
Faïez GARGOURI (MIRACL, Université de Sfax, Sfax, Tunisie)

Comité de lecture du numéro

ABED Hafida, (LRSI Blida Algérie)
ADAM Frederic, (BIS University College Cork Ireland)
AHMED NACER Mohamed, (USTHB Alger Algérie)
ALIMAZIGHI Zaia, (USTHB Alger Algérie)
ALMARIMI Abdelsalam, (Higher Institute of Electronics Beniwalid Libya)
ATAMANI Baghdad, (Université d'Oran Algérie)
BADACHE Nadjib, (CERIST Alger Algérie)
BADARD Thierry, (CRG Université Laval Canada)
BADRI Abdelmajid, (FST Université Hassan II Maroc)
BEDARD Yvan, CRG (Université Laval Canada)
BELDJILALI Bouziane, (Université Oran Algérie)
BELLAFKIH Mostafa, (INPT Rabat Maroc)
BELLATRECHE Ladjel, (ENSMA Poitiers France)
BEN ABDALLAH Hanêne, (MIRACL Sfax Tunisie)
BEN BLIDIA Nadjia, (LDRSI Blida Algérie)
BEN MESSAOUD Riadh, (UTC Tunis Tunisie)
BENHARKAT Nabila, (LIRIS Lyon France)
BENMOHAMED Mohamed, (LIRE Constantine Algérie)
BENSLIMANE Djamel, (LIRIS Lyon France)
BENTAYEB Fadila, (ERIC Lyon France)
BERRADA Ilham, (ENSIAS Rabat Maroc)
BOUAZIZ Rafik, (MIRACL Sfax Tunisie
BIMONTE Sandro, (Cemagref, Clermont-Ferrand France)
BOUCELMA Omar, (LSIS Marseille France)
BOUFAIDA Mahmoud, (LIRE Constantine Algérie)
BOUFARES Faouzi, (LIPN Paris France)
BOUKERRAM Abdellah, (Université Sétif Algérie)
BOULMAKOUL Abdeljabbar, (HP Bristol Angleterre)
BOULMAKOUL Azedine, (FST Mohammedia Maroc)
BOUSSAID Omar, (ERIC Lyon France)
DARMONT Jérôme, (ERIC Lyon France)
EL HEBIL Farid, (INPT Rabat Maroc)
FAVRE Cécile, (ERIC, Lyon France)

FEKI Jamel, (MIRACL Sfax Tunisie)
GARGOURI Faïez, (MIRACL Sfax Tunisie)
HARBI Nouria, (ERIC Lyon France)
JANATI Mohamed, (ENSIAS Rabat Maroc)
KHOLLADI Med-khireddine, (LIRE Constantine Algérie)
LALAM Mustapha, (Université Tizi-Ouzou Algérie)
LEMIRE Daniel, (UQ Montréal Canada)
LOUDCHER Sabine, (ERIC Lyon France)
MAHBOUBI Hadj, (ERIC, Lyon France)
MAHIEDDINE Mohamed, (LDRSI Blida Algérie)
MALKI Mimoune, (USB Sidi Bel Abbes Algérie)
MARGHOUBI Rabia, (INPT Rabat Maroc)
MELIT Ali, (LAMEL Jijel Algérie)
MISSAOUI Rokia, (LARIM U.Q. Outaouais Canada)
MOHAMED Kerada, (LTPHU Jijel Alglérie)
MOUSSAOUI Abdelouaheb, (Université de Sétif Alglérie)
OUKID Saliha, (LDRSI Blida Alglérie)
PINET François, (Cemagref, Clermont-Ferrand France)
RAMDANI Mohamed , (FSTM Mohammedia, Maroc)
RAVAT Frank , (IRIT, Toulouse, France)
REGUIEG F. Zohra, (LDRSI Blida Alglérie)
SIDHOM Sahbi , (LOREA, Nancy, France)
TESTE Olivier, (IRIT, Toulouse, France)
ZEITOUNI Karine, (PRiSM, Versailles, France)

Comité d'organisation

Melit Ali (Université de Jijel)
Kerada Mohamed (Université de Jijel)
Lemouari Ali (Université de Jijel)
Boukraâ Doulkifli (Université de Jijel)
Rouibah Said (Université de Jijel)
Hemioud Mourad (Université de Jijel)
Kerada Ouidad (Université de Jijel)
Yahiaoui A/Baki (Université de Jijel)
Boussetoua Riad (Université de Jijel)
Ouaret Zoubir (Ecole Sup. d'informatique, Alger)
Djaoui Chafika (Université de Jijel)
Fezani Mustapha (Université de Jijel)
Lehtihet Raja (Université de Jijel)
Ghebghoub Ouafia (Université de Jijel)
Lahoulou Atidel (Université de Jijel)
Bouainah Madiha (Université de Jijel)
Triki Salah (Université de Sfax)

TABLE DES MATIÈRES

Vers une réorganisation des cubes de données par une approche neuronale dédiée à la visualisation

Nabil Zanoun, Baghdad Atmani, Fawzia.Zohra Abdelouhab

Equipe de recherche « Simulation, Intégration et Fouille de données »
Département d'Informatique, Faculté des Sciences
Université d'Oran BP 1524 El-M'Naouer 31000 ,oran Algérie
zanoun.nabil@gmail.com
atmani.baghdad@univ-oran.dz
fzabdelouhab@yahoo.fr

Résumé. La fouille de données est l'analyse des observations d'un ensemble de données dont le but est d'identifier des relations non soupçonnées, et de résumer la connaissance incluse au sein de ces données sous de nouvelles formes à la fois compréhensibles et utiles pour l'expert de ces données; les études montrent que cette analyse est d'autant plus facile et plus explicite lorsqu'elle utilise une composante visuelle. Dans cet article, nous proposons une nouvelle approche qui permet d'apporter une solution au problème de la visualisation des données engendré par l'éparsité dans les cubes de données et ceci en utilisant une technique basée sur l'apprentissage automatique par réseaux de neurones à partir d'exemples. Notre travail, appuyé par les résultats de Ben Messaoud et al., s'inscrit dans une approche générale de couplage entre fouille de données et analyse en ligne. Il consiste à éliminer les variables exogènes non pertinentes par minimisation des connexions et à sélectionner les individus non applicables afin d'atténuer l'effet négatif de l'éparsité en organisant différemment les cellules d'un cube de données. Notre but consiste à construire un nouvel espace de représentation, regroupant toutes les cellules pleines et se prêtant mieux à l'analyse et à l'exploitation des données.

Mots-clés : Cube de données, fouille de données, éparsité d'un cube, réseaux de neurones, OLAP, visualisation.

1 Introduction

Dans les systèmes décisionnels, les données sont représentées selon un modèle en étoile, autour d'une table de faits centrale contenant une ou plusieurs mesures à observer, il existe plusieurs tables de dimensions comprenant des descripteurs. Ces tables sont alors représentées dans une structure multidimensionnelle adaptée à l'analyse qu'on appelle les *cubes* de données (Ben messaoud et al, 2007). Un fait est représenté par un ensemble de modalités provenant des dimensions d'un cube et observé par une ou plusieurs mesures ayant des propriétés d'additivité plus ou moins fortes.

L'analyse en ligne OLAP (On Line Analytical Processing) est un outil, basé sur la visualisation, permettant la navigation et l'exploration dans les cubes de données (Ben messaoud et al, 2007). Son objectif est l'observation des faits, à travers une ou plusieurs

mesures, en fonction de différentes dimensions afin de fournir à l'utilisateur des opérateurs pour résumer et naviguer dans les données et y découvrir des informations pertinentes. Par exemple, observer les niveaux de ventes en fonction des produits, des périmètres commerciaux (localisations géographiques) et de la période d'achat.

Cependant, la difficulté de la navigation dans les données s'accroît avec l'augmentation de la dimensionnalité du cube. En effet, la représentation multidimensionnelle engendre une éparsité menant à un effet négatif. A titre d'exemple, pour un cube dont les faits représentent des images, les axes d'analyse possibles s'avèrent très nombreux. En effet, une image peut être décrite par une large gamme de variables tels que les descripteurs sémantiques, visuels ou les métas données. Les modalités des dimensions sont généralement représentées selon un ordre naturel: ordre chronologique pour les dates, alphabétique pour les libellés, etc. Dans la plupart des cas, cet ordre entraîne une distribution aléatoire des points représentant les faits observés (les cellules pleines) dans l'espace des dimensions.

Devant cette complexité dimensionnelle des cubes de données, l'analyse en ligne doit disposer d'outils d'assistance afin de mieux guider l'utilisateur vers les axes répondant au mieux à ses objectifs d'analyse. Ces outils doivent aussi créer de nouveaux espaces de représentation, permettent de produire une meilleure visualisation, et, ajustant au mieux le nuage des faits, en mettant, en amont les points de vue intéressants pour l'analyse.

En exploitant les techniques avancées de l'intelligence artificielle (Breiman, 1984) (Quinlan, 1986 et 1993), en général, et des réseaux de neurones en particulier, notre contribution consiste à élaborer un outil d'aide à la construction de cubes de données ayant de meilleures caractéristiques pour la visualisation. L'amélioration de la visualisation des données dans les cubes permet de réduire l'éparsité du cube et de regrouper les cellules pleines pour mieux cerner les individus une fois condensés.

Cet article est structuré comme suit. La section 2 présente un état de l'art sur des travaux similaires. Les différentes étapes de construction de notre projet sont présentées dans la section 3. La problématique qui est la réorganisation, la classification et la minimisation des connexions par les réseaux de neurones est abordée dans la section 4. Dans la section 5 on va faire une étude de cas. Enfin, nos perspectives seront données en guise de conclusion.

2 Travaux connexes

Plusieurs travaux ont été menés pour améliorer la visualisation et la navigation dans le cube de données. Nous pouvons dire que ce domaine a suscité l'intérêt de plusieurs chercheurs et a donné des résultats riches et prometteurs. Parmi eux nous pouvons citer les travaux de Vitter et al. (1999) qui ont proposé un algorithme pour construire un cube de données compact. Les résultats obtenues à l'aide de cet algorithme nous semblent meilleurs que ceux obtenues par l'approximation des histogrammes ou par échantillonnage aléatoire (voir Vitter et al., 1998).

Une autre approche dite Quasi-Cube a été proposée par Barbara et Sullivan (1997). Son principe est au lieu de matérialiser la totalité d'un cube on s'intéresse à seulement une partie de ce dernier en se basant sur une description incomplète mais suffisante de ses données. Les données non matérialisées sont ensuite approximées par une régression linéaire. Une autre technique de compression basée sur la modélisation statistique de la structure des données d'un cube a été proposée par Shanmugasundaram et al. (1999). Ces auteurs construisent une représentation compacte capable de supporter des requêtes d'agrégation

dans le cas des cubes qui ont des dimensions continues, après une estimation de la densité de probabilité des données.

Sismanis et al. (2002) proposent une méthode de compression (Dwarf) qui permet de réduire l'espace de stockage dans un cube de données et identifie les n-uplets redondants dans la table de faits, où cette redondance de données est remplacée par un seul enregistrement. Wang et al. (2002) proposent un seul n-uplet qui s'appel BST (Base Single Tuple). En se basant sur le BST, les auteurs construisent un cube de données de taille minimal MinCube (Minimalcondensed BST Cube). Feng et al. (2004) ont repris cette approche en introduisant une nouvelle structure de données PrefixCube. Les auteurs utilisent un seul BST par dimension, ils proposent un algorithme qui est une version améliorée de l'algorithme BUC (Bottom Up Computation algorithm) proposé à l'origine par Beyer et Ramakrishan (1999), cet algorithme appelé BU-BST (Bottom Up BST algorithm) permet de construire un cube de données compressé.

Lakshmanan et al. (2002) proposent pour la compression du cube de données la méthode Quotient Cube qui permet de résumer le contenu sémantique du cube de données et en le structurant sous forme de partitions de classes en réduisant la taille du cube, cette technique nous permet de conserver une structure de treillis valide donnant la possibilité de naviguer avec les opérations d'agrégation (Roll-Up) et de spécification (Drill-Down) dans le cube réduit mais l'inconvénient de cette méthode réside dans la mise à jour des données.

En 2003 Lakshmanan et al. (2003) ont proposé une nouvelle version améliorée QC-Tree (Quotient Cube Tree) qui permet de rechercher les structures compactes de données dans un cube, d'extraire et de construire les cubes intéressants à partir des données mises à jour.

On trouve aussi une autre méthode Range CUBE qui a été proposée par Feng et al. (2004), pour faire la compression des cube de données en se basant sur les corrélations entre les cellules du cube, on obtiendra un arrangement des cellules d'un cube ce dernier permet de produire une nouvelle structure du cube plus compacte et moins coûteuse en stockage et en temps de réponse.

D'autres travaux ont été menés, en utilisant les règles floues, tels que Choong et al. (2003, 2004). Ils utilisent la combinaison d'un algorithme de règles d'association avec la théorie des sous-ensembles flous. Le but de cette approche, consiste à identifier et à construire des blocs de données similaires au sens de la mesure du cube mais cette approche ne prend pas en compte le problème d'éparsité du cube.

Dans Ben messaoud et al. (2007) une nouvelle approche qui permet d'apporter une solution au problème de visualisation des données engendré par l'éparsité est proposée. En se basant sur les résultats d'une analyse en correspondances multiples (ACM), cette méthode cherche à construire un nouvel espace de représentation dans le but d'atténuer l'effet négatif de l'éparsité en organisant différemment les cellules d'un cube de données. Notre but consiste à construire par une approche neuronale un nouvel espace de représentation, regroupant toutes les cellules corrélées et se prêtant mieux à l'analyse et à l'exploitation des données.

3 Démarches conceptuelles du projet

La vocation de l'OLAP est de fournir à l'utilisateur des opérateurs pour résumer et naviguer dans les données afin d'y découvrir des informations pertinentes. Cependant, la navigation dans les cubes de données ne permet pas souvent de trouver des résultats intéressants. Rappelons que la difficulté de la navigation dans les données s'accroît avec l'augmentation de la dimensionnalité du cube. Les moyens actuels de l'analyse en ligne ne permettent pas d'évaluer les niveaux de pertinence des axes d'analyse dans un cube pour cela les chercheurs font appel à d'autres issues dans d'autres disciplines citons par exemple les travaux de Ben messaoud et al. (2007), ligne directrice de notre approche, qui a fait recours aux méthodes factorielles afin de parvenir à réduire la dimensionnalité d'un cube.

Pour illustrer notre démarche, nous supposons une société de chaussures qui désire suivre l'évolution de ses ventes, par exemple en fonction des mois (MIS), modèles de chaussures (MDL) et magasins de vente (MGS). Dans la terminologie OLAP, ces critères d'analyse sont des dimensions. ou encore d'axes**.**

Prenons l'exemple du ***d***-cube des magasins de la figure 1. Ce cube est constitué de trois dimensions (***d***=3) et deux mesures (Nombre et Prix HT). Pour l'instant, les deux indicateurs surveillés par les responsables de la société sont le nombre de chaussures vendues et les prix de vente hors taxe. Ces indicateurs sont appelés des mesures**.** La figure 2 illustre le passage réalisé par Ben Messaoud du ***d-cube*** au ***r-cube*** en prenant en considération qu'une mesure : Prix HT (PHT).

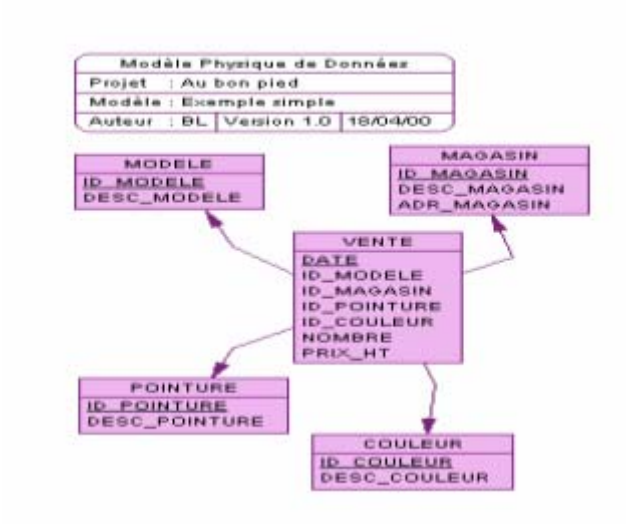


FIG. 1 – *Schémas en étoile*

La **première étape** consiste à aplatir le ***d***-cube et à générer un tableau "Individus-Variables" où les faits sont transformés en individus et les dimensions en variables (Figure 2).

Les ***m*** mesures ne seront pas considérées comme variables mais plutôt comme des poids que nous pouvons affecter aux individus. Classiquement, les dimensions sont formées de modalités symboliques. Par exemple, une dimension géographique contient des noms de magasins. Dans ce cas une telle dimension se transforme en une variable qualitative. Dans certains cas, nous pouvons aussi trouver des dimensions ayant des modalités numériques qui se transforment en variables quantitatives. La **deuxième étape** est consacrée à une modélisation par réseau de neurones multicouches. Dans la **troisième étape**, nous récupérons les coordonnées des cellules fortement corrélées par ordre décroisant. Et enfin dans **la quatrième étape** nous obtenons un nouvel espace de représentation pour les faits provenant du ***d-cube*** initial : ***r-cube***.

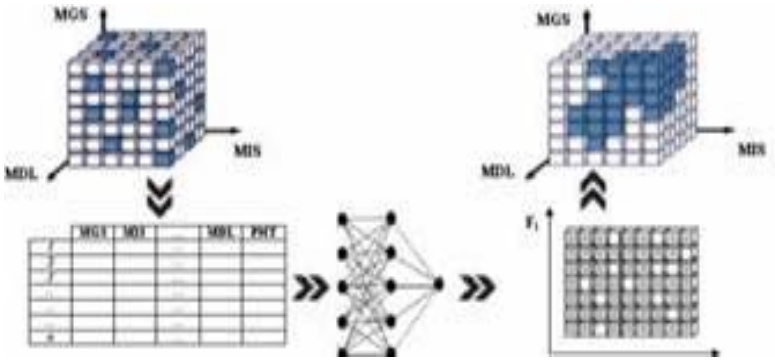


FIG. 2 – *Démarche pour réorganiser le* **d**-*cube.*

4 Réorganisation par réseau de neurones

Comme nous l'avons déjà signalé dans les sections précédentes, le prétraitement et en particulier la phase de sélection des données fortement corrélées est une phase importante puisque c'est du choix des descripteurs et de la connaissance précise de la population que va dépendre l'espace final de visualisation. L'information nécessaire à la construction d'un bon modèle de visualisation peut être disponible dans les données mais un choix inapproprié de variables ou d'individus peut faire échouer l'opération. Cette étape de mise en forme dépend fortement de la nature du domaine traité et a pour principe d'extraire des caractéristiques, descripteurs ou encore variables exogènes.

Dans cette section nous allons présenter le système **PMC**$^{\mathbf{SN}}$ (Atmani et al,2007) qui permet, à partir d'une base de cas pratique, de construire un modèle de prédiction. En détectant et en proposant à l'élimination les individus non applicables et les variables non pertinentes, le réseau de neurones **PMC**$^{\mathbf{SN}}$ optimise la table individus/variables qui peut contribuer à la réduction de l'éparsité et améliore la construction du cube de données.

4.1 Classification par réseau de neurones

La classification consiste à examiner les caractéristiques d'un individu (exemple) et lui attribuer une classe. La classe est un champ particulier à valeurs discrètes. En général les modèles de classification sont constitués en plusieurs étapes. L'étape la plus importante consiste à élaborer des règles de classification à partir des connaissances disponibles à priori ; il s'agit de la phase d'apprentissage. Cette dernière utilise soit l'apprentissage déductif, soit l'apprentissage inductif (Atmani et Beldjilali, 2007). Les algorithmes d'apprentissage inductif dégagent un ensemble de règles de classification à partir d'un ensemble d'exemples déjà classés. Le but de ces algorithmes est de produire des règles de classification afin de prédire la classe d'affectation d'un nouveau cas. Parmi les méthodes de classification utilisant ce type d'apprentissage, citons les méthodes des *k* plus proches voisins, la méthode bayésienne, la méthode d'analyse discriminante, l'approche d'arbres de décision et l'approche des réseaux de neurones.

Le processus général d'apprentissage comporte généralement trois étapes que nous récapitulons ci-dessous :

Elaboration du modèle : C'est l'étape qui fait appel à un échantillon d'apprentissage (table individus/variables) noté Ω_A dont tous les individus sont décrits dans un espace de représentation noté *D* et appartiennent à l'une des *m* classes notées *Cj* ; *j*=1,…,*m*. Il s'agit de construire l'application Φ qui permet de calculer la classe à partir de la représentation.

Validation du modèle : Il s'agit de vérifier, sur un échantillon test Ω_T dont nous connaissons, pour chacun de ses individus, la représentation et la classe, si le modèle de prédiction issue de l'étape précédente donne bien la classe attendue.

Généralisation du modèle : C'est l'étape qui consiste à étendre l'application du modèle à tous les individus de la population Ω.

Ainsi, le processus général que notre système **PMC^{SN}** applique à un cube de données est organisé sur cinq étapes :

1. Aplatir le ***d-cube*** et générer un tableau "Individus-Variables" où les faits sont transformés en individus et les dimensions en variables ;

2. Elaboration d'un modèle de classification numérique et traitement des données par un perceptron multi-couches PMC^{SN} (sélection des variables fortement corrélées et des individus non applicables) ;

3. Validation ;

4. Généralisation ;

5. Enfin, récupération des paramètres fournies par le **PMC^{SN}** et construire un nouvel espace de représentation pour les faits provenant du ***d-cube*** initial.

4.2 Minimisation des connexions par réseau de neurones

Pour l'élaboration du modèle φ^{RN} nous avons adopté le perceptron à trois couches qui a été expérimenté par Atmani et Beldjilali (2007). Le nombre de nœuds dans la couche d'*Entrée* correspond à la dimension des exemples (*vecteurs*) d'*Entrée* après codification. ***m*** unités de Sortie avec un codage binaire sont employées pour distinguer les ***m*** classes et ***H*** neurones dans la couche *Cachée*. Sachant qu'il n'existe aucune règle générale pour déterminer le nombre ***H*** de nœuds cachés à inclure dans le réseau, la littérature propose deux approches différentes pour surmonter ce problème et déterminer le nombre optimal de nœuds

cachés. La première approche commence par un réseau minimal et ajoute des nœuds dans la couche cachée seulement quand ils sont nécessaires pour améliorer les performances du réseau (Setiono et al., 1995). La deuxième approche commence par un réseau surdimensionné puis, sélectionne et élimine des connexions cachés entre les couches du réseau (Atmani et Beldjilali, 2007) (Reed, 1993).

Soit w_{hl} le poids synaptique affecté à la connexion allant du neurone d'entrée *l* vers le neurone caché *h*. Etant donnée un exemplew_i, $i \in \{1, 2, \dots, a\}$, où *a* désigne le nombre d'exemples dans l'échantillonΩ_A^{RN}, la valeur d'activation du *h*-ème neurone caché est : $\delta_h = g(\sum_{l=1}^{q}(w_{hl}x_l^i) - \theta_h)$, où *q* est le nombre de neurones d'entrée, *g* (.) est la fonction d'activation et θ_h le seuil du neurone caché *h*. Dans notre cas, *g* est la fonction d'activation tangente hyperbolique qui est définie par :

$$g(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Une fois les valeurs d'activation de tous les neurones cachés sont calculés, la sortie y^i du réseau pour l'individu ω_i est calculée par :

$$y^i = \sigma(\sum_{h=1}^{H}(\delta_h v_h))$$

Où v_h est le poids de connexion entre le nœud caché *h* et le nœud de sortie, et *H* représente le nombre de nœuds cachés dans le réseau. La fonction d'activation utilisée pour la sortie est la fonction sigmoïde normale définit par :

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Un individu ω sera correctement classé si $|y_i - y_i^d| \leq \beta 1$, où y_i^d est le résultat désiré (cible) et $\beta 1$ un seuil positif. L'objectif final de la phase d'apprentissage est d'obtenir un ensemble de poids, *w* et *v*, qui vont permettre au réseau de classer correctement les individus d'entrée. Pour mesurer l'erreur de classification, une fonction d'erreur, $E_{app}(w, v)$, est nécessaire de tel sorte que le processus d'apprentissage devient un simple processus pour ajuster les poids (*w, v*), et ainsi réduire au minimum cette fonction. En outre, pour faciliter la phase d'élagage du réseau, nous désirons avoir beaucoup de connexions avec des valeurs très petites de sorte qu'elles puissent être à zéro. Ceci est réalisé en ajoutant un terme de pénalité $E_{min}(w, v)$, proposé par Atmani et Beldjilali (2007), à la fonction d'erreur.

Généralement, la phase d'apprentissage commence par un premier ensemble de poids (*w, v*) initialisé aléatoirement en utilisant un générateur, et met à jour itérativement ces poids pour réduire au minimum la fonction globale $E_{app}(w, v) + E_{min}(w, v)$. La méthode de descente de gradient à segment nul proposée et expérimentée par Atmani et Beldjilali (2007) est employée à cette fin. L'apprentissage du réseau est terminé quand le gradient de la fonction est suffisamment petit.

Pour mener nos expérimentations nous avons suivi les principes analogues à ceux de Atmani qui propose de produire des règles d'un réseau de neurones multicouche suivant plusieurs étapes :

1. Une première phase d'apprentissage permet de déterminer les poids de raccordements d'un réseau possédant seulement une couche cachée avec un nombre arbitraire de neurones.

2. Par une méthode de minimisation, le réseau obtenu est simplifié en éliminant les raccordements (connexions) avec les plus petits poids, tout en maintenant l'exactitude du modèle (bassin solution).
3. Une nouvelle formation (apprentissage) est faite pour les raccordements utiles restants.
4. Le processus d'optimisation s'arrête selon un ou plusieurs critères de satisfaction qui seront guidés par l'expérimentation et fixés, en général, par l'utilisateur.
5. Validation, après élagage, du réseau optimal qui contient seulement les raccordements qui sont estimés appropriés (pertinents).

4.3 Les fonctions d'erreurs utilisées par le système

Nous avons utilisé dans la phase d'apprentissage de notre réseau de neurones PMC avec ***H*** unités cachées la fonction simple $E_{app}(w,v)$ nulle sur un segment. Dans le cas où la cible est 0, la fonction d'erreur $E_{app}(w,v)$ proposée par Atmani et Beldjilali (2007) est la fonction quadratique définie par :

$$y^i \in \left[0,\frac{1}{2}\right] \rightarrow E_{app}(w,v) = 0$$

$$y^i \in \left[\frac{1}{2}+\epsilon,1\right] \rightarrow E_{app}(w,v) = \sum_{i=1}^{a}\left(\left(\frac{1}{\frac{1}{2}-\epsilon}\right)^2 (y^i-\epsilon)^2\right) \text{ Avec } 0 < \epsilon \ll \frac{1}{2}$$

Dans le cas où la cible est 1, nous avons utilisé la fonction symétrique par rapport à 1/2.

La première étape du processus consiste, tout en réduisant au minimum la fonction d'erreur$E_{app}(w,v)$, à conduire le réseau vers une solution optimale qui permet de réaliser la fonction désirée. L'ensemble des solutions constitue un bassin dans l'espace des poids synaptiques. La deuxième étape consiste, en utilisant cet ensemble de solution, à minimiser le nombre de raccordements du réseau. Le principe est simple : après avoir convergé dans un bassin solution, on va rechercher le point du bassin pour lequel on a un maximum de connexions à zéro. Il s'agit d'intégrer la diminution progressive des valeurs poids, au processus de convergence. Ceci est réalisé en ajoutant un second terme$E_{min}(w,v)$, qui ne dépend que des valeurs des poids, à la fonction d'erreur$E_{min}(w,v)$. Le rôle de $E_{min}(w,v)$ est de faire tendre les poids vers zéro lors du processus de convergence. Nous avons adoptés la fonction d'erreur de pénalité définie par (Atmani et Beldjilali, 2007) :

$$E_{min}(w,v) = k \times \sum_{i=1}^{a}\left(f(w)+f(v)\right)$$

Avec :

$$f(x) = \frac{(\eta x)^2}{1+(\eta x)^2}$$

5 Implémentation et expérimentation

Nous avons consacré cette section à la présentation de notre système **OLAP-PMC^{SN}** qui permet, à partir d'un cube, de faire coopérer un navigateur **OLAP** (Excel, dans ce cas) avec un réseau de neurone multicouches (**PMC^{SN}**) pour la construction d'un modèle de prédiction qui permet de contribuer à améliorer l'espace de représentation. En détectant et en éliminant

les individus non applicables et en réorganisant les variables fortement corrélées, le réseau de neurones **PMC^{SN}** optimise la table individus/variables initiale.

5.1 Outils de développement du système OLAP- PMC^{SN}

La réalisation de notre logiciel prototype a nécessité l'utilisation de plusieurs modules à différents niveaux. L'ensemble de ces modules forme l'unité Opératrice du logiciel dont la figure 3 ci-dessous illustre le schéma général. Les cinq modules du **OLAP-PMC^{SN}**, comme vous prouvez le constater, sont :

Access : ce module sert à créer la base de données.

Ms query : ce module transforme la base de données créée par access en base de donnée multidimentionel sous forme d'un cube de données.

Excel : ce module joue le rôle du navigateur OLAP pour l'exploitation du cube de données.

WEKA : Ce module est un logiciel gratuit de DATA MINING destiné à l'enseignement et à la recherche. Il implémente une série de méthodes de fouilles de données issues du domaine de la statistique exploratoire, de l'analyse de données, de l'apprentissage automatique et des bases de données. Il permet de prétraiter des données (onglet ***Preprocess*** dans l'interface graphique), faire de la classification supervisée (***Classify***) et non-supervisée (***Cluster***), des régressions (***Select Attributes***), rechercher des règles d'association (***Associate***), et de visualiser différentes représentations graphiques des données (***Visualize***).

PMC^{SN} : Ce module, qui est actuellement intégré dans la plateforme WEKA, est utilisé pour réorganiser par ordre décroissant les faits les moins corrélés.

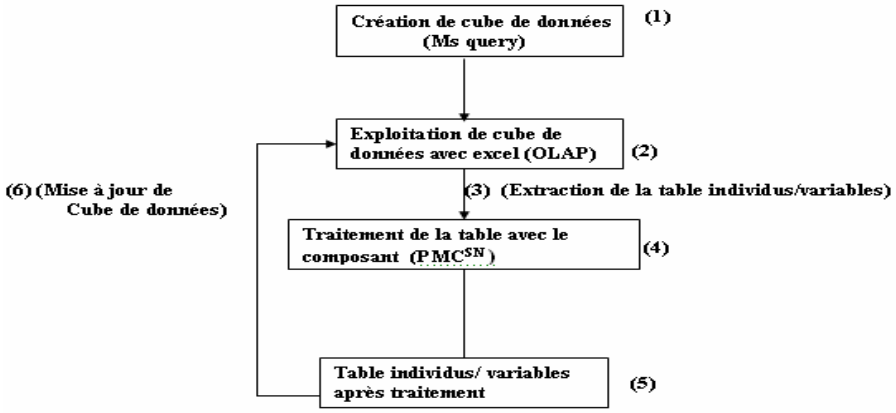


FIG. 3 – *Architecture générale du système OLAP- PMC^{SN}*

5.2 Etude de cas et résultats expérimentaux

Par la figure 4, nous illustrons le cube de données vente initial. La représentation bi-dimensionnelle du cube nous donne la table individus/variables de 532 individus et 6 variables : Référence, Temps, Magasin, Nombre, Total_HT, Classe. La table est illustrée par la figure 5.

En utilisant la plateforme Weka nous importons la table individus/variables du cube vente initiale comme le montre la figure 6. Nous sélectionnons ensuite le classifieur **PMC**^{SN} et nous fixons les paramètres d'apprentissage (voir figure 7).

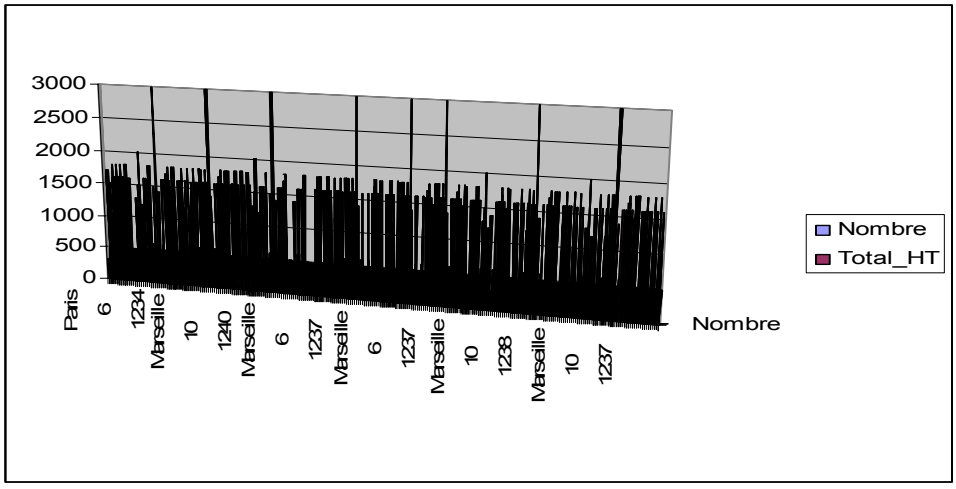
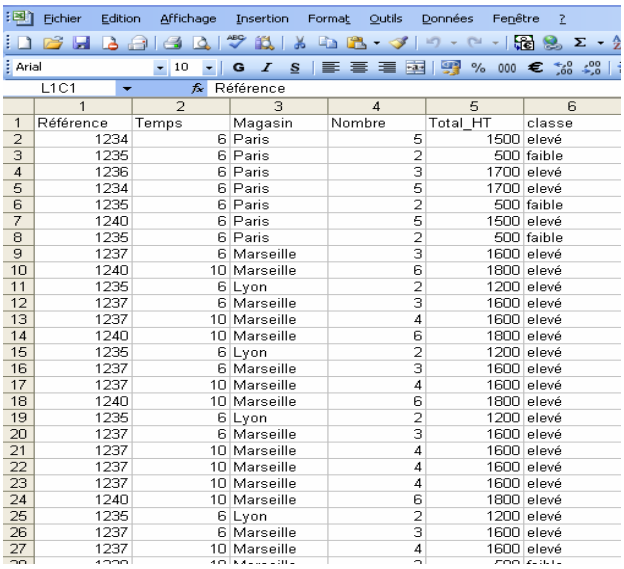


FIG. 4 – *Cube de données vente initial*



	1	2	3	4	5	6
1	Référence	Temps	Magasin	Nombre	Total_HT	classe
2	1234	6	Paris	5	1500	elevé
3	1235	6	Paris	2	500	faible
4	1236	6	Paris	3	1700	elevé
5	1234	6	Paris	5	1700	elevé
6	1235	6	Paris	2	500	faible
7	1240	6	Paris	5	1500	elevé
8	1235	6	Paris	2	500	faible
9	1237	6	Marseille	3	1600	elevé
10	1240	10	Marseille	6	1800	elevé
11	1235	6	Lyon	2	1200	elevé
12	1237	6	Marseille	3	1600	elevé
13	1237	10	Marseille	4	1600	elevé
14	1240	10	Marseille	6	1800	elevé
15	1235	6	Lyon	2	1200	elevé
16	1237	6	Marseille	3	1600	elevé
17	1237	10	Marseille	4	1600	elevé
18	1240	10	Marseille	6	1800	elevé
19	1235	6	Lyon	2	1200	elevé
20	1237	6	Marseille	3	1600	elevé
21	1237	10	Marseille	4	1600	elevé
22	1237	10	Marseille	4	1600	elevé
23	1237	10	Marseille	4	1600	elevé
24	1240	10	Marseille	6	1800	elevé
25	1235	6	Lyon	2	1200	elevé
26	1237	6	Marseille	3	1600	elevé
27	1237	10	Marseille	4	1600	elevé

FIG. 5 – *Table individus/variables initiale*

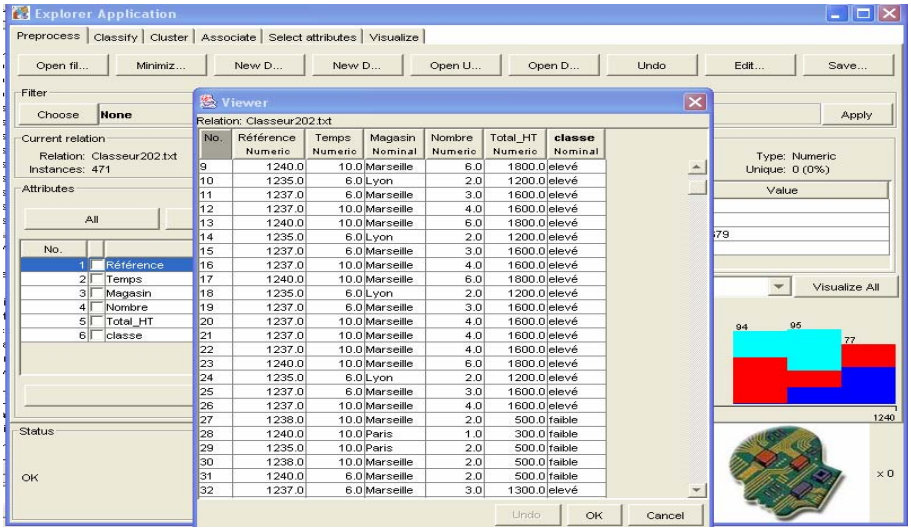


FIG. 6 – *Importation de la Table individus/variables initiale dans Weka*

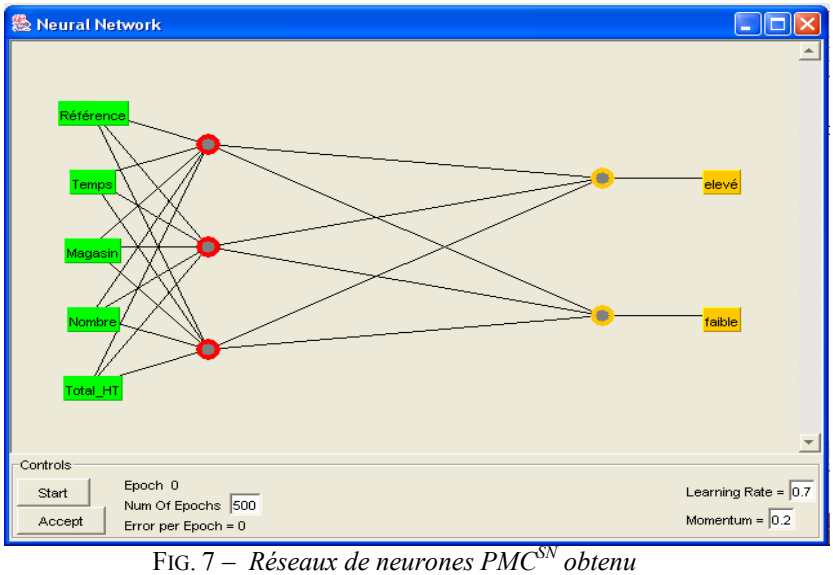


FIG. 7 – *Réseaux de neurones PMC^{SN} obtenu*

Enfin, par la figure 8 nous exposons les résultats obtenus après convergence du réseau de neurones vers un bassin solution. Le résultat obtenu est une table individus/variables de 310 individus et 6 variables classées par ordre croissant de corrélation que nous utiliserons pour la réorganisation du cube de données. La configuration finale du cube est présentée dans la figure 9.

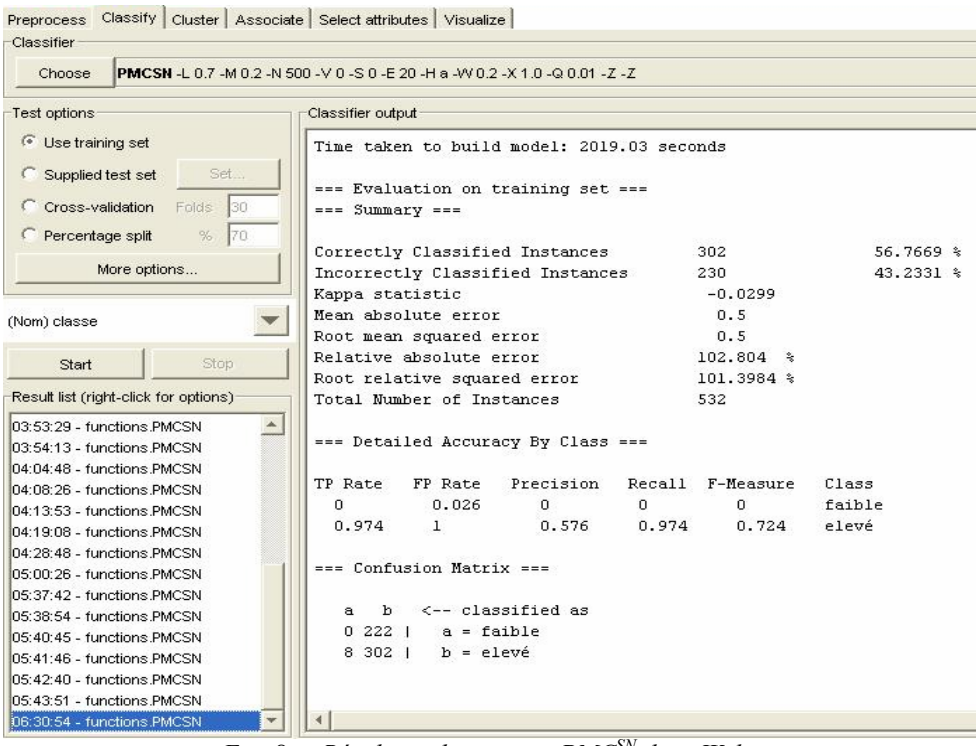


FIG. 8 – *Résultats obtenus par PMC^{SN} dans Weka*

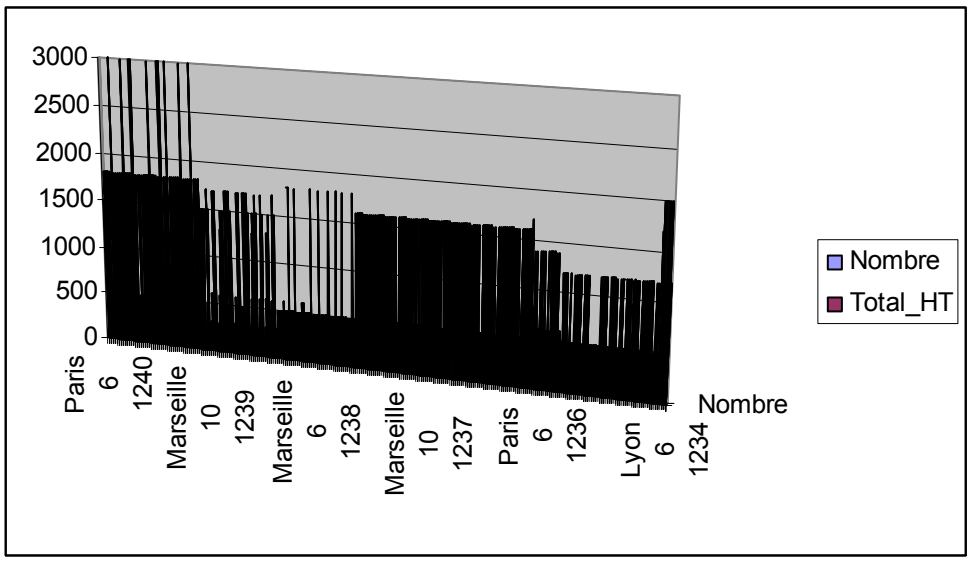


FIG. 9 – *Cube de données vente final*

Enfin et comme illustrer par la configuration finale du cube nous remarquons qu'il y a une forte homogénéité dans le cube final par rapport au cube initial et cela nous donne une bonne représentation du cube de données qui constitue l'objectif de notre approche.

6 Conclusion

Dans cet article, nous avons proposé une approche neuronale, inspirée des travaux de Ben Messaoud, pour apporter une solution au problème de la visualisation des données dans un cube éparse. En réduisant l'éparsité, nous cherchons à organiser l'espace multidimensionnel des données afin de regrouper géométriquement les cellules pleines (corrélés) dans un cube. La recherche d'un arrangement optimal du cube est un problème complexe et coûteux en temps de calcul. Nous avons opté pour les réseaux de neurone avec fonction de minimisation de connexion comme outil pour réduire cette complexité.

Références

Atmani, B. et Beldjilali, B. (2007). Neuro-IG : A Hybrid System for selection and elimination of predictor variables and non relevant individuals, INFORMATICA, International Journal, 2007, Vol. 18, No. 2.

Barbara, D. et Sullivan, M. (1997). Quasi-Cubes: Exploiting Approximations in Multidimensional Databases, SIGMOD Record, 26(3), 12-17.

Ben Messaoud, R . Aouiche, K. et Favre, C. (2007). Une approche de construction d'éspace de représentation multifonctionnelle dédié à la visualisation. arXiv: 0707.1288 v1 [cs.DB] 9 Jul.

Beyer, K. et Ramakrishnan, R. (1999). Bottom-Computation of Sparse and Iceberg CUBEs, In Proceedings of ACM SIGMOD Record, pages 359-370.

Breiman, L. J. H. Friedman, R. A. Olshen, and C. J. Stone (1984). *Classification And Regression Trees*. New York: Chapman and Hall.

Choong, Y.W. Laurent, D. et Marcel, P. (2003). Computing Appropriate Representations for Multidimensional Data, Data & knowledge Engineering Journal .

Feng, J. Fang, Q. et Ding, H. (2004). PrefixCube : Prefixsharing Condensed Data Cube. In Proceedings of the 7th ACM international workshop on Data warehousing and OLAP (DOLAP 04), pages 38-47, Washington D.C., U.S.A.

Feng, Y. Agrawal, D. El Abbadi, A. Metwally, A. (2004). Range CUBE : Ecient Cube Computation by Exploiting Data Correlation. In Proceedings of the 20th International Conference on Data Engineering, pages 658-670.

Laks, V.S. Lakashmanan, Pei, J. Han, J. (2002). Quotient Cube: How to Summarize the Semantics of a Data Cube,In Proceedings of International Conference of Very Large Data Bases, VLDB'02.

Laks, V.S. Lakashmanan, Pei, J. Zhao, Y. (2003). QC-TreesAn Efficient Summary Structure for Semantic OLAP,In ACM Press, editor, Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, pages 64-75, 2003.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning* 1, 81–106.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann.

Reed, R. (1993). Pruning algorithms, A survey, IEEE transaction on Neural Networks, vol.4, pp, 740-747.

Setiono, R. Liu, H. (1995). Understanding Neural Networks via Rule Extraction, Proceeding of tea 14th Internatianals Joint Conference one Artificial Intelligence, 1995.

Shanmugasundaram, J. Fayyad, U. M . Bradley, P. S.(1999). Compressed Data Cubes for OLAP Aggregate Query Approximation on Continuous Dimensions, In Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 223-232, August.

Sismanis, Y. Deligiannakis, A. Roussopoulos, N. Kotidis, Y. (2002). Dwarf : Shrinking the PetaCube. In Proceedings of the 2002 ACM SIGMOD international conference on Management of data, pages 464-475. ACM Press.

Vitter, J. S. Wang, M. (1999). Approximate Computation of Multidimensional Aggregates of Sparse Data Using Wavelets, In Proceedings of the 1999 ACM SIGMOD international conference on Management of Data, pages193-204, Philadelphia, Pennsylvania, U.S.A., ACM Press.

Vitter, J. S. Wang, M. Iyer, B.(1998). Data cube approximation and histograms via wavelets, In Proceedings of the 7th ACM International Conferences on Information and Knowledge Management (CIKM'98), pages 96-104, Washington D.C., U.S.A.,. Association for Computer Machinery.

Wang, W. Lu, H. Feng, J. Xu Yu, J. (2002). Condensed Cube An Effective Approach to Reducing Data Cube Size, In Proceedings of the 18th IEEE International Conference on Data Engineering (ICDE'02).

Wei Choong, Y. Laurent, A . Laurent, D. Maussion, P. (2004). Résume de cube de données multidimensionnelles a l'aide de règles floues. In Revue des Nouvelles Technologies de l'Information, editor, 4 èmes Journees Francophones d'Extraction et de Gestion des Connaissances (EGC 04) Clermont-Ferrand, France .

Summary

The data research is the analysis of the observations of a set of data that aims to identify non-suspected relations, and to summarize the knowledge included within these data under a new form, at once understandable and useful for the expert of these data; the studies show that this analysis is all the easier and more explicit when it uses a visual constituent. In this article, we propose a new approach which provides a solution to the problem of visualization

of the data generated by the dispersal in the data cubes and this by using a technique based on the automatic learning by networks of neurons from examples.

Our work joins in a general approach of coupling between the search of data and on-line analysis. It consists in eliminating the irrelevant exogenous variables by minimizing the connections and in selecting the non- applicable individuals to ease the negative effect of the dispersal by organizing differently the cells (units) of a data cube. Our purpose consists in building a new space of representation, grouping together all the cells (units) for a better analysis and operation of the data.

Keywords: Data cube, dispersal of a cube, neural networks, OLAP, visualization.

Les histogrammes pour une fragmentation dynamique dans les entrepôts de données

Hacène Derrar*, Mohamed Ahmed-Nacer*, Omar Boussaid**

* Laboratoire LSI, Département Informatique, USTHB Alger
Bp 32 El Alia, bab Ezzouar/ Ager
Hderrar@yahoo.fr, Anacer@mail.cerist.dz
** Laboratoire ERIC, Université Lyon2
Campus Porte des Alpes, 69676 Bron Cedex
Omar.Boussaid@univ-lyon2.fr

Résumé. La conception d'un schéma de fragmentation optimal se base sur des techniques d'analyse des informations statistiques des requêtes les plus fréquentes. Ces techniques, conçues initialement pour les bases de données transactionnelles, permettent de concevoir des approches de fragmentation statiques sans tenir compte des changements survenus sur les informations relatives à l'exploitation des données. Cependant, dans le contexte des entrepôts de données, où la charge de travail est en perpétuel changement, de telles techniques deviennent inefficaces. Dans cet article, nous proposons une approche d'exploitation des statistiques des accès aux données basée sur les histogrammes pour une fragmentation dynamique des entrepôts de données.

1 Introduction

Les entrepôts de données se caractérisent par un volume de plus en plus important et des requêtes d'interrogation de plus en plus complexes. Pour une meilleure exploitation de cette masse de données, plusieurs techniques d'optimisation ont été proposées et adaptées. On peut citer les vues matérialisées (Chuan Zhang et Xin Yao, 2001), les index (Chaudhuri, 2004), la fragmentation des données (Bellatreche et al., 2005; Noaman et Barker, 1999), le groupement (Jagadish et al., 1999), le traitement distribué (Bernadiro et al,. 2002) et le traitement parallèle (Furtado, 2004).

En effet, le fait de partitionner les tables, les index et les vues matérialisées en fragments stockés et consultés séparément apporte des améliorations considérables en termes de gestion des données et de coût d'exécution des requêtes.

La technique de fragmentation des données a été proposée par Eswaran (1974), ce qui a donné par la suite naissance à diverses approches de fragmentation, en l'occurrence l'approche verticale (Navathe et al., 1984), horizontale (Ceri et al., 1982) et mixte (Sacco, 1986; Zhang et Orlowska, 1994).

Ces approches sont conçues à partir d'une analyse statistique des requêtes les plus fréquentes en se basant sur des informations tant qualitatives que quantitatives. Deux types

d'algorithme sont généralement utilisés : les algorithmes dirigés par la complétude et la minimalité des prédicats (Noaman et Barker, 1999) et les algorithmes dirigés par l'affinité (Navathe et al., 1989).

Ces algorithmes de fragmentation sont des algorithmes statiques. Ils se basent dans leurs entrées sur des informations statistiques recueillies à partir de l'exploitation des données. Si un changement intervient dans les entrées de ces algorithmes, ces derniers doivent être réexécutés afin de déterminer un nouveau schéma de fragmentation. De plus, ces algorithmes s'appuient sur le principe de groupement (clustering) qui est considéré comme étant un problème combinatoire qui nécessite, pour sa résolution, à faire appel à des heuristiques (Bellatreche, 2000). Ainsi, nous pouvons constater qu'en cas d'évolution des modèles et/ou des changements de la charge de travail ces algorithmes deviennent très complexes voire impraticables.

Dans le cadre des bases de données relationnelles ou objets et quelque soit l'environnement (centralisé, parallèle, réparti) une grande partie de la littérature a traité cette problématique. Les travaux se sont focalisés sur les techniques de redistribution des données ou de réallocation des fragments en cas de détérioration des performances. Dans ce contexte, on a considéré que la solution réside au niveau physique en appliquant des stratégies d'équilibrage de charge des traitements et des données entres les nœuds. Le volet logique, à savoir la conception de l'approche de fragmentation, elle-même, demeure adapté étant donné que la charge de travail est pratiquement stable.

A l'inverse, l'évolution du schéma et de la charge de travail, dans le contexte des entrepôts de données, est dynamique. Ceci est dû à l'évolution perpétuelle des systèmes, des applications et des données et plus particulièrement aux caractéristiques spécifiques des requêtes décisionnelles. Cette évolution engendre non seulement une évolution dans le modèle et des structures redondantes, dans le cas de la maintenance des vues matérialisées, mais aussi une évolution de la stratégie d'optimisation elle-même (Cécile et al., 2008).

De ce fait, l'adaptation de ces approches de fragmentation aux entrepôts de données s'avère délicate. Pour une exploitation efficiente de ces techniques d'optimisation, il ne s'agit pas seulement d'analyser les fréquences d'accès aux données pour choisir un schéma de fragmentation optimal, mais de rendre ce choix dynamique et adapté aux changements de la charge de travail.

Notre approche à ce problème porte sur une réfragmentation des données en se basant sur des informations statistiques récentes. Nous proposons, dans ce papier, un algorithme itératif basé sur la sélectivité des données et nous introduisons pour sa mise œuvre l'utilisation des histogrammes. Le choix de cette structure de données pour la sauvegarde des informations statistiques est justifié par le fait que les histogrammes :

- Peuvent contenir assez d'informations issues des modèles des accès ;
- Permettent un traitement efficace des mises à jour, sachant que dans les entrepôts de données ces dernières seront très fréquentes ;
- Ne nécessitent pas beaucoup d'espace mémoire ;
- Permettent de manipuler avec souplesse les informations conservées, c'est-à-dire permettent de supprimer facilement l'ancienne historique et de garder uniquement le nouveau historique ;
- Sont faciles à implémenter.

Cet article est organisé comme suit, dans la deuxième section, nous présentons un état de l'art sur les approches de fragmentation et la notion d'évolution de la charge de travail. Dans la section 3, nous décrivons d'une manière générale le problème généré par les techniques de conception statiques des approches de fragmentation et nous positionnons notre approche. Dans la section 4, nous présentons notre approche et nous formalisons le problème. Dans la sections 5, nous décrivons l'algorithme de réfragmentation des données ainsi que les opérations de manipulation des histogrammes. Dans la section 6 nous présentons les premiers résultats expérimentaux obtenus en utilisant le benchmark APB-1 release II sous Oracle10g. Enfin, nous terminons cet article par une conclusion et des perspectives.

2 Etat de l'art

Dans le contexte des entrepôts de données, plusieurs travaux ont été menés pour une meilleure adaptation et exploitation des techniques de fragmentation. On cite plus particulièrement les travaux de Bellatreche et al. (2006), qui ont proposé une technique pour partitionner la table de faits en fonction des schémas de fragmentation des tables de dimension. Les auteurs considèrent que l'application du partitionnement horizontal aux entrepôts de données rend l'espace de recherche très important pour la sélection du schéma de fragmentation et peut générer un nombre important de partitions qui sera difficile à gérer. Pour remédier à ces problèmes, les auteurs formalisent le problème de sélection de schéma de fragmentation comme un problème d'optimisation et proposent pour sa résolution une approche qui combine un algorithme génétique et un recuit simulé.

Dans le cadre des entrepôt de données distribuées, Bernardino et al., considèrent que la table de faits est généralement de taille importante et les tables de dimension plus petites. Ils proposent un algorithme basé sur la technique du DWS (Data Warehouse Striping) pour la distribution d'un grand entrepôt de données à travers un cluster de PC. Les tables de dimensions sont répliquées dans chaque nœud et les données de la table de faits sont distribuées à travers tous les nœuds en utilisant la fragmentation horizontale selon une méthode circulaire (Round robin) (Bernadiro et al. , 2002).

S'agissant des travaux proches de notre problématique, Alexandre et al. (2004) proposent pour le traitement efficace des requêtes OLAP une approche appelée partitionnement adaptative virtuelle qui permet d'ajuster dynamiquement les tailles des partitions, sans faire appel aux informations relatives à la base de données. D'autres travaux ont proposé l'automatisation de la conception logique de la fragmentation des données. Ainsi, Papadomanolakis et Ailamaki (2004) ont proposé l'automatisation de la phase de conception logique de la fragmentation à partir des informations récentes sur l'exploitation des données. Leur outil dénommé AutoPart, fragmente les tables selon une charge de travail actualisée. AutoPart reçoit en entrée les informations de la charge de travail et conçoit un nouveau schéma de fragmentation selon une approche verticale. Dans le même contexte, Stöhr (2001) propose, pour l'automatisation de la fragmentation et le placement des données d'un entrepôt dans un environnement parallèle, un outil dénommé WARLOCK. Son principe consiste à déterminer une allocation des disques qui optimise les entrées sorties lors des accès à la table des faits et ce par l'utilisation du traitement parallèle des requêtes. D'autre part, Karahoca et al (2002) ont proposé un algorithme non linéaire basé

sur les réseaux de neurones afin de détecter automatiquement l'approche de fragmentation la plus adaptée à une base de données distribuée.

3 Positionnement

Les algorithmes de conception d'un schéma de fragmentation optimal, développés dans le contexte relationnelle et objet, se déclinent en deux catégories : 1) les algorithmes basés sur la minimalité et la complétudes des prédicats (Ceri et al., 1982), et 2) les algorithmes dirigé par l'affinité des prédicats (Navathe et al.,1989).

Ces algorithmes s'appuient sur le principe de groupement (clustering) en se basant sur une analyse statistique des requêtes fréquentes afin de déterminer un schéma de fragmentation optimal.

Ainsi, les algorithmes dirigés par la complétude et la minimalité des prédicats nécessitent un calcul combinatoire des probabilités d'accès. Dans le contexte des entrepôts de données, ces algorithmes deviennent pratiquement inapplicables du fait que les requêtes décisionnelles possèdent un grand nombre de prédicats de sélection. En ce qui concerne les algorithmes dirigés par l'affinité, leur inconvénient réside dans le fait que la matrice d'affinité est exprimée uniquement ente des paires de prédicats. De plus, ces algorithmes sont statiques, si un changement intervient dans les entrées de ces algorithmes, ces derniers doivent être réexécutés (Bellatreche, 2000).

L'adaptation de ces algorithmes aux entrepôts de données, s'avère donc plus délicate en raison de l'évolution du modèle de données et plus particulièrement à la nature des requêtes analytiques. Ces requêtes sont longues, complexes et nécessitent parfois un grand nombre d'opération de sélection, de jointure et d'agrégat et peuvent manipuler des centaines voir des milliers de tuples. Les requêtes analytiques sont extrêmement variables, elles sont généralement composées en interactif dont le code n'est pas connu à l'avance et peuvent être exécutées une ou plusieurs fois (Derrar et al., 2008). Ce type de requête appelé aussi requêtes ad hoc correspond à des requêtes saisies en ligne sans une longue réflexion préalable. La « dynamicité » de ce type de requête rend, avec le temps, le schéma de fragmentation inapproprié du fait qu'il a été conçu à partir des informations statistiques instables.

Par ailleurs, l'aspect automatisation des phases de conception logique des entrepôts de données qui a été considéré, dans certain travaux, comme étant une solution à l'évolution de la charge de travail c'est focalisé, à l'instar de ce qui a été fait dans les bases de données transactionnelles, sur les algorithmes relatifs à la distribution de données pour un meilleur équilibrage de charges et ce en gardant toujours le même schéma de fragmentation des données.

De ce fait, la solution consiste, à notre sens, non pas à rechercher de nouvelles approches spécifiques pour les entrepôts de données ou d'une adaptation pure et simple des algorithmes de fragmentation existants, mais de transformer les algorithmes statiques de fragmentation des données en algorithmes dynamiques permettant, en cas de changement de la charge de travail, de réfragmenter les données de l'entrepôt. Pour ce faire, nous proposons un algorithme de type itératif basé sur l'observation des accès et la sélectivité

des données pour déterminer un schéma de fragmentation selon un historique récent des accès.

4 La réfragmentation des données : notre approche

Notre approche de réfragmentation des données s'effectue en trois phases :

1- fragmenter la table des faits selon une approche horizontale;

2- mise en ouvre d'un modèle d'observation des accès aux différents fragments ;

3- réfragmenter la table des faits sur la base des statistiques recueillies.

Le principe de notre approche se base sur l'observation et l'enregistrement des informations statistiques relatives aux accès aux données. La sauvegarde de ces informations s'effectue d'une manière continue, en écartant périodiquement les anciennes informations pour que la réfragmentation ne porte que sur des informations récentes.

Les composants essentiels de notre approche sont : le critère d'évaluation et la structure permettant de conserver et d'observer les accès aux données.

Le critère d'évaluation permet d'estimer l'opportunité d'une exécution ou non de la réfragmentation. Pour une raison d'adaptabilité, ce critère pourra être la minimisation d'une fonction objective qui portera sur le cout d'accès aux données, le coût de transfert des données entre fragments ou dans un contexte distribué sur le coût de communication. Il peut être également une valeur d'un seuil alloué au temps d'exécution d'une requête décisionnelle.

En ce qui concerne la conservation et l'exploitation des informations statistiques, elle est réalisée par l'utilisation des histogrammes. Ces derniers sont un moyen flexible pour la construction de structures sommaire pour les grandes base de données. Leur utilité a été vérifiée dans de nombreux domaines, tel que l'optmitisation des requêtes (Bruno et al., 2002), les réponses approchées aux requêtes (*Approximative Query Answering*)(Acharya et al., 1999), l'estimation des tailles des vues (Hass et al., 1995) et dans les bases de données distribuées (Donjerkovic et al., 2000). Dans le cadre des entrepôts de données, les histogrammes ont été principalement utilisés pour l'optimisation des requêtes OLAP (V.poosala et al. 1999) ou pour donner des réponses rapprochées aux requêtes complexes (Todd EAvis et al., 2008).

D'une manière générale, le principe de base de l'utilisation des histogrammes dans le cadre de l'optimisation des requêtes repose sur la connaissance des statistiques portant sur les objets manipulés. Les histogrammes permettent d'offrir une vision réelle de la distribution des données d'une colonne par une meilleure estimation de leurs sélectivité. Ces estimations sont utilisées par l'optimiseur pour évaluer les plan de requêtes.

Exemple de l'utilisation des histogrammes pour l'estimation des cardinalités. Considérant la requête suivante : *Select * from R where R.a < 20* et supposant qu'on a un histogramme en *R.a*. Pour estimer la cardinalité de chaque requête, on considère, alternativement, tous les buckets (appelé également groupe, plage ou fraction) de l'histogramme qui sont complètement ou partiellement couvert par le prédicat et on agrège par la suite tout les résultats intermédiaires. Cette procédure est illustrée ci-dessous :

Considérons quatre buckets de l'histogramme sur le l'attribut *R.a*. le bucket b_1, par exemple, couvre $0 \leq x < 10$ et sa fréquence est 100 (qui représente 100 tuples dans

l'ensemble de données). De la même manière, les buckets b_2, b_3 et b_4, représentent respectivement 50, 80 et 100 tuples. Supposons qu'on souhaite estimer la cardinalité du prédicat $P=R.a < 20$. Etant donné que P est totalement inclus dans le bucket b_1, on peut garantir que les 100 tuples de b_1 vérifie le prédicat P. De plus, P est disjoint du groupe b_3 et b_4 , de ce fait, il n'existe aucun tuple dans b_3 et b_4 vérifiant le prédicat P. Et enfin, le prédicat P est partiellement couvert par le bucket b_2 (P est vérifié par 50 % des valeurs distinctes uniformément propagé dans b_2. on utilisant l'hypothèse, citée si-dessus, on estime que 50% des tuples de b_2 vérifies P. en conclusion, le nombre de tuples vérifiant le prédicat $P = R.a < 20$ est estimé à : 100+50/2 = 125.

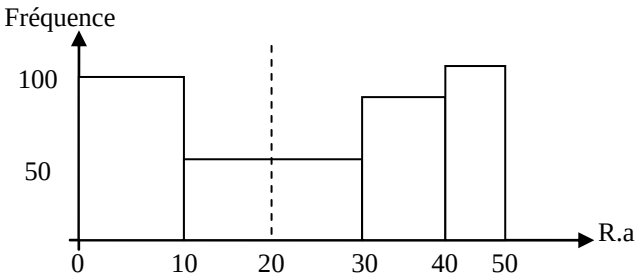


FIG. 1- Estimation de la sélectivité d'une requête utilisant les histogrammes.

Les histogrammes sont utilisés par la plupart des SGBDs commerciaux. Ainsi, dans la version 10g du SGBD Oracle, c'est le package *DBM_STAS*, composé de plusieurs procédures telle que : *GATHER_TABLES_STAT, GATHER_SHEMA_STAT, INDEX_STAT*, qui permet d'avoir des statistiques sur les objets de la base de données. Les statistiques qui portent sur une colonne d'une table sont stockées sous forme d'histogrammes. Oracle utilise deux types d'histogrammes : les histogrammes de hauteur équilibrée (*height-balanced histograms*) et les histogrammes de fréquence (*frequency histograms*).

4.1 Formulation

4.1.1 L'approche de fragmentation

Considérons un entrepôt de données constitué d'un schéma en étoile avec une table des faits F reliée à d tables de dimensions $\{d_1, \ldots, d_d\}$. Pour des raisons de simplification, on fragmente uniquement la table des faits F selon une approche horizontale par intervalle. Le fragment i de la table des faits est noté f_i. Soit D le domaine de valeur de l'attribut de fragmentation. Chaque fragment couvre un intervalle du domaine de l'attribut, qu'on va le dénommer Domaine de la Valeur du Fragment *DVF*. Le *DVF* d'un fragment f_i est : $DVF(f_i) = f_i\ [min_i, max_i]$. Aussi, deux fragments f_i , f_j sont adjacents si leur *DVF* sont contigus, c'est-à-dire :

$$Adj\left(f_i, f_j\right) \Longrightarrow max_i = min_j \vee max_j = min_i$$

D'autre part, deux ou plusieurs fragments adjacents peuvent être regroupés en un nouveau fragment f_{nouv} si le *DVF* du nouveau fragment égale à la somme des *DVF* des fragments précédents :

$$f_{nouv} = \cup_{i=1}^{n} f_i$$
$$\forall f_i \in \{f_1, \dots, f_n\} \exists \left(f_j \in \{f_1, \dots, f_n\} \right) Adj\,(f_i, f_j)$$

4.1.2 Les histogrammes

Un histogramme sur l'attribut de fragmentation se compose d'un ensemble de buckets. Chaque bucket b_k correspond à un fragment, c'est-à-dire chaque bucket couvre un domaine de valeur de fragment $DVF(f_i)$. Dans notre approche nous avons opté pour l'utilisation de l'histogramme de largeur égale (*equi-width histogram*). Ces derniers, sont plus simple à accéder et à mettre en œuvre. La distribution des données pour ce type d'histogramme est uniforme. Les buckets ont la même taille, ce qui cadre parfaitement avec l'approche de fragmentation que nous avons utilisée à savoir l'approche horizontale par intervalle dans laquelle les fragments ont pratiquement la même taille. De plus, ceci permettra de faciliter les opérations sur les deux type d'histogrammes utilisés, pour stockés les anciens et les récents historiques des accès, étant donné qu'ils sont bien alignés.

Pour la manipulation des histogrammes, nous utilisons les notations suivantes : l'histogramme est noté H_i contenant b_k buckets le nombre de ces buckets est $B_i[b_k]$. T_b est la taille du bucket. Elle correspond au nombre de ligne que couvre le bucket. La limite des valeurs des intervalles se débute et termine sur un multiple de T_b. Cela signifie que la valeur de l'intervalle couverte par le bucket $B_i[b_k]$ est :

$$[b_k * T_b\ , (b_{k+1}) * T_b]$$

Afin de limiter l'usage de la mémoire, il existe un nombre maximum MAX_B de buckets à sauvegarder.

Pour permettre de prendre en compte, lors de la réfragmentation des données, uniquement l'historique récent, on utilise deux jeux d'histogrammes : l'ancien histogramme (H_{anc}) et le récent histogramme (H_{rec}). Toutes les informations sont enregistrées dans l'histogramme récent.

5 Algorithme de réfragmentation des données

L'exploitation de l'entrepôt des données se traduit par une séquence d'accès des requêtes. Le coût d'accès à un tuple à l'instant *t* dépend du schéma de fragmentation Sf_t.

Le but de l'algorithme de réfragmentation est de déterminer un schéma de fragmentation optimal en se basant sur l'historique récent des accès aux données.

D'une manière formelle, on suppose un schéma de fragmentation Sf qui regroupe un ensemble de fragment f_i avec $DVF\,(f_i) = [\,min_i, max_i]$. Il s'agit de déterminer un nouveau schéma de fragmentation Sf_{nouv} optimal composé d'un ensemble de fragments $f_m, \dots, f_n$, tel que $\cup f_m, \dots, f_n = F$.

L'algorithme est exécuté à un intervalle de temps régulier. Selon le critère d'évaluation de la réfragmentation, le résultat sera soit : de garder le schéma de fragmentation tel qu'il est; ou bien déterminer un nouveau schéma de fragmentation Sf_{nouv} avec $DVF\,(f_{nouv}) = f_{nouv}\,[\,min_{nouv}, max_{nouv}]$.

Dans un premier temps, les deux histogrammes, récent et ancien, sont construits sur la valeur de l'attribut de fragmentation. A chaque accès aux données d'un fragment le bucket correspondant est mis à jour par l'exécution de la fonction *HistogramUpdate()* décrite dans la section 5.1.1. A un instant donné, le critère d'évaluation est estimé. Si ce critère n'est pas satisfait, on garde toujours le schéma de fragmentation actuel. Sinon, la réfragmentation est exécutée et on aura un nouveau schéma de fragmentation composé de nouveaux fragments dont le nombre ne devra pas dépasser le seuil MAX_B.

A l'issue d'une réfragmentation des données, la fonction *HistogramOrganize()*, décrite dans la section 5.1.2, est exécutée pour redimensionner les buckets des deux histogrammes selon le nouveau schéma de fragmentation.

Algorithme 1 : Réfragmentation

Entrée *: - une table des faits fragmentée selon un schéma de fragmentation SF.*
- Deux histogrammes : $H_{anc}[F]$, $H_{rec}[F]$

Sortie *: un schéma de fragmentation optimal SF_{opt}*

Début
Pour tout $B_i[b_k] \in H_{rec,}$ **faire**
HistogramUpdate () // calcul de la sélectivité //
Si *critère d'évaluation est satisfait* **Alors**
Réfragmentation de F
HistogramRorganize() // réorganisation et redimensionnement des bcukets//
Fin si
Fin pour
End

5.1 Opérations sur les histogrammes

5.1.1 Mise à jour des histogrammes

A chaque accès aux données, la sélectivité du bucket concerné est mise jour. Supposons un fragment f_i avec $DVF(f_i) = f_i[min_i, max_i]$ et un tuple t_j et v_j la valeur qui correspond au numéro de la ligne de ce tuple dans la table des faits. Etant donné que l'intervalle de valeurs des buckets b_k est :

$$[b_k * T_b , (b_{k+1}) * T_b]$$

Le traitement consiste alors à déterminer le bucket b_k qui regroupe le tuple qui a été accédé selon la valeur v_j et d'incrémenter son nombre d'accès noté $SEL[b_k]$. La formule est alors : $b_k = v_j / T_b$.

Exemple. Supposons que T_b =5 et la valeur v_j = 30, alors le bucket concerné par la sélectivité est : 30/5 = 6 (cette valeur est toujours arrondie).

5.1.2 Réorganisation et redimensionnement des l'histogrammes

A l'issue d'une réfragmentation des données et dans le cas ou on veut garder le nombre et la taille des anciens fragments inchangé, c'est-à-dire $DVF(f_i) = f_i\ [min_i, max_i] = DVF(f_{nouv}) = f_{nouv}\ [min_{nouv}, max_{nouv}]$, alors la fonction de réorganisation des histogrammes *HistogramReorganize()* consistera uniquement à altérer les deux jeux d'histogramme, c'est-à-dire, l'ancien devient récent et le récent jeu est mis à zéro (vidé). Avec :

$$H_{anc}[F] \leftarrow H_{rec}[F] \text{ et } H_{rec}[F] \leftarrow 0.$$

Dans le cas $DVF(f_{nouv}) \neq DVF(f_i)$ alors le nombre de buckets et leur taille pourra diminuer ou augmenter selon le nombre de buckets maximum MAX_B. Ceci par l'utilisation d'un facteur Z_T permettant de modifier la taille du bucket. Ce qui permettra de s'assurer que le nombre de buckets correspond toujours au nombre de fragments et ce pour une meilleure capture de l'historique des accès aux données des fragments.

5.1.3 Etude de complexité de l'algorithme de réfragmentation des données

Pour évaluer notre algorithme de réfragmentation, nous avons procédé par l'étude de complexité des différentes phase qui le compose, il en ressort que :

- Le traitement de la de mise à jour des histogrammes est O(1) donc une complexité constante. Les histogrammes sont gardés en mémoire principale, cette mise à jours n'engendre aucun accès disque.
- L'évaluation de la taille des fragments, pour le redimensionnement des histogrammes, est de complexité polynomiale $O(n^2)$ ou n représente le nombre de buckets.
- La détermination d'un nouveau schéma de fragmentation est de complexité polynomiale.

6 Etude expérimentale : premiers résultats

Pour l'évaluation de notre approche, nous avons utilisé le benchmark APB-1 release II Council (1998) implémenté sous Oracle 10g. Ce benchmark, utilise un schéma en étoile composé de quatre tables de dimensions (Prodlevel de 9000 tuples, Custlevel de 900 tuples, Timelevel de 24 tuples et Chanlevel de 9 tuples) et une table de faits (Actvars de 24 000 000 tuples). Pour le calcul des temps d'exécution des requêtes, nous avons utilisé l'utilitaire Aqua Data Studio 2.0.7. Pour mener nos tests, nous avons utilisé un ensemble de 50 requêtes englobant différents opérateurs : opérations de jointure, de sélection et des fonctions de calcul et d'agrégations (SUM, COUNT, AVG, MIN, MAX).

La première phase de nos tests a porté sur la fragmentation de la table des faits selon une approche horizontale par intervalle sur l'attribut *TIME_LEVEL* en utilisant l'instruction *SQL PARTITION BY RANGE (TIME_LEVEL).* Nous avons par la suite procéder à la

création des histogrammes et ce en utilisant les procédures de DBMS_STATS.GATHE, décrite ci-dessous.

```
BEGIN
   DBMS_STATS.GATHER_TABLE_STATS
   (OWNAME=>'OE',TNAME='ACTAVRS',
    METHOD_OPT=>'FOR COLUMNS SIZE 10 TIME_LEVEL');
END
```

Pour des raisons de simplification de nos tests, nous avons choisi le critère d'évaluation de réfragmentation des données, l'uniformité en terme des fréquences des accès aux données, toute en gardant une distribution uniforme des données entre fragments. De plus, après chaque réfragmentation les deux types d'histogramme ancien et récent sont recrées. Nous n'avons pas pris en considération les procédures de réorganisation des histogrammes.

Conformément au critère d'évaluation choisi, nous avons procédé à la réfragmentation des données selon une approche par hachage.

Les expérimentations ont été menées afin de mesurer les consommations en terme de temps et d'espace mémoire, lors de l'exploitation des statistique et la réfragmentation des données. Les premiers résultats obtenus sont satisfaisant en terme d'espace mémoire étant données que les statistiques sont directement extraites du dictionnaire données et la manipulation des histogrammes ne nécessite pas d'espace mémoire. Pour tester la validité de notre approche de réfragmentation nous avons utilisée une fonction hachage classique $t_i \ mod \ B_i[b_k]$ avec t_i le $i^{ème}$ tuple et $B_i[b_k]$ le nombre de buckets. Cette fonction permet de garder le même nombre de fragments que l'ancien schéma de fragmentation avec un temps d'exécution acceptable.

Cependant, si cette fonction de hachage permet d'avoir une uniformité entre les fragments en terme de nombre de tuples, elle n'assure pas toujours une redistribution uniforme des données selon leurs fréquence d'accès.

7 Conclusion et perspectives

Dans cet article nous avons présenté une approche de réfragmentation des données basée sur l'utilisation des histogrammes. Cette approche consiste d'abord à fragmenter la table des faits selon une approche horizontale par intervalle et de procéder par la suite à l'observation des fréquences des accès des requêtes aux données. La sauvegarde des informations statistiques s'effectue par l'utilisation des histogrammes. Ces derniers sont une structure de données et un moyen flexible permettant l'estimation des sélectivités des données des différents fragments. Pour déclencher la réfragmentation des données nous avons pris comme critère d'évaluation l'uniformité des fréquences des accès entre les fragments. Ainsi, la table des faits est réfragmentée horizontalement selon une approche par hachage. Le choix de cette approche est justifié par le fait qu'une définition d'une fonction de hachage adaptée pourra assurer une uniformité des données entre fragments selon leur fréquence d'accès.

Pour nos travaux futurs, nous envisageons de continuer les tests expérimentaux qui porteront sur la définition d'une fonction de hachage plus adaptée et sur la dynamicité de notre approche par l'application des procédures relatives à la réorganisation des

histogrammes. De plus, ces tests porteront sur la variation du critère d'évaluation de la réfragmentation en vue de permettre de déterminer le temps minimal de réfragmentation des données. Nous envisageons également, d'appliquer notre approche aux algorithmes de fragmentation basés sur la minimalité et la complétude par la création des histogrammes sur les prédicats.

Références

Acharya.S, P.Gibbons, V.Pooaala and S. Ramaswamy. (1999). *The Aqua Approximate Query Answering System*. Proceedings of ACM SIGMOD Philladephia PA, pages 574-578, june 1999.

Alexandre A. B. Lima, M.Mattoso, and P.Valduriez. (2004). *Adaptive Virtual partitioning for OLAP Query Processing in a Database Cluster*. In 19 th proceeding on SBBD, Brazil, 2004.

Bellatreche,L., K.B. (2005). An Evolutionary Approch to Schema Partitioning Selection in a Data Warehouse Environment, Proceeding of the International Conference on Data Warehousing and Knowledge Discovery (DAWAK'05).

Bellatreche,L., Boukhelfa.K, H.I.Abdalla.Sage, (2006). A combinaison of genetic and simulated annealing algorithms for physical data warehouse design. In 23rd British National Conference on Database.

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

Bernardino,J., P.Furtado, P, H. Madeira (2002). Approximate Query Answering Using Data Warehouse Striping. Journal of Intelligent Information Systems- Integrating Artificial Intelligence and Database Technologies, Volume 19, Issue 2, Elsevier Science Publication.

Bruno.N and S. Chaudhuri. (2002). Exploiting statistics on query expressions for optimization. ACM SIGMOD'2002, june 4-6, Madison, Wisconsin, USA.

Cecile. F F. Bentayeb, O. Boussaid, *Maintenance de charge pour l'optimisation des entrepôts de données évolutifs : aide à l'administrateur*. 8èmes journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA 08).

Ceri.S, M.Negri, and G.Pelagatti. (1982). *Horizontal data partitioning in data base design*.Proceeding of the ACM SIGMOD, International Conference on on Management of Data. SIGPLAN Notices, page 128-136, 1982.

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

Chuan Zhang, Xin Yao. (2001).*An Evolutionary Approach to Materialized Views Selection in a Data Warehouse Environment*. IEEE Transactions on Systems, Man, and Cybernetics, part c: applications and reviews, vol. 31, no. 3, august 2001.

Derrar.H, O.Boussaid, M.A.Nacer. (2008). Une approche de répartition des données d'un entrepôt basé sur l'Optimisation par Essaim Particulaire. 8èmes journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA 08).

Donjerkovic.D, Y. Ioannidis, and R. Ramakrishnan. (2000). *Dynamic histograms: Capturing evolving data sets*. In Proceedings of ICDE, 2000.

Eavis.T, G.Dimitrov, I.Dimitrov, D.Cueva, A.Lopez, and A.Taleb. (2008). *Parallel OLAP with the Sidera server*, Science direct, October 2008.

Eswaran, K.P. (1974). *Placement of Records in a File and File Allocation in a Computer Network*. Proceedings of IFIP Congress on Information Processing, Stockholm, Sweden, pp: 304-307.

Furtado.P, 'E Experimental Evidence on Partitioning in Parallel Data Warehouses, DOLAP'04, November 12–13, 2004, Washington, DC, USA.

Haas.P, J.Naughton, S.Seshadri, S, and L.Stokes. (1995). *Sampling-based estimation of the number of distinct values of an attribute*. In VLDB'95, pages 311–322.

Jagadish, H., L. V. S. Lakshmanan, et D. Srivastava (1999). *Snakes and sandwiches : Optimal clustering strategies for a data warehouse*. Proceedings of the ACM SIGMOD International Conference on Management of Data, 37–48.

Karahoca.A, N.Osman, and D. Erkan. (2002). *Random Neural Network Approach in Distributed Database Management Systems*. Technical report (2002).

Navathe,S., S. Ceri, G.Wierhold, et J.Dou (1984). Vertical Partitioning Algorithms for Database Design. ACM Transactions on Database Systems, Vol. 9, No. 4, Decembre 1984, pages 680-710.

Navathe,B., M. Ra (1989). Vertical Partitioning for Database Design: A Graphical Algorithm. ACM SIGMOD International Conference on Management of Data, 1989, pp. 44-450, Conference on Management of Data, 1989, .pp. 44-450,

Noaman, A. Y. et K. Barker (1999). *A horizontal fragmentation algorithm for the fact relation in a distributed data warehouse*. In the 8th International Conference on Information and Knowledge Management (CIKM'99), 154–161.

Papadomanolakis.S, A.Ailamaki, (2004). *AutoPart*: *Automating Schema Design for Large Scientific Databases Using Data Partitioning*. Scientific and Statistical Database Management, 2004

Poosala.V V.Ganti. (1999). *Fast approximate answers to aggregate queries on a data cube.*In Proceedings of the 11th International Conference on Statistical and Scientific Database Management, Cleveland, OH, USA, July 1999, pp.24–33.

Sacco. G.(1986). *Fragmentation: A Technique for Efficient Query Processing*. ACM Transactionon Database Systems, 11: 113-133.

Stöhr. E.R. (2001), *WARLOCK : A Data Allocation Tool for Parallel Warehouses*. Proceedings of the 27th VLDB Conference, Roma, Italy, 2001.

Zhang, Y. and M.E. Orlowska, (1994). *On Fragmentation Approaches for Distributed Database Design*. Information Sci., 1: 117-132.

Summary

The conception of an optimal fragmentation schema bases on statistical analysis of most frequent requests. These techniques allow to conceive a static fragmentation approaches without taking into account changes on information relative to the data exploitation. However, in the data warehouses context, the former fragmentation techniques become ineffective. The dynamicity which characterizes the workload requires to revise conception phases of fragmentation by making it dynamic and adaptable to the data warehouses context. We propose in this article an approach for exploiting the data accesses statistics based on histograms for a dynamic fragmentation.

Sécurisation des entrepôts de données : Etat de l'art et proposition d'une architecture

Salah TRIKI, Jamel FEKI, Hanene BEN-ABDALLAH, Nouria HARBI

Laboratoire Mir@cl
Département d'Informatique, Faculté des Sciences Economiques et de Gestion de Sfax,
Route de l'Aéroport Km 4 – 3018 Sfax, BP. 1088

{Salah.Triki, Jamel.Feki, Hanene.BenAbdallah}@fsegs.rnu.tn, Nouria.Harbi@univ-lyon2.fr

Résumé. Les entrepôts de données intègrent des données provenant de sources hétérogènes et sont utilisés par les dirigeants pour prendre des décisions stratégiques. Etant souvent propriétaires, ces données peuvent être sensibles et doivent être contrôlées à l'accès d'où la nécessité de leur sécurisation. Dans cet article, nous présentons d'abord une synthèse des travaux de recherche relatifs à la sécurité des entrepôts de données, ensuite nous exposons les grandes lignes d'une proposition pour leur sécurisation.

1 Introduction et Motivations

Les entrepôts de données sont alimentés par plusieurs sources de données qui peuvent être hétérogènes ; ils permettent aux utilisateurs décisionnels d'orienter leurs requêtes vers une seule cible, c'est-à-dire un seul espace de stockage. Cela évite de gérer l'hétérogénéité des sources au moment de l'expression et l'évaluation des requêtes. En collectant et consolidant les données de sources différentes**,** les entrepôts de données permettent aux dirigeants de prendre des décisions stratégiques et d'établir des prévisions. En aval des entrepôts, des extraits orientés sujet et réorganisés selon un modèle multidimensionnel sont construits. Ces extraits sont dits des magasins de données ; ils visent à faciliter les opérations d'analyse décisionnelles. Des outils dédiés du marché d'entreposage de données (« Data Warehousing ») offrent de nombreuses opérations pour les traitements analytiques en ligne (OLAP : « On-Line Analytical Processing »). Les entrepôts de données occupent ainsi une place centrale dans les systèmes d'information décisionnels des organisations.
Les entrepôts de données visent à avoir une vue commune de l'ensemble des données du système opérationnel, permettant ainsi la prise de décision. Cependant ils créent un conflit (Ralph K. (1997)). D'une part, les entrepôts de données doivent permettre un accès facile aux données et, d'autre part, les organisations doivent s'assurer que ces données ne sont pas divulguées sans contrôle. En effet certaines données sont personnelles et peuvent porter préjudice à leurs propriétaires quand elles sont divulguées comme, par exemple, les données médicales, les croyances religieuses ou idéologiques (Eduardo F. et al (2006)). Ainsi, plusieurs gouvernements ont promulgué des lois pour la protection des vies privées de leurs citoyens. Parmi ces lois, HIPPA (« Health Insurance Portability and Accountability Act » HHS (1996)) vise à protéger les données médicales des patients américains en obligeant les

établissements du secteur des soins de la santé à suivre des règles de sécurité strictes. De même, GLBA (« Gramm-Leach-Bliley Act » GPO (1999)) oblige les organismes financiers américains à protéger les données de leurs clients ; quant à Safe Harbor (Export (2008)) permet aux entreprises s'y conformant de transférer et d'utiliser les données concernant les internautes européens ; Sarbanes-Oxley (Soxlaw (2002)) garantit la fiabilité des données financières des entreprises. Les organismes doivent utiliser des règles de sécurité strictes pour être conformes à ces lois, autrement ils seront sanctionnés. Malgré la présence de ces lois, les aspects de sécurité sont quasiment absents dans les entrepôts de données.

Le présent travail vise deux objectifs. Premièrement, il examine l'état de l'art de la sécurité des entrepôts de données, et deuxièmement propose une solution pour la sécurisation d'un entrepôt.

2 Etat de l'art

L'étude de l'état de l'art des travaux sur la sécurité des entrepôts de données nous a permis d'identifier deux classes d'approches:

- les approches portant sur *la sécurisation des opérations* : ces travaux permettent de répondre aux questions *Qui a le droit d'accès* et *A quoi a-t-il le droit* ?
- les approches portant sur la *prévention contre les problèmes d'inférence* ; elles permettent de répondre à la question *Comment interdire à un utilisateur d'inférer des données protégées à partir des données accessibles* ?

Bien évidemment, les deux classes d'approches se complémentent dans les services de sécurité qu'elles offrent. Dans la suite de cette section, nous exposons les caractéristiques de chacune de ces deux classes afin de dégager leurs points forts et leurs insuffisances. Ensuite, nous enchainons sur une proposition pour la sécurisation des entrepôts de données.

2.1 Approches de sécurisation des opérations

Priebe et Pernul (2000) proposent une méthodologie multi-phases pour la conception de la sécurité des entrepôts de données. Les phases de la méthodologie sont : analyse préliminaire, conception, modélisation logique, modélisation physique et implantation. En se focalisant sur la phase d'analyse préliminaire les auteurs définissent deux catégories de besoins en sécurité, *les besoins basics* et *les besoins avancés*. Les besoins basics consistent à cacher un cube, les faces d'un cube, les détails des données, et/ou les dimensions. Tandis que les besoins avancés consistent à cacher les détails de certaines faces d'un cube, et/ou définir des règles de sécurité dépendant des données elles mêmes.
Les besoins définis par les auteurs couvrent toutes les données existantes dans un entrepôt de données, par contre ils n'ont pas proposé une démarche pour leur identification.

Rosenthal et Sciore (2000) proposent une méthode basée sur le langage SQL. Elle permet de sécuriser un entrepôt de données en considérant les règles de sécurité définies par les administrateurs des sources de données. Cette méthode se base sur les trois règles suivantes :

- L'accès à une table nécessite deux autorisations :

  - le droit d'information, accordé par l'administrateur de sources des données : *Qui a accès et à quelle information ?* (Par exemple, le salaire des employés est accessible au président).
  - le droit physique, accordé par l'administrateur de l'entrepôt de données : *Qui a accès et à quelle table physique ?* (Par exemple, les analyste-décideurs ont le droit d'interroger la table des ventes).

  Ainsi un utilisateur n'a le droit d'accéder à une table que s'il possède simultanément ces deux droits, information et physique.
- Si un utilisateur a le droit d'exécuter une requête *Q*, alors cet utilisateur aura le droit d'exécuter toute requête *Q' équivalente à Q*. Ainsi, cette règle permet d'inférer de nouvelles autorisations ; par conséquent, la tâche de l'administrateur sera allégée. Par exemple, si un utilisateur a le droit d'interroger une vue définie entre deux tables *T1* et *T2* mais ne dispose pas du droit d'interroger chacune de ces deux tables, alors il sera autorisé d'exécuter une requête de jointure entre *T1* et *T2*.

La méthode proposée par les auteurs se base sur le langage SQL qui est très répandue et très simple, cependant elle présente les inconvénients suivants :

- La nécessité de récupérer les droits d'information par l'utilisateur de la source. Généralement, ces droits ne sont pas explicitement définis par des commandes LMD (« Langage de Manipulation de Données ») mais sont implicitement exprimés à travers des vues relationnelles d'où la nécessité d'une analyse de toutes les vues.
- La difficulté de faire correspondre une information sécurisée avec son homologue dans l'entrepôt de données surtout si cette dernière a subit une transformation significative (*e.g.*, calcul) lors du processus ETL (« Extract/Transform/Load ») de son chargement dans l'entrepôt de données.

Les entrepôts de données utilisent des modèles de contrôles d'accès. Parmi ces modèles, RBAC (Role-Based Access Control, Sandhu et al. (1996)) qui se base sur le rôle qu'occupe l'utilisateur dans l'organisation pour prendre les décisions d'accès. Les entrepôts de données doivent répliquer les contraintes de sécurité qui existent dans les sources. Si par exemple, les sources de données exigent une contrainte de sécurité qui indique que les analyste-décideurs n'ont accès qu'à la table *Vente* alors une contrainte du même type doit exister dans l'entrepôt de données. Lorsque l'administrateur de l'entrepôt veut augmenter les privilèges des utilisateurs, il faut qu'il informe l'administrateur de la source de données. Celui-ci peut accepter ou refuser. **Thuraisingham et al. (2007)** proposent le modèle E-RBAC (Extended-Role Based Access Control) qui permet de gérer ce cas. C'est une combinaison du modèle RBAC et du modèle de contrôle d'usage UCON (Usage Control Model, Jaehong P. R. S. (2004)). UCON se base sur des conditions des obligations et des droits pour prendre les décisions d'accès. Les obligations assurent le respect des droits d'accès définis dans les sources tandis que les conditions définies à partir des règles de sécurité permettent de gérer les conflits entre ces dernières. Un conflit se produit dans le cas de règles contradictoires, par exemple les analyste-décideurs ont le droit d'accéder à la table *Vente* dans la source *A* mais ne possèdent pas ce droit dans la source *B*.

Le modèle proposé par les auteurs se base sur deux modèles très répandus facilitant ainsi son intégration dans les systèmes existants, néanmoins son implantation est difficile et nécessite un temps d'exécution important.

Villarroel et al. (2006) proposent un profil UML (OMG (2005)) pour modéliser la sécurité ainsi qu'une extension du langage OCL (« Object Constraint Language ») (OMG (2006)) afin de spécifier les contraintes de sécurité lors de la phase de conception d'un entrepôt de données. Le profil UML, nommé SECDW (« Secure Data Warehouse »), comprend de nouveaux types, stéréotypes et valeurs étiquetées. Il permet de prendre en compte les contrôles d'accès obligatoire (« Mandatory Access Control » USDoD (1985)) et RBAC (*cf.* Tab 1 et Tab 2).
Le Tableau 1 indique les nouveaux types de données définis dans SECDW et le Tableau 2 montre leurs valeurs correspondantes. SECDW comporte quatre stéréotypes :

- les stéréotypes de classes sécurisées et entrepôt de données sécurisés comprenant des valeurs étiquetées associées aux attributs, niveaux de sécurité, rôles et compartiments,
- les stéréotypes d'attributs et d'instances contenant des valeurs étiquetées associées aux *niveaux de sécurité*, *rôles* et *compartiments,*
- les stéréotypes permettant de représenter les contraintes de sécurité, les règles d'autorisation et les règles d'audit.
- le stéréotype UserProfile concernant la création de contraintes dépendantes des informations des utilisateurs.

Nom	Classe de base	Description
Niveau	Enumeration	Le type Niveau est composé de tous les niveaux de sécurité considérés
Niveaux (de sécurité)	Primitif	Le type Niveaux (de sécurité) contient le niveau de sécurité le plus bas et le niveau de sécurité le plus haut.
Rôle	Primitif	Le type Rôle représente une hiérarchie de rôles d'utilisateurs
Compartiment	Enumeration	Le type Compartiment est composé de tous les compartiments d'utilisateurs
Privilège	Enumeration	Le type Privilège est composé de tous les privilèges considérés
TentativeAccés	Enumeration	Le type TentativeAccés est composé des différentes tentatives d'accès

Tab 1 - Les nouveaux types de données de SECDW (Villarroel et al. 2006)

Nom	Type	Description	Valeur par défaut
Classes	Set(OclType)	Spécifie toutes les classes du modèle. Cette valeur étiquetée permet de naviguer à travers toutes les classes du modèle.	Ensemble vide.
Attributs	Set(OclType)	Spécifie tout les attributs d'une classe. Cette valeur étiquetée permet de naviguer à travers tous les attributs du modèle.	Ensemble vide.
NiveauxSécurité	Niveaux	Spécifie la valeur de l'intervalle de niveau de sécurité, qu'une instance d'une classe peut recevoir.	Le niveau le plus bas.
RôlesSecurité	Set(Rôle)	Spécifie un ensemble de rôles	L'ensemble est

		d'utilisateur. Chaque rôle est la racine d'un sous arbre de hiérarchie de rôles définis dans l'organisation.	composé d'un rôle qui est la hiérarchie de rôles définis pour l'organisation.
Compartiments-Sécurité	Set(Compartiment)	Spécifie un ensemble de compartiments.	Ensemble vide de compartiments
ClassesImpliquées	Set(OclType)	Spécifie les classes qui sont impliquées dans une requête.	Vide
LogType	TentativeAccés	Spécifie les accès qui doivent être journalisés	Vide
AutorisationSign	+,-	Spécifie l'autorisation d'accès à une classe pour un utilisateur ou un groupe d'utilisateurs. – autorisé (+) – interdit (-)	+
AutorisationPrivilège	Set(Privilège)	Spécifie les privilèges attribués ou retirés à un utilisateur.	Lecture
estTemps	Booléen	Spécifie si une dimension représente le temps ou non.	Faux
DérivationRôle	Chaine de caractères	Si l'attribut est dérivé, cette valeur étiquetée représente la règle de dérivation.	Vide

Tab 2 : Les valeurs étiquetées de SECDW (Villarroel et al. 2006)

Il existe plusieurs types de modèle RBAC. Dans le type hiérarchique les rôles sont définis en suivant la hiérarchie des rôles des utilisateurs dans l'organisation. Ainsi un rôle est constitué d'un sous ensemble des droits d'accès appartenant à celui qui le précède dans la hiérarchie. Par exemple, les employés auront un sous ensemble des droits d'accès inférieur à celui du directeur. Villarroel et al. (2006) ont ajouté un nouveau type appelé *Tree* permettant de représenter la hiérarchie des rôles.
Le profil proposé par les auteurs est basé sur le langage UML implanté par plusieurs outils ; cependant, ils se sont limités au niveau conceptuel et n'ont pas proposé une démarche pour l'élaboration d'un modèle d'entrepôt de données sécurisé.

Soler et al. (2006) proposent une extension du CWM (« Common Warehouse Metamodel ») (OMG 2003) (*cf.* Fig 1) afin de définir les règles de sécurité et d'audit au niveau logique.

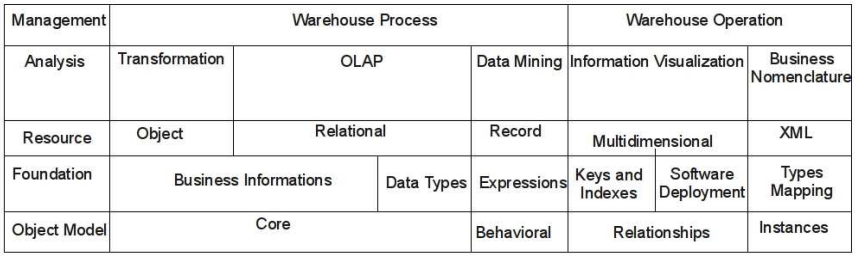


FIG 1 : Le métamodèle CWM (OMG 2003)

Dans cette extension, les auteurs se sont focalisés sur le package *Relational* de la couche *Resource* du CWM. Les extensions apportées sont (*cf.* Fig 2) :

- l'ajout de la métaclasse *SecurityProperty* ainsi que ses sous-metaclasses *SecurityLevels*, *SecurityCompartments* et *SecurityRole*. Celles-ci permettent d'attribuer des catégories aux données et aux utilisateurs et d'affecter à ces derniers des rôles.
- l'ajout de la metaclasse *SecurityConstraint* et de ses sous-metaclasses *AuditRule*, *AuthorizationRule* et *SecurityRule* permettant de spécifier les contraintes de sécurité et les règles d'audit.

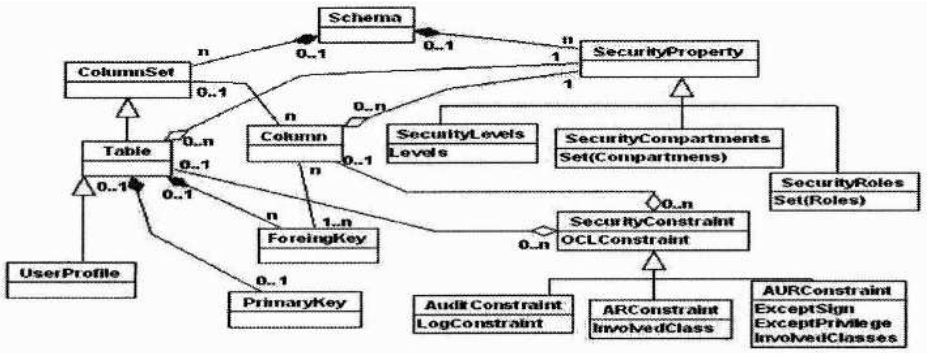


FIG 2 : Modélisation Relationnelle Sécurisée d'un entrepôt de données (Soler et al. 2006)

L'extension proposée par les auteurs prend en compte les modèles de sécurité RBAC et obligatoire mais elle n'est valable que pour le modèle de données relationnel.

Soler et al. (2007) proposent une approche MDA (« Model Driven Approach ») (*cf.* Fig 3) pour le développement d'un entrepôt de données sécurisé. Cette approche exploite le langage QVT (« Query View Transformation ») pour automatiser le passage du niveau conceptuel (Villarroel et al. (2006)) au niveau logique (Soler et al. (2006)).
L'approche MDA proposée par les auteurs traite la sécurité au niveau conceptuel et au niveau logique ; toutefois, ces auteurs n'ont pas prévu une méthode pour vérifier que les contraintes de sécurité définies au niveau conceptuel sont respectées au niveau logique.

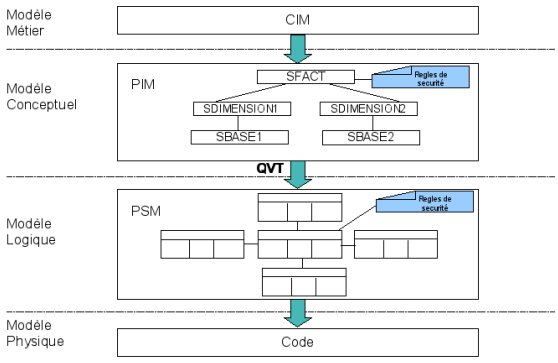


FIG 3 : Approche MDA pour la sécurisation d'un entrepôt de données (Soler et al. 2007)

*i** est un framework (Eric S. K. Y. (1997)) permettant de modéliser les besoins lors de la phase d'analyse d'un projet informatique. **Soler et al. (2008)** proposent un profil basé sur *i** qui permet de modéliser les besoins en sécurité d'un entrepôt de données au niveau métier d'une approche MDA. Ce profil permet de :

- représenter l'acteur *Gestionnaire de sécurité* qui est la personne chargée de la sécurité dans une organisation,
- modéliser les besoins en sécurité, et
- définir les *compartiments*, les *niveaux* et les *rôles* des utilisateurs en tant que ressources.

Les auteurs proposent un profil qui prend en compte le modèle RBAC et le modèle obligatoire par contre ils n'ont pas proposé une démarche pour identifier les besoins en sécurité.

Nouria H. et al (2008) ont étudié l'état de l'art et présenté une étude comparative concernant quelques travaux de recherche sur la sécurisation des entrepôts. De même ils ont détaillé l'aspect sécurité des outils commerciaux OLAP. L'étude comparative est basée sur les quatre critères suivants :

- Confidentialité : les données ne sont accessibles qu'aux utilisateurs autorisés,
- Intégrité : les données ne sont pas corrompues,
- Disponibilité : les données sont accessibles en permanence, et
- Authenticité : l'origine et l'intégrité des données sont garanties.

Ces critères correspondent aux quatre services de sécurité retenus pour les entrepôts de données.
L'étude a montré que :

- la majorité des travaux concluent que la sécurité physique (des infrastructures, des serveurs…) est insuffisante pour garantir la sécurité d'un entrepôt de données,
- l'essentiel des travaux a porté sur la gestion de la confidentialité,
- aucun travail ne propose une démarche pour la sécurisation de toutes les phases du processus d'entreposage (Nouria H. 2008),
- quelques outils commerciaux fournissent des mécanismes pour la gestion des droits d'accès, mais
- aucun outil ne propose un standard de contrôle d'accès dans le monde multidimensionnel.

Les auteurs ont défini les exigences en sécurité d'un entrepôt de données ; cependant, ils n'ont pas pris en compte le cas des inférences.

Après cette présentation des principaux travaux sur la sécurisation dans les entrepôts de données exprimée au niveau des opérations, nous passons à la deuxième classe d'approches, qui porte sur la prévention des inférences.

2.2 Prévention des inférences

Sung et al. (2006) définissent la sécurisation des données comme un moyen pour préserver la confidentialité des données des cellules d'un cube tout en fournissant les réponses aux requêtes avec une exactitude élevée et en respectant les trois objectifs de :

- Sécurité : les données sensibles ne doivent pas être divulguées,
- Exactitude : les résultats des requêtes doivent avoir un degré d'exactitude élevé, et

- Accessibilité : les restrictions ne doivent pas interdire les requêtes légitimes.

La méthode *zero-sum*, qu'ils proposent, prend en compte uniquement les requêtes de sommation. Elle consiste à ajouter des valeurs aléatoires aux cellules afin d'altérer leur contenu. Les sommes des valeurs aléatoires par lignes et par colonnes sont égales à 0.
La méthode *zero-sum* préserve la sécurité des données tout en répondant aux requêtes avec un degré d'exactitude important ; toutefois, elle nécessite un temps de calcul important et ne traite que les requêtes de type somme.

Cuzzocrea et al. (2008) proposent un framework qui permet de générer à partir d'un cube *A* un cube *A'* respectant les contraintes de sécurité. Celles-ci se basent sur des métriques dont la fiabilité est reconnue par le domaine de la sécurisation des données (Dwork, C. (2008)).
Ce framework permet de sélectionner :

- une fraction des dimensions,
- les régions des données ayant une distribution biaisée, et
- les données qui satisfont aux contraintes de sécurité pour chaque région.

Le framework ne nécessite pas un temps de calcul important ; mais il présente l'inconvénient de ne pas traiter les requêtes de type *SUM* et *AVG*, très fréquentes dans les entrepôts de données.

3 Synthèse de l'état de l'art

Les travaux sur la sécurisation des opérations traitant la sécurité aux niveaux conceptuel et physique sont les plus nombreux. Les travaux sur la prévention des inférences ne traitent qu'un seul type de requêtes.
Le Tableau 3 récapitule les avantages et les limites des travaux que nous avons présentés dans cet article.

	Travaux	Avantages	Limites
Sécurité des Opérations	Priebe et Pernul (2000)	Utilisation de langage naturel pour définir les besoins.	Absence de démarche pour l'identification des besoins en sécurité. Difficulté de validation/vérification des besoins.
	Rosenthal et Sciore (2000)	Simplicité car basé sur le langage SQL.	La sécurité est traitée au niveau physique seulement.
	Thuraisingham et al. (2007)	Comble les manques du modèle RBAC qui est très utilisé.	Difficile à implanter.
	Villarroel et al. (2006)	Prise en compte des modèles RBAC et obligatoire.	– Uniquement le niveau conceptuel a été pris en compte, – Pas de démarche pour la conception d'un entrepôt de données sécurisé.
	Soler et al. (2006)	Prise en compte du modèle RBAC et mandataire.	Ne traite pas le modèle objet.
	Soler et al. (2007)	Traitement de la sécurité au niveau conceptuel et au niveau logique.	Pas de méthode pour vérifier que les concepts de sécurité, définis au niveau

			conceptuel ont été respectés au niveau logique.
	Soler et al. (2008)	Prise en compte du modèle RBAC et obligatoire.	Pas de démarche pour identifier les besoins.
	Nouria et al (2008)	Définition des exigences en sécurité d'un entrepôt de données.	Le cas des inférences n'a pas été pris en compte.
Prévention des Inférences	Sung et al. (2006)	Préservation de la sécurité des données tout en répondant aux requêtes avec un degré d'exactitude important.	– Nécessite un temps de calcul important, – Traite uniquement les requêtes de types somme.
	Cuzzocrea et al. (2008)	Introduction de la notion de compression qui permet de réduire le temps de calcul.	Pas de prise en compte des requêtes fréquentes dans les entrepôts de données de type SUM, AVG.

Tab 3 : Synthèse des travaux sur la sécurité des entrepôts de données

4 Proposition d'une approche pour la sécurisation des entrepôts de données

Afin de sécuriser un entrepôt de données, nous proposons une approche qui permet d'empêcher d'inférer des données cachées à partir des données accessibles tout en prenant en compte la sécurité des opérations. Ainsi, notre approche se base sur une combinaison des propositions actuelles permettant une sécurisation globale.

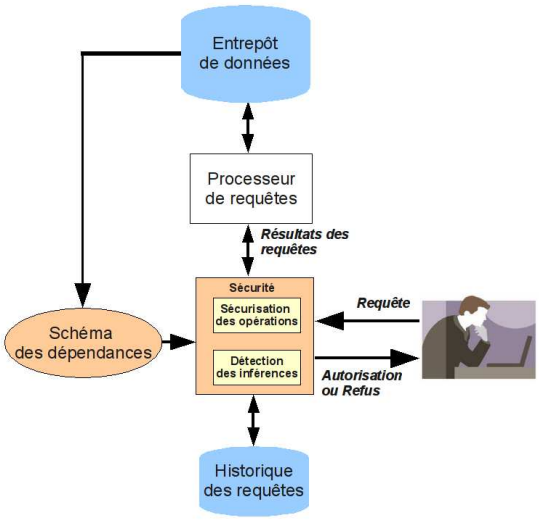


FIG 4 : Approche proposée pour la sécurisation d'un entrepôt de données

La Figure 4 décrit cette approche qui comporte principalement les trois volets suivants :

- Détection des dépendances entre les données.
- Création d'un réseau Bayésien en se basant sur les dépendances entre les données. Un réseau Bayésien est un graphe dans lequel les relations de cause à effet entre les nœuds sont probabilisées.

Par exemple, un avion ne peut atterrir que sur une piste ayant une certaine longueur minimale. Sachant que la longueur de la piste ne doit pas être divulguée, si nous savons que l'avion Av_1 a atterri sur la piste P_1 alors nous pouvons déduire la longueur minimale de la piste. La dépendance entre l'attribut *Avion* et l'attribut *Piste d'atterrissage* peut être représentée par le réseau Bayésien de la Figure 5. Le Tableau 4 contient les probabilités d'inférer la longueur d'une piste ; où l'on distingue quatre cas :

1- Longueur minimale de la piste d'atterrissage exigée par un avion est inconnue ; la probabilité concernant la longueur de la piste est la même pour les trois types d'atterrissage ; courte, moyenne et longue ; elle est égale à 0,33

2- Longueur minimale de la piste d'atterrissage exigée par un avion est courte ; après atterrissage de l'avion, la probabilité est aussi la même pour les trois types ; sa valeur est de 0,33

3- Longueur minimale de la piste d'atterrissage exigée par un avion est moyenne ; dans ce cas après atterrissage de l'avion la probabilité concernant le type de piste est 0 pour la courte, 0,5 pour la piste moyenne et de 0,5 pour la piste longue.

4- Longueur minimale de la piste d'atterrissage exigée par un avion est longue ; dans ce cas, après atterrissage la probabilité est de 0 pour la courte et la moyenne et elle est de 1 pour la piste longue.

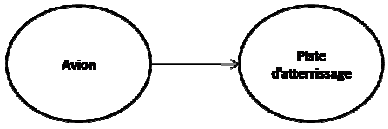


FIG 5 : Réseau Bayésien modélisant la dépendance entre Avion et Piste d'atterrissage

	Avion (longueur minimale de la piste d'atterrissage)				
		Inconnue	Courte	Moyenne	Longue
Longueur de la piste	Courte	0,33	0,33	0	0
	Moyenne	0,33	0,33	0,5	0
	Longue	0,33	0,33	0,5	1

Tab 4 : Les probabilités d'inférer la longueur d'une piste

- Création d'un module de sécurité. Le fonctionnement de ce module concerne l'autorisation ou le refus de nouvelles requêtes. Il se compose de deux parties. La première partie est relative à la sécurisation des opérations, la deuxième porte sur la détection des inférences à l'aide de l'historique des requêtes et du réseau Bayésien : après lancement d'une requête, trois étapes sont à franchir :

  1- la partie sécurisation des opérations vérifie que la requête, ne viole aucune règle de sécurité,

  2- la partie détection des inférences consulte l'historique des requêtes de l'utilisateur et calcule les probabilités du réseau Bayésien,

  3- après vérification, sur la base des probabilités calculées la requête demandée est autorisée ou refusée.

5 Conclusion

Dans cet article, nous avons mené une étude et présenté une synthèse de l'état de l'art sur la sécurité des entrepôts de données. Les travaux, jusque là proposés ont abordé la problématique de la sécurité des entrepôts de données selon deux approches : la sécurisation des opérations et la prévention des inférences. Toutefois, la majorité de ces travaux sont relatifs à la première classe sans pour autant prendre en compte les cas de sources structurellement hétérogènes ou ayant des modèles de sécurité différents.
Quant aux quelques travaux dans le cadre de la prévention des inférences, nous constatons qu'ils ne prennent en compte qu'un seul type de requêtes utilisant les fonctions d'agrégation : *MAX* ou *SUM*.
A la lueur de cet état de l'art, nous avons présenté une approche pour la sécurisation des entrepôts de données basée sur les réseaux Bayésiens et qui combine les deux axes de la sécurité des entrepôts : les opérations et les inférences de données cachées. Cette proposition est en cours de finalisation pour être implantée et testée.

Références

Cuzzocrea, A., Russo V., et Saccà D. (2008). A robust sampling-based framework for privacy preserving olap. In DaWaK, pp. 97–114. Springer.

Dwork, C. (2008). Differential privacy : A survey of results. In TAMC, pp. 1–19.

Eduardo F., Juan T., Rodolfo V., et Mario P. (2006). Access control and audit model for the multidimensional modelling of data warehouses. In Decision Support Systems Volume 42 , Issue 3, pages: 1270 – 1289.

Eric S. K. Y. (1997). Towards modelling and reasoning support for early-phase requirements engineering. Proceedings of the Third IEEE International Symposium on Requirements Engineering.

Export (2008). http://www.export.gov/safeharbor/

Jaehong P. R. S. (2004). The UCON usage control model. In Proceedings of ACM Trans. Inf. Syst. Secur., vol 7, pages 128--174.

HHS (1996). http://www.hhs.gov/ocr/privacy/index.html

Nouria H. (2008). Cours: Sécurité des entrepôts de données, Master 2 ECD, Université Lumière Lyon2.

Nouria H., Maaroufi G., Omar B. (2008) Sécurité des entrepôts de données - état de l'art -. Troisième Atelier sur les Systèmes Décisionnels. 10-11 octobre 2008, Mohammedia, Maroc.

OMG (2003). Common Warehouse Metamodel (CWM), Version 1.1.

OMG (2006) Object Constraint Language (OCL), Version 2.0.

OMG (2005) Unified Modeling Language (UML),Version 2.0.

Priebe, T. et Pernul G. (2000). Towards OLAP security design - survey and research issues. In International Workshop on Data Warehousing and OLAP.

GPO (1999). http://www.gpo.gov/fdsys/pkg/PLAW-106publ102/content-detail.html

Ralph K. (1997). Hackers, Crackers, and Spooks; ensuring that your data warehouse is secure. DBMS Magazine.

Rosenthal, A. et Sciore E. (2000). View security as the basis for data warehouse security. In CAiSE Workshop on Design and Management of Data Warehouses.

Sandhu, R. S., Coyne E.J., Feinstein H.L. et Youman C.E. (1996), "Role-Based Access Control Models", IEEE Computer 29(2): 38-47, IEEE Press, 1996.- proposed a framework for RBAC models

Soler, E., V. Stefanov, J.-N. Mazón, Trujillo J., Fernández-Medina E., et Piattini M. (2008). Towards comprehensive requirement analysis for data warehouses : Considering security requirements. In ARES, pp. 104–111. IEEE Computer Society.

Soler, E., Trujillo J., Fernández-Medina E., et Piattini M. (2007). A framework for the development of secure data warehouses based on mda and qvt. In ARES, pp. 294–300. IEEE Computer Society.

Soler, E., Villarroel R., Trujillo J., Fernández-Medina E., et Piattini M. (2006). Representing security and audit rules for data warehouses at the logical level by using the common warehouse metamodel. In ARES, pp. 914–921. IEEE Computer Society.

Sung, S., Y. Liu, et P. Ng (2006). Privacy preservation for data cubes. Knowledge and Information Systems 9, 38–61.

Soxlaw (2002). http://www.soxlaw.com/

Sung S. Y., Liu Y., Xiong H., Peter A. (2006). Privacy preservation for data cubes. In Knowledge and Information Systems 9, pp 38-61.

USDoD (1985). Trusted Computer System Evaluation Criteria. United States Department of Defense. December 1985. DoD Standard 5200.28-STD.

Thuraisingham, B., Kantarcioglu M., et Iyer S.(2007). Extended rbac-based design and implementation for a secure data warehouse. IJBIDM 2, 367–382.

Villarroel, R., Fernández-Medina E., Piattini M., et Trujillo J. (2006). A uml 2.0/ocl extension for designing secure data warehouses. Journal of Research and Practice in Information Technology 38.

Summary

Data warehouses integrate data from heterogeneous sources and are used by decisional users to make strategic decisions. These data may be sensitive and should not be accessed without controls thus the need for their security is highly required. In this article we present a synthesis of research in the field of data warehouse security and we propose an approach for their secure.

Une approche basée sur les Modèles d'Entreprise pour l'intégration des services de l'e-Business

Soumia Bendekkoum, Mahmoud Boufaida

Laboratoire LIRE, Université Mentouri de Constantine, Algérie
{Bendekkoums, boufaida_mahmoud}@yahoo.fr

Résumé. Avec l'adoption des modèles d'Architectures Orientées Services (AOS[1]), la plupart des systèmes non orientés services sont devenus des systèmes patrimoniaux[2]. Ainsi la réingénierie de ces systèmes est devenue nécessaire pour les rendre capables de survivre dans un environnement distribué orienté services. Dans cet article, nous proposons une approche basée sur les techniques de modélisation en entreprise (EM[3]), permettant de réutiliser et d'intégrer les composants des systèmes non orientés services comme des services web. Cette approche propose d'intégrer l'utilisation des modèles hiérarchiques de l'architecture fonctionnelle de l'entreprise, fondés sur la granularité de ses applications dans les phases d'analyse et d'identification des services sur lesquelles nous focalisons les idées de notre contribution. Les services adéquats sont définis selon la phase d'identification des services, qui est basée sur une projection des modèles fonctionnels et des modèles d'objectifs, dans la mesure où, la modélisation hiérarchique des applications et l'analyse des flux de données entre eux est très utile pour récupérer l'information nécessaire à la détection et à l'intégration des services adéquats. Une étude de cas a été réalisée afin de valider la rentabilité de l'approche proposée, et quelques interfaces ont été présentées.

Mots clés : Architecture Orientée Services (AOS), Système Patrimonial, Réingénierie des systèmes, Modélisation en Entreprise (EM), granularité de l'application.

1 Introduction

La plupart des entreprises cherchent à évoluer pour rester au premier rang du monde industriel, en adoptant le modèle e-Business, qui permet de contrôler les chaînes d'achat et d'approvisionnement, les relations avec les clients et de fournir des applications et des services basées sur le web. Ceci force l'entreprise à fusionner son système traditionnel avec

[1] En Anglais, Service Oriented Architecture.
[2] En Anglais, legacy systems.
[3] Enterprise Modeling

d'autres entreprises étrangères, réorganiser sa structure interne, et adapter de nouvelles technologies et plateformes pendant qu'elle essaye d'obtenir des avantages concurrentiels (Izza et al, 2004). Cependant, l'adaptation des systèmes traditionnels aux nouvelles technologies est une tâche très difficile, les entreprises ont donc pensé au remplacement de leurs systèmes par d'autres plus avancés (Oracle1, 2008), ce qui est impossible d'un point de vue, technologique à cause du manque d'une documentation détaillée de ces systèmes et de l'absence des anciennes compétences, et aussi d'un point de vue économique dans la mesure où ce genre de projets nécessitent un énorme investissement pour le développement et la maintenance de ces systèmes (Oracle1, 2008). Une autre solution consiste à restructurer les anciens systèmes en une architecture de services web. Ces derniers permettent d'assurer et de faciliter les échanges entre les différentes applications anciennes ou avancées à travers le web (Oracle2, 2008). Dans cet article, la deuxième et la troisième section, présentent respectivement quelques travaux basés sur les architectures orientées services et les principales caractéristiques des modèles d'entreprises utilisés pour comprendre, améliorer et adapter le métier des anciens systèmes aux nouvelles technologies. Dans la section 4, nous présentons une description globale de notre approche proposée. La section 5 détaille les différentes étapes de l'approche. Enfin la section 6, présente une étude de cas permettant l'application des étapes de l'approche sur une entreprise commerciale.

2 Quelques travaux basés sur les AOS

L'Architecture Orientée Services apparaît depuis des années comme un paradigme efficace de développement des systèmes distribués constituant l'e-Business et l'intégration dynamique des ces services à travers le web. Mais cette notion n'est effectivement utilisée qu'après l'apparition et le succès des modèles d'architecture des services web (Tonic et al, 2006). Par définition (Guimnich, 2008), les systèmes basés sur l'Architecture Orientée Services différent des autres architectures par leur style architectural qui offre une solution flexible d'intégration en termes d'encapsulation, de réutilisation, de composition, de réduction de couplage entres les composants applicatifs distribués et leur interopérabilité ouverte basée sur les standards du web. L'Architecture Orientée Services fournit un degré élevé de flexibilité aux systèmes informatiques des entreprises et une composition sur mesure de leurs processus métiers, et bien d'autres avantages, qui sont vues éprouvés non seulement par la théorie mais aussi par la pratique dans beaucoup de travaux comme Sneed (1996,2006), (Arsanjani, 2004), Ziemann et al. (2006).

Le concept de l'architecture orientée services offre la possibilité d'adapter les applications des systèmes existants et les publier via les interfaces des services web Tonic et al. (2006). La majorité des entreprises ont donc décidé de structurer leurs systèmes informatiques en une architecture de services. Cependant, la transition vers un environnement distribué de services web, nécessite une compréhension totale et approfondie du métier de ces systèmes Ziemann et al. (2006). Les méthodes de modélisation en entreprise offre le meilleur moyen permettant de surmonter la complexité de l'architecture des anciens systèmes, de faciliter la compréhension de leurs applications et de fournir une documentation claire sur leur structure interne Martin et Andersson (1996).

3 Modélisation en Entreprise

Le concept de « la Modélisation en Entreprise » a été introduit pour la première fois au milieu de l'année 1970 dans le domaine du génie logiciel pour la conception et le développement des systèmes d'information Fox et Gruninger (1998). Ce paradigme a été introduit une autre fois vers la fin de l'année 1990 dans le domaine de réingénierie pour la restructuration et l'amélioration des systèmes d'entreprises existants. A partir de cette date plusieurs définitions de ce concept ont été proposées Martin et Anderson (1996) : « La Modélisation en Entreprise est l'art de capturer et d'externaliser des connaissances définissant une entreprise pour ajouter de la valeur métier à l'entre prise, partager l'information entre les différents partenaires et pour représenter toutes les entités de l'entreprise : les informations, les ressources, et la structure de l'organisation ».

La modélisation de métier de l'entreprise et de son environnement permettant de faciliter la création, la compréhension et l'amélioration de ses systèmes et de ses processus métiers et plus particulièrement les relations entre eux. Elle permet la modélisation de la structure interne d'une partie ou de système entier ainsi que la structure de ces partenaires. Ceci inclut la modélisation de toutes les entités de l'entreprise : composants, ressources, processus métiers, objectifs, besoins, utilisateurs, fonctions …etc.

La majorité des entreprises d'aujourd'hui, et particulièrement les entreprises algériennes commencent une évolution particulière avec l'avènement des modèles d'architecture de services web. Cependant il n'existe pas une approche permettant de faciliter la transition vers la nouvelle architecture orientée services et plus précisément de surmonter la complexité des anciennes architectures. Dans cet article, nous proposons une approche pour l'intégration des services e-Business, basée sur les techniques de réingénierie des systèmes d'information. Cette approche permet l'intégration et la réutilisation des systèmes traditionnels constituant l'e-Business dans une architecture orientée services, en utilisant les services web pour définir les échanges entre les différentes applications anciennes ou avancées.

4 Description globale de l'approche proposée

Notre processus de migration est constitué de 4 phases principales. La figure 1 montre ces phases numérotées de 1 à 4. (voir FIG 1)

La première phase d'Etude de faisabilité de migration consiste à la réingénierie et l'évaluation des systèmes existants et à la modélisation de leurs comportements en des modèles hiérarchiques, fondés sur la granularité des fonctions et/ou des applications qui les constituent. Et aussi la modélisation des besoins attendus en un modèle hiérarchique d'objectifs de niveau d'abstraction élevé. La deuxième phase consiste à l'identification des services. De ces derniers, en appliquant la procédure de projection de modèle d'arbre fonctionnel et de modèle des besoins nous distinguons les services qui existent dans le système et qui doivent être extrait du code existant et les autres qui n'existent pas dans le système et qui doivent être implémentés à nouveau. Après la détection du code qui implémente les services, vient la dernière phase d'adaptation et de publication qui produit à la fin des services web prêts à être intégrés et réutilisés dans un système e-Business orienté services.

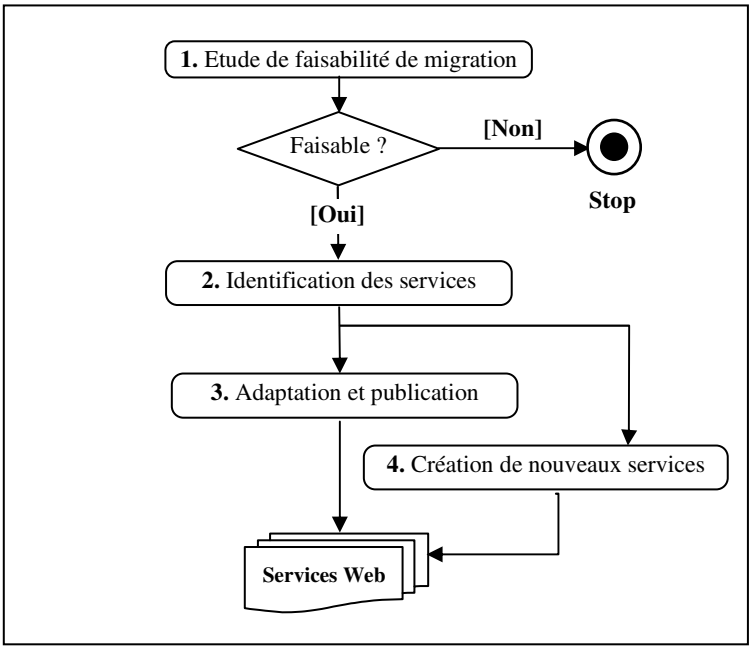


FIG. 1 – *Différentes phases de l'approche proposée.*

5 Différentes phases de l'approche

Afin de détecter, extraire et intégrer les services à partir de code source des systèmes d'information traditionnels constituant l'e-Business, nous avons utilisé les modèles d'entreprises. Ces derniers sont intégrés dans les phases de notre approche de la manière suivante :

5.1 Etude de faisabilité

Nous avons choisi de commencer notre approche par une phase d'étude de faisabilité, après avoir évalué les systèmes existants constituant l'e-Business et analyser les besoins de futur système intégré. Cette phase a pour objectif de vérifier s'il est réellement possible de faire migrer un système existant vers une architecture orientée services. Et si elle est possible, est-elle rentable ?

A l'inverse des approches de Zhang (1996), Lewis et al. (2005, 2008), qui utilisent des techniques classiques d'aide à la décision[4], permettant d'analyser et d'appliquer quelques critères pour décider si un code mérite d'être un service ou non. Nous avons proposé d'utiliser une méthode de réingénierie d'aide à l'analyse et à la décision, basée sur les mo-

[4] Comme l'arbre de décision et l'analyse des options pour la technique de réingénierie (OAR OAR : the Option Analysis for Rengineering).

dèles d'entreprise : le modèle d'arbre fonctionnel de système patrimonial[5] et le modèle d'objectifs de futur système. L'idée de base de notre contribution, vise à transformer les applications existantes en des modèles hiérarchiques fondés sur le degré de granularité de leurs fonctions, et l'analysé pour décider sur la faisabilité de migration.

5.1.1 Modèle d'arbre fonctionnel

Le modèle fonctionnel de système d'information représente une structuration abstraite et hiérarchique de toutes les fonctions (code source) des applications qui constituent le système. La construction de modèle d'arbre fonctionnel d'un système se fait de la manière suivante (voir TAB. 1):

Modèle d'arbre fonctionnel	Code source (Système/Application)
La racine de modèle	Application/fonction principale ayant le niveau de granularité le plus élevé dans un système
Les nœuds du modèle	Application/fonctions réalisant une application de niveau hiérarchique supérieur
Chaque sous nœuds de modèle	Sous fonction qui implémentent la fonction principale de niveau de granularité supérieur
Les feuilles de modèle	Instructions qui interviennent dans l'implémentation des fonctions à faible valeur métier

TAB. 1 - *Modélisation de code source en un modèle d'arbre fonctionnel.*

La figure 2 représente un modèle d'arbre fonctionnel d'un système d'une entreprise commerciale écrit en langage Basic, cette présentation de code fondée sur la granularité des fonctions qui le constituent (plus de détail est omis ici en raison de contraintes d'espace) (voir Fig. 2).

5.1.2 Modèle d'objectifs

Le modèle d'objectif représente une description hiérarchique des besoins de futur système dans un e-Business. Ce modèle est similaire au modèle d'arbre fonctionnel de système existant, où le nœud de ce modèle représente un objectif, et les sous-nœuds représentent les sous objectifs réalisant l'objectif principal. De ce modèle, la tâche consiste à analyser et sélectionner les composants, permettant d'offrir les fonctions réalisant les besoins.

[5] En Anglais Function Tree Model for Legacy System.

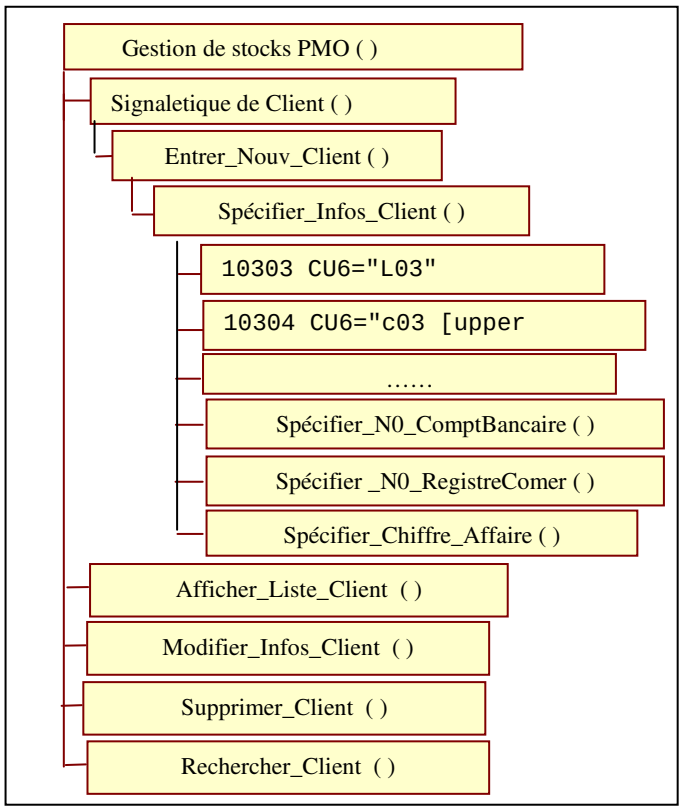


FIG. 2 – *Modèle d'arbre fonctionnel.*

5.1.3 Analyse de faisabilité

Nous proposons pour vérifier la faisabilité d'intégrer un système dans un environnement distribué comme étant des services autonomes, d'analyser les modèles établis dans cette phase, en associant les nœuds de modèle d'objectifs à la fonction de modèle d'arbre de fonctions permettant la réalisation de cet objectif.

A l'issue de cette phase, nous obtenons les composants qui contiennent les services web éventuels. En effet, dans cette définition d'une première liste de services web éventuels, seuls les composants permettant d'offrir ces services sont détectés. On ne précise pas quelles fonctions ou tâches (actions) réalisent effectivement une activité d'un objectif donné, ou comment un processus métier est réalisé ?

5.2 Identification des services

Cette phase consiste à détecter et extraire les bons candidats de services web[6] (Xebia, 2008). Elle est constituée de deux étapes successives, sont : le raffinement des modèles. Ceci est effectué par la spécification des entrées ainsi que de la sortie finale de chaque fonction d'un composant dans le modèle fonctionnel. Ensuite l'application de l'approche de projection des modèles, pour l'extraction de code minimal adéquat. Celle-ci est simple et réalisable dans le domaine de services web.

Les étapes de l'approche de détection des services web basée sur la procédure de projection que nous proposons, sont résumées comme suit :

1. La première tâche est la sélection d'une activité d'un processus métier élémentaire, cette dernière est associée à un composant qui contient les fonctions réalisant cette activité.
2. Commençons par les feuilles (instruction) de composant sélectionné, en vérifiant si cette instruction influe sur le résultat final retourné par l'activité de processus sélectionnée.
3. Si l'instruction sélectionnée retourne peu de variables par apport à l'activité, il est nécessaire de sélectionner le nœud qui se trouve à un niveau hiérarchique supérieur.
4. Dans le cas contraire, c'est-à-dire les variables retournées sont équivalentes à celles de l'activité sélectionnée, la fonction est donc définie comme étant le futur service public qui implémente l'activité sélectionnée de processus élémentaire.
5. Répéter les étapes 1-4 pour tous les nœuds de modèle de processus élémentaire.

5.3 Adaptation et intégration des services

Cette phase est nécessaire pour qualifier les fonctions détectées dans la phase précédente, pour être des services web fonctionnels. Elle consiste à l'adaptation technique (technical wrapping) des fonctions identifiées par des services web.

5.3.1 Définition de l'interface de description des services "WSDL"

Le but de cette phase est la définition des fonctionnalités des services web en WSDL, chaque service web expose une interface qui définie les types de messages et les modes de leurs échanges. Pour cela il faut d'abord spécifier une interface bien précise pour chaque service détecté. Ensuite, nous passons à la définition des fonctionnalités des services web en WSDL.

5.3.2 Création du mécanisme de transport "SOAP"

Une fois l'interface du service web est définie, nous passons à la création et l'intégration du mécanisme d'accès et de transport des messages qui sera le processeur SOAP, qui garantit facilement l'échange des messages sur le web.

5.3.3 Implémentation du « Interface_Service »

Le code de « Interface_Service » est une partie du code ajouté au dessus du service web, en tant qu'interface. Il joue le rôle de la réception des messages d'appels, l'extraction des

[6] Un service est un bon candidat, s'il a une valeur ajoutée métier équivalente à un objectif donné.

paramètres d'entrées d'un service web et l'empaquetage des résultats de l'exécution sur service dans des messages et les envoyer vers les clients.

5.4 Création de nouveaux services

Dans cette étape, le développement des services qui sont apparus dans la phase d'identification des services et ne peuvent pas être extrait du système existant même par la composition des services web définis, est effectué. Ces nouveaux services seront intégrés par la suite dans les futurs systèmes orientés services pour améliorer la logique d'affaires de ces derniers.

6 Quelques aspects d'implémentation

Afin de valider la rentabilité de notre processus, nous avons choisi un exemple d'une entreprise commerciale de production de machines outils. Le système d'information de cette entreprise est écrit en langage Basic, en utilisant le générateur de programme « Top Key », il est utilisé de puis 30 ans. L'objectif de notre étude de cas est de faire migrer ce système en détectant, extrayant et intégrant ces fonctionnalités dans un environnement e-Business distribué de services de bon niveau de granularité.

Dans notre étude de cas et avant de commencer l'étape de modélisation de système, nous avons choisi d'amblé de faire une phase de prétraitement permettant de supprimer les branchements existants dans le code de système, notamment l'instruction « GOTO » et les deux autres « GOSUB et RETURN». En effet, les modifications effectuées à ce niveau sont faites manuellement, en appliquant le processus d'élimination du « GOTO » proposé dans (Zang et al., 2005).

La figure 3 (Fig. 3) représente la phase de détection et de définition des services web, en appliquant la procédure de projection sur le modèle fonctionnel (a) (voir Fig.2) et le modèle d'objectif (b), en se basant sur les sorties finaux des fonctions.

Après la détection des parties de code patrimoniales qui méritent d'être publiées comme des services web, nous pouvons dire qu'un grand pas dans le processus de migration vers les services web est effectué, et il ne nous plus reste que l'adaptation et l'intégration de code dans des processus métiers, où nous allons créer une description de nos services web selon le format WSDL (Web Service Description Language).

La réalisation de notre système orienté service doit commencer par la création des interfaces permettant la communication du code extrait de l'application patrimoniale avec le service web.

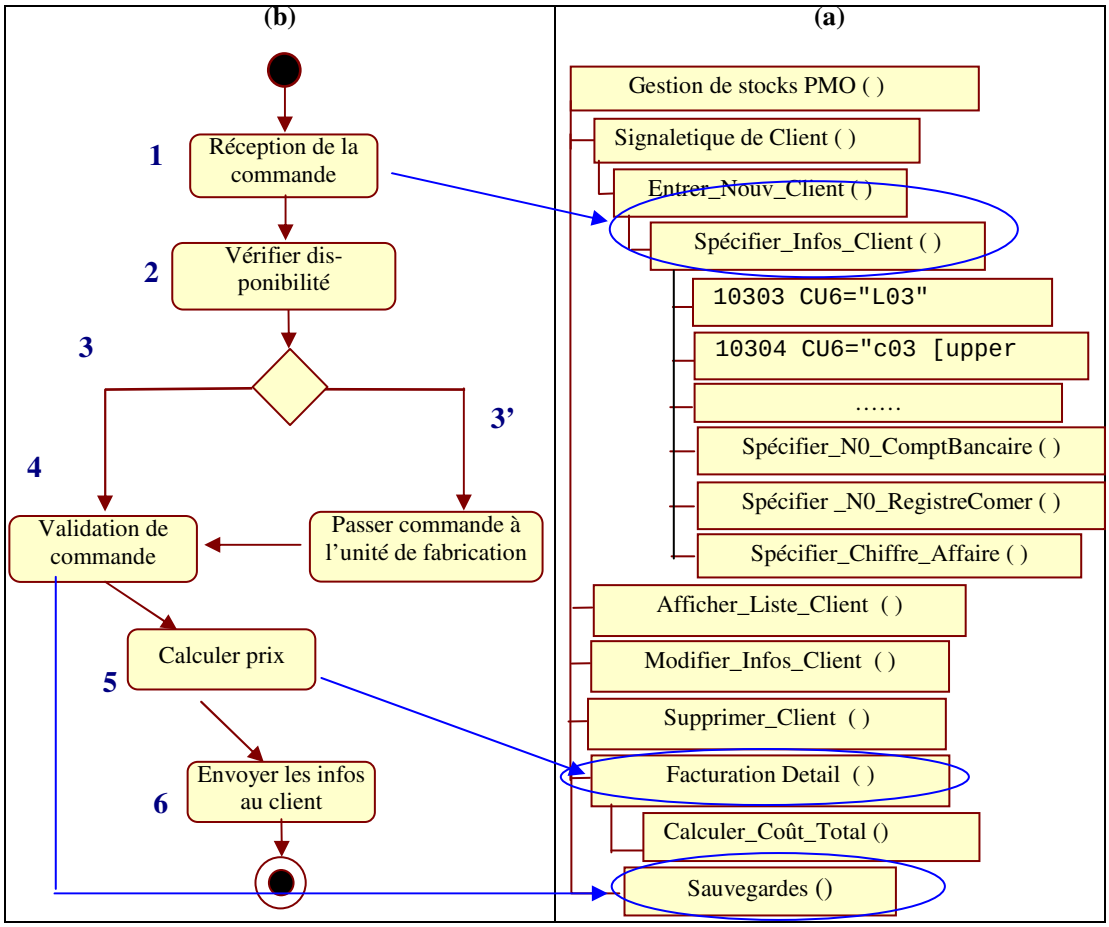


FIG. 3 – *Identification des services web.*

En ce qui concerne notre étude de cas, le code (c'est-à-dire les services) extrait de l'application patrimoniale n'a pas d'API (Application Programing Interface) et d'interface utilisateur. Pour cela, nous avons choisi la méthode des interfaces COM (Component Object Model) pour exposer ce code aux autres applications de différentes plateformes. Pour ce faire nous utiliserons le système de programmation Visual Basic 6.0, qui permet de générer des programmes autonomes et aussi des composants ActiveX pour des programmes. L'application ActiveX rend le code de services extraits de l'application patrimoniale sous forme de module de classes. Chaque classe implémente et expose les méthodes d'un service par VB sous l'extension « .cls ». Après la création des interfaces permettant la communication avec le code patrimonial, vient l'étape de publication et d'adaptation du code via les interfaces de services web. Nous avons créé un nouveau module en utilisant le langage VB dans la plateforme .Net (VB.Net), ce module contient les interfaces "Interface_Services" des services web. La partie de code présentée dans la figure 4 présente une de ces interfaces « Interface_Code », qui charge la requête xml "RequestXML" envoyée par l'application

client au service web **(1)**. La classe "FiltreXml" est chargée de chercher et extraire les arguments et le service en question. Par la suite, le service proposé est utilisé comme n'importe quelle autre méthode. Le résultat est ensuite envoyé sous forme d'un document xml "xmlResult" **(2).**

```
… … …
Public Class FiltrerResultatXML
        … …
Public Function FiltreXML(ByVal attribut As String) ← (1)
   Return xmlDoc.SelectSingleNode("Response/ReturnValues/Item
   [Name='"&attribut&"']").SelectSingleNode("@value").InnerXml
End Function
… … …
Public Sub New(ByVal requestXML As String)
   xmlDoc = New XmlDocument
   xmlDoc.LoadXml(requestXML)
   dim idImployee as String=lecture.FiltreXML(xml.Doc,"CArt")
   dim Art as new Interop.applicatio_patrimoniale.LectureDonne
   XMLResult= Art.lecture(Cart)  ← (2)
… … …
        End Sub
… … …
    End Class
End Namespace
… … …
```

FIG. 4 – *Implémentation des interfaces des services web.*

7 Conclusion

Dans cet article, nous proposons une approche pour l'intégration des services de l'e-Business. Cette approche repose sur des itérations successives des phases de réingénierie des systèmes existants, d'identification et d'extraction de code des services et de génération et de publication des services. Nous avons appliqué les étapes de notre approche sur une entreprise commerciale de production de machine outils, dont notre objectif principal est la décomposition de son système en services web ayant une valeur ajoutée métier élevée et adéquate dans l'ensemble des entreprises intégrées constituant l'e-Business, et donc aider le concepteur à maîtriser la complexité de système e-Business.

Références

Arsanjani, A. (2004). *Service-Oriented Modeling and Architecture: How to identify, specify and realize services for your Service Oriented Architecture*. http://www-128.ibm.com/developerworks/web-services/library/ws-soa-design1/

Fox, M. S., et M. Gruninger (1998). *Enterprise Modeling*. American Association for Artificial Intelligence Magazine, 109-122.

Gimnich, R. (2008). *Migration approaches and Experiences.* Proceedings of the 10th IEEE European Conference on Softeware Maintenance and Reengineering.

Izza, S., L. Vincent, P. Brulat, H. Solignac et P. Lebrun (2004). *Intégration d'Applications: Etat de l'art et perspectives.* Blida (Algérie): Les 2 èmes Journées d'Informatique pour l'Entreprise.

Lewis, G., E. Morris, D. Smith, L. Wrage et L. O'Brien (2005). *Service-Oriented Migration and Reuse Technique (SMART).* In: 13th IEEE International Workshop on Software Technology and Engineering Practice.

Lewis, G. A., A. J. Morris, D. B. Smith et S. Simanta (2008). *SMART: Analyzing the reuse potential of migrating legacy components to a service-oriented architecture.* In: 10th IEEE European Conference on Software Maintenance and Reengineering. Istanbul (Turkey)

Martin, R., et M. Andersson (1996). *Reverse Engineering of Legacy Systems: from value-based to object-based models*. Thèse de doctorat, Ecole Polytechnique Fédérale de Lausanne (EPFL).

Oracle1 (2008). *Why Modernize?* Oracle IT Modernization Series: An Oracle White Paper. http://www.oracle.com/technologies/modernization/docs/whymodernize.pdf

Oracle2 (2008). *The Types of Modernization*. Oracle IT Modernization Series: An Oracle White paper. http://www.oracle.com/technologies/modernization/docs/typesofmoder.pdf

Sneed, H. (1996). *Wrapping Legacy Software for Reuse in a SOA.* Proceeding of the 4th IEEE International Workshop on high Perfomance Computing. reengineering and Knowledge Engeneering. Berlin.

Sneed, H. (2006). *Integrating legacy Software into a Service-Oriented Architecture*. Proceedings of the Conference on Software maintenance and Reengineering, 3-14.

Tonic, F., M. Boulier, B. Paroissin, J. Clune, F. Bernnard et M. Gardette (2006). *SOA: votre nouvelle architecture*. Magasine de développement: Programmez!, 78, 679-696.

Xebia France (2008). *Mise en oeuvre d'une SOA: les clés de success*.

Ziemann, J., K. Leyking, T. Kahl, et D. Werth (2006). *Enterprise Model driven Migration from Legacy to SOA.* In: IEEE Software Reengineering and Services Workshop.

Zachman, J. (2005). *A Framework for Informations-Systems Architecture*. Ibm System Journal 3, 276-292.

Zhang, Z., R. Liu, et H. Yang (2005). *Service Identification and Packaging in Service Oriented Reengineering.* Proceeding of the 17th International Conference on Software Engineering and Knowledge Engeneering.

Une approche basée sur les modèles pour l'intégration des services de l'e-Business

Summary

With the adoption of services oriented architecture models, the majority of systems non oriented services became legacy systems, where reengineering these systems became necessary in order to make possible survive in a distributed oriented services environment. In this paper, we propose an approach based Enterprise Modelling techniques allowing reusing and integrating the components of non oriented services systems as web services. This approach proposes to incorporate the use of function hierarchical models of the company, founded on applications granularity into analysis and identification services phases, on which we focus the ideas of our contribution. The adequate services are defined according to the identification services phase, which is based on the mapping of functional model and goals model, insofar as, the hierarchical modelling of the applications and the analysis of the data flows between them are very useful to capture information necessary to the detection and the integration of the adequate services. A case study was realized to validate the approach proposed, and some interfaces were presented.

Fouille dans les documents XML : Etat de l'art

Amina MADANI*, Omar BOUSSAID**
Hafida ABED*, Sabine LOUDCHER**

*Université SAAD DAHLAB de Blida (Algérie)
a_madani@univ-blida.dz, hafidabouarfa@hotmail.com

**Université Lumière Lyon2 (France)
{Omar.Boussaid, Sabine.Loudcher}@univ-lyon2.fr
http://eric.univ-lyon2.fr/~boussaid/

Résumé. Dans cet article, nous parcourons un ensemble de travaux concernant la fouille dans les documents XML (fouille dans la structure et fouille dans le contenu). Nous exposons ces approches en les classifiant selon cinq modèles de présentation des documents XML. Nous dressons des tableaux comparatifs des travaux en utilisant quelques critères de comparaison. Nous concluons notre étude par des remarques qui pourraient être importantes et utiles pour des futurs travaux dans ce domaine.

1 Introduction

XML s'est imposé comme le langage permettant de représenter et d'échanger des données non seulement sur le web mais aussi de façon générale dans les entreprises (Gardarin et al., 2003). Face aux quantités d'informations qui ne cessent d'augmenter dans les documents XML, l'extraction automatique des connaissances à partir de ces documents et les techniques de visualisation des résultats sont devenues indispensables. C'est la raison d'être de la fouille dans les documents XML (*XML mining*).

La figure *FIG.1* montre les résultats d'un sondage effectué en 2008[1] pour les différents types de données sur lesquels le *data mining* est appliqué. Le XML est classé parmi les derniers avec seulement 5,6 %, ce qui montre que les documents XML sont très peu pris en compte dans le domaine du *data mining*.

La fouille dans les documents XML comprend deux catégories (Nayak et al., 2002) : la fouille dans la structure et la fouille dans le contenu. Les balises et leurs imbrications représentent la structure d'un document XML. La fouille dans la structure XML (*XML Structure Mining*) est essentiellement la fouille des schémas (Nayak et al., 2002). Elle s'intéresse à l'extraction d'informations à partir de la structure des documents XML (Garofalakis et al., 1999). Dans un document XML, le contenu est le texte entre chaque balise ouvrante et fermante. La fouille dans le contenu XML (*XML Content Mining*) est essentiellement la fouille

[1] Data Types Analyzed / Mined, http://www.kdnuggets.com/polls/2008/data-mining-applications.htm, September 2008.

des valeurs (une instance d'une relation) (Nayak et al., 2002). La fouille sur le contenu utilise habituellement les techniques de *Text Mining* pour extraire l'information contenue dans le document.

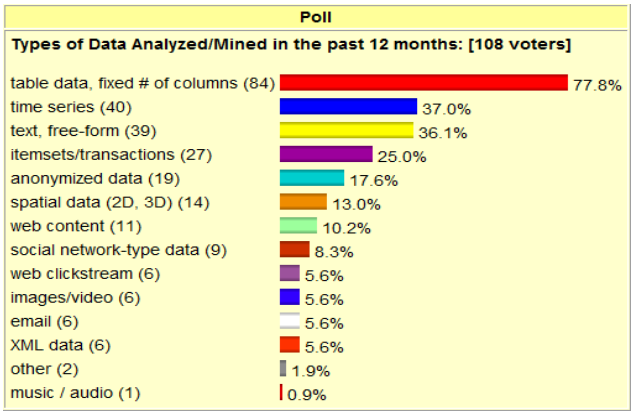


FIG. 1 – *Sondage sur les types de données sur lesquels le data mining est appliqué* [2].

L'association de la fouille dans la structure et celle dans le contenu devrait permettre d'élargir et d'améliorer les techniques de fouille des documents XML pour améliorer la qualité sémantique des résultats obtenus (Boussaid et al., 2004) (Candillier et al., 2007).

Dans cet article, nous proposons une classification des différentes approches de fouille dans la structure et dans le contenu proposées dans la littérature. Cette classification est effectuée selon cinq modèles (voir *Fig.2*). Nous ajoutons un autre modèle aux quatre premiers cités dans (Denoyer et al., 2006) :

- Modèle basé sur les vecteurs ;
- Modèle basé sur la similarité ;
- Modèle basé sur les réseaux de neurones ;
- Modèle d'arbres fréquents ;
- Modèle basé sur les réseaux bayésiens.

A l'aide de quelques critères, nous dressons des tableaux comparatifs de ces travaux. Nous concluons notre étude par des remarques qui pourraient être importantes et utiles pour des futurs travaux dans ce domaine.

2 Fouille dans la structure

2.1 Modèle d'arbres fréquents

Les approches de ce modèle utilisent généralement un ensemble fréquent d'items d'un arbre XML. (Termier et al., 2002) proposent de faire le *clustering* des documents XML représentés par des arbres étiquetés. Chaque arbre est représenté par un itemset. Un item est une paire d'étiquettes de deux nœuds ancêtre et descendant. Les itemsets fréquents sont cal-

[2] http://www.kdnuggets.com/polls/2008/data-mining-applications.htm.

culés avec l'algorithme *A-Priori* (Agrawal et al., 1994) avec un seuil de fréquence prédéfini. Des ensembles contenant des arbres XML dans lesquels les itemsets fréquents apparaissent forment les clusters. Pour chaque cluster, l'arbre maximal est calculé en utilisant une méthode empruntée à la programmation logique inductive et en calculant une LGG (*Least General Generalization*) (Plotkin, 1970) sur une représentation adaptée des arbres de chaque cluster, qui sont des formules relationnelles. La LGG est limitée par son calcul qui est très coûteux. L'algorithme ne prend pas l'ordre des fils en considération. Il préserve la relation d'ancestralité plutôt que la relation de parenté. Par contre, il ignore les polysémies (une même étiquette dénotant différents concepts).

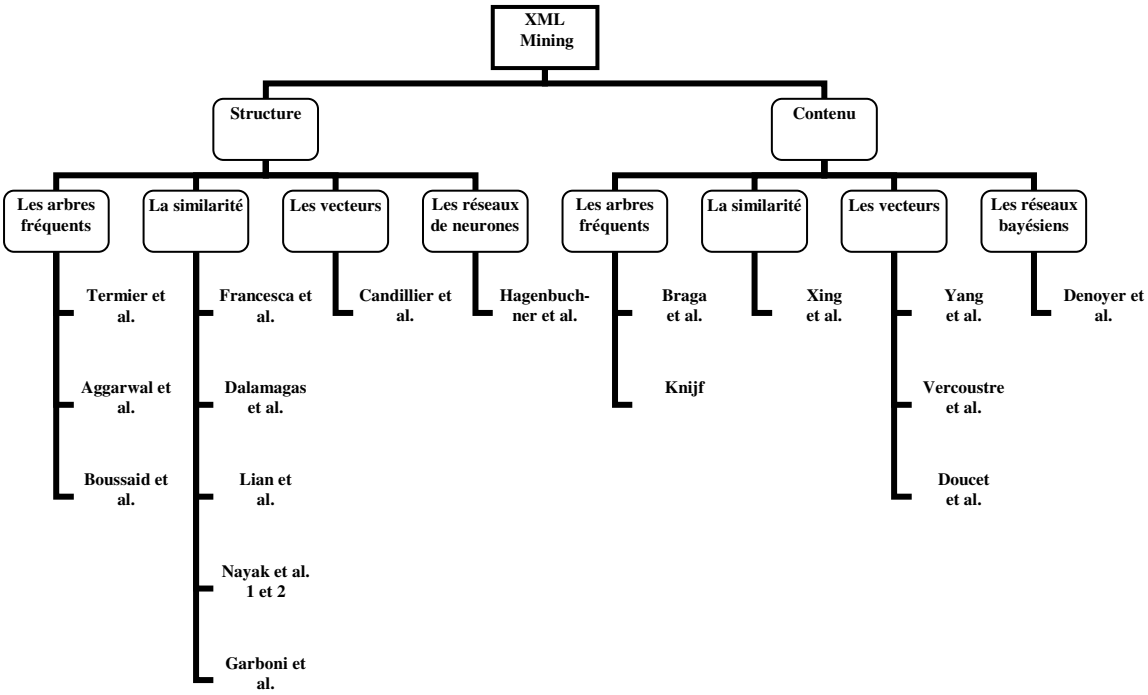


FIG. 2 – *Classification des approches de XML mining.*

Dans (Aggarwal et al., 2003), il est proposé un classifieur basé sur les règles structurelles qui utilisent des sous-structures. Une règle structurelle est de la forme $T \Rightarrow c_i$, avec T un arbre enraciné étiqueté ordonné, et c_i une des k classes dans laquelle l'arbre T est associé. La classification comporte deux phases : apprentissage et test. L'apprentissage consiste à créer le modèle de classification avec l'algorithme *XMiner*. Ce dernier prend en entrée un ensemble d'arbres avec leurs listes des scopes (la position d'un nœud et la position de son nœud descendant droit le plus bas) et une liste de seuils de supports minimaux pour chaque classe. En sortie, il génère un ensemble de règles fréquentes pour chaque classe. Ces règles sont ordonnées et élaguées en vérifiant le seuil de confiance minimum et le rapport de probabilité minimum définis par l'utilisateur. Plusieurs paramètres sont à spécifier dans cette phase, tels que les paramètres du choix des arbres fréquents, ceux de l'élagage des règles construites, etc. La phase doit aussi traiter une grande collection d'exemples afin de faciliter

la prédiction des classes dans la phase de test. Cette dernière prend en entrée le modèle de classification et un ensemble de données d'exemples avec des classes inconnues pour prédire la classe pour chaque exemple de test.

Une nouvelle méthode d'extraction des règles d'association à partir de la structure des documents XML est proposée par (Boussaid et al., 2004). Elle consiste d'abords à pré-formater les documents XML. collectant des informations sur le nom du document, son chemin (en local ou sur internet) et le nombre de balises qu'il contient (son nom et son nombre d'occurrences ainsi que ses parents et ses fils). Une matrice booléenne est constituée indiquant si une balise est renseignée ou non dans un document, elle permet d'accéder rapidement aux balises composant un document ou aux documents comprenant une balise donnée. La méthode permet de mettre en place des structures adéquates pour la gestion de la hiérarchie entre les balises, la DTD minimale en l'occurrence. Enfin, l'algorithme *A-Priori* est utilisé pour la recherche d'itemsets fréquents, les règles d'association sont extraites et les résultats sont présentés sous forme de documents XML. Cette méthode permet de représenter les liens existants entre les balises d'un ensemble de documents XML de même structure. Donc, elle n'est pas adaptée à une collection de documents XML possédant des structures hétérogènes.

2.2 Modèle basé sur la similarité

Différents travaux, (Francesca et al., 2003), (Dalamagas et al., 2004), (Lian et al., 2004), (Nayak et al., 2005), et (Nayak et al., 2007), proposent une méthodologie de *clustering* regroupant ensemble toutes les données de structures similaires en définissant des mesures de similarités. Les deux premiers travaux utilisent la distance d'édition pour mesurer la similarité proposée dans (Isert, 1999). La distance d'édition entre les arbres calcule le coût minimum (*Minimum-Cost Sequence*) des opérations requises pour convertir un arbre donné en un autre arbre (Zhang et al., 1989). Les autres utilisent d'autres mesures pour estimer la similarité.

La notion du représentatif du cluster XML est proposé dans (Francesca et al., 2003). C'est un document XML prototype qui regroupe les composants les plus pertinents d'un ensemble de documents XML dans le cluster. Il capture toutes les spécificités structurelles en utilisant la notion d'appariement structurel entre deux arbres XML. Initialement, les clusters sont définis et chaque arbre XML est placé dans son propre cluster. Une matrice contenant les distances des paires d'arbres est calculée en adoptant la distance d'édition. L'algorithme *XRep* fusionne les clusters les moins dissimilaires d'une façon itérative et la matrice des distances est mise à jour pour refléter l'opération de fusion.

Un seul arbre représentatif peut être très limité pour inclure tous les composants d'un groupe de documents XML. Les stratégies basées sur la découverte d'un pattern fréquent sont mieux appropriées. La notion d'un arbre représentatif peut être améliorée avec l'utilisation d'une forêt de sous arbres fréquents.

Dans (Dalamagas et al., 2004), des résumés structurels des arbres XML sont définis pour accélérer le calcul de la distance. Un résumé structurel est obtenu en faisant la réduction des imbrications et des répétitions dans un arbre. Un algorithme proche de celui de (Chawathe, 1999) est utilisé pour calculer la distance d'édition entre les résumés structurels des arbres. Les clusters sont définis à partir d'un graphe MST (*Minimum Spanning Tree*) (Gower et al., 1969) avec l'algorithme SLHM (*Single Link Hierarchical Method*) (Rasmussen, 1992) (Halkidi et al., 2001). Les résumés structurels représentent les nœuds du graphe MST et les arcs sont les poids représentés par les distances. Cette approche est plus appropriée pour une

collection de documents XML de structures homogènes. Pour ceux qui ont des structures différentes, il est possible que l'approche ne puisse pas donner des résultats performants.

La notion du graphe de structure (*s-graph*) est proposée dans (Lian et al., 2004). Un *s-graph* est un graphe direct qui contient les éléments et leurs relations de parenté d'un document XML. Les *s-graphs* des documents XML sont calculés et entreposés dans une structure appelé SG. Un *s-graph* est représenté par des bits qui encodent les arcs du graphe. Chaque entrée dans le SG a deux champs : des bits représentant les arcs du *s-graph* et un ensemble contenant les identifiants de tous les documents dont leurs *s-graphs* sont représentés par ces bits. Le *clustering* est alors appliqué sur les ensembles de bits par l'algorithme *S-GRACE* qui applique l'algorithme *ROCK* (Guha et al., 1999) en utilisant une métrique de distance. Les expériences ont montré que le *clustering* basé sur la notion de distance d'édition entre les arbres est très couteux par rapport au *clustering* avec *S-GRACE*. Ce dernier peut découvrir les clusters qui ne peuvent pas être facilement trouvés par une inspection manuelle.

L'algorithme *XCLS* (the *XML documents Clustering with Level Similarity*) (Nayak et al., 2005) est basé sur le calcul de *LevelSim* (*Level Similarity*) pour quantifier les similarités structurelles entre des documents XML et des clusters prédéfinis. Il groupe aussi les documents XML dans les clusters de similarité de niveau maximal. Les documents sont représentés par des arbres étiquetés qui sont transformés en des structures niveaux. Ensuite, les similarités structurelles entre l'arbre XML et chaque cluster existant sont mesurées avec *LevelSim*. Une structure niveau est une matrice qui montre les éléments dans chaque niveau de l'arbre en préservant l'hiérarchie et le contexte des éléments dans le document. Elle se concentre sur les chemins des éléments avec les valeurs du contenu. *XCLS* est plus approprié pour le *clustering* des données hétérogènes, car en plus de la mesure de similarité structurelle entre la relation père-fils il inclut la relation ancêtre des données. En calculant la similarité, il suppose que l'ordre des nœuds frères n'est pas important. *XCLS* ignore la similarité sémantique entre les balises, bien que l'aspect sémantique puisse améliorer le clustering.

XMine (Nayak et al., 2007) est une méthodologie adaptée seulement pour les schémas XML (DTD ou schéma XSD) où un schéma XML est transformé en un arbre. Ce dernier est représenté par un ensemble d'expressions de chemins. Le coefficient de similarité linguistique est mesuré en comparant chaque paire d'éléments de deux schémas en supposant que les mêmes noms ont la même sémantique et en utilisant *WordNet*. De plus, il fait appel aux experts de domaine pour définir des dictionnaires. Par contre *XMine* permet de faire le *clustering* des schémas hétérogènes ; Il peut exister des éléments de mêmes noms mais avec une sémantique différente (par exemple avocat : fruit et métier), ce qui n'est pas géré par *XMine*. Parmi un ensemble d'expressions des chemins dans deux arbres, il faut trouver les chemins similaires maximaux en utilisant une version modifiée de l'algorithme de (Agrawal et al., 1996). Les similarités entre deux schémas sont calculées et mappées en une matrice de similarités des schémas. En utilisant cette matrice, le *clustering* est appliqué et un ensemble de clusters est construit.

(Garboni et al., 2006) proposent une méthode de classification supervisée où un ensemble de clusters d'une collection doit être prédéfini à l'avance. Les nœuds des arbres étiquetés qui représentent les documents XML d'un cluster sont transformés en des identifiants. Et ils sont générés en un ensemble de séquences. Un algorithme traditionnel d'extraction de pattern séquentiel capable d'extraire les séquences fréquentes est appliqué pour pouvoir extraire à partir de chaque cluster un pattern structurel fréquent. Il effectue ensuite la classification en faisant l'appariement des documents XML avec les patterns en mesurant la distance entre les documents et les clusters. Un problème d'appariement peut se poser quand deux clusters ont

des patterns séquentiels similaires. De plus, ce n'est pas assez judicieux de choisir la première classe arbitrairement, quand deux scores sont égaux lors du calcul de la distance pour un même document XML.

2.3 Modèle basé sur les vecteurs

Dans (Candillier et al., 2005) un arbre d'un document XML est transformé en des ensembles de paires d'attribut-valeur en utilisant les différentes relations entre les nœuds de l'arbre : un ensemble de balises et leurs occurrences ; un ensemble de relations père-fils et leurs occurrences ; un ensemble de relations frères suivants et leurs occurrences ; un ensemble de chemins et sous chemins distincts commençant de la racine et leurs occurrences ; un ensemble de positions des nœuds et le nombre de leurs fils.

Des méthodes de classification et de *clustering* peuvent êtres appliquées sur les ensembles d'attribut-valeur. Un nouvel algorithme de *Subspace Clustering* SSC (*Statistical Subspace Clustering algorithm*) (Candillier et al., 2005a) est utilisé. Il est basé sur l'algorithme EM (*Expectation Maximization*) (Ye et al., 2003) en ajoutant l'hypothèse que les données ont été générées selon des distributions indépendantes sur chaque dimension. SSC nécessite un paramètre *k* qui est le nombre de clusters recherchés mais qui requiert un choix judicieux. Lorsque *k* est supérieur au nombre réel des clusters recherchés, les règles associées aux clusters se chevauchent. L'algorithme *boosted* C5 (Quinlan, 2004) est utilisé pour faire la classification sur l'ensemble de données mais en ignorant quelques attributs pour pouvoir être appliqué car un nombre très élevé d'attributs est construit.

2.4 Modèle basé sur les réseaux de neurones

(Hagenbuchner et al., 2005) présentent le *clustering* de documents XML avec deux modèles basés sur SOM (*Self Organizing Maps*) (Kohonen, 1990). SOM-SD (SOM for *Structured Data*) fait le clustering de données pour les documents XML qui sont représentés par des graphes de données structurées et CSOM-SD (*Contextual SOM-SD*) ajoute la notion du contexte (l'état des parents, des fils, des voisins,…) La limite majeure de ce modèle est qu'un bon apprentissage du réseau de neurones nécessite toujours un échantillon important. De plus, l'interprétation des groupes constitués est souvent difficile à faire.

2.5 Synthèse

Nous présentons un tableau comparatif (TAB.1) entre les approches, en nous basant sur certains critères tels que :

- Le type des documents XML utilisé (instances XML ou schémas XML)
- Le type d'entrée qui représente le document XML
- La tâche de fouille accomplie par l'approche (*clustering*, classification ou association)
- À priori : par exemple la spécification des paramètres par les experts du domaine ;
- La sémantique des éléments telle que la gestion des mêmes mots qui ont une sémantique différente, les mots différents qui peuvent avoir la même signification, etc.
- Échantillonnage : si un échantillon est nécessaire pour assurer la phase d'apprentissage.
- Sortie : définir les résultats obtenus à la fin.

	Approche	Documents XML	Entrée	Tâche de fouille	Modèle	A priori	Sémantique	Échantillonnage	Sortie
Structure	**Termier et al. 2002**	Instances	Arbre étiqueté	Clustering	Arbres fréquents	*			Clusters
	Francesca et al. 2003	Instances	Arbre enraciné étiqueté	Clustering	Similarité				Clusters
	Aggarwal et al. 2003	Instances	Règles structurelles	Classification	Arbres fréquents	*		*	Classes
	Boussaid et al. 2004	Instances	Matrice booléenne et DTD minimale	Association	Arbres fréquents	*			Doc. XML
	Dalamagas et al. 2004	Instances	Graphe connexe	Clustering	Similarité				Clusters
	Lian et al. 2004	Instances	Structure SG de bits	Clustering	Similarité	*			Clusters
	Candillier et al. 2005	Instances	Ensembles de paires d'attribut-valeur	Classification	Vecteurs	*		*	Arbre de décision
				Clustering					Hiérarchie de clusters
	Hagenbuchner et al. 2005	Instances	Vecteurs	Clustering	Réseaux de neurones			*	Clusters
	Nayak et al. 2005	Instances	Structures niveaux	Clustering	Similarité	*			Clusters
	Garboni et al. 2006	Instances	Vecteurs de séquences	Classification	Similarité			*	Clusters
	Nayak et al. 2007	Schémas	Matrice de similarité	Clustering	Similarité	*	*		Arbre de clusters

TAB. 1 – *Comparaison des approches de fouille dans la structure XML.*

3 Fouille dans le contenu XML

3.1 Modèle d'arbres fréquents

(Braga et al., 2002) développent l'opérateur *XMINE* basé sur *XPath* et inspiré de la syntaxe *XQuery* et des travaux de (Meo et al. 1998). Il permet d'extraire les règles d'association à partir de la structure et du contenu des documents XML. Une table relationnelle est générée à partir d'une collection de fragments XML définis par les expressions *XPath*. Avec la table relationnelle et les contraintes spécifiées pour chaque fragment XML. Les règles d'association sont extraites avec *A-Priori* et elles sont mappées en une représentation XML.

FAT-CAT (*Frequent Attribute Trees based Classification*) (Knijf, 2007) est une méthode de classification qui calcule d'abord les patterns fréquents des arbres enracinés, étiquetés, ordonnés pour les différentes classes sur l'ensemble d'apprentissage en utilisant l'algorithme *FAT-Miner*. Il recherche tous les sous-arbres qui apparaissent au moins *n%* (*n* est la contrainte du support minimum prédéfini) dans les données en prenant en considération les attributs. *FAT-CAT* sélectionne le pattern émergeant (Dong et al., 1999) dont son support augmente significativement d'une classe à une autre. Le pattern émergeant est utilisé pour

apprendre à faire le modèle de classification sur l'ensemble de test en construisant des composants binaires pour chaque document XML. Chaque composant indique la présence ou l'absence d'un pattern émergeant dans un enregistrement. L'implémentation de (Borgelt, 1998) est utilisée, pour construire des arbres de décisions. À chaque nœud de l'arbre de décision, un nombre d'attributs qui discrimine bien entre les classes est choisi.

3.2 Modèle basé sur la similarité

Dans (Xing et al., 2006), à partir d'un ensemble d'apprentissage contenant *n* classes prédéfinies, un seul schéma est généré pour chaque classe. Chaque schéma est converti en une grammaire NRHG (*Normalized Regular Hedge Grammar*). Le but est de faire la classification basée sur le calcul de la distance d'édition entre chaque document XML représenté par un arbre étiqueté ordonné et les NRHG des classes en calculant un vecteur de distances. Pour le contenu texte, un autre vecteur est calculé en mesurant la similarité du contenu texte des documents XML avec des techniques standards telles que : le modèle d'espace vectoriel (VSM), et LSI (*Latent Semantic Indexing*). Les vecteurs de la structure et du contenu sont concaténés pour représenter un document XML. Le schéma qui représente un groupe de documents XML dépend non seulement de la méthode d'extraction du schéma, mais aussi des documents sélectionnés pour l'extraction du schéma. Comment sélectionner les documents qui peuvent capturer les propriétés structurelles de chaque classe est une tâche intéressante. Ce système fonctionne très bien pour deux classes, cependant, la performance du système se dégrade considérablement dès que le nombre de classes grandit.

3.3 Modèle basé sur les vecteurs

(Yang et al., 2002), (Vercoustre et al., 06), (Doucet et al., 2006) sont des travaux qui représentent la structure et le contenu des documents XML par des vecteurs et appliquent sur ces vecteurs l'algorithme de *clustering k-means*.

(Yang et al., 2002) propose le modèle SLVM (*Structured Link Vector Model*), où l'information dans la structure et les liens du document sont effectivement exploités. Un vecteur SLV (*Structured Link Vector*) représente un document tel que les éléments du vecteur sont déterminés par les termes, la structure du document, et les documents voisins.

Vercoustre et al. (Vercoustre et al., 2006) ont proposé de représenter les documents XML par des ensembles de chemins générés à partir de leurs arbres en ajoutant le contenu textuel et les attributs comme des nœuds de l'arbre. Une phase de prétraitement est nécessaire afin de réduire le nombre de chemins générés, en essayant de diminuer successivement le nombre de balises, de mots vides, de suffixes... L'approche suppose une connaissance d'une sémantique implicite des éléments pour pouvoir les sélectionner. Du fait de la taille du vocabulaire généré, il n'est pas possible de faire le *clustering* basé sur la structure et le contenu d'une grande collection de documents XML.

Pour (Doucet et al., 2006), un document XML est représenté par un vecteur de *N*-dimensions dans un espace vectoriel (VSM : *Vector Space Model*), *N* est le nombre de composants du document. Le vecteur contient le poids spécifié par *TF-IDF* de chaque composant dans le document. L'approche combine le texte (*bag of words*) en supprimant les mots vides, et les suffixes des mots avec l'algorithme de (Porter, 1997), et les balises (*bag of structure*) en un seul vecteur en fusionnant les mots et les noms de balises « *text+tags* ». La qualité du *clustering* augmente avec l'utilisation des composants texte tandis que l'ajout du vecteur

structurel ne permet pas d'améliorer de manière significative les résultats du *clustering*. Une faiblesse de cette technique est que la structure arborescente des éléments est ignorée et les mots ne sont pas reliés à leur chemin dans un arbre XML.

3.4 Modèle basé sur les réseaux bayésiens

(Denoyer et al., 2004) propose un modèle d'apprentissage général pour la classification de documents structurés permettant de prendre en compte simultanément la structure et le contenu. Ils définissent un modèle génératif pour chaque document décrit par un réseau bayésien. Ils transforment ce modèle en un modèle discriminant en utilisant la méthode du noyau de Fisher. Le modèle peut être étendu facilement pour prendre en compte des types différents d'information comme la classification multimédia. La longueur du document et la profondeur sont omises dans ce modèle.

3.5 Synthèse

Le tableau ci-dessous illustre une comparaison entre les approches de fouilles dans le contenu XML en utilisant les mêmes critères cités dans la section (2.5).

	Approche	Documents XML	Entrée	Tache de fouille	Modèle	A priori	Sémantique	Échantillonnage	Sortie
Contenu	**Yang et al. 2002**	Instances	Structured Link Vector	Clustering	Vecteurs	*			Clusters
	Braga et al. 2002	Instances	Table relationnelle	Association	Arbres fréquents				Doc. XML
	Denoyer et al. 2004	Instances	Arbres	Classification	Réseaux bayésiens			*	Classes
	Vercoustre et al. 2006	Instances	Ensembles de chemins	Clustering	Vecteurs	*	*		Clusters
	Xing et al. 2006	Instances Schémas	Arbre étiqueté ordonné et grammaire NRHG	Classification	Similarité			*	Classes
	Doucet et al. 2006	Instances	Vecteurs	Clustering	Vecteurs	*			Clusters
	Knijf 2007	Instances	Arbre d'attributs enraciné étiqueté ordonné	Classification	Arbres fréquents	*		*	Arbre de décision

TAB. 2 – *Comparaison des approches de fouille dans le contenu XML.*

4 Conclusion et perspectives

Dans cet article, nous avons fait une analyse approfondie de l'état de l'art du domaine de *XML mining*. Des 18 travaux étudiés, dont 11 concernent la fouille dans la structure et 7 concernent la fouille dans le contenu XML, nous remarquons que la fouille est généralement appliquée sur les instances des documents XML bien que les schémas XML sont mieux adaptés pour la fouille dans la structure. Dans cette dernière, le modèle basé sur la similarité

est le plus utilisé avec des mesures de similarité telles que la distance d'édition même si elle est très couteuse par rapport à d'autres distances. La fouille dans le contenu reste encore moins traitée. Cependant, les techniques de *Text Mining* peuvent étendre les techniques de data mining pour le contenu textuel. Quelques aspects importants sont ignorés dans la majorité des travaux étudiés, telles la sémantique et la structure arborescente des éléments où ces derniers ne sont pas reliés à leur chemin dans un arbre XML.

L'évaluation expérimentale de ces différentes approches sur des collections de documents XML n'est pas traitée dans notre étude.

Nous réfléchissons actuellement à une approche de fouille dans la structure et le contenu des documents XML en même temps. Nous pensons utiliser les connaissances extraites de la fouille dans la structure des documents XML pour mieux conduire celle dans le contenu. Les techniques du *Text Mining* peuvent nous être utiles bien qu'elles soient insuffisantes car elles ne tiennent pas compte de la structure des documents qui véhicule des informations pertinentes.

Références

Aggarwal, C. C., et M. J. Zaki (2003). *XRules: An Effective Structural Classifier for XML Data*, SIGKDD 03.

Agrawal, R., et R. Srikant (1994). *Fast algorithms for mining association rules*. In Proceedings of 20th VLDB Conference, Santiago, Chile.

Agrawal, R., et R. Srikant (1996). Mining *Sequential Patterns: Generalizations and Performance Improvements*. Paper presented at the fifth International Conference on Extending Database Technology (EDBT'96), France.

Borgelt, C. (1998). *A decision tree plug-in for data engine*. In: Proc. 6th European Congress on Intelligent Techniques and Soft Computing.

Boussaid, O., A. Duffoux, S. Lallich, et F. Bentayeb (2004). *Fouille dans la structure de documents xml*. EGC 04, Revue des Nouvelles Technologies de l'Information, volume 2, pages 519-524, Clermont-Ferrand, France.

Braga, D., A. Campi, S. Ceri, M. Klemettinen, et PL. Lanzi (2002). *A Tool for Extracting XML Association Rules from XML Documents*, Research paper in Proceedings of IEEE-ICTAI 2002, Washington DC, USA, Nov.

Candillier, L., I. Tellier, et F. Torre (2005). *Transforming XML trees for efficient classification and clustering*. In Proceedings of the 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX'05.

Candillier, L., I. Tellier, F. Torre, et O. Bousquet (2005a). *SSC : Statistical Subspace Clustering*. In Perner, P., Imiya, A., eds.: 4th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM'2005). Volume LNAI 3587 of LNCS. Leipzig, Germany, Springer Verlag, 100-109.

Candillier, L., L. Denoyer, P. Gallinari, M. C. Rousset, A. Termier, et A.-M. Vercoustre (2007). *Mining XML Documents*, in Data Mining Patterns: New Methods and Applications Information Science Reference (Ed.) pp. 198-219.

Chawathe, S. S. (1999). *Comparing hierarchical data in external memory*, in: Proceedings of the VLDB Conference, Edinburgh, Scotland, UK, pp. 90-101.

Dalamagas T., T. Cheng, K. Winkel, et T. Sellis (2004). *Clustering XML Documents using Structural Summaries*, In Proc. of ClustWeb - International Workshop on Clustering Information over the Web in conjunction with EDBT 04, Crete, Greece.

Denoyer, L., et P. Gallinari (2004). *Bayesian Network Model for Semi-Structured Document Classification*. Revue Information Processing & Management, Special Issue on Bayesian Networks and Information Retrieval, Elsevier.

Denoyer, L., P. Gallinari, et A.-M. Vercoustre (2006). *Report on the XML Mining Track at INEX 2005 and INEX 2006 - Categorization and Clustering of XML Documents*. In: Proceedings of INEX.

Dong, G., et J. Li (1999). *Efficient mining of emerging patterns: Discovering trends and differences*. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 43–52.

Doucet, A., M. Lehtonen (2006). *Unsupervised classification of text-centric XML document collections*. In: Workshop of the INitiative for the Evaluation of XML Retrieval.

Francesca, D. F., G. Gordano, R. Ortale, et A. Tagarelli (2003), *Distance-based Clustering of XML Documents*, ECML'03 and PKDD'03 Cavtat-Dubrovnik, Croatia.

Garboni, C., F. Masseglia , B. Trousse (2006). *Sequential pattern mining for structure based XML document classification*. In: INEX 2005 Workshop of the INitiative for the Evaluation of XML Retrieval, pp. 458–468.

Gardarin, G., et T. T. Dang-Ngoc (2003). *Architecture de médiation "tout-XML" Conception et évaluation*. Ingénierie des Systèmes d'Information 8(5-6): 11-25.

Garofalakis, M., Rastogi, S. Seshredi, et K. Shim (1999). *Data mining and the web : past, present and future*. Kansas city,USA. 2nd international workshop on web information and data management.

Gower, J. C., et G. J. S. (1969). Ross, *Minimum spanning trees and single linkage cluster analysis*, Applied Statistics 18, 54-64.

Guha, S., R. Rastogi, et K. Shim (1999). *ROCK: A Robust Clustering Algorithm For Categorical Attributes*, Proc. 15th Int'l Conf. Data Eng., pp. 512-521.

Hagenbuchner, M., F. Trentini, A. Sperduti, A. Tsoi, F. Scarselli, et M. Gori (2005). *Clustering XML Documents using Self-Organizing Maps for Structures*. INEX 2005 Workshop on Mining XML documents.

Halkidi, M., Y. Batistakis, et M. Vazirgiannis (2001), *Clustering algorithms and validity measures*, in: SSDBM Conference, Virginia, USA.

Isert, C. (1999). *The editing distance between trees*, Technical report, Ferienakademie, for course 2: Bume: Algorithmik Und Kombinatorik, Italy.

Knijf, J. De.,(2007). *FAT-CAT: Frequent Attributes Tree Based Classification*. In Fuhr, N., Lalmas, M., and Trotman, A., editors, Comparative Evaluation of XML Information Retrieval Systems, INEX 2006, pages 485–496.

Kohonen, T (1990). *Self-Organisation and Associative Memory*. Springer, 3rd edition.

Lian, W., et D. W. Cheung (2004). *An Efficient and Scalable Algorithm for Clustering XML Documents by Structure*, in IEEE Transactions on Knowledge and Data Engineering.

Meo, R., G. Psaila, et S. Ceri (1998). *An extension to SQL for mining association rules*. Data Mining and Knowledge Discovery, 2(2):195 . 224.

Nayak, R., W. Rebecca, et T. Anton (2002). *Data mining and XML documents*. In Proceedings International Conference on Internet Computing, IC'2002 3, pages pp. 660-666, Las Vegas, Nevada.

Nayak, R., et S. Xu (2005). XML *documents clustering by structures with XCLS*. In: Workshop of the INitiative for the Evaluation of XML Retrieval INEX, pp. 432–442.

Nayak, R., et W. Iryadi (2007). *XML schema clustering with semantic and hierarchical similarity measures*, Knowledge-Based Systems, Volume 20, Issue 4, Pages 336-349.

Plotkin, G. (1970). *A note on inductive generalisation*. Machine Intelligence, 5:153-163.

Porter, M. F. (1997). *An algorithm for suffix stripping. In Readings in information retrieval*, San Francisco, CA, USA, pp. 313–316. Morgan Kaufmann Publishers Inc.

Quinlan, R. (2004). *Data mining tools see5 and c5.0*.

Rasmussen, E. (1992). *Clustering algorithms*, in: W. Frakes, R. Baeza-Yates (Eds.), Information Retrieval: Data Structures and Algorithms, Prentice Hall.

Termier, A., M. C. Rousset, et M. Sebag (2002). *TreeFinder : a First Step towards XML Data Mining*. International Conference on Data Mining ICDM'02, Maebashi, Japon.

Vercoustre, A.-M., M. FEGAS , S. Gul, et Y. Lechevallier (2006). *A flexible structured-based representation for XML document mining*. INEX'05, Schloss Dagstuhl, Germany, LNCS (3977), Springer.

Xing, G., et Z. Xia (2006). *Classifying XML documents based on structure/content similarity*. In: Workshop of the INitiative for the Evaluation of XML Retrieval.

Yang, J., et C. Xiaoou (2002). *A semi-structured document model for text mining*, Journal of Computer Science and Technology archive, Volume 17(5), 603-610.

Ye, L., M. Spetsakis (2003). Clustering *on unobserved data using mixture of gaussians*. Technical report, York University, Toronto, Canada.

Zaki, M. J. (2002). *Efficiently Mining Frequent Trees in a Forest*. SIGKDD.

Zhang, K. et D. Shasha (1989). *Simple fast algorithms for the editing distance between trees and related problems*, SIAM J. Comput., 18(6):1245–1262.

Summary

In this paper, we are going to browse many works achieved in the literature concerning the XML documents mining (XML structure mining and XML content mining). We classify these approaches according to five models of presentation of XML documents. We present a comparison between the different works while using some criterias of comparison. At the end, we concluded our overview by remarks that could be important and useful for future works in this domain.

Extraction de structure d'un document XML : Modélisation Booléenne

Fawzia Zohra Abdelouhab, Baghdad Atmani

fzabdelouhab@yahoo.fr
atmani.baghdad@univ-oran.dz

Equipe de recherche « Simulation, Intégration et Fouille de données »
Département Informatique, Faculté des Sciences, Université d'Oran
BP 1524, El M'Naouer, Es Senia, 31 000 Oran, Algérie

Résumé. Notre travail s'insère dans le cadre de l'entreposage de données en proposant, une nouvelle voie de recherche qui est née à partir de la rencontre des automates cellulaires et des bases de données décisionnelles. L'un des challenges intéressant, consiste à combiner les fonctions booléennes et les techniques mathématiques formelles de la machine cellulaire CASI (Atmani et Beldjilali, 2007) avec les performances des bases de données décisionnelles, afin, de concevoir une intégration automatique des données semi-structurées des sources hétérogènes dans un entrepôt de données. Cette intégration est entièrement guidée par la machine cellulaire CASI. De cette dernière, nous avons dérivé une nouvelle machine nommée BICS-XML en trois processus cellulaires : le premier processus consiste à extraire le schéma des documents sources XML représenté sous forme d'un graphe, appelé schéma spécifique. Ce dernier sera soumis au deuxième processus qui, à la base du schéma global de l'entrepôt, appelé schéma générique, en extrait le schéma le plus ressemblant au schéma spécifique. Le troisième processus détermine les correspondances entre les schémas génériques et le schéma spécifique. Nous nous positionnons, dans cet article, dans le cas de l'extraction d'un schéma spécifique d'un document XML, seulement, et nous montrons comment se fait l'intégration automatique des données en utilisant une modélisation booléenne.

Mots clés: Entrepôts de données, Intégration des données hétérogènes, Automate cellulaire, Document XML, Règles de production.

1 Introduction

Le traitement de données complètement hétérogènes structurées, semi-structurées et non structurées s'avère être un volet de recherche récent et assez peu exploré. Dans un tel contexte, le besoin d'intégration devient un concept incontournable mais en même temps com-

pliqué car il se voit contraint de composer avec la répartition des sources, l'hétérogénéité de leurs structures et la complexité de leurs données Boussaid et al. (2006) et Nassis et al. (2004).

Il existe deux méthodes parallèles pour l'intégration des données : l'entrepôt (Inmon, 1992) et Kimball (1998) et le médiateur (Goasdoué et al., 2000), Lamarre et al. (2004) et Huang et al. (2000). La première, qui fait l'objet de notre étude, consiste à construire une base de données réelle et centralisée, selon un schéma particulier. Celle-ci contient les données intégrées à partir des différentes sources de données et elle est prête à supporter le processus d'analyse en ligne OLAP (Choquet et Boussaid, 2007). Cette approche est caractérisée par sa performance en termes de temps de réponse des requêtes. D'un autre côté, un entrepôt de données, qu'il soit homogène ou hétérogène, nécessite d'être maintenu. Il doit aussi évoluer en fonction de l'évolution des sources aussi bien au niveau des structures qu'au niveau des données**.** Le problème de la maintenance est également très complexe. Les algorithmes proposés dans la littérature traitent essentiellement des données de sources homogènes (Laurent et al., 2001), (O'Gorman et al., 1999) et (Zhuge et al., 1995,1996). Par contre, concernant les données sources hétérogènes la question reste à étudier...

Au cours de cet article, nous allons présenter notre contribution dans cette approche combinatoire de l'intelligence artificielle, plus précisément, les automates cellulaires et les bases de données décisionnelles. Cette contribution est concrétisée par la spécification, la conception et la réalisation d'un nouveau système cellulaire, ***BICS-XML,*** pour l'intégration booléenne des données semi structurées en XML dans un entrepôt de données. Nous détaillerons plus précisément la partie extraction de schéma spécifique d'un document XML.

Nous avons organisé notre article comme suit : la section 2 présente les travaux existant dans le domaine de l'intégration des données et plus précisément l'extraction de structure des documents XML. La section 3 donne une idée sur l'architecture générale du système *BICS-XML*. La section 4 présente notre approche d'extraction automatique booléenne du schéma spécifique. La conclusion et les perspectives sont discutées dans la section 5.

2 Etat de l'art

Ces dernières années ont été témoin de l'émergence de l'intégration des données et des challenges inhérents à cette problématique. Parmi ces derniers sont concernés la gestion des schémas, l'évolution des schémas, le Mapping et le Matching des schémas (voir Sellami et al., 2007).

En analysant les travaux réalisés dans ce sens, nous avons répertorié deux catégories d'approches : le Matching et le Mapping des schémas XML :

1. **Le Matching :** est un processus qui effectue des correspondances sémantiques entre les éléments et les attributs des schémas, et retourne comme résultat les valeurs de similarités sémantiques entre les deux schémas. Nous parlons d'un appariement structurel (Madhavan et al., 2001), Similarity Flooding proposé par S. Melnik et al. (2002) et utilisée dans Lamolle et Zerdazi, (2007).
2. **Le Mapping :** se sont des expressions décrivant le moyen dont les instances du schéma cible (final) sont dérivées à partir des instances de schéma source (initial). Elles décrivent la correspondance sémantique entre les instances de schémas en complémentarité avec le Matching. Et là nous parlons d'un appariement ontologique qui dépend du domaine d'application à traiter (Boussaid et al., 2006) et Sellami et al. (2007).

Ces deux processus qui se suivent sont des pré-requis à l'intégration et à la transformation des schémas XML. Avec ces deux descriptions nous pouvons restructurer les données d'une représentation à une autre ce qui permet de passer d'un schéma source à un schéma cible et vis-et-versa. Le problème essentiel dans la mise au point de ces techniques est lié à l'hétérogénéité et à la diversité des sources d'informations (voir Abdelouhab et Atmani, 2008).

Le projet XYLEM (Sorlin et Solnon, 2004) est un système d'entrepôt dynamique ayant pour but de stocker et d'intégrer de manière semi automatique toutes les ressources XML du Web. Ce stockage permet à l'utilisateur final d'avoir un accès unique et transparent à toutes les données hétérogènes. L'utilisation d'un système à base d'arbre par Delobel et al. (2003), contribue à faire de XYLEM un système efficace pour l'évaluation des requêtes, l'intégration des données et leur maintenance.

Au regard des travaux récents nous avons conclu qu'il n'existe aucune solution au problème de l'intégration des données combinant les automates cellulaires avec les bases de données décisionnelles, d'où notre contribution.

3 Architecture générale de BICS-XML

La première motivation de notre approche est de fournir une plateforme d'intégration qui enchaîne les processus comme suit : figure1.

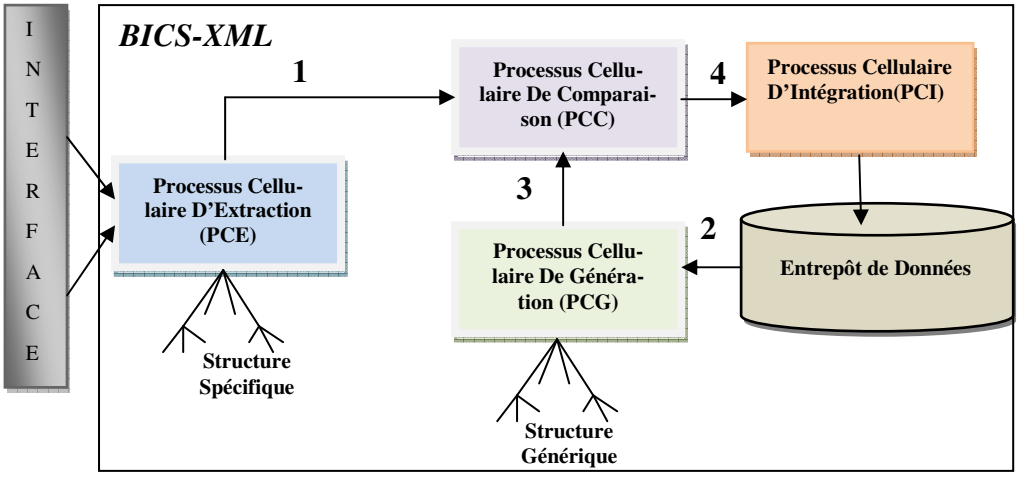


FIG. 1 – *Architecture générale de BICS-XML*

1. Le Processus Cellulaire d'Extraction (PCE)

L'entrée de cette couche est un document XML saisi par l'utilisateur et sera qualifié de spécifique. Ce document est traité par la machine cellulaire développée par Atmani et Beldjilali (2007) afin d'en extraire la structure spécifique (le schéma du document). La reconnaissance de structures passe par l'identification de chaque granule d'information contenu dans le document source. Cette identification est considérablement facilitée par la présence des balises des documents structurés XML.

Le résultat de cette phase est une structure arborescente ordonnée et étiquetée. Chaque nœud de cette arborescence peut être mono-valué (cardinalité " ", c'est la valeur par défaut), ou multi-valué (cardinalités "+" : un-plusieurs ou "*" : zéro-plusieurs).

Les étiquettes d'une arborescence correspondent aux différents noms des balises extraites du document.

2. Le Processus Cellulaire de Génération (PCG)

Ce processus permet d'extraire, à partir l'entrepôt de données, la structure, dite générique, la plus ressemblante à la structure spécifique, en utilisant la machine cellulaire présentée dans Abdelouhab et Atmani, (2008).

3. Le processus Cellulaire de Comparaison (PCC)

L'entrée de cette couche étant deux structures sous forme de graphes orientés et étiquetés. Le principe consiste à utiliser le moteur d'inférence cellulaire (Atmani et Beldjilali, 2007) afin de mesurer la similarité ou la distance entre les structures des deux documents manipulés, c'est-à-dire, pouvoir quantifier les points communs et les différences entre les deux structures.

4. Le processus Cellulaire d'Intégration (PCI)

L'entrée de cette couche est le résultat de la comparaison de la couche précédente qui est donné sous forme d'un coefficient de similarité. Si le système ne trouve aucun schéma générique répondant aux critères de la structure spécifique, on en déduit qu'il s'agisse d'une nouvelle structure à insérer dans l'entrepôt. Sinon, le système procède à un appariement d'abords structurel en utilisant les règles de Matching ensuite sémantique en utilisant les règles de Mapping. Suite à cet appariement deux cas peuvent se présenter :

Cas 1. La structure spécifique est identique ou entièrement incluse dans la structure générique. Dans ce cas, nous pouvons rattacher la structure spécifique du document à cette structure générique ;

Cas 2. La structure spécifique n'est pas entièrement identique ou entièrement incluse dans la structure générique. Dans ce cas, il faudra adapter l'une en fonction de l'autre.

4 Processus cellulaire d'extraction de schéma

4.1 Modèle de données

Il est nécessaire dans un premier temps d'utiliser un modèle commun pour assurer une bonne compréhension et une bonne intégration des données échangées. Ce modèle devra être indépendant des sources et permettre de construire une vue commune sur ces différentes sources. Le but d'une telle représentation logique est de capturer les caractéristiques structurelles et sémantiques des schémas de départ pour faciliter leur compréhension tout en offrant plus de souplesse et plus de flexibilité lors de leur comparaison et de leur intégration. Le choix s'est porté sur les schémas XML (voir Abdelouhab et Atmani, 2008).

Dans notre approche de modélisation, nous présentons la structure logique d'un document sous forme d'un ensemble, imbriqué et ordonné d'éléments logiques spécifiques, représenté par des balises (Figure 2). Chaque élément Balise représente une partie du document, qui peut-être lui-même décomposé en sous éléments Balises. Les attributs spécifiques permettent de décrire les éléments Balises tels que X1, X2 et X3.

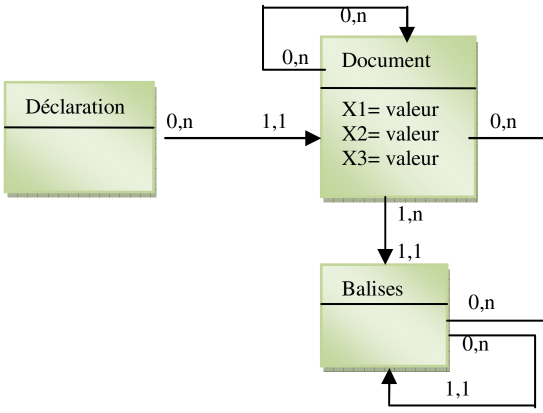


FIG. 2 – *Modélisation de la structure logique d'un document*

4.2 Domaine d'application

Comme cet article est dédié à présenté, seulement, la première couche d'extraction et la transformation des schémas, notre application se limite à saisir ou à charger un dossier de type XML dans un entrepôt de données cellulaire. Pour cette raison nous avons opté pour un exemple simple de document XML décrivant un article Audio (Figure 3) de taille conséquente afin de montrer toutes les facettes du projet.

```
<Doc_Audio>
<Section>
<Tour>
<Locuteur Nom = Loc1, Genre = 'M', Accent ='Native'>
<Date_Deb> 0 </Date_Deb> <Date_Fin> 172</Date_Fin>
<Transcription> Bonjour …</Transcription>
<Transcription> Dans ce …</Transcription>
<Transcription> … </Transcription>
</Locuteur>
<Effet_Spécial Description = 'Rire'>
<Date_Deb> 172 </Date_Deb>
<Date_Fin> 175 </Date_Fin>
</Effet_Spécial>
<Effet_Spécial Description = 'Applaudissement'>
<Date_Deb> 175 </Date_Deb>
<Date_Fin> 200 </Date_Fin>
</Effet_Spécial>
<Locuteur Nom = Loc2, Genre = 'M'>
<Date_Deb> 200 </Date_Deb>
<Date_Fin> 334 </Date_Fin>
<Transcription> Merci …</Transcription>
<Transcription> … </Transcription>
</Locuteur>
</Tour>
<Tour>
<Locuteur Nom = Loc1, Genre = 'M'>
<Date_Deb> 334 </Date_Deb>
<Date_Fin> 402 </Date_Fin>
<Transcription> On va passer …</Transcription>
</Locuteur>
</Tour>
<Tour> … </Tour>
</Section>
<Section> … </Section>
</Doc_Audio>
```

FIG. 3 – *Fichier correspondant au Document Audio.XML*

4.3 Conception du processus d'extraction

Afin de faciliter la génération des règles de structures qui régissent le document XML en entrée, une phase de réécriture du document est nécessaire pour en extraire les informations

pertinentes. Cette réécriture du document consiste à générer une table de correspondance de quatre colonnes décrite comme suit:

1- *NOM* : désigne les noms des classes (balises du document XML) définissant les sommets exemple : **<Doc_Audio>**
<Section>
<Tour>
<Locuteur Nom = Loc1, Genre = 'M', Accent ='Native'>
Sur cette partie du Document Audio, nous avons *Doc_Audio*, *Section*, *Tour* et *Locuteur* qui définissent les sommets avec *Doc_Audio* étant la racine du futur arbre.

2- *NIVEAU* : désigne les niveaux des sommets dans l'hiérarchie
EXEMPLE : Doc*_Audio* est un sommet de niveau 1. Il contient une sous balise '*Section* ' qui elle sera de niveau 2 et ainsi de suite ;

3- *CARDINALITE* : désigne le nombre d'occurrence de chaque sommet. Initialement tous les sommets ont une cardinalité égale à 1 ;

4- *ATTRIBUTS* : désigne la liste des attributs de chaque sommet
EXEMPLE : dans **<Locuteur Nom = Loc1, Genre = 'M', Accent ='Native'>,** *Nom*, *Genre* et *Accent* sont des attributs du sommet *Locuteur* ;

Le résultat de la réécriture est une table multivaluée donnée comme suit :

Nom	Niveau	Cardinalité	Liste sommet et attribut
Doc_Audio	1	1	Doc_Audio ,
Section	2	1	Section ,
Tour	3	1	Tour ,
Locuteur	4	1	Locuteur ,Nom,genre,accent
Date_Deb	5	1	Date_Deb ,
Date_Fin	5	1	Date_Fin ,
Transcription	5	1	Transcription ,
Transcription	5	1	Transcription ,
Transcription	5	1	Transcription ,
Effet_Spécial	4	1	Effet_Spécial ,description
Date_Deb	5	1	Date_Deb ,
Date_Fin	5	1	Date_Fin ,
Effet_Spécial	4	1	Effet_Spécial ,description
Date_Deb	5	1	Date_Deb ,
Date_Fin	5	1	Date_Fin ,
Locuteur	4	1	Locuteur ,Nom,genre
Date_Deb	5	1	Date_Deb ,
Date_Fin	5	1	Date_Fin ,
Transcription	5	1	Transcription ,
Transcription	5	1	Transcription ,
Tour	3	1	Tour ,
Locuteur	4	1	Locuteur ,Nom,genre
Date_Deb	5	1	Date_Deb ,
Date_Fin	5	1	Date_Fin ,
Transcription	5	1	Transcription ,
Tour	3	1	Tour ,
Section	2	1	Section ,

TAB. 1 – *Les informations pertinentes.*

A partir de la table1, nous générons la base de connaissance initiale suivante où chaque règle est exprimée comme suit : Si *Prémisse* Alors *Conclusion*, où *Prémisse* est une conjonction de Balises établies et *Conclusion* une conjonction de proposition de type Balise ou Attribut=valeur.

R1 : SI Doc_Audio ALORS Section;
R2 : SI Doc_Audio, Section ALORS Tour;
R3 : SI Doc_Audio, Section, Tour ALORS Locuteur;
R4 : SI Doc_Audio, Section, Tour, Locuteur ALORS Date_Deb;
R5 : SI Doc_Audio, Section, Tour, Locuteur ALORS Date_Fin;
R6 : SI Doc_Audio, Section, Tour, Locuteur ALORS Transcription;
R7 : SI Doc_Audio, Section, Tour, Locuteur ALORS Transcription;
R8 : SI Doc_Audio, Section, Tour, Locuteur ALORS Transcription;
R9 : SI Doc_Audio, Section, Tour ALORS Effet_Spécial;
R10 : SI Doc_Audio, Section, Tour, Effet_Spécial ALORS Date_Deb;
R11 : SI Doc_Audio, Section, Tour, Effet_Spécial ALORS Date_Fin;
R12 : SI Doc_Audio, Section, Tour ALORS Effet_Spécial;
R13 : SI Doc_Audio, Section, Tour, Effet_Spécial ALORS Date_Deb;
R14 : SI Doc_Audio, Section, Tour, Effet_Spécial ALORS Date_Fin;
R15 : SI Doc_Audio, Section, Tour ALORS Locuteur;
R16 : SI Doc_Audio, Section, Tour, Locuteur ALORS Date_Deb;
R17 : SI Doc_Audio, Section, Tour, Locuteur ALORS Date_Fin;
R18 : SI Doc_Audio, Section, Tour, Locuteur ALORS Transcription;
R19 : SI Doc_Audio, Section, Tour, Locuteur ALORS Transcription;
R20 : SI Doc_Audio, Section ALORS Tour;
R21 : SI Doc_Audio, Section, Tour ALORS Locuteur;
R22 : SI Doc_Audio, Section, Tour, Locuteur ALORS Date_Deb;
R23 : SI Doc_Audio, Section, Tour, Locuteur ALORS Date_Fin;
R24 : SI Doc_Audio, Section, Tour, Locuteur ALORS Transcription;
R25 : SI Doc_Audio, Section ALORS Tour;
R26 : SI Doc_Audio ALORS Section;
R27 : SI Locuteur ALORS Nom, Genre, Accent;
R28 : SI Effet_Spécial ALORS Description;
R29 : SI Effet_Spécial ALORS Description;

FIG. 4 – *La base de connaissance générée*

4.3.1 Le Moteur d'inférence cellulaire de BICS-XML

La base de connaissance telle qu'elle est générée n'est pas optimale et peut présenter beaucoup de redondantes.

Définition1 *: on appelle des règles redondantes un ensemble de règles ayant les mêmes prémisses. Ex :* ***R30 : SI Locuteur ALORS Nom, Genre et R31 : SI Locuteur ALORS Nom, Genre.***

Cependant, cette redondance est utile car elle montre que des règles peuvent être soit incluses l'une dans l'autre, soit identiques.

Définition2 *: on appelle des règles incluses l'ensemble de règles redondantes ayant des conclusions différentes. Ex :* ***R4 : SI Doc_Audio, Section, Tour, Locuteur ALORS Date_Deb et R5 : SI Doc_Audio, Section, Tour, Locuteur ALORS Date_Fin.***

Définition3 *: on appelle des règles identiques l'ensemble de règles redondantes ayant les mêmes conclusions. Ex :* ***R30 : SI Locuteur ALORS Nom, Genre et R31 : SI Locuteur ALORS Nom, Genre.***

Pour optimiser la base de connaissance, nous allons utiliser le moteur d'inférence cellulaire (Abdelouhab et Atmani, 2008) dans *BICS-XML* pour éliminer toutes les redondances et la simplifier au maximum tout en préservant la cohérence et la sémantique du document. Deux types de simplification seront envisagés : Pour les règles Redondantes Incluses : l'ensemble de toutes les Règles Redondantes Incluses sera remplacé par une seule Règle dont la conclusion sera formée par l'union de toutes les conclusions de l'ensemble des règles qu'elle remplace. Les attributs et/ou les sommets qui y figurent seront enregistrés avec la cardinalité "*".
Exemple : R4 et R5 pour obtenir R4'

R4 : SI Doc_Audio, Section, Tour, Locuteur ALORS Date_Deb
R5 : SI Doc_Audio, Section, Tour, Locuteur ALORS Date_Fin

R4' : SI Doc_Audio, Section, Tour, Locuteur ALORS Date_Deb*, Date_Fin*

Pour les règles Redondantes Identiques : l'ensemble de toutes les Règles Redondantes Identiques sera remplacé par une seule Règle. Les attributs et/ou les sommets qui figurent dans la partie conclusion seront enregistrés avec la cardinalité "+".
Exemple : R30 et R31 pour obtenir R30'

R30 : SI Locuteur ALORS Nom, Genre
R31 : SI Locuteur ALORS Nom, Genre

R30' : SI Locuteur ALORS Nom+, Genre+

Le moteur cellulaire de la machine BICS-XML, simule le fonctionnement du cycle de base d'un moteur d'inférence en utilisant deux couches finies d'automates finis. La première couche, *CELFACT*, pour la base des faits et, la deuxième couche, *CELRULE*, pour la base de règles. Chaque cellule au temps *t+1* ne dépend que de l'état de ses voisines et du sien au temps *t*. Dans chaque couche, le contenu d'une cellule détermine si elle participe et comment elle participe à chaque étape d'inférence. A chaque étape, une cellule peut être active (1) ou passive (0), c'est-à-dire participe ou non à l'inférence. Le principe est simple :
– Toute cellule *i* de la première couche *CELFACT* est considérée comme fait établi si sa valeur est 1, sinon, elle est considérée comme fait à établir.
– Toute cellule *j* de la deuxième couche *CELRULE* est considérée comme une règle candidate si sa valeur est 1, sinon, elle est considérée comme une règle qui ne doit pas participer à l'inférence.

La configuration initiale de la machine est donnée par l'état initial des cellules de bases *CELFACT* et *CELRULE*. Les états des cellules chacun représenté par des vecteurs se composent de trois états Entrée, Interne et de Sortie comme le note le tableau 2 suivant:

Couches	Entrée	Interne	Sortie
CELFACT	EF	IF, ID, IN	SF
CELRULE	ER	IR	SR

TAB. 2 – *Couches de la machines BICS-XML.*

Le vecteur ID gère les règles redondantes identiques (indique la cardinalité '+')
Le vecteur IN gère les règles redondantes incluses (indique la cardinalité '*')
Le vecteur IF : indique le rôle du *Fait* dans le graphe : Si IF = 0, le Fait est du type sommet ; et Si IF = 1, le Fait est du type *attribut=valeur*. Initialement, toutes les entrées des cellules dans la couche *CELFACT* sont passives (*EF* = 0), exceptées celles qui représentent la base de faits initiale (*EF*(1) = 1). A partir de ces notations nous construisons les couches *CELFACT*

et *CELRULE*. En plus de ces deux couches, la machine BICS-XML utilise deux autres matrices d'incidence R_E et R_S représentant la correspondance d'entrée et de sortie des faits par rapport aux règles.

–la relation d'entrée, notée iR_Ej, est formulée comme suit : $\forall i \in [1, l], \ \forall j \in [1, r]$, si (le Fait $i \in$ à la *Prémisse* de la règle j) alors $R_E(i, j) \leftarrow 1$.

–la relation de sortie, notée iR_Sj, est formulée comme suit : $\forall i \in [1, l], \ \forall j \in [1, r]$, si (le Fait $i \in$ à la *Conclusion* de la règle j) alors $R_S(i, j) \leftarrow 1$.

La configuration initiale de la machine est générée comme suit (voir figure 5) :

Faits	EF	ID	IN	IF	SF
Doc_Audio	1	0	0	0	0
Section	0	0	0	0	0
Tour	0	0	0	0	0
Locuteur	0	0	0	0	0
Nom	0	0	0	1	0
Genre	0	0	0	1	0
Accent	0	0	0	1	0
Date_Deb	0	0	0	0	0
Date_Fin	0	0	0	0	0
Transcription	0	0	0	0	0
Effet_Spécial	0	0	0	0	0
Description	0	0	0	1	0

REGLES	ER	IR	SR
R1	0	1	1
R2	0	1	1
R3	0	1	1
R4	0	1	1
R5	0	1	1
⋮			
R17	0	1	1
R18	0	1	1
R19	0	1	1
⋮			
R29	0	1	1
R30	0	1	1
R31	0	1	1

FIG. 5 – *Couches CELFACT et CELRULE*

R_E

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
Doc_Audio	1	1	1	1	1	1	1	1	1	1
Section	0	1	1	1	1	1	1	1	1	1
Tour	0	0	1	1	1	1	1	1	1	1
Locuteur	0	0	0	1	1	1	1	0	0	
Nom	0	0	0	0	0	0	0	0	0	0
Genre	0	0	0	0	0	0	0	0	0	0
Accent	0	0	0	0	0	0	0	0	0	0
Date_Deb	0	0	0	0	0	0	0	0	0	0
Date_Fin	0	0	0	0	0	0	0	0	0	0
Transcription	0	0	0	0	0	0	0	0	0	0
Effet_Spécial	0	0	0	0	0	0	0	0	0	1
Description	0	0	0	0	0	0	0	0	0	0

R_S

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
Doc_Audio	0	0	0	0	0	0	0	0	0	0
Section	1	0	0	0	0	0	0	0	0	0
Tour	0	1	0	0	0	0	0	0	0	0
Locuteur	0	0	1	0	0	0	0	0	0	0
Nom	0	0	0	0	0	0	0	0	0	0
Genre	0	0	0	0	0	0	0	0	0	0
Accent	0	0	0	0	0	0	0	0	0	0
Date_Deb	0	0	0	1	0	0	0	0	0	1
Date_Fin	0	0	0	0	1	0	0	0	0	0
Transcription	0	0	0	0	0	1	1	1	0	0
Effet_Spécial	0	0	0	0	0	0	0	0	1	0
Description	0	0	0	0	0	0	0	0	0	0

FIG. 6 – *Configuration des matrices d'entrées/sorties R_E et R_S des 10 premières règles*

4.3.2 Mécanismes d'inférences

Le cycle du moteur d'inférence effectue son travail par séquence. Il passe d'une configuration à l'autre en appliquent des fonctions de transitions ***δrule*** **et** ***δfact*** définies comme suit :

- La fonction de transition δ_{fact} :

δ_{fact} (EF, IN, ID, IF, SF, ER, IR, SR)=(EF, IN, **ID+(R_S.Tj)**, IF, EF, **ER+Tj**, IR, SR)

- La fonction de transition δ_{rule} :

δ_{rule} (EF, IN, ID, IF, SF, ER, IR, SR)=(**EF+(R_S.ER)**, **IN+(R_S.Tj)**, ID, IF, SF, ER, Tj, **SR⊕H**)

Le déroulement du moteur se fait en appliquant le programme suivant :

Procédure Analyser Séquence (***CELFACT*** **[][5],** ***CELRULE*** **[][3], R_E [][], R_S [][]) ;**

```
Début
j := 0 ;
Tantque (j<M) faire si (ER (j==0) alors
                        Pour (i=0 à N) faire
                        TRj :=RE
                        Δ := δrule ◦ δfact ;
                        Fin pour ;
                        Fin si ;
j :=j+1 ;
Fin tant que ;
Fin.
```

Pendant chaque séquence il effectue les opérations suivantes :

1- Détection des règles redondantes
2- Filtrage des règles

Ceci peut être traduit par les algorithmes suivants :

Algorithme de transition

Entrée : La couche *CELRULE*
Début
1. Faire Pour i=1 à n
2. Si ER(Ri)=0 Alors LancerInférence(Ri)
3. Transition*CELRULE* (Ri) /* permet de déterminer les règles Redondantes
4. Transition*CELFACT* (Ri)/* permet de séparer les incluses des identiques
5. Optimiser (Ri) /* désactive les règles redondantes à Ri
6. FinPour

Fin

Procédure LancerInférence(R)

Début
1. TR : vecteur de n éléments (n=nombre de règles) initialisé à 0
2. Pour j=1 à n
3. Si R_E (Rj) = R Alors TR(j)=1

Fin

Procédure Transition*CELRULE* (R)

```
Début
HR : vecteur de n éléments (n=nombre de règles) initialisé à 1 sauf pour l'indice de la règle R
Pour i=1à n
        IR[i] = TR[i]
        ER[i] = ER[i] +TR[i]
        SR[i] =SR[i] ⊕ HR[i]
FinPour
Fin
```

Procédure Transition*CELFACT* (R)

```
Début
Pour j=1 à n
   Som=Som+TR[j]
   Si Som<>1 Alors G[i] =RS[i] [i] * TR[i]
                     Sinon G[i]=0
Pour i=1à n
        EF[i] = EF+ (RS[i] [i]*ER[i])
        IN[i] = IN[i] + (RS[i] [i]*TR[i]) /* IN=1ٹ  le Fait est en Inclusion (Cardinalité=*)
        ID[i] = ID[i] +G[i] /* ID=1 ٹ  Le Fait est identique (Cardinalité=+)
        SF[i] = SF[i] + EF[i] /* le Fait est établi en sortie
FinPour
Fin
```

Après plusieurs itérations la machine converge vers une configuration finale. A partir de cet état final, le système valide les règles pertinentes et désactive les règles redondantes. En appliquant ce principe à notre exemple illustratif nous obtenons la base de connaissances optimale suivante :

R1 : SI Doc_Audio ALORS Section(+).
R2 : SI Doc_Audio, Section ALORS Tour(+).
R3 : SI Doc_Audio, Section, Tour ALORS Locuteur(*), Effet_Spécial(*).
R4 : SI Doc_Audio, Section, Tour, Locuteur ALORS Date_Deb(*), Date_Fin(*), Transcription(*).
R5 : SI Doc_Audio, Section, Tour, Effet_Spécial ALORS Date_Deb(*), Date_Fin(*).
R6 : SI Locuteur ALORS att:Nom(+), att:Genre(+), att:Accent(*).
R7 : SI Effet_Spécial ALORS att:Description(+).

4.3.3 Construction de l'arbre

A partir de la base de connaissances optimale, le système génère l'arbre de structure spécifique comme illustrer dans la figure 7 :

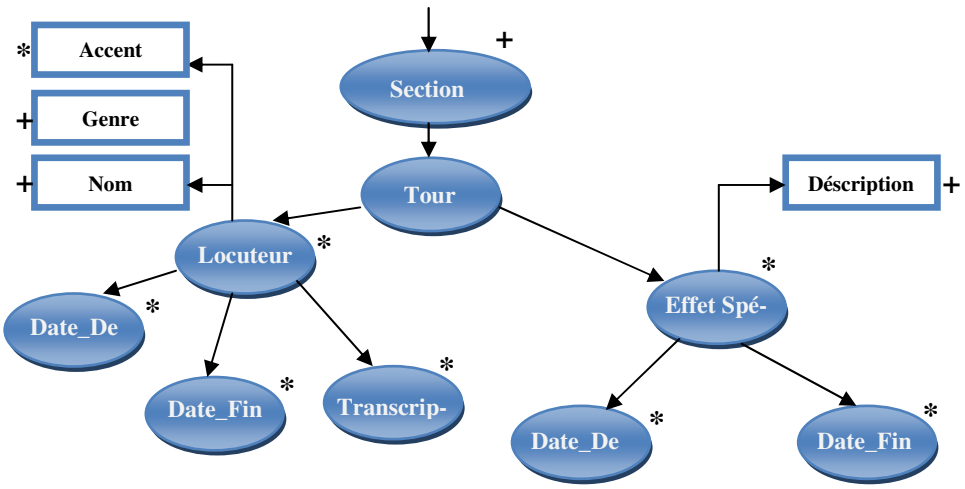


FIG. 7 – *L'arbre de la structure spécifique*

4.3.4 Représentation et stockage

Pour sauvegarder les données, nous avons, d'abords, modélisé notre arbre de sortie sous forme d'un schéma en étoile où la table de fait est construite à partir des noms des attributs et des balises. Chacun d'eux représente une clé vers une dimension. La dimension représente l'état des cellules de la couche *CELFACT*, et des matrices R_E et R_S pour la clé concernée (voir figure 8).

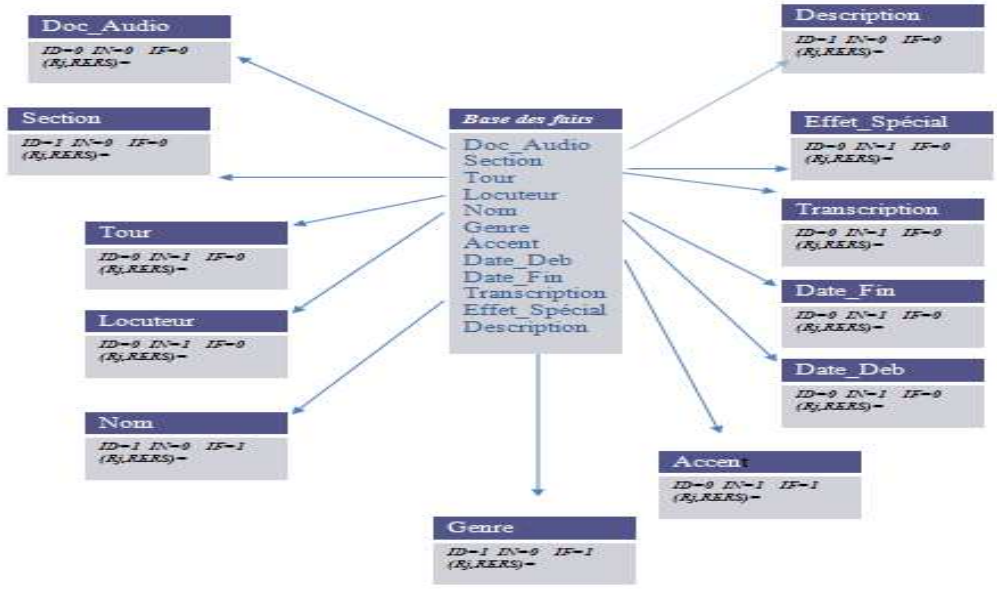


FIG. 8 – *Schéma en étoile de l'arbre final.*

Au niveau physique chaque document XML, repéré par son nom, est représenté par une liste d'éléments pouvant être des noms d'attributs ou des noms de balises. Chaque élément est représenté par son état interne, à savoir, les valeurs de ID, IN et IF de *CELFACT*, suivi par la liste de toutes les règles auxquelles il participe.

5 Conclusion et perspectives

Sur un thème très porteur en bases de données et en intelligence artificielle, cet article présente notre apport au problème de l'entreposage virtuel des données semi structurées. Nous l'avons dédié uniquement au problème de l'extraction et de transformation des schémas XML. En se basant sur l'architecture générale du système, nous avons présenté un aperçu sur la spécification et la conception de chaque module constituant le système. Nous avons présenté, en détaille, la partie extraction de schéma de structure en se basant sur le formalisme mathématique de la machine CASI.

D'après (Euzenat, 2008) l'ensemble des travaux d'intégration réalisés restent, jusqu'à nos jours, en manque d'évaluation. Dans le cas présent, ces travaux doivent être évalués que par l'expérimentation. Dans notre cas, une solution logicielle est en gestation ne faisant appel à aucun produit préconçu. Nous envisageons, par la suite, de compléter l'implémentation de l'approche sur un cas réel complexe pour pouvoir comparer sa précision avec les approches existantes.

En dépit de nombreux outils disponibles sur le marché, le problème d'intégration des données demeure entier en particulier dès que l'on passe à un contexte fortement dynamique et à large échelle.

Références

Abdelouhab, F., B. Atmani (2008). Intégration automatique des données semi-structurées dans un entrepôt cellulaire, Troisième atelier sur les systèmes décisionnels, pp. 109-120. Mohammadia – Maroc10 et 11 octobre 2008

Atmani, B., B. Beldjilali (2007a). Knowledge Discovery in Database: Induction Graph and Cellular Automaton, Computing and Informatics Journal, V.26, N°2 (2007) 171-197.

Boussaid O., R. Ben Messaoud, R. Choquet, S. Anthoard (2006). *Conception et construction d'entrepôts en XML.* Dans la RNTI correnspondant à la 2ième journée francophone sur les entrepôts de données et l'analyse en ligne EDA'06 Versaille 19.

Choquet, R., O. Boussaïd (2007). *Interrogation OLAP d'un entrepôt de données XML.* ECG2007.

Delobel, C., C. Reynaud, M.C. Rousset, J.P. Sirot, D. Vodislav (2003). *Semantic integration in Xyleme : A uniform tree-based approach*. Data and Knowledge Engineering 44, pp 267-298.

Euzenat, J. (2008). *Quelques pistes pour une distance entre ontologies.* 8èmes Journées Francophones Extraction et Gestion des Connaissances Sophia Antipolis 29 janvier 2008.

Goasdoué, F., V. Lattès et M.C. Rousset (2000). The use of carin language and algorithms for information integration : The picsel system. *Int. J. Cooperative Inf. Syst. 9*(4), 383–401.

Hamdoun, S, F. Boufarès et M. Badri (2007). *Construction et maintenance des entrepôts de données hétérogènes*.. e-TI - la revue électronique des technologies d'information, Numéro 4, 23juin, http://www.revue-eti.netdocument.php?id=1331.

Huang, H.-C., J. M. Kerridge et S.-L. Chen (2000). A query mediation approach to interoperability of heterogeneous databases. In *Australasian Database Conference*, pp. 41–48.

Inmon, W. H. (1992). *Building the Data Warehouse*. New York, NY, USA : John Wiley & Sons, Inc.

Kimball, R. (1998). The operational data warehouse. *DBMS 11*(1), 14–16.

Lamarre, P., S. Cazalens, S. Lemp et P. Valduriez (2004). A flexiblemediation process for large distributed information systems. In *CoopIS/DOA/ODBASE (1)*, Volume 3290 of *Lecture Notes in Computer Science*, pp. 19–36. Springer.

Lamolle, M. et A. Zerdazi (2007), *Matching of enhanced XML schemas with a measure of context similarity* Workshop EGC 2007.

Laurent, D., J. Lenchtenboer-Ger, N. Spyratos et G. Vossen (2001). *Monotonic Complements for Independent Data Warehouses.* The International Journal of Very Large Data Base VLDB, volume 10 issue 4, pp 295-315.

Madhavan, J, P.A. Bernstein et Rahm (2001). *Generic Schema Matching with Cupid.*Proceedings of the 27th VLDB Conference, Rome, Italy, pp.49-58.

Maiz, N., O. Boussaid et F. Bentayeb (2008). *Fusion automatique des ontologies par classification hiérarchique pour la conception d'un entrepôt de données. EGC2008*

Melnik, S., H.Garcia-Molina et E. Rahm (2002). *Similarity Flooding: A versatile Graph Matching approaches*. Proceeding (ICDE), San Jose, Californie, USA.

Nassis, V., R. Rajugan, T.S. Dillon et W. Rahayu (2004). *Conceptual Design of XML Document Warehouses*. Data Warehousing and Knowledge Discovery: 6th International Conference, DaWaK Zaragoza, Espagne, Septembre 1-3.

O'Gorman, K., D. Agrawal et A. El Abbadi (1999). *Posse: A Framework for Optimizing Incremental View Maintenance at Data Warehouses*. Data Warehousing and Knowledge Discovery. pp 106-115, Italie.

Sorlin, S. et C. Solnon (2004). *Mesurer la similarité de graphes pour la recherche d'informations* . Dans http://www-clips.imag.fr/mrim/User/catherine.berrut

Sellami, S., N. Benharkat, R. Rifaieh et Y. Amghar (2007). *Vers une plateforme de gestion des schémas XML.* Workshop EGC 2007

Zhuge, Y., H. Garcia-Molina et J.L. Wiener (1996). *The Strobe Algorithms for Multi-Source Warehouse Consistency*. Parallel and Distributed Information Systems, pp 146-157 Décembre.

Zhuge, Y., H. Garcia-Molina, J. Hammer et J. Windom (1995). *View Maintenance in a Warehousing Environment*. Proc. of the ACM SIGMOD, pp 316-327, Mai. California

Summary

Our work is part of data warehousing by providing a new avenue of research that is born from the encounter of cellular automata and databases decision. One of the interesting challenges is to combine Boolean functions and formal mathematical techniques of the cellular machine CASI with performance databases decision, order, design an automatic integration of semi-structured data from heterogeneous sources in a data warehouse.

This integration is guided entirely by the cellular machine CASI. We derived a new machine called BICS XML in three cellular processes: the first process is to extract the schema of XML source documents represented as a graph, called the specific diagram. The latter will be submitted to the second process at the base of the overall pattern of the warehouse, called generic schema, extract the pattern most resembling the specific schema. The third process determines the correspondences between generic patterns and the specific pattern. We are positioning ourselves in this paper, in the case of extracting a specific schema of an XML document, only, and we show how the data automatic integration is using a Boolean model.

Keywords: Data Warehouse, Heterogeneous data integration, Cellular automaton, XML Document, Production rules.

Extraction de connaissances à partir de données textuelles : Application aux avis des consommateurs

Zoubida Agha Benlalam *, Lynda Zaoui**
Département d'Informatique, Université des Sciences et de la Technologie d'Oran
* aghabenlalamz@yahoo.fr
**zaoui_lynda@yahoo.fr

Résumé. Nous présenterons dans cet article une démarche de fouille de texte appliquée aux avis des consommateurs. Le but de ce travail est de réaliser un outil permettant d'extraire de l'information utile à partir de ces opinions en s'appuyant sur l'extraction de règles d'association. Cet outil sera très bénéfique aux entreprises dans la mesure où il leur permettra de connaître ce que pensent les consommateurs de leurs produits sans être obligé de lire toutes les opinions.

Mots clés : fouille de texte, règle d'association, avis des consommateurs, marketing.

1 Introduction

L'analyse et le traitement d'informations textuelles est devenu un enjeu majeur, avec l'explosion du Web où environ 80% de l'information accessible l'est sous forme textuelle (bibliothèque électronique, pages HTML, forums de discussion, …). En effet, au vu du flot d'information que nous connaissons, accéder aujourd'hui à l'information textuelle utile est devenu un vrai « casse-tête » pour l'utilisateur en quête d'information textuelle réutilisable. De ce fait, il est nécessaire de mettre en œuvre des systèmes permettant d'aller rapidement à l'essentiel dans un document textuel, d'analyser les contenus des documents, de les organiser et de les représenter automatiquement.

La fouille de texte, ou text mining, est un domaine qui vise à résoudre de tels problèmes, il permet également de résoudre les problèmes de surcharge d'informations et à faciliter la recherche des connaissances cachées dans les documents.

Deux approches très différentes sont utilisées en text mining : la première repose sur la classification et la seconde approche s'appuie sur l'extraction de règles d'association.

La classification des textes, fait partie du processus d'extraction d'éléments d'information dans des données textuelles. Les processus de classification des textes visent à : structurer des textes selon des thèmes communs, construire des ensembles homogènes de textes selon un ou plusieurs points de vue, rechercher des ensembles de paragraphes liés selon une mesure de similarité, etc. Il existe deux grandes approches pour la classification de textes : l'approche supervisée (text categorization) et non supervisée (clustering). (Garrouste 2002, Jaillet 2005)

L'extraction de règles d'association est une méthode assez répandue en fouille de données. Appliquée aux textes, elle vise à extraire d'un corpus des liens entre les termes caractérisant les textes. Ces liens sont exprimés à travers des règles du type A⇨ B.

Cet article porte sur la problématique d'extraction de connaissances à partir de données textuelles, en utilisant les règles d'association. Notre processus de fouille de textes cherche donc à extraire, d'un ensemble d'opinions de consommateurs, des règles d'association portant sur les termes contenus dans ces textes.

Notre choix est lié au fait que nous voulons fouiller dans les textes de façon non supervisée, sans imposer de contraintes *a priori* à notre processus mais en ayant un objectif de fouille défini par l'analyste, qui est avoir une idée sur ce que pensent les consommateurs de son produit.

D'autre part, une règle d'association est généralement composée de peu de termes. Cela facilite le travail d'interprétation de l'analyste qui, par rapport à ses connaissances, cherche à relier seulement quelques notions pour chaque règle (c'est-à-dire, par exemple, identifier la relation entre les termes présents par rapport à son domaine de spécialité). Ceci est à opposer à des classes volumineuses issues d'une classification de textes contenant des relations diverses entre les termes. Enfin, une autre motivation est le fait que les règles d'association peuvent être pondérées par une mesure de validité appelée *confiance*. Si la règle n'est pas valide, elle sera pondérée par la confiance et interprétée par : « Dans x % des cas, les textes qui possèdent A possèdent B ».

Cet article est organisé comme suit : en première partie, nous définissons le processus de fouille de texte et ses principales caractéristiques, ensuite nous présenterons l'apport du textmining au domaine du marketing et à l'entreprise de façon générale et enfin nous exposant notre expérimentation sur l'extraction de connaissances à partir d'un corpus constitué des avis des consommateurs d'un antivirus.

2 Text mining

Le text mining est défini comme « le processus non trivial d'extraction d'informations implicites, précédemment inconnues, et potentiellement utiles, à partir de données textuelles non structurées dans de grandes collections de textes » (Lehmam 2006). C'est une extension des techniques traditionnelles du Data Mining à des données non structurées.

Le Text Mining s'adresse de manière générale à des personnes qui cherchent à décrire, classer et analyser des textes pour leurs recherches ou leurs études (chercheurs, entreprises). Une des spécificités du Text Mining est que les documents analysés sont écrits pour des lectures par l'homme et non pas pour un traitement par la machine.

Le processus de fouille de texte nous permet entre autre de faire les études suivantes :

- L'étude lexicométrique, qui juge de la richesse du vocabulaire
- L'analyse descriptive, qui permet de définir des proximités entre les formes et les documents et de réaliser des typologies
- La construction de modèles statistiques, capable de prédire l'appartenance d'un nouveau document à une catégorie déjà définie

Les textes étudiés en fouille de texte peuvent être de différents types. Ceci nous amène à évoquer la typologie des textes traités avant de détailler le processus de text mining.

2.1 Typologie de texte

Une typologie de textes, dans laquelle deux catégories de textes sont distinguées, a été définie dans (Cherfi 2004, chapitre 3) **:**

1. *les textes scientifiques et techniques :* dans cette catégorie, s'inscrivent les textes dont l'univers du discours est limité à un domaine de spécialité, tels que : les articles scientifiques, les documentations techniques, les bulletins météorologiques, etc.
2. *les textes de la langue commune :* dans cette catégorie, l'univers du discours des textes est plus ouvert, tels que : les romans, les dictionnaires ou les articles de presse.

La distinction entre ces deux catégories ne se fait pas sur la longueur du texte puisqu'une documentation technique peut être longue et un article de presse peut être bref. La distinction se fait donc sur l'étendue de l'univers du discours.

Nous pouvons ajouter un autre type de texte qui peut constituer une nouvelle catégorie vu les spécificités qu'il apporte. Il s'agit des textes issus des forums de discussion, les opinions et les avis des consommateurs des produits sur le net. Ceux sont des textes qui peuvent être sur le même sujet mais qui sont de structures et vocabulaires différents.

Un texte d'opinion présente un avis argumenté, positif ou négatif, sur un sujet donné. Les domaines faisant l'objet de textes d'opinions sont nombreux : critiques de films ou de livres, jugements qualitatifs de produits, controverses sur un projet politique...etc.

Dans le tableau suivant (Tab 1), nous présenterons une comparaison entre ces trois catégories. Elle porte sur la structure des textes (la hiérarchie) et sur le vocabulaire utilisé.

Cette comparaison, nous permet de voir clairement la difficulté de traitement des opinions par rapport aux autres catégories, car ces opinions n'ont aucune structure, un internaute peut donner son avis en quelques mots seulement par contre un autre préfère exprimer son avis sur plusieurs lignes. L'autre difficulté réside dans le vocabulaire très varié et qui est composé de mots le plus souvent absents du dictionnaire. Ce type de texte nécessite donc un traitement spécifique avant et après l'extraction de connaissances.

Le travail présenté dans cet article porte sur le traitement d'opinions de consommateurs. Ce choix a été motivé par deux points : le premier, est le fait que peu de travaux s'intéresse à ce type de texte et le second, c'est l'importance de la masse d'informations que peuvent apportés ces opinions aux marketers d'une entreprise pour aider à la prise de décision.

	Texte Scientifique	Article de presse	Opinions
Structure	Titre, résumé, mots clé, introduction, résultats, figures, références bibliographiques, annexes…	L'attaque Le corps La chute	Aucune structure prédéfinie.
Vocabulaire	Mots et termes spécifiques selon le choix de type de texte scientifique (mots techniques). -Moins de fautes d'orthographe.	Légende, édito, interview, Dépêche, news, accroche……	Dépend de celui qui rédige, mélange d'abréviations, de termes structurés, de mots composés, d'argots……. - Présence de fautes d'orthographes.

TAB. 1– *Comparaison entre les catégories du texte.*

2.2 Processus du text mining

Le text mining comprend une succession d'étapes permettant de passer des documents au texte, du texte au nombre, du nombre à l'analyse, de l'analyse à la prise de décision. La fouille de textes débute par la modélisation des textes en vue de leur préparation pour l'étape de *fouille de données* et s'achève par l'interprétation des résultats de la fouille pour l'enrichissement des connaissances d'un domaine.

L'ensemble de ces trois tâches constitue une chaîne appelée "processus de fouille de textes". Les différentes étapes de ce processus sont représentées dans la figure 1.

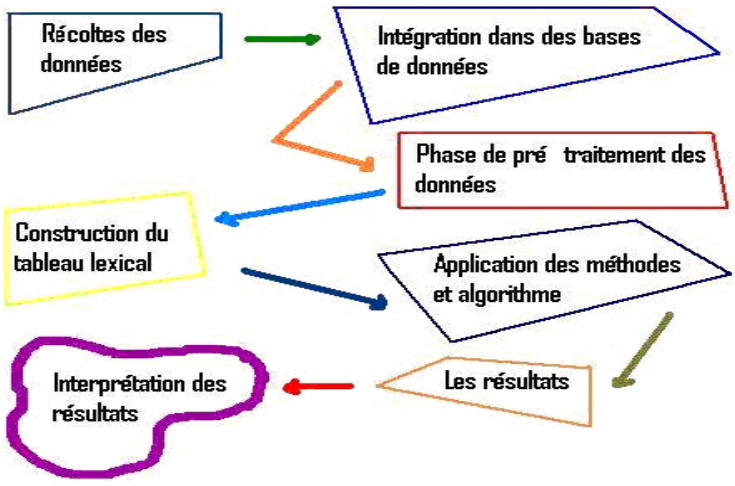


FIG. 1–*Les étapes du processus de Text Mining.*

Les données. La première étape consiste en la récupération des documents. L'étude à effectuer porte toujours sur un ensemble de textes ayant une caractéristique commune. Cette caractéristique peut concerner le sujet, l'auteur, l'année de publication, etc. Cette première étape consiste donc à effectuer une simple recherche au sein des ressources disponibles, en général à partir du Web et de bases de données bibliographiques ou textuelles, pour trouver les documents ayant cette caractéristique. La modélisation du contenu des textes permet d'extraire les données à partir des textes. Elle s'appuie, en général, sur une représentation de type : un texte = {un ensemble de mots-clés} qui est une représentation également communément utilisée en *recherche d'information.* Cette représentation permet d'utiliser, par la suite, des outils de fouille de texte.

Le prétraitement : L'ensemble de documents à fouiller n'est pas structuré (du moins au sens informatique, car le contenu des documents possède bien sûr une structure sémantique). Mais, de nombreuses méthodes, essentiellement issues du domaine du traitement automatique des langages, permettent de pallier à ce problème. La plus couramment utilisée consiste à rechercher, et si nécessaire à filtrer, les mots-clés ou les phrases-clés contenus dans les documents, et éventuellement les relations existantes entre ces divers éléments clés. Cette procédure d'identification indispensable se déroule en plusieurs phases qui peuvent se résumer dans les points suivants :

- Normalisation : traitement des incohérences (exp. fautes d'orthographes) et des ambiguïtés (exp. lexiques d'abréviation).
- Lemmatisation : consiste à ramener les mots à une forme de base, et à reconnaître toutes les variations liées à cette forme.
- Numérisation des formes textuelles.
- Segmentation : découper le texte en unités minimales (phrases ou mots).
- Motifs fréquents : suites de formes graphiques qui se répètent plusieurs fois dans le corpus.

Il est commode de ranger les décomptes des occurrences de chacune des *V* formes dans chacune des *N* parties du corpus dans un tableau. On appelle ce tableau : le *Tableau Lexical Entier* (TLE).

L'ensemble des phases précédentes relèvent de ce qu'on appelle l'analyse avant traitement (ou prétraitement), à la fin de laquelle nous avons un document qu'on a transformé. Si le document initial était fait pour les yeux de l'humain, le document après modification est fait pour un traitement par les machines.

Le traitement : dans cette étape du processus on applique une des méthodes du datamining aux données textuelles. Comme nous l'avons déjà évoqué plus haut, la méthode que nous avons choisie est l'extraction de règles d'association. Dans ce qui suit, nous introduirons quelques définitions concernant cette méthode. Ces définitions seront utilisées plus loin.

Définition d'une règle d'association (Agrawal 1994) **:** Une règle R : B ➔ H est constituée d'un ensemble de termes B (prémisse) impliquant un ensemble de termes H (conséquent). Une règle d'association est notée :

$$t_1 \Lambda t_2 \Lambda \ldots \Lambda t_k \rightarrow t_{k+1} \Lambda \ldots. \Lambda t_n$$

Où $\{t_1 \ldots, t_k\}$ et $\{t_{k+1}, \ldots, t_n\}$ sont deux ensembles non vides de termes et l'opérateur Λ exprime la présence simultanée des ensembles de termes de la règle.

L'interprétation intuitive de la règle R en fouille de texte est : si des textes contiennent les termes t_1 et t_2 … et t_k alors ces textes ont tendance à contenir également, avec une probabilité P, les termes t_{k+1} et t_{k+2}…. et t_n.

Soit un ensemble fini $\Gamma = \{t_1, t_2,\ldots, t_n\}$ de termes caractérisant un ensemble fini de textes $D = \{d_1,d_2,\ldots,d_m\}$. Un motif T est un sous-ensemble de Γ.

Définition du support d'un motif : Le support d'un motif T est défini par la fréquence d'apparition du motif dans l'ensemble D.

Définition d'un motif fréquent : Un motif est dit fréquent s'il apparaît un nombre de fois supérieur à un seuil de support dans l'ensemble de textes D, i.e. support(T) $\geq$ minsup où minsup est un seuil (ou support minimal).

Définition d'un motif fermé fréquent : Un motif est dit fermé si et seulement s'il est le motif maximal commun aux textes qui possèdent ce motif.

Définition de la confiance d'une règle : C'est le degré de validité de la règle. La confiance de la règle B➔ H est la probabilité conditionnelle de H sachant B.

Ces définitions et d'autres détails sur la façon de calculer le support et la confiance sont donnés dans (Guillaume 2000, Bastide 1999 et 2000, Cherfi 2004).

Lecture et interprétation des résultats : L'interprétation des résultats revient à l'expert du domaine qui doit être en mesure d'analyser les résultats obtenus afin d'en tirer de l'information explicite pouvant servir à comprendre et mieux gérer son entreprise.

3 Apport du Text Mining au Marketing

Le développement des nouvelles technologies et, en particulier, la pénétration croissante d'Internet dans les foyers offrent des opportunités quasiment sans limite d'exploitation de données. Le problème est de savoir comment exploiter toutes ses données qui s'amassent sans limites pour en dégager de l'information potentiellement utile. Le Text Mining propose une solution à ces besoins. Nous avons choisi d'illustrer notre problématique par un exemple édifiant qu'est celui de l'apport du processus de fouille de texte au marketing. Sachant que les avis des consommateurs sur les forums et autres blogs revêtent d'une grande importance pour les marketers afin de mieux connaître leur clientèle et de mieux cibler ceux-ci.

Deux principaux axes d'étude marketing peuvent bénéficier des techniques de text mining (Gauzente 2006). Le premier concerne la compréhension du comportement du consommateur : ses critères de choix, ses processus de recherche d'information, ses modes d'évaluation des produits et services, l'usage qu'il fait des produits / services, les critères de satisfaction et d'insatisfaction, les recommandations (bouche à oreille positif ou négatif sur la toile), agrément suscité par les marques la communication des marques et enseignes.

Le second axe relève plus d'une visée stratégique d'anticipation et de détection d'évolutions en germe. Cette dimension à visée de veille est peut être plus délicate à mettre en œuvre.

4 Expérimentation

Nous présenterons dans ce qui suit notre expérience qui vise à appliquer le processus de fouille de texte sur un corpus constitués d'avis de consommateurs écrits en français. Le but de cette expérience est de réaliser un outil permettant de faciliter la tâche à l'expert (marketer) en lui présentant l'essentiel de ces avis à travers un ensemble de règles, sans qu'il soit obligé de les lire tous. Pour tester cet outil, il fallait désigner un produit pour pouvoir cueillir les opinions. Nous avons choisi, donc, l'antivirus Kaspersky.

4.1 Pourquoi un antivirus ? Et pourquoi Kaspersky ?

Le choix d'un antivirus comme produit à évaluer a été motivé par le fait que les antivirus sont parmi les produits informatiques les plus vendus dans le monde. De tels produits sont devenus obligatoires vu les risques que peuvent courir nos ordinateurs s'ils ne sont pas protégés. Ces risques peuvent variés d'un inoffensif message clignotant jusqu'à la perte de toutes les données contenues sur le disque dur. Nous avons choisi en particulier l'antivirus Kasperky car c'est l'un des antivirus les plus populaires ces dernières années.

4.2 La recherche de données

Nous avons donc orienté notre analyse sur le produit antivirus Kaspersky, en faisant dans un premier temps la récolte des avis des consommateurs de ce produit sur les forums de discussion. La collecte des données a pris en considération plusieurs critères que sont :

- l'identification de sources différentes : commerciales et non commerciales,
- l'identification d'un forum contenant suffisamment de commentaires (inutile d'avoir 5 avis consommateurs),
- l'identification d'un corpus de taille modeste permettant d'évaluer rapidement l'apport du logiciel de Textmining.

Pour valider notre méthode et tester les performances de notre prototype, nous avons importé 60 commentaires issus de différents forums de discussion (forum de discussion ciao, comment ça marche et casafree).

4.3 Le prétraitement

Après avoir importé les avis des forums différents les uns des autres, par la taille et la structure du texte, une préparation sur l'ensemble du corpus est effectuée en vue d'en extraire les règles d'association :

Normalisation : Pour cette étape nous avons envisagé les opérations suivantes : correction orthographique, suppression des points superposés et virgules, suppression des points d'exclamation ou autre d'interrogation, élimination des parenthèses et accolades, correction des abréviations et enfin marquer la fin de chaque phrase par un point. Nous obtiendrons alors des commentaires dépourvus d'ambiguïtés pour faciliter l'extraction des unités lexicales sur lesquelles nous allons pouvoir appliquer notre analyse. Dans cette étape, seules les corrections d'orthographe et d'abréviations sont réalisées manuellement, toutes les autres opérations sont faites automatiquement.

Segmentation du corpus : Le prototype réalise une première passe sur le texte et offre une liste de mots candidats à l'analyse. Ces mots sont sélectionnés selon la taille minimale choisie, ainsi les mots ayant une taille (nombre de caractères du mot) inférieure ou égale à la taille minimale seront automatiquement éliminés.

Par défaut nous avons choisie une taille minimale de 4 éliminant ainsi les mots comme (est, les, la, des ….), ceux-ci ne servant à rien dans notre analyse.

Lemmatisation : cette étape consiste à lemmatiser le corpus visant à diminuer fortement le nombre de mots analysés, en éliminant toutes les flexions et les dérivations grammaticales. Pour la partie lemmatisation nous nous sommes basés sur un algorithme de lemmatisation dans la langue française (voir le site http://snowball.tartarus.Org/algorithms/french/ stemmer.html), nous avons pu faire une fonction qui transcrit assez fidèlement cet algorithme mais les résultats que nous avons obtenus sur le corpus n'étaient pas satisfaisants. Actuellement nous sommes entrain d'intégrer l'outil de lemmatisation Tree Tagger qui permet de lemmatiser tous les mots d'un fichier donné en entrée.

Numérisation des unités textuelles : Pour chaque unité textuelle, un numéro est affecté. Sur cette base, le logiciel va dresser une première liste de mots sur lesquels l'analyse sera effectuée, cette liste est appelée liste de mots signifiants (unités textuelles). Un calcul sur la fréquence d'apparition de chaque unité est fait pour ne garder que les mots signifiants qui ont une fréquence supérieure à la fréquence minimale fixée. Ceci donnera la liste finale des mots signifiants.

Le tableau lexical : Le tableau lexical est construit sur la base de n lignes (le nombre de mots signifiants) contenant les unités textuelles traitées et prêtes à l'analyse. Chaque ligne représente une unité ayant comme propriétés (colonnes du tableau) :

- Son numéro dédié
- Le mot associé
- Fréquence d'apparition du mot dans le corpus

4.4 Traitement des données

La construction des règles d'association se décompose en deux étapes. La première consiste à déterminer l'ensemble des motifs fréquents. Dans la deuxième étape, on génère pour chaque motif fréquent, toutes les règles d'association dont la confiance est supérieure à un certain seuil minconf (Guillaume 2000).

Pour réaliser la première étape, on s'est basé sur l'algorithme *Close* (Bastide 1999). Cet algorithme est inspiré de l'algorithme *Apriori* (Agrawal 1994) pour la recherche de motifs fréquents par lecture et comptage des données en entrée.

L'algorithme *Close* est composé de quatre étapes. Les trois premières étapes concernent la recherche de tous les motifs fermés fréquents. Ces trois étapes ont la plus grande complexité calculatoire. La quatrième étape est un calcul simple, sans accéder à la base de données, qui découle des calculs faits durant les trois premières étapes.

Le nombre de motifs fermés fréquents est très inférieur au nombre de motifs fréquents, même dans le pire cas, lorsque les données sont fortement corrélées. Ce qui rend la recherche de motifs fermés fréquents (*Close*) moins coûteuse que la recherche de motifs fréquents (*Apriori*). L'idée de *Close* est de calculer les motifs fermés fréquents puis de trouver

l'ensemble des motifs fréquents sans recours à la lecture et au comptage des données. De plus, *Close* utilise une technique itérative, dite *par niveaux* :

- Au niveau 1 : calcul du support de chaque 1-motif (*i.e.* la fréquence d'apparition de chaque terme de T); suppression des termes non fréquents (*i.e.* dont le support est strictement inférieur à **minsup**), calcul de leurs fermés;
- Au niveau k : calcul des k-motifs générateurs candidats ; calcul de leurs fermés et leurs supports, suppression des k-motifs non fréquents et des k-motifs non générateurs.
- Au niveau k+*1* : les k-motifs générateurs sont utilisés pour générer les (k+1)-motifs candidats ; puis le traitement fait au niveau *k* est renouvelé.

En sortie de l'algorithme *Close*, nous avons besoin de garder une trace des motifs fermés fréquents et des générateurs pour la génération des règles d'association.

L'extraction des règles d'association est une technique de fouille de données qui a fait ses preuves pour la fouille dans de grandes masses de données. Cependant, elle présente aussi certaines limites qui sont : le nombre de règles qui peut être très important et la présence de règles redondantes et/ou de règles non significatives. Une règle est dite redondante par rapport à d'autres règles d'association si l'information qu'elle apporte est présente dans d'autres règles. Une règle est dite non significative si elle n'apporte aucune information utile.

4.5 Résultats obtenus et discussion

Nous présenterons dans les deux tableaux suivants les résultats de notre analyse sur le corpus constitué des 60 avis importés des différents forums. Dans le premier tableau, nous avons fixé la taille minimale des unités textuelles à retenir à 4 caractères et la fréquence minimale d'apparition requise pour ces unités à 3. Tandis que dans le deuxième, la taille minimale est égale à 3 et la fréquence minimale est égale à 2.

Le nombre de forme du corpus (respectant la taille minimale)	499
Le nombre de hapax (unité textuelle possédant une fréquence=1) dans le corpus	5
La fréquence maximale est celle de la forme	Antivirus (=64)
Le nombre de formes respectant les contrainte de recherche soit taille et fréquence	134
Le nombre de fréquents fermés	139
Le nombre de règles générées	144

TAB. 2– *Les résultats pour une fréquence minimale=3 et une taille minimale=4.*

Le nombre de forme du corpus (respectant la taille minimale)	555
Le nombre de hapax (unité textuelle possédant une fréquence=1) dans le corpus	5
La fréquence maximale est celle de la forme	Est (=74)
Le nombre de formes respectant les contrainte de recherche soit taille et fréquence	550
Le nombre de fréquents fermés	196
Le nombre de règles générées	201

TAB. 3– *Les résultats pour une fréquence minimale=2 et une taille minimale=3.*

Notre discussion va porter essentiellement sur le nombre de règles générées et sur la signification de ces règles.

Pour le nombre de règles, nous pouvons voir clairement qu'il a augmenté lorsqu'on a diminué les valeurs des paramètres taille et fréquence. Ceci montre que le bon paramétrage des valeurs en entrée permet de résoudre en partie le problème du nombre de règles générées. Après une succession de tests sur le même corpus et sur d'autres en modifiant ces paramètres d'entrée, nous avons pris comme valeurs par défaut 4 pour la taille minimale et 3 pour la fréquence minimale, car ces deux valeurs ont donné les meilleurs résultats par rapport aux autres valeurs. Le choix de ces paramètres peut aussi influencer la pertinence des mots signifiants retenus. En effet, en comparant les formes les plus fréquentes dans les deux tableaux nous constatons que dans le premier la forme *Antivirus* est plus significative à notre domaine d'analyse que la forme *Est* du second. Cependant, le choix d'une taille minimale supérieure à 2 pose le problème de l'élimination de la négation (ne, n'), ce qui va changer complètement le sens de l'opinion. C'est l'une des difficultés rencontrées lors du traitement de ce type de données. Enfin, nous tenons à préciser que le nombre de règles donné dans les deux tableaux est le nombre total de toutes les règles (significatives et non significatives).

Exemples de règles significatives	
mémoire =>Kaspersky,ralentit,ordinateur	Confiance : 100.00%
installer =>Kaspersky,Pour,chevaux,Troie	Confiance : 100.00%
protection =>très	Confiance : 66.67%
cher =>très	Confiance : 66.67%
Exemples de règles non significatives	
votre =>vous	Confiance : 60.00%
alors =>version,trouve,dans,page	Confiance : 100.00%

TAB. 4– *Exemples de règles d'association significatives et non significatives*

En prenant la première règle significative, nous pouvons facilement comprendre que tous les utilisateurs (confiance = 100%) qui ont parlé de mémoire ont trouvé que Kaspersky ralentit l'ordinateur. Par contre, en examinant la règle votre =>vous, nous constatons qu'elle n'apporte aucune information utile. Donc, parmi les règles extraites il existe un nombre assez important de règles non significatives. Pour cette raison il est indispensable de filtrer toutes ces règles pour ne garder que les plus significatives au domaine étudié. Ceci n'a pas été pris en charge, pour le moment, par notre application car ça demande tout un autre travail faisant

appel à des méthodes portant sur la sémantique et les modèles de connaissances, mais il existe des travaux dans ce domaine appliqués à d'autres types de données textuelles (Cherfi 2004).

Pour les règles significatives, l'expert du domaine peut facilement les interpréter, ce qui est considéré comme point fort de cette méthode. Ainsi, il peut comprendre ce que pensent les utilisateurs de son produit. Ceci permettra à l'entreprise d'améliorer ses produits et services.

5 Conclusion et perspectives

En conclusion, nous pouvons dire que cette étude nous a été très bénéfique dans la mesure où elle nous a permis de toucher de plus près aux difficultés d'extraction d'informations utiles à partir de données textuelles. Elle nous a permis aussi de montrer l'intérêt d'un outil de fouille de texte pour les entreprises. En effet, par rapport à l'évolution du marketing vers une prise en compte plus fine et personnalisée du client, le text mining est susceptible de permettre l'élaboration de réponses plus pertinentes et subtiles. Le processus nous évite alors de se faire submergés par le flux de données. En plus il permettra aux fabricants et distributeurs de produits de prendre conscience des critères sur lesquels leurs produits sont jugés, de connaître la performance de leurs produits sur ces critères, d'identifier les concurrents avec lesquels les consommateurs les mettent en comparaison et de connaître les failles de leurs produits.

Cela étant, les premiers résultats obtenus par notre étude attendent d'être améliorés à l'avenir. Cette amélioration va toucher principalement au prototype car vu les résultats que nous avons obtenu, notre démarche est efficace. Il faudra donc trouver des solutions aux limites de notre prototype qui sont le nombre important de règles d'association générées malgré l'utilisation de l'algorithme Close et la présence de règles non significatives. Ceci en approfondissant la phase de prétraitement par l'intégration d'un outil de lemmatisation qui va nous permettre de réduire le nombre d'unités textuelles (par exemple : détecte, détection, détecté seront tous remplacés par détecter) ce qui réduira le nombre de motifs fréquents et par conséquent le nombre de règles extraites. Et pour augmenter le nombre de règles significatives, nous envisageons de prendre en considération la négation et d'ajouter des mots composés tels que : mise-à-jour à la place de "mise à jour" qui sont considérés comme trois mots séparés. Enfin, il sera nécessaire d'ajouter un outil pour éliminer les règles non significatives.

Références

Agrawal, R., Ramakrishnan, S. (1994). Fast algorithms for mining association rules in large databases. Dans Proc. of the 20th -Int'l Conf. on Very Large Databases (VLDB'94), pages 478–499, Santiago, Chile.

Bastide, Y., Pasquier. N., R. Taouil et Lakhal. L. (1999). "Efficient mining of association rules using closed itemset lattices". *Information Systems*, 24(1):25–46.

Bastide, Y. (2000). "Data mining : algorithmes par niveau, techniques d'implantation et applications". Thèse de doctorat, Université Blaise Pascal, Clermont-Ferrand II.

Candillier, L. (2005). "Classification non supervisée contextuelle". Lille Univ.

Cherfi, H. (2004). "Etude et réalisation d'un système d'extraction de connaissance à partir de textes". Thèse de Doctorat. Université Henri Poincaré-Nancy.

Garrouste, D. (2002). "Introduction au Text Mining". Atelier Technique SAS

Gauzente, C. (2006). "E-marketing et textmining". JADT: 8es Journées internationales d'Analyse statistique des Données Textuelles

Guillaume, S. (2000). *Traitement des données volumineuses : Mesures et algorithmes d'extraction de règles d'association et règles ordinales*. Thèse de doctorat, Université de Nantes,.

Jaillet, S. (2005). "Catégorisation automatique de documents textuels d'une représentation basée sur les concepts aux motifs séquentiels. University Montpellier 2.

Lehmam, A. (2006). "Solutions de traitement du document textuel avec prise en charge de ressources linguistiques"

Forum de discussion "ciao" http://www.ciao.fr/Avis/Kaspersky_Anti_Virus_Personal_Pro_Ensemble_complet__263188.

Forum de discussion "casafree" www.casafree.com

Forum de discussion "comment ça marche" www.commentcamarche.net/forum

Summary

In this article, we will present a gait of text mining applied to the opinions of the consumers. The goal of this work is to achieve a tool permitting to extract the useful information from these opinions while leaning on the extraction of association rules. This tool will be very beneficial to the enterprises insofar as it will allow them to know what the consumers of their products think without being obliged to read all opinions.

Une architecture à base des profils pour la reformulation contextuelle des requêtes utilisateur

Abdelkrim Bouramoul*, Khireddine Kholladi **, Bich-Liên Doan***

* Département Informatique, Université 8 mai 1945-Guelma,
BP 401 Guelma 24000 - Algérie
a.bouramoul@yahoo.fr
** Département Informatique, Université de Constantine,
B.P. 325, Constantine 25017 - Algérie
kholladi@yahoo.fr
***SUPÉLEC, Département Informatique
Plateau de Moulon, 3 rue Joliot-Curie, 91192 Gif sur yvette, France
bich-lien.doan@supelec.fr

Résumé. Cet article s'inscrit dans le domaine de la recherche d'information sur le web et propose une architecture basée sur les profils utilisateurs pour la prise en compte du contexte dans la reformulation des requêtes. Il s'agit de capitaliser l'ensemble d'informations caractérisant chaque utilisateur sous forme d'éléments contextuels afin de les utiliser par la suite lors de la reformulation de la requête. Après un bilan sur les approches classiques pour la reformulation des requêtes, nous présentons la notion du contexte et celle du profil utilisateur. Nous décrivons par la suite l'architecture proposée toute en montrant comment les contextes statique et dynamique sont capturés et la manière selon laquelle ces deux types de contexte sont utilisés dans notre proposition. Nous présentons enfin le prototype développé et nous concluons.

1 Introduction

La Recherche d'Information (RI) peut être définie comme une activité dont la finalité est de localiser et délivrer un ensemble de documents à un utilisateur en fonction de son besoin (Hernandez, 2006). Afin d'améliorer la performance de ce type de systèmes, le domaine de la RI contextuelle est apparu récemment comme une priorité (Allan 2003), son objectif est de replacer l'utilisateur au cœur des modèles en rendant explicites certains éléments du contexte qui peuvent influencer les performances des systèmes.

D'autre part, les utilisateurs d'un système de recherche d'information, ne sont pas des professionnels de la documentation (Lin et Wang, 2006), ils ne savent pas choisir les bons termes qui expriment le mieux leurs besoins d'information, une reformulation de requête s'impose alors, cette reformulation est motivée par le fait que la requête initiale retourne rarement un résultat qui satisfait ce dernier. Il s'agit en particulier de modifier la requête initiale de l'utilisateur en lui rajoutant des termes significatifs afin de retourner un résultat plus pertinent.

Dans cet article nous proposons un système de reformulation contextuelle des requêtes utilitaires, cette reformulation est qualifiée de contextuelle car elle prend en compte la notion du contexte via les profils utilisateurs pour modifier leur requête initiale. L'ajout de la notion

du contexte lors de la reformulation vise à augmenter l'efficacité des SRI en améliorant leur pertinence et en permettant ainsi de prendre en compte les caractéristiques personnelles de l'utilisateur, ses intérêts, ses préférences et l'historique de ses interactions avec le système de recherche. Ces éléments sont capitalisés dans notre système sous forme du contexte statique et dynamique pour être utilisés par la suite dans la reformulation contextuelle.

L'organisation retenue pour cet article s'articule autour de quatre sections. La première sera consacrée à la présentation des différentes approches pour la reformulation de requêtes. Dans la section suivante nous présentons la notion du contexte et celle du profil ainsi que leur apport en recherche d'information. Les deux dernières sections présentent notre système de reformulation contextuelle de requêtes, elles décrivent respectivement l'architecture proposée et le prototype développé. Enfin nous concluons.

2 Approches classiques pour la reformulation des requêtes

L'utilisateur est souvent incapable de formuler son besoin exact en information. Par conséquent, parmi les documents qui lui sont retournés, certains l'intéressent moins que d'autres. Compte tenu des volumes croissants des bases d'information, retrouver celles qui sont pertinentes en utilisant seulement la requête initiale de l'utilisateur est une tâche quasi impossible.

La reformulation de la requête consiste donc à modifier la requête de l'utilisateur par ajout de termes significatifs, cette idée d'affinement de requêtes n'est pas nouvelle. Plusieurs approches utilisent différentes techniques pour sélectionner les termes à ajouter à une requête. Nous distinguons trois types d'approches pour la reformulation de requête, la différence entre ces approches réside, soit dans la source des termes utilisés dans la reformulation et qui peuvent provenir des résultats de recherches précédentes (réinjection de pertinence) ou d'une ressource terminologique (réseau sémantique, thesaurus, ontologie), soit dans le mécanisme qui permet de sélectionner les termes à ajouter à la requête initiale (probabiliste, lien sémantique).

Un premier type d'approches repose sur l'analyse globale de la collection de documents considérée, la plus répandue d'entre elles est basée sur des analyses statistiques de corpus de documents (Cui et al, 2002). L'objectif est de relever la fréquence des termes apparaissant conjointement sur un même document et de sélectionner les termes avec le plus grand coefficient. Les informations ainsi extraites sont généralement utilisées pour reformuler automatiquement une requête par ajout des termes liés aux termes initialement présents dans la requête. Les termes ainsi ajoutés sont issus des documents et permettent donc une meilleure adéquation entre le besoin d'information et la collection.

Un deuxième type d'approches basé sur le principe de réinjection de pertinence vise également à reformuler une requête initiale pour qu'elle corresponde mieux au contenu de la collection. Le principe est le suivant : l'utilisateur soumet sa requête initiale, le système restitue un premier ensemble de documents que l'utilisateur doit juger (pertinent, non pertinent). La connaissance de la pertinence des documents initialement restitués est utilisée pour sélectionner des termes à ajouter à la requête initiale. Nous citons dans cette catégorie les travaux de (Lin et Wang, 2006) dans lesquels le système propose, suite à une requête, un ensemble de documents et suivant ceux visualisés par l'utilisateur, le système met à jour son index de termes concordant par des méthodes d'apprentissage automatique.

Le dernier type d'approches figurant dans la littérature utilise des ressources terminologiques telles que des ontologies ou des thésaurus contenant le vocabulaire servant à l'enrichissement des requêtes. Les approches de ce type utilisent des ontologies avec des relations d'équivalence et de subsomption (Navigli et Velardi, 2003) afin d'extraire les termes à rajouter a la requête initiale.

L'approche de reformulation que nous proposons dans cet article est basée sur la prise en compte du contexte utilisateur via son profil, elle offre un double avantage par rapport aux approches présentées précédemment, d'une part, et contrairement aux deux premières classes d'approches elle est utilisable directement sans phase d'analyse ou d'apprentissage, d'une autre part elle n'est pas contrainte du problème présent dans la troisième classe d'approche, ces dernières n'utilisent que les relations d'équivalence et de subsomption et n'exploite pas toutes les relations sémantiques offertes par une ontologie. Dans la section suivante nous présentons la notion du contexte et celle du profil utilisateur, nous donnons par la suite une classification des profils et de leur utilisation dans les systèmes de recherche d'information.

3 Aspects du profil utiles pour la capture du contexte

3.1 Définition du contexte

Le contexte n'est pas un concept nouveau en informatique : dès les années soixante, systèmes d'exploitation, théorie des langages et intelligence artificielle exploitent déjà cette notion. Avec l'émergence des systèmes de recherche d'information, le terme est redécouvert et placé au cœur des débats sans pour autant faire l'objet d'une définition consensuelle claire et définitive (Gaëtan, 2006). Toutefois, l'analyse des définitions présentes dans la littérature conduit à ce double constat :

- « Il n'y a pas de contexte sans contexte » (Brézillon, 2006). Autrement dit, le contexte n'existe pas en tant que tel. Il émerge, ou se définit, pour une finalité ou une utilité précise.
- « Le contexte est un ensemble d'informations. Cet ensemble est structuré, il est partagé, il évolue et sert l'interprétation » (Winograd, 2001). La nature des informations, de même, l'interprétation qui en est faite, dépendent de la finalité.

En recherche d'information, le contexte est défini comme « l'ensemble des facteurs cognitifs et sociaux ainsi que les buts et intentions de l'utilisateur au cours d'une session de recherche », (Calabretto et Egyd-Zsigmond, 2006). D'une manière générale, le contexte regroupe des éléments de natures divers qui délimitent la compréhension, le champ d'application ou les choix possibles. Les éléments les plus couramment invoqués concernent des données spatiotemporelles (lieu, heure, jour.) ou des connaissances spécifiques en relation avec le domaine étudié. Plus rarement nous observons l'utilisation des éléments concernant les émotions, des états d'esprit, des données culturelles (Brézillon, 2006). Ainsi certain éléments du contexte peuvent être difficile à cerner car nous les utilisons inconsciemment, d'autres se trouvent hors d'atteinte des périphériques d'entrée des machines et donc difficile à mettre en œuvre dans des systèmes de recherche d'information.

3.2 Définition du profil

Un profil utilisateur est défini comme « une source de connaissance qui contient des acquisitions sur tous les aspects de l'utilisateur qui peuvent être utiles pour le comportement du

système » (Wahlster et Kobsa, 1986). Cette proposition, bien que générale, correspond à nos orientations, elle met en évidence trois aspects du profil qui s'exploitent ainsi :

- *Source de connaissance :* le profil utilisateur peut regrouper des informations très diverses selon la tâche à accomplir, en recherche d'information le contenu d'un profil utilisateur se résume en : ses caractéristiques personnelles, ses intérêts et ses préférences, ses compétences, son but courant et enfin l'historique des ses interactions avec le système (Belkin et al, 2004). Nous signalons que la notion du contexte, présenté précédemment, est une extension du profil utilisateur. Le contexte contient des informations complémentaires permettant une meilleure adaptation du profil.
- *Acquisitions :* le contenu du profil utilisateur est une connaissance à récupérer, selon le degré d'adaptation du système, les données du profil utilisateur peuvent être soit, renseignées par l'utilisateur lui-même, soit récupérées par sélection d'un profil pré-existant créé par des experts du domaine, ou encore capturées par le système de recherche d'information au cours de l'utilisation.
- *Utile pour le comportement du système :* en recherche d'information l'apport du profil utilisateur est de permettre une personnalisation ou une adaptation des services pour améliorer les performances du système, ou encore pour filtrer les résultats retournés par un moteur de recherche.

3.3 Classification des profils et de leurs utilisations en RI

Nous présentons dans cette section les différents types de profils abordés dans la littérature et qui sont en lien avec la tâche de recherche d'information, à cet effet nous avons défini quatre classes de profils en se basant sur des critères de regroupement, ces critères s'articulent autour du degré de l'implication de l'utilisateur, du moment de l'utilisation du profil, du contenu de profil en information et enfin de la complexité des informations capitalisées par le profil.

3.3.1 Selon l'implication de l'utilisateur

Il s'agit de mesurer le degré de l'implication de l'utilisateur dans le processus de capture de son profil, le travail de (Benammar et al, 2002), distingue deux types de gestion de profils :

- *Indirecte :* c'est le cas où la gestion des profils est transparente à l'utilisateur, autrement dit l'utilisateur n'intervient pas dans la gestion de ses profils.
- *Directe :* à l'opposé, dans la gestion directe des profils, l'utilisateur doit intervenir dans toutes les étapes du processus de recherche pour gérer ses profils.

3.3.2 Selon le moment de la reformulation

Nous nous intéressons ici au moment de l'utilisation du profil dans un système de recherche d'information, deux possibilités sont à retenir :

- *Pré-recherche :* un profil peut être utilisé dans une étape de pré-recherche pour aider l'utilisateur à formuler ou reformuler son besoin. Il peut s'agir par exemple d'affiner l'expression d'une requête proposée par l'utilisateur en fonction de son profil.
- *Post-recherche :* un profil peut également être utilisé dans une étape post-recherche pour filtrer les résultats d'une recherche.

3.3.3 Selon la complexité

Dans cette catégorie le focus est mis sur le degré de la complexité des informations présentes dans le profil, différents formats de ce type de profils ont été étudiés dans (Korfhage, 1997), les plus répandus d'entre eux sont :

- *Simple :* un profil simple se présente sous la forme d'un ensemble de mots-clés et éventuellement des poids associés, un poids traduit l'importance de chaque terme dans le profil.
- *Étendu :* un profil étendu inclut, en plus des mots-clés et de leur poids, une série d'informations qui décrivent le contexte de la recherche.

3.3.4 Selon la nature d'information

Enfin, la dernière classe distingue les profils utilisateurs en se basant sur la nature d'information qu'ils contiennent, les travaux de (Benammar et al, 2002) exploitent les profils suivants :

- *Profil d'identification* : Cette première composante du profil sert à identifier un utilisateur à travers une série d'informations. Il est défini à la première connexion au système des profils et est mis à jour par incrémentation à chaque création d'un profil d'interrogation.
- *Profil d'interrogation* : Il peut être assimilé à une requête. Il traduit le besoin en information de l'utilisateur et il facilite l'association de la recherche faite par un utilisateur à son contexte.

Les différentes classes de profils que nous avons identifiées peuvent être exploitées simultanément dans un même système et l'utilisation d'un type de profils n'implique pas l'isolement des autres, néanmoins un système à base de profils doit définir les caractéristiques de chaque type de profils utilisé ainsi que les liens les reliant. Dans la suite de cet article nous présentons notre système et nous justifions le choix de ses paramètres en terme du profil, nous décrivons également l'architecture proposée et le prototype développé.

4 Une architecture à base des profils pour la reformulation contextuelle des requêtes

4.1 Choix des paramètres du système en termes de l'utilisation du profil

Dans les sections précédentes nous avons dressé une étude comparative des éléments nécessaires à la définition de notre système, à savoir, la reformulation, le contexte et l'utilisation des profils dans les systèmes de recherche d'information, cette étude nous a permis de catégoriser séparément les caractéristiques de ces éléments et de cerner les limites de chaque catégorie. Nous présentons dans cette section les différents paramètres caractérisant notre proposition et nous éclairons nos choix par rapport aux approches étudiés précédemment.

Notre choix s'est fixé sur l'utilisation du contexte pour la reformulation de la requête utilisateur. Nous avons présenté dans la section 3 les différentes classes du profil et les particularités relatives à chaque classe. Le tableau 1 présente les paramètres caractérisant notre système en termes du profil utilisateur.

Selon l'implication de l'utilisateur	Directe	Indirecte
	+	
Selon le moment de la reformulation	Pré-recherche	Post-recherche
	+	
Selon le degré de complexité	Simple	Étendu
		+
Selon la nature d'informations	Profil d'identification	Profil d'interrogation
	+	+

TAB. 1 – *Choix des paramètres du système en terme de l'utilisation du profil.*

Ce tableau présente les choix que nous avons adoptés pour les profils utilisateurs, et donc la manière selon laquelle le contexte est utilisé pour aider à la reformulation contextuelle des requêtes, ces choix s'interprètent comme suit :

- *Implication de l'utilisateur :* l'utilisateur intervient en partie dans la définition de son profil, l'implication est donc directe, le système récupère automatiquement des informations qu'on suppose pertinentes pour enrichir le contexte de l'utilisateur pour des éventuelles prochaines recherches, et les propose à l'utilisateur qui valide par la suite celles qu'il juge réellement pertinentes parmi l'ensemble de propositions.
- *Moment de la reformulation :* dans notre cas il s'agit d'un profil pré-recherche, le système reformule le besoin de l'utilisateur en affinant l'expression de sa requête en fonction de son contexte.
- *Degré de complexité :* le profil est étendu, il inclut, en plus des mots-clés, une série d'information qui décrivent le contexte de la recherche, ces information sont stockées dans une table sous la forme de couples attribut-valeur où chaque couple représente une propriété du profil.
- *Nature d'informations :* nous utilisons à la fois et d'une façon complémentaire, un profil d'identification et un profil d'interrogation. Le premier sert à identifier un utilisateur à travers une série d'informations défini à la première connexion au système, le deuxième est issu de l'historique des recherches faite par le même utilisateur dans des sessions ultérieures, donc son contenu se développe à chaque fois que l'utilisateur formule une nouvelle recherche.

Synthèse : pour rendre l'utilisation des profils utile pour la reformulation contextuelle des requêtes et utilisable dans un système de recherche d'information, nous regroupons nos choix en terme de l'utilisation du profil pour modéliser le contexte en deux grandes classes :

1. ***Contexte statique :*** il prend les caractéristiques d'un profil d'identification étendu qui sera capturé dans une étape de pré-recherche et qui se caractérise par une implication directe de l'utilisateur
2. ***Contexte dynamique :*** Il constitue l'élément principal de notre système, il regroupe les caractéristiques d'un profil d'interrogation étendu qui est utilisé dans une étape pré-recherche et qui nécessite une implication directe de l'utilisateur

4.2 Présentation de l'architecture

Notre système s'articule autour de quatre modules afin de permettre la reformulation contextuelle de la requête utilisateur en se basant sur son profil, il s'agit dans un premier temps de capturer les deux types de contexte nécessaires à la catégorisation de l'utilisateur (contexte statique et contexte dynamique), puis de les utiliser par le module de reformulation pour générer une nouvelle requête à partir de la requête initiale, enfin le module de recherche prend en charge la délivrance d'un résultat qui se rapproche le mieux des besoins de l'utilisateur. Nous décrivons dans les sections suivantes chacun de ces modules en donnant ses différents composants et son principe de fonctionnement. Le regroupement de ces quatre modules nous a permis par la suite de définir notre architecture pour la reformulation des requêtes à base de profils.

4.2.1 Module pour la capture du contexte statique

Cette première composante du contexte sert à identifier un utilisateur à travers une série d'informations afin de catégoriser l'utilisateur. Le contexte statique est défini à la première connexion au système, à cet effet nous avons défini quatre catégories d'informations relatives au contexte statique, ces informations se résument en :

- *Les paramètres de connexion :* e-mail, mot de passe.
- *Les caractéristiques personnelles :* nom, prénom, pays, langue ...
- *Les intérêts et préférences :* domaine, domaine secondaire, spécialité…
- *Les compétences et niveau de savoir-faire :* profession, niveau d'étude...

La figure 1 présente les éléments du contexte statique, et la manière selon laquelle les informations composant ce type de contexte sont capturées.

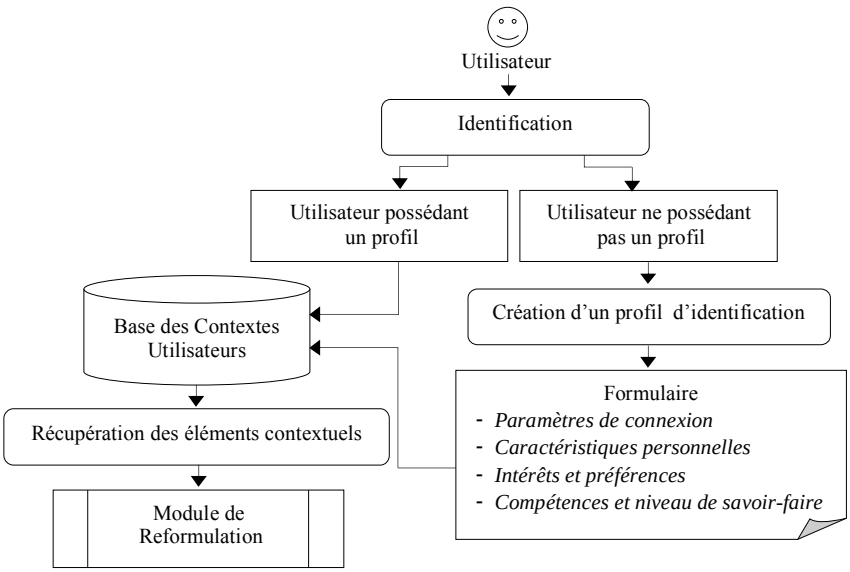


FIG. 1 – *Module pour la récupération du contexte statique.*

4.2.2 Module pour la capture du contexte dynamique

Dans le but d'optimiser la réutilisation des profils et faciliter leur compréhension, cette deuxième composante du contexte consiste en l'association de la recherche au contexte de l'utilisateur. A la fin de chaque session de recherche le module de capture du contexte dynamique procède à l'extraction automatique d'un ensemble d'éléments relatifs au contexte de l'utilisateur, il les organise sous forme de compte (attribut, valeur) et les propose à l'utilisateur, ce dernier valide par la suite ceux qu'il juge réellement pertinents. Ces informations seront enfin stockées dans la base des contextes utilisateurs. La figure 2 présente la manière selon laquelle les éléments du contexte dynamique sont capturés.

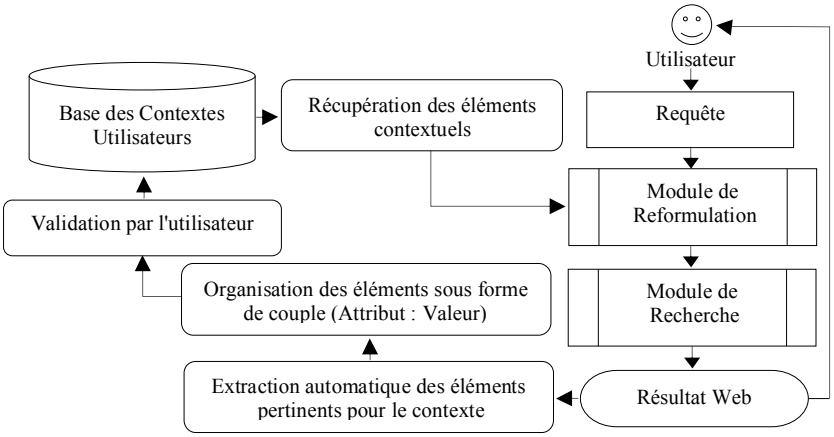


FIG. 2 – *Module pour la récupération du contexte dynamique*

4.2.3 Module de reformulation

Le module de reformulation a pour objectif de produire une nouvelle requête à partir de la requête initialement formulée par l'utilisateur et cela en rajoutant des termes issus de son contexte de recherche actuelle. Dans un premier temps l'utilisateur formule sa requête en utilisant ses propres termes, par la suite le système procède à l'extraction de l'ensemble des termes à rajouter afin de produire une nouvelle requête, ces termes sont extraits de la base des contextes utilisateurs. Une fois la requête reformulée elle sera envoyée au module de recherche qui prend en charge la délivrance des résultats à l'utilisateur. La figure 3 présente le processus de reformulation de la requête à base du contexte utilisateur.

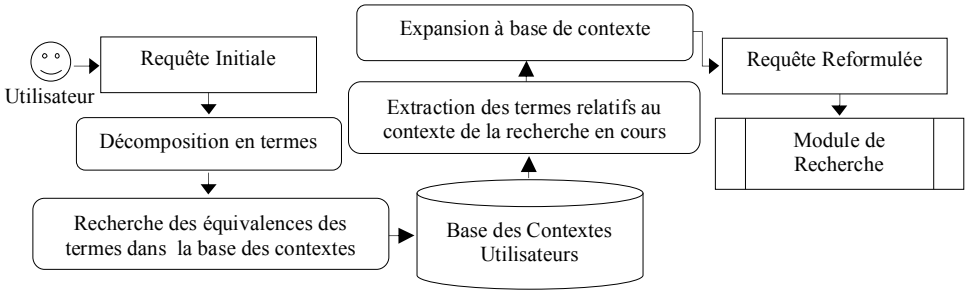


FIG. 3 – *Module de reformulation contextuelle des requêtes*

4.2.4 Module de recherche

Notre système offre une recherche ouverte sur internet, l'utilisateur exprime son besoin en information sous forme de requête, et le module de reformulation procède par à son expansion en rajoutant des termes issus du contexte utilisateur et renvoie la requête au module de recherche. Ce dernier prend en entrée la requête reformulée et offre à l'utilisateur la possibilité de choisir l'un des trois moteurs de recherche que le système propose (Google, Yahoo, Msn), le résultat obtenu est enfin communiqué à l'utilisateur. La figure 4 montre le principe de fonctionnement du module de recherche.

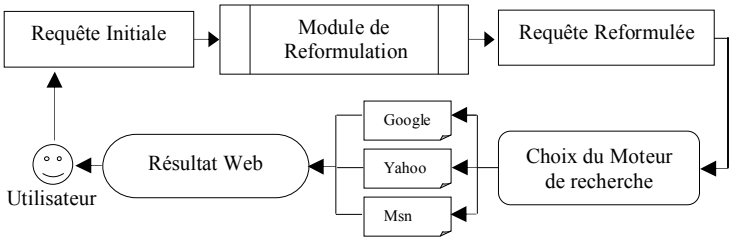


FIG. 4 – *Module de Recherche*

4.2.5 Architecture Générale

La composition des quatre modules décrit précédemment nous a permis de définir l'architecture générale de notre système, nous signalons que le fonctionnement des quatre modules est étroitement lié dans le sens où les sorties de chaque module sont les entrées du module suivant. La figure 5 présente l'architecture générale de notre système pour la reformulation contextuelle des requêtes utilisateur.

Avant de lancer sa requête, l'utilisateur s'identifie dans le système qui procède alors à la récupération de son *Contexte Statique*, il s'agit de ses caractéristiques personnelles pouvant influencer le contexte de recherche. Ces renseignements ont été enregistrés dans la *Base des Contextes Utilisateurs* lors de la première connexion au système. Dans le cas d'un utilisateur qui ne possède pas un profil, le système lui demande de remplir ses préférences et la *Base des Contextes Utilisateurs* sera mise à jour pour une éventuelle utilisation dans des prochaines sessions de recherche.

Une fois le *Contexte Statique* récupéré, l'utilisateur peut alors formuler sa requête et le système procède à la reformulation contextuelle, il se charge de générer la nouvelle requête en sélectionnant les termes relatifs au contexte de la recherche en cours, cette sélection est faite à partir de la *Base des Contextes Utilisateurs*, les deux types de contexte (*Statique* et *Dynamique*), contribuent donc mutuellement à l'opération de reformulation. Par la suite le système lance une recherche ouverte sur internet en utilisant la requête reformulée et en appelant selon le choix de l'utilisateur l'un des trois moteurs de recherche qu'il propose (Google, Yahoo ou Msn). Le résultat de la recherche est retourné enfin à l'utilisateur, il sera stocké également dans l'historique des recherches pour être utilisé par la suite dans la capture du *Contexte Dynamique*.

A la fin de chaque session de recherche et en se basant sur l'historique des recherches faites, le système récupère automatiquement des informations (*Eléments Contextuels*) qu'il suppose pertinentes pour enrichir le *Contexte Dynamique*. Il les propose à l'utilisateur qui valide par la suite celles qu'il juge réellement pertinentes parmi l'ensemble de propositions.

Une architecture à base des profils pour la reformulation contextuelle des requêtes utilisateur.

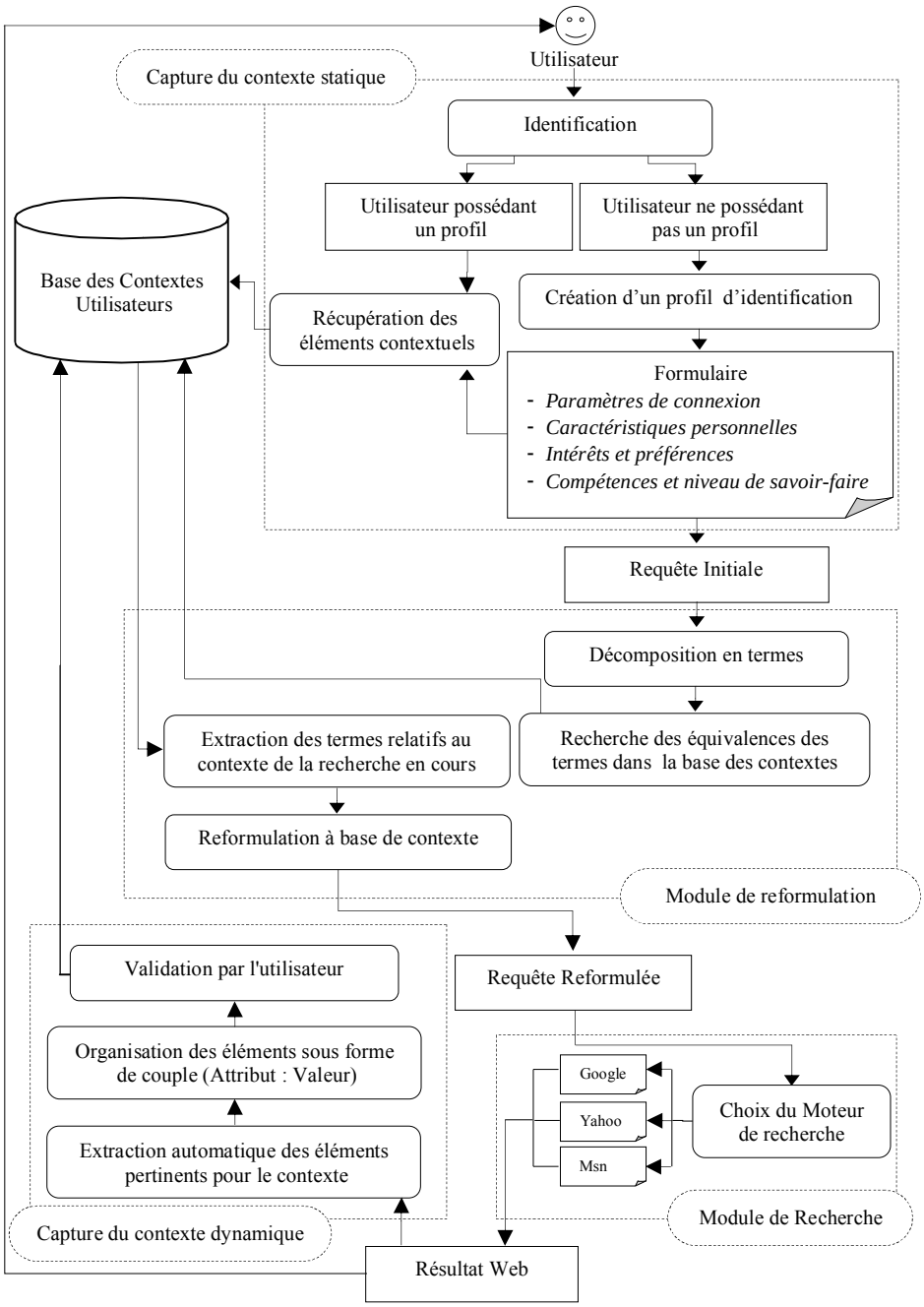


Fig. 5 – Architecture générale d'un système pour la reformulation contextuelle des requêtes utilisateur à base des profils

5 Prototype développé

Afin de montrer l'applicabilité de l'architecture proposée, nous avons mis en place un prototype d'un système pour la reformulation contextuelle des requêtes. L'application communique avec une base de données pour stocker les éléments contextuels. Cette base de données contient deux tables, la première sert à garder les préférences de l'utilisateur (contexte statique) ainsi que les éléments contextuels (contexte dynamique), la deuxième table contienne les historiques des recherches servant à la capture du contexte dynamique. Les figures 6, 7 et 8 présentent respectivement la fenêtre principale de l'application avec les différente fonctionnalités offertes par le système, le mécanisme pour la capture du contexte dynamique et enfin la manière selon laquelle la requête initiale est reformulée.

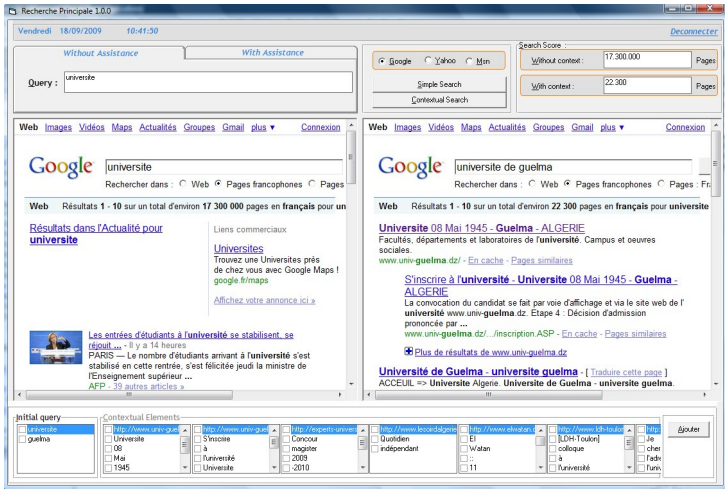


Fig. 6 – Fenêtre principale de l'application.

Cette fenêtre propose à gauche un résultat de recherche en utilisant la requête sans reformulation, et à droite un résultat de recherche de la même requête après reformulation. L'interface offre également la possibilité de choisir le moteur de recherche souhaité et donne le score de pertinence en terme de nombre de page retournée dans les deux cas, recherche avec et sans reformulation. L'analyse du résultat retourné pour la requête 'université de guelma' en utilisant le moteur de recherche 'Google' montre que le nombre de pages retournées sans reformulation était de '5.800.000 pages' tandis que dans le cas de la recherche avec reformulation le nombre de pages est réduit à '143.000 pages', ce dernier score est nettement plus satisfaisant par rapport à l'utilisateur car le résultat retourné ne contient que les pages qui l'intéressent réellement.

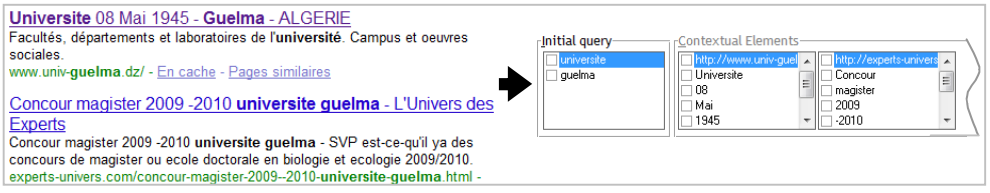


FIG. 7 – Mécanisme pour la capture du contexte dynamique

Pour la récupération du contexte dynamique, le système analyse automatiquement le contenu de la page web résultante. En utilisant les titres de chaque résultat il procède à la segmentation des phrases récupérées puis à l'élimination des mots vident en utilisant un anti-dictionnaire. Le système propose enfin les mots obtenus à l'utilisateur, ce dernier sélectionne ceux qu'il juge réellement pertinents et qui seront donc rajouter à sont contexte dynamique.

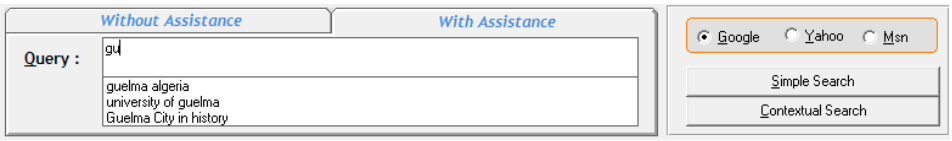


FIG. 8 – Reformulation de la requête initiale

Afin d'assurer la reformulation de la requête initiale, le module de reformulation récupère la saisie de l'utilisateur, il la compare avec le contenu de sa base de contextes et il propose à l'utilisateur des possibilités de reformulation selon le contexte de la recherche en cours. Le système propose également un autre type de reformulation sans intervention de l'utilisateur dans lequel la requête est reformulée après la fin de sa saisie.

6 Conclusion

Dans cet article nous avons proposé un système de reformulation des requêtes basé sur le contexte de l'utilisateur. La contribution essentielle de notre travail consiste en la proposition d'une architecture qui s'articule autour de quatre modules afin de permettre la récupération des deux types de contexte (statique et dynamique) permettant la catégorisation de chaque utilisateur, puis leurs utilisations dans la reformulation de la requête initiale. Le prototype que nous avons implémenté démontre l'applicabilité de l'architecture proposée et montre que le résultat obtenu avec une requête reformulée est plus pertinent que celui rendu en utilisant une requête sans reformulation.

Ce travail ouvre la voie vers diverses perspectives qui se situent sur deux plans : un plan d'approfondissement de la recherche réalisée et un plan d'élargissement de son domaine d'application. Pour ce qui est de l'approfondissement, il serait intéressant d'améliorer le prototype développé pour que la capture des éléments contextuels soit faite sans l'intervention de l'utilisateur. En ce qui concerne l'élargissement du domaine d'application, il serait intéressant d'expérimenter l'architecture proposée sur plusieurs moteurs de recherche afin de mesurer l'apport de la reformulation contextuelle sur chacun d'entre eux et focaliser par la suite le prototype sur le moteur de recherche qui s'adapte au mieux à notre proposition.

Références

Allan, J. (2003). *Challenges in information retrieval and language modeling*, SIGIR Forum, 37(1), pp 31-47.

Benammar, A., J. Mothe et G. Hubert (2002). *Automatic profile reformulation using a local document analysis*. European colloquium on IR research, Glasgow. Springer-Verlag , pages 124-134.

Belkin, N., G. Muresan et Zhang, X. (2004). *Using User's Context for IR Personalization*. Proceedings of the ACM/SIGIR Workshop on Information Retrieval in Context.

Brézillon, P. (2006). *Expliciter le contexte dans les objets communicants*. C. Kintzig, G. Poulain, G. Privat, P.-N. Favennec (Eds.), Hermès, chapitre 21, 2002, p. 295-303.

Calabretto, S., E. Egyd-Zsigmond (2006). *Recherche d'Information en contexte*. EARIA'06, France.

Cui, H., J-R. Wen et Nie, J-Y. *Probabilistic query expansion using query logs*. (2002). 11th international conference on World Wide Web, p. 325-332, Honolulu, Hawaii.

Gaëtan, R. (2006). *Méthode pour la modélisation du contexte d'interaction*. RSTI - ISI – 11/2006. Adaptation et contexte, pages 141 à 166.

Hernandez N. (2006). *Ontologie de domaine pour la modélisation du contexte en recherche d'information*, thèse de doctorat en informatique, Université Paul Sabatier.

Korfhage, R. (1997). *Information storage and retrieval*. Wiley Computer Publishing 0-471-14- 338 3.

Lin, H et L. Wang. (2006). *Query expansion for document retrieval based on fuzzy rules and user relevance feedback techniques*. Expert Systems with Applications, 31(2), 397-405.

Navigli, R et P. Velardi. (2003). *An analysis of ontology-based query expansion strategies*. Proceeding of the Workshop on Adaptive Text Extraction and Mining, Croatia.

Wahlster, Wet A. Kobsa, (1986). *Dialogue-based user models*. Proceedings of IEEE, Vol. 74(7), pp. 948-960.

Winograd, T. (2001). *Architectures for context, Human-Computer Interaction*,, p.402-419.

Summary

This paper falls under the field of information retrieval on the Web and proposes an architecture based on the user's profiles to take context into account in the queries reformulation. It is to make information of each user in the form of contextual elements to be used later in the process of query reformulation. After an assessment of classic query reformulation approach, we introduce the notion of context and user's profile. Next we describe the proposed architecture and we show how static and dynamic contexts are captured and how these two types of context are used in our proposal. Finally, we present the developed prototype, and we conclude.

Une structure logicielle distribuée pour la découverte des règles d'association spatiales

Azedine Boulmakoul, Abdelfettah Idri, Mohamed Bendaoud, Rabia Marghoubi
Département informatique, FST Mohammedia, B.P. 146 Mohammedia Maroc
azedine.boulmakoul@yahoo.fr

Résumé. Dans ce travail nous présentons l'architecture et les fonctionnalités d'un système de data mining spatial exploitant les transactions spatiales. Le système proposé met en œuvre des techniques de découverte des règles d'association spatiales basées sur la fermeture de Galois parallèle, distribuée et déployée sur un bus CORBA. La spécification de l'architecture du système ainsi que les diverses composantes logicielles seront développées. Le module d'extraction des transactions spatiales constitue la pièce maîtresse du système. IL fait recours à des éléments structurants de voisinage de type *grille*, *buffer* et *polygone de voronoï*.

1 Introduction générale

Dans le domaine de l'informatique décisionnelle, une évolution importante de ces dernières années a été constatée pour le développement rapide du data mining spatial (Koperski et al., 1993), (Shekhar et al., 2003), (Malerba, 2008). Ce nouveau domaine de recherche se situe au croisement des bases de données spatiales, des statistiques spatiales et de l'intelligence artificielle. Ces recherches s'étendent aujourd'hui à des données complexes et notamment aux données spatio-temporelles. Dans le contexte des bases de données géographiques ou spatiales, la nature et le volume de données dépassent les capacités humaines en matière d'analyse décisionnelle. D'où le besoin d'utilisation des méthodes avancées en matière de découverte de connaissances. Les règles d'association spatiales constituent une solution prometteuse parmi la panoplie de méthodes offertes par le data mining spatial. Ces dernières, appliquées dans plusieurs domaines d'activité, présentent des résultats clairs, faciles à interpréter. Cependant, deux problèmes majeures liés à l'utilisation des règles d'association spatiales à savoir , le problème des temps d'extraction des règles d'association à partir du jeu de données et le problème de la pertinence et de l'utilité des règles d'association extraites. Par ailleurs, l'extraction de connaissances des données géographiques répond à un besoin réel des applications en géomatique pour tirer profit de la disponibilité croissante de données localisées et de la richesse potentielle en informations de ces données. Ce travail de recherche propose d'apporter des solutions concernant le problème d'extraction de la connaissance spatiale à l'aide des règles d'association spatiales dans divers domaines, tels que, les télécommunications, le géomarketing, l'analyse des risques, etc. Ce travail a été finalisé par le développement d'un prototype logiciel en phase de *pré-industrialisation*. Dans le cadre d'une étude réalisé pour le compte d'un opérateur de télécommunication, cet outil décisionnel a permis de suivre et de conduire en particulier l'évolution du service universel. La mise en œuvre de la solution dans le domaine du géomarketing, a permis aussi de mettre en évidence les corrélations existantes entre les comportements d'achat (selon le type de service : fixe, mobile, LS, ADSL, etc.) des clients et certaines de leurs caractéristiques individuelles. La localisation géographique est pratiquement toujours la caractéristique la mieux renseignée.

Faisant référence à la problématique suscitée, nous pouvons classer les objectifs de ce projet en deux catégories : la première facette est stratégique ; elle concerne l'extraction des transactions spatiales, pour permettre l'usage des algorithmes de mining pour la génération des règles d'association spatiales. La seconde est spécifique ; elle se focalise sur le développement des algorithmes d'extraction de la connaissance spatiale qui se basent sur les règles d'associations spatiales. Les approches proposées doivent améliorer les solutions existantes, en particulier le problème des temps d'extraction des règles d'associations à partir du jeu de données et le problème de la pertinence et de l'utilité des règles d'associations extraites.

La structure de ce papier est donnée comme suit : la section 2 introduit l'architecture globale du système de data mining spatial. Ce système exploite les transactions spatiales afin de mettre en œuvre des techniques d'extraction des règles d'association spatiales basées sur les treillis de Galois. La section 3 traite l'extraction des transactions spatiales. Ce module constitue la pièce maîtresse du système. L'objectif de l'extracteur spatial est de fournir un mécanisme pour représenter des relations spatiales implicites sous une forme appropriée pour les algorithmes de mining. Cette proposition exploite l'organisation thématique de l'ensemble de données spatiales, en employant des éléments structurants de voisinage de type *grille, buffer, polygone de voronoï*. Ces éléments structurants constituent les sélecteurs des requêtes spatiales. Chaque instance de l'élément structurant identifie les objets des couches qui lui sont reliés par une relation spatiale (intersection, intérieur, etc.). Dans la section 4 sont développés les composants liés au mining et qui sont basés sur les fermetures de Galois. L'algorithme de mining utilisé dans ce travail est de nature parallèle, distribuée et déployé sur un bus CORBA. La section 5, aborde les aspects liés à l'interface homme machine, et donne quelques interfaces matérialisant les fonctionnalités du système. Ce travail est complété par une conclusion et trace le bilan ainsi que les perspectives.

2 Architecture globale du système cible

L'architecture globale du système se présente sous la forme d'un système modulaire intégrant les fonctionnalités et les données essentielles à la mise en œuvre des analyses spatiales complexes (figure 1). Elle décrit d'une manière symbolique et schématique les différents composants, leurs interrelations et leurs interactions. Cette architecture met en exergue cinq couches fondamentales : les serveurs de données spatiales et sémantiques, le module d'extraction spatiale, le module de découverte et structuration, l'interface homme machine, et le *web mapping*. Dans ce papier seront abordés en priorité l'extraction des transactions spatiales, la découverte des règles d'association spatiales, et l'interface homme machine. Les autres modules seront traités dans d'autres papiers.

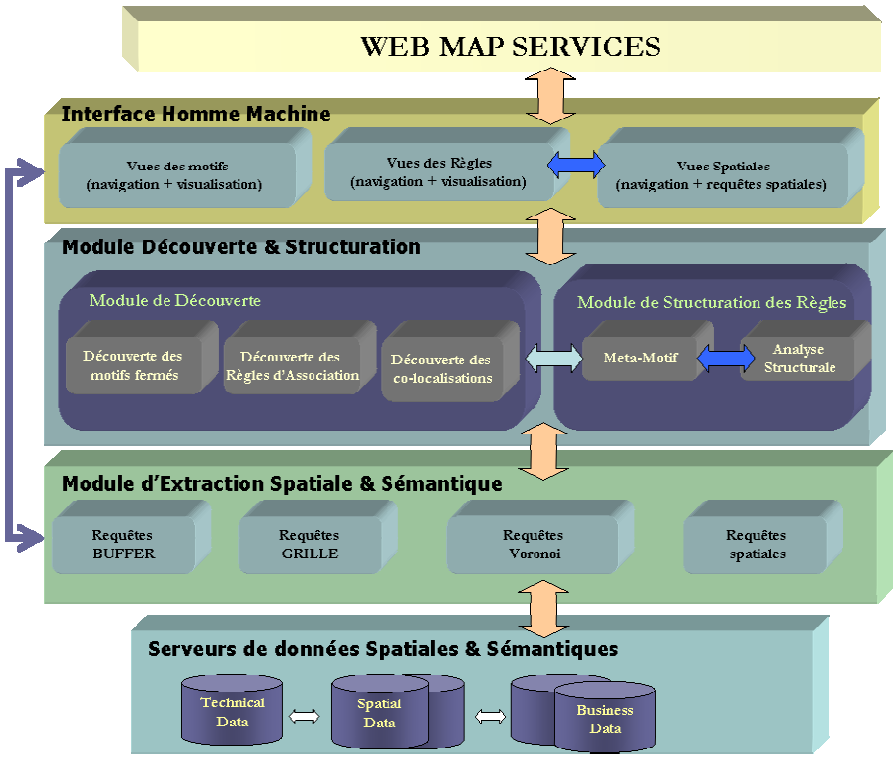


FIG. 1 – *Architecture globale du système*

3 Module d'extraction des transactions spatiales

3.1 Requêtes et analyse spatiales

Les requêtes spatiales permettent d'exécuter une requête répondant à un ou des critères portant sur la position des objets et leurs relations spatiales avec d'autres objets. Ces relations peuvent être soit intra-thèmes (la même couche thématique) ou inter-thèmes (différentes couches thématiques). Par exemple lors d'une intervention sur le réseau télécoms, nous pourrions identifier à l'aide d'une requête spatiale tous les types de localités qui sont concernés par cette intervention. Elles permettent de choisir une ou plusieurs couches à sélectionner, un "opérateur spatial" et une couche de référence. La position théorique générale de l'analyse spatiale consiste à proposer une explication partielle, et des possibilités de prévision, quant à l'état et à l'évolution probable des objets/unités géographiques, à partir de la connaissance de leur situation par rapport à d'autres objets géographiques. La plupart des fonctions d'analyse spatiale donnent lieu à la création d'une nouvelle classe d'entité ou d'un fichier de formes : calcul de zones de proximité (zones tampons ou buffers), jointures spatiales, croisement de couches, extraction suivant des critères géométriques, fusion d'entités en fonction d'un attribut, fusion de couches (ou combinaison), distance Point/Nœud. Ces données peuvent être également classées selon leurs formes géométriques : les points (carre-

fours, gares), les lignes (routes, rivières, frontières), les zones (occupation du sol, commune), les images (cartes numérisée, images satellites). De même que les données internes et socio-démographiques se déclinent sous des formes différentes, il existe deux formats principaux pour représenter les données cartographiques : le format *raster* et le format *vectoriel*.

3.2 Extraction des transactions spatiales

C'est une phase très importante dans le processus d'extraction et de visualisation de la connaissance spatiale. Cette opération est une partie complexe et longue du processus spatial de découverte de la connaissance, dû à la taille des objets spatiaux et à la complexité de l'extraction spatiale des relations entre les objets spatiaux. Nous proposons un mécanisme général pour représenter les transactions spatiales d'une manière à permettre l'utilisation des méthodes de découvertes de connaissances de type règles d'association. Notre solution permet à l'analyste de sélectionner une couche de référence à partir d'un système d'information géographique, de définir le contexte d'extraction spatiale (ensemble de couches concernées par la découverte) et de spécifier les relations spatiales appropriées par l'intermédiaire du choix d'un élément structurant de voisinage. Etant donné une couche de référence, il est possible de décrire le voisinage des objets appartenants à la couche, en considérant l'attribut de l'objet lui-même et les objets connexes par l'élément structurant choisi (*buffer, grille, polygone de voronoi*). Les transactions spatiales en résultant peuvent être considérées comme des transactions « traditionnelles », qui pourront être exploitées par un processus de génération de règles d'association. La figure 2 décrit une vue d'ensemble de la phase de prétraitement. Les étapes ultimes pour l'extraction des transactions spatiales sont aussi succinctement présentées.

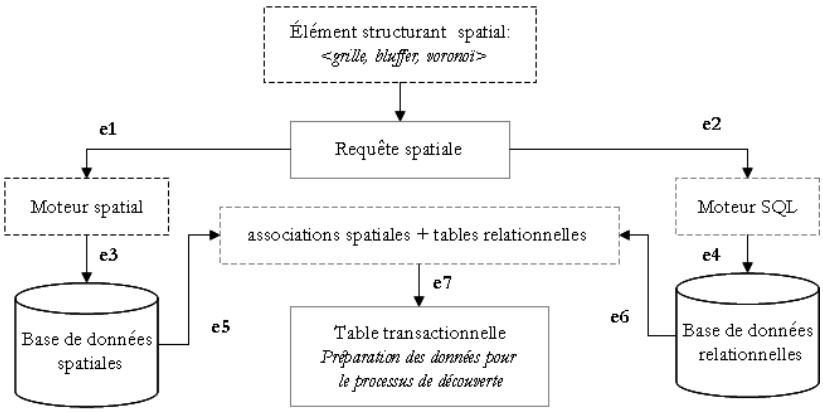


FIG. 2 – *Processus d'extraction des données transactionnelles.*

E1 : Cette étape utilise le moteur spatial de base de données pour recevoir des requêtes spatiales de l'utilisateur, selon l'élément structurant spatial: grille, buffer ou polygone de voronoï.

E2 : Cette étape utilise le moteur de serveur de SQL pour recevoir des requêtes métiers de l'utilisateur.

E3 : Cette étape emploie le moteur spatial pour exploiter les algorithmes géométriques (*MapObject ESRI©*) afin de trouver des relations spatiales entre deux objets spatiaux selon les contraintes d'utilisateur.

E4 : Cette étape utilise le moteur de serveur de SQL pour rechercher une donnée métier d'un objet spatial indiqué et pour transformer une telle donnée en format qualitatif. Par exemple, la donnée point de vente pou laquelle tout le revenu de ses clients égal à 500 000 DH/mois. Une telle information métier est transformée en format qualitatif "haute", elle est stockée dans une base de données relationnelle.

E5 : Cette étape stocke le résultat de l'étape E3 dans une table provisoire comme montré dans le schéma.

E6 : Cette étape stocke le résultat de l'étape E4 dans une table provisoire comme montré dans le schéma.

E7 : Cette étape utilise le moteur SQL pour joindre des tables obtenues à partir des étapes E4 et E5, comme montré dans le schéma.

Dans suite nous allons décrire brièvement les types de requêtes qui valorisent la composante spatiale. Ces requêtes peuvent être classées comme suit :

3.3 Requête de type buffer

Ce type de requête répond à un besoin spécifique de l'utilisateur. Elle permet à ce dernier de créer ou délimiter une zone dans l'espace (généralement 50 mètres) pour savoir l'environnement ou l'entourage d'un service donné (figure 3). Autrement dit que, cette zone va rassembler un ensemble de points potentiellement différents mais dont on peut considérer qu'ils sont liés vis-à-vis d'une propriété sur laquelle on souhaite baser un raisonnement.

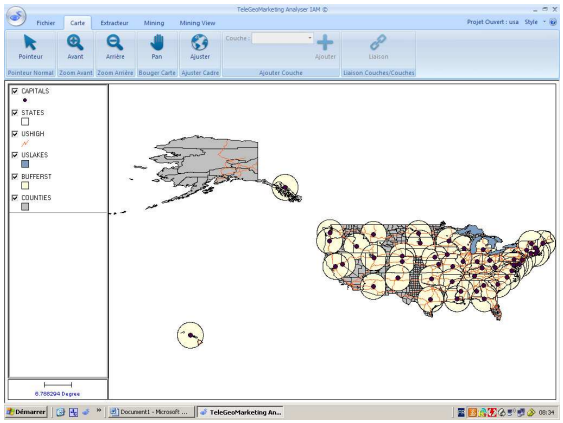


FIG. 3 – *Extraction par voisinage de type buffer.*

3.4 Requêtes de type grille

Ce type de requêtes produisent un découpage de l'espace en cellules de taille uniforme (méthode *fixed grid*) ou en cellules de taille variable (méthode *grid file*) dans le cas où la distribution des objets est non uniforme (figure 4) ; elles ont pour inconvénient que le nombre de cellules formées par ce partitionnement peut croître très rapidement pour des grands volumes de données. Elles peuvent être appropriées si la distribution et la taille des objets spatiaux sont uniformes, par contre pour des objets de dimensions variables, la méthode des arbres quaternaires est plus appropriée puisque le découpage s'adapte à la densité de l'information spatiale. Ainsi, cela évite la présence d'un grand nombre de subdivisions vides, inutiles pour décrire la répartition des objets, ou la présence de cellules comportant au contraire trop d'information et qui devraient être divisées plus finement.

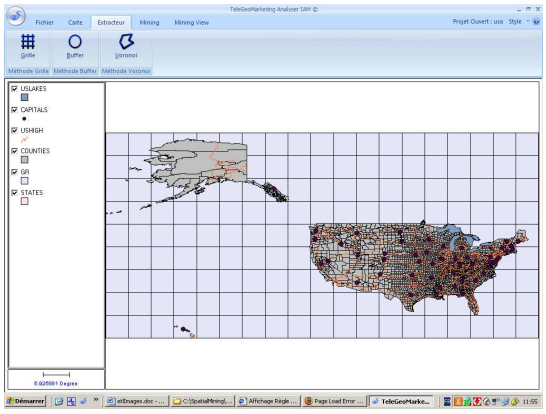


FIG. 4 – *Extraction par voisinage de type grille.*

3.5 Requêtes de type Voronoï

En géométrie algorithmique, un diagramme de Voronoï est une décomposition particulière d'un espace métrique déterminée par les distances à un ensemble discret d'objets de l'espace, en général un ensemble discret de points. On se place dans un espace euclidien E. soit S un ensemble fini de n points de E; les éléments de S sont appelés centres, sites ou encore germes. On appelle région de Voronoï ou cellule de Voronoï associée à un élément p de S l'ensemble des points qui sont plus proches de p que de tout autre point de S (figure 5).

$$Vor_s(p) = \{x \in E \ / \ \forall q \in S \ d(x,p) \leq d(x,q)\}$$

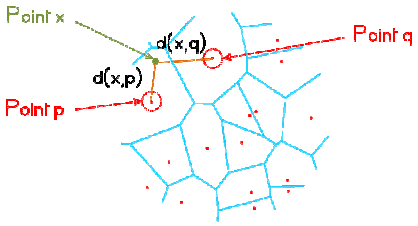


FIG. 5 – *Partitionnement en régions de Voronoi.*

Contrairement au requêtes de type *buffer*, qui nécessite d'abord le paramètre de la distance (le rayon) pour délimiter la zone tampon, les types de requêtes *polygone de voronoi* permettent à l'utilisateur de faire une sélection spatiale des différentes couches thématiques, inter thème ou intra thème en se positionnant seulement sur un seul point de l'espace. Elles permettent de rechercher les plus proches voisins. Par exemple, dans le domaine de télécommunications, et plus particulièrement dans la modélisation des réseaux téléphoniques. Aujourd'hui, quand nous téléphonons avec notre portable, notre demande est (en première approximation) transmise à la station la plus proche. Toutes les stations de la région sont reliées entre elles par un réseau filaire qui permet de faire circuler la communication jusqu'à la station la plus proche de notre correspondant. Où doit-on ajouter des stations et combien si le réseau grandit ? La zone précise où chaque abonné est plus proche d'une station que de n'importe quelle autre est la cellule de *Voronoï* de cette station (figure 6).

En résumé, on peut dire que cet extracteur spatial a un double avantage, le plus évidement c'est la préparation du contexte spatial pour le module de découverte que nous allons détailler dans la section 4. Le deuxième avantage est qu'il produit des résultats qui peuvent être exploités directement par l'utilisateur final et ce, grâce à l'utilisation des différents types de requêtes.

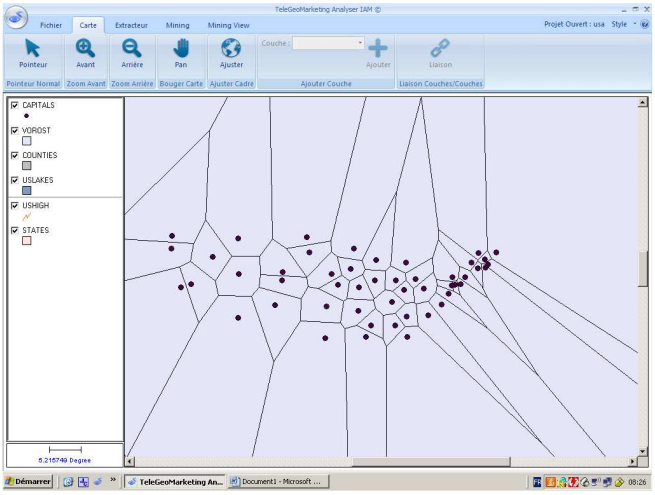


FIG. 6 – *Extraction par voisinage de type polygone de voronoi.*

3.6 Formalisme d'extraction des transactions spatiales

Dans cette section nous proposons un formalisme pour représenter le processus d'extraction. Deux possibilités sont offertes : la première concerne les éléments structurants de voisinage de type *buffer ou polygone*; la seconde est relative à la *grille* ou structure uniforme équivalente. Pour la première catégorie le processus est basé sur la définition des rela-

tions spatiales entre une couche de référence et un ensemble de couches candidates. Dans une application SIG, un ensemble de données spatiales se compose d'un ensemble de couches, où chaque couche apporte l'information sur un aspect particulier du monde réel. Une couche est un ensemble d'entités spatiales, où chaque entité f est associée à une géométrie, représentant sa localisation dans l'espace, et un ensemble d'attributs, qui décrivent l'état de l'entité. Dans la suite, nous nous référons aux entités géographiques en tant que qu'éléments spatiaux et en tant qu'objets spatiaux. Ce qui caractérise une région géographique est l'union de toute l'information dans toutes les couches, c.-à-d. l'information apportée par toutes les entités dans toutes les couches, situées dans la région considérée. Cette manière d'organiser des données spatiales soulève un nouveau défi en définissant une transaction spatiale. En fait, une transaction est un tuple des attributs rassemblés par toutes les couches et liés à une géométrie représentative (c.-à-d. la géométrie où les localisations de tuples). En général, une des couches disponible est choisie comme couche de référence (selon le type d'éléments structurant de voisinage), et chaque entité dans cette couche est employée pour choisir les entités dans les autres couches. Cette formalisation exploite les travaux de Rinzivillo et al. (2007) et ceux de Estivill-Castro et al. (2001). Le choix du voisinage de type *polygone de voronoi* est proposé pour la première fois dans ce travail.

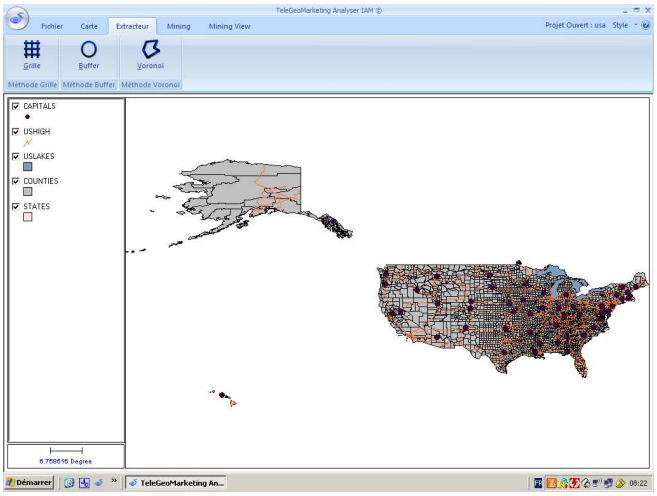


FIG. 7 – *Contexte spatial d'extraction de transactions spatiales.*

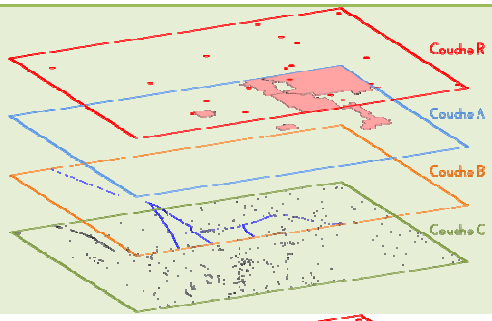
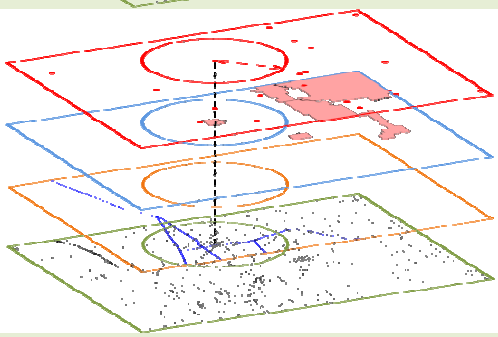
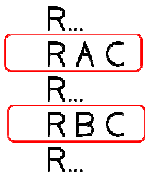
Formellement, soit $L = \{L_1, l_2. . . , L_n\}$ un ensemble de couches, L_r dénote une couche de référence, et S_R un ensemble d'élément structurant de voisinage (figure 7). Chaque couche possède un ensemble d'attributs non-spatiaux qui décrivent l'état de chaque objet dans la couche. Pour la clarté de la présentation, nous supposons que chaque couche L_i possède seulement un attribut catégorique Al_i. Pour chaque objet o de L_i la fonction VL_i (o) renvoie la valeur de l'attribut Al_i pour o. Etant donné un objet o_r de L_r, Une transaction spatiale de l'objet o_r est un ensemble de la forme $To_r = \{<R_i, VL_j (o_i)>|R_i \in S_R \wedge VLj \in ALj, j = 1. . . , n \wedge R_i(o_r, o)\}$, c.-à-d. l'ensemble de tous les objets dans les couches $L_1, L_2. . . , L_n$ qui satisfont une relation de S_R avec l'objet de référence o. La paire $<R_i, VL_j (o_i)>$ s'appelle un item spatial.

Intuitivement, un item spatial peut être représenté par l'attribut (R·, o) et son support est donné par le nombre d'objets dans la couche de référence qui la satisfont. Un ensemble d'items spatiaux s'appelle l'itemset spatial. Le support d'un itemset spatial est le nombre d'objets dans la couche de référence qui satisfont la relation spatiale dans la transaction. Etant donné une séquence de couches $L_1 . . . ,L_n$, un ensemble de relations spatiales (éléments structurants de voisinage) $R_1 . . . ,R_n$ où la relation R_i est associée à la couche L_i pour chacun i = 1. . . , n, et une couche L_r de référence, un ensemble spatial de transaction est défini comme ensemble de tuples : $STS(L_r, (L_1, R_1) . . . , (L_n, R_n)) = \{o_i \mid To_i \in L_r\}$.

Dans le contexte de ce travail, la relation contexte spatial, peut être simplement définie par : $\lambda_{i,j} = \partial_{\left\{V_{S,i}\left\langle f_{ref}^{i} \right\rangle \cap L_j \neq \varnothing\right\}}$, où ∂ désigne une relation spatiale (intersect, touche, contient, etc.) $V_{S,i}$ désigne un élément structurant de voisinage relatif à la couche de référence de type {*cellule, buffer, polygone de voronoi*}.

	L_{ref}	L_1	..	L_j		L_{k-1}
$V_{S,0}$	1					
	1					
	1					
$V_{S,i}$	1			$\lambda_{i,j}$		
…	1					
$V_{S,N}$	1					

TAB. 1 – *Base de données des transactions.*

1		*Chargement des couches*
2		*Projections*
3		*Evaluation des prédicats spatiaux.*
4		*Génération du fichier de transactions*

TAB. 2 – *Processus de génération des transactions spatiales.*

4 Module de découverte

Pour ce module, nous allons détailler les concepts fondamentaux pour la découverte des connaissances à partir de la base de données spatiales transactionnelle (Agrawal et al., 1993). En ce qui concerne la phase relative à la découverte nous allons présenter notre approche pour découvrir les motifs fréquents ainsi les règles d'associations spatiales. Quant à la phase relative à la structuration, elle ne sera pas décrite dans ce papier.

4.1 Module de découverte des règles d'association spatiales

La théorie des treillis a été employée avec succès dans le contexte d'extraction des motifs fréquents fermés en « data mining ». En particulier, les fondements théoriques basés sur les treillis de Galois ont été utilisés dans l'élaboration des algorithmes efficaces pour la découverte des motifs fréquents dans les bases de données transactionnelles. Les graphes bipartites ont été également employés en tant que théorie formelle pour énumérer tous les motifs fréquents maximaux. Rappelons que l'extraction des itemsets fermés fréquents est une méthode en rapport avec l'analyse de concept formel (treillis de concepts), basée sur la fermeture de la connexion de Galois (Zaki et al., 1998). Ces itemsets sont les itemsets fréquents qui sont fermés selon l'opérateur de fermeture de la connexion de Galois (Boulmakoul et al., 2007), (Idri et al., 2008). Les itemsets fermés fréquents, selon cet opérateur de fermeture, constituent un ensemble générateur non redondant minimal pour tous les itemsets fréquents et leurs supports. Tous les itemsets fréquents et leurs supports, et donc toutes les règles d'association ainsi que leurs supports et leurs confiances, peuvent donc être déduits efficacement, sans accéder au jeu de données, à partir des itemsets fermés fréquents et leurs supports. Cette propriété découle du fait que le support d'un itemset fréquent est égal au support de sa fermeture et que les itemsets fréquents maximaux sont des itemsets fermés fréquents maximaux. Les itemsets fermés fréquents forment un treillis dont la taille est bornée par la taille du treillis des itemsets fréquents. Toutefois, en pratique, la taille de ce treillis est en moyenne bien inférieure à la taille du treillis des itemsets. Le principe de base de ces algorithmes est de déterminer les itemsets fermés fréquents afin de déduire les itemsets fréquents, ce qui réduit le temps d'extraction et produit des règles non redondantes. Ces algorithmes considèrent un ensemble de générateurs candidats d'une taille donnée, et déterminent leurs supports et leurs fermetures en réalisant un balayage du contexte lors de chaque itération. Les fermetures des générateurs fréquents sont les itemsets fermés fréquents extraits lors de l'itération. Les générateurs candidats sont construits en combinant les générateurs fréquents extraits durant l'itération précédente.

4.1.1 Découverte des règles d'associations spatiales

L'information géo-référencée ne cesse de croître chaque jour, et les systèmes d'information géographiques deviennent cruciaux dans beaucoup de procédés de décision. Par conséquent, extraire la connaissance à partir des bases de données spatiales (SGBDS) peut avoir un impact important et influencer les choix décisionnels. Le présent travail présente une approche nouvelle pour extraire des ensembles de transactions spatiales à partir des SGBDS, et pour ensuite appliquer des algorithmes de mining à ces données. Le processus que nous présentons aboutira à l'extraction des règles d'association spatiales.

Rappelons qu'une règle d'association d'une manière générale est de la forme suivante :

« si A alors B », (s), (c) ; où A et B sont des ensembles attributs (Itemset), (s) est le support de la règle et (c) est la confiance de la règle. Le but de ces règles est de trouver les différentes corrélations existantes entre A et B. Pour chaque règle, deux paramètres sont associés : le support et la confiance. Le support, mesurant l'utilité d'une règle, c'est la probabilité absolue P(A&B) pour que la règle soit vérifiée dans une transaction, c'est-à-dire que A et B

soient présents. La confiance, mesurant la pertinence d'une règle, indique le pourcentage de transactions qui vérifient la conclusion d'une règle parmi celles qui vérifient la prémisse. Les règles d'association spatiales sont une extension des règles d'association classiques. Elles opèrent sur plusieurs couches thématiques pour permettre d'expliquer un phénomène suivant les propriétés de son voisinage et ce, en introduisant les données et les critères spatiaux. Elles sont de la forme suivante :

R2 : « $P : P1 \wedge P2 \wedge P3 \wedge ... \wedge Pn \Rightarrow Q : Q1 \wedge Q2 \wedge Q3 \wedge .... \wedge Qm(s,c)$ » où l'un des prédicats P1, P2….Pn, Q1, Q2….Qm est un prédicat spatial.

Exemple de règle :

R2 : *<estUn, pointDeVente>* $\wedge$ *<présDe, Gare>* $\rightarrow$ *<HAUT, chiffreAffaire> (100%,100%).* La règle R2 exprime que tous les points de vente qui sont prés de la gare présentent un chiffre d'affaire important (HAUT).

4.2 Extraction des règles d'association spatiales

La construction de treillis de Galois a intéressé plusieurs chercheurs, spécialement dans les domaines d'analyse de concepts formels d'une part (Ganter, 1999), (Bordat, 1986), (Chein, 1969) (Stumme, 1999) et la fouille de données d'autre part (Zaki et al., 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 al., 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 al., 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 cette section nous présentons 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. 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.

4.2.1 Architecture du système

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.

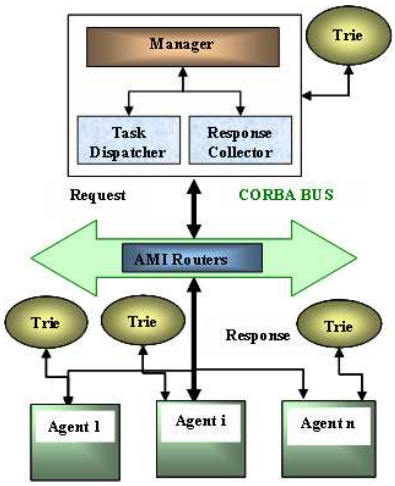


FIG. 8 – *Architecture du système de découverte*

La démarche globale adoptée pour la conception de cette architecture est décrite dans ce qui suit. Pour être en mesure de concevoir une architecture parallèle de treillis de Galois, il faut en premier lieu savoir 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 (Njiwoua et al., 1997), (Bordat, 1986), (Choi, 2006) et (Ganter et al., 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 ce ci 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 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. Dans le cas contraire, une copie de cet arbre doit être maintenue et gérée par chaque agent ce qui rend la gestion de cet arbre très complexe. 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 (figure 8). 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) et gestion des tâches (distribution, collection, synchronisation).

Agent : Génération des concepts enfants, contrôle de fermeture de l'intention ou l'extension d'un concept.

4.2.2 L'algorithme parallèle de construction de treillis de Galois

```
Initialise concept queue Q= {C};  C = (O,attr(O))
 Initialise Context, Agenda, Trie
 Initialise and Synchronise Agents
while Q is not empty or Agenda is not empty
                    or waiting for reponses do
  if Q is not empty then
   C= dequeue(Q)
       Enqueue C to Agenda
    end if
    for each available Agenti do
      if Agenda is not empty then
            C = dequeue(Agenda)
            Send_request(generateChildren(C)) to Agenti
            Mark Agenti as busy
      end if
    end for
    for each buzy agenti do
      if response of agenti is available then
              get response
              for each child in response do
              if child not exists in Trie then
                  insert child into Trie
                      Identify child as successor of C
                      Enqueue child to Q
             else
                      Identify child as successor of C
            end if
          end for
              mark agenti as free
      end if
  end for
end while
```

TAB. 3 – *Algorithme parallèle Manager*

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 ré-

ponses 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.

4.2.3 Implémentation

Dans ce paragraphe, on traite les aspects d'implémentation de l'algorithme. En fait, notre implémentation est applicable aussi bien dans le domaine de l'analyse formelle des concepts que dans le domaine de la fouille de données. Après la définition des données d'entrée et de sortie ; nous présentons les structures de données principales utilisées dans l'algorithme. Ensuite, on discutera la communication entre le Manager et les Agents ainsi que l'outil CORBA.

Les entrées et sorties : Nous avons 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 XML et un format interne spécifique pour des fins d'analyse et de recherche.

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. L'ensemble des objets et l'ensemble des attributs sont implémentés comme *set<int>* avec toutes les opérations nécessaires de manipulations des ensembles. La relation binaire comporte sa table de données sous forme d'un *set<reltype>* où, *reltype* est une structure comportant un objet, un attribut et une booléenne indiquant leur relation. En plus, la relation binaire possède un ensemble de listes adjacentes objets et un ensemble de listes adjacentes attributs. C'est utile de les calculer auparavant et de les charger en mémoire pour la bonne performance de l'algorithme. 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 qui nous permet de remplir les listes adjacentes sans passer par la structure *reltype*.

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). Le nœud de l'arbre comporte entre autres une booléenne indiquant s'il s'agit d'un concept, l'ensemble des identités des parents du concept ainsi que l'ensemble des identités de ses enfants.

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.

4.2.4 CORBA et communication Manager/Agent

Pour couvrir la partie communication entre le Manager et les Agents on a choisi d'utiliser CORBA pour les raisons mentionnées ci-dessus. Dans ce qui suit, on discutera l'intégration de CORBA dans notre implémentation.

Le package CORBA : La version CORBA utilisée est celle d'Orbacus 4.3 (2008). L'installation du package nécessite une compilation et une configuration de l'environnement de développement (Visual Studio).

L'interface IDL (Agent/Server) : La conception de CORBA exige la création d'une interface IDL (*Interface Definition Language*). Cette interface expose les fonctionnalités du serveur qui vont être sollicitées par le client. Dans notre modèle Manager/Agent, les rôles du contexte client/serveur sont inversés ; et par conséquent, l'Agent se comportera comme serveur et le Manager jouera le rôle du client conformément à l'architecture du système. Dans notre contexte, l'interface IDL comporte au moins la signature de la fonction qui se chargera de la génération des concepts enfants d'un concept donné. L'implémentation de cette interface, qui est d'ailleurs utilisée et par le Manager et par les Agents, englobe les fonctions de conversion de base pour commuter entre les structures de données CORBA et les structures de données Manager/Agent.

Le modèle de communication distribué : Pour implémenter une telle architecture basée sur CORBA, plusieurs alternatives se présentent selon notre contexte.

Multithreading du Manager : Cette méthode consiste à créer une version du Manager qui supporte le multithreading. Du fait que le manager (client), par défaut, communique avec l'agent (serveur) d'une manière synchrone, le multithreading permettra au manager de lancer simultanément ses requêtes vers les agents. La gestion de la mémoire partagée s'avère nécessaire pour conserver la consistance des données (l'arbre des concepts). Pour se faire on peut se servir des Mutex (sémaphores).

Callback mécanisme : Cette une technique de CORBA qui repose sur le fait d'envoyer dans la requête destinée pour l'agent, une référence de l'objet du manager afin que l'agent puisse avertir le manager une fois les calculs terminés. Ça peut servir dans notre cas afin de minimiser le temps de contrôle de disponibilité des réponses.

Gestion des événements distribués : C'est, en fait, le mécanisme idéal pour implémenter une telle architecture. Seulement, ce service n'est pas offert en standard dans les implémentations de CORBA. Cette technique gère les événements des composants CORBA d'une manière transparente comme si c'était local. Le modèle le plus adapté à notre cas est celui du *canonical event push model.*

AMI (Asynchronous Method Invocation) : Cette méthode permet la conception d'un client (Manager) asynchrone non bloqué mais sans aucune modification du serveur (*AMI Poller*). Le Manager peut donc envoyer des requêtes aux Agents et retourner immédiatement ce qui sert bien notre architecture. On associe à *AMI* l'utilisation de routeurs *AMI* qui prennent en charge le routage des requêtes vers leur destination et leurs renvois ultérieurement en

cas d'échec du serveur (destination). Il existe deux types d'implémentation pour cette approche : *AMI Reply Handler* et *AMI Poller*.

AMI Reply Handler exige l'utilisation d'un objet callback qui fera partie de la requête. Cet objet est utilisé par le serveur (Agent) pour avertir le client (Manager) de l'état de la réponse. Par contre, *AMI Poller* retourne un *poller* au Manager que ce dernier peut interroger pour savoir l'état de la réponse. Une configuration du serveur et du client est nécessaire pour profiter de cette technique.

Pour notre implémentation, on a choisi la technique *AMI Poller*.

Avantages : Le plus intéressant dans cette approche, c'est que l'Agent n'a pas besoin de subir des modifications. Un deuxième avantage réside dans le fait que le routeur AMI Poller est persistent : dans le cas d'un plantage du Manager, le routeur AMI s'occupe de renvoyer la requête du client une fois le serveur rétabli. Le troisième avantage c'est le mode asynchrone de la méthode qui permet au Manager de retourner immédiatement après l'envoi de la requête et continuer sa propre exécution.

Inconvénients : On s'est aperçu en aval que cette méthode, pour le moment, ne supporte qu'un seul routeur AMI persistent à la fois. Pour garantir la scalabilité du système, un multithreading du Manager devrait s'ajouter à l'architecture en plus du mode asynchrone. En contrepartie, *AMI Reply Handler* supporte plusieurs routeurs par ordre de priorité et surmonte cette difficulté.

5 Interface Homme Machine

L'interface homme machine prend en charge la sélection des couches spatiales ainsi que les données sémantiques destinées au mining spatial. Des composants seront aussi spécialisés pour la structuration des règles d'association spatiales générées. Des procédures de navigation dans l'espace des règles sont aussi proposées. Le système de visualisation offre une ouverture sur le web (voir figure 9, figure 10).

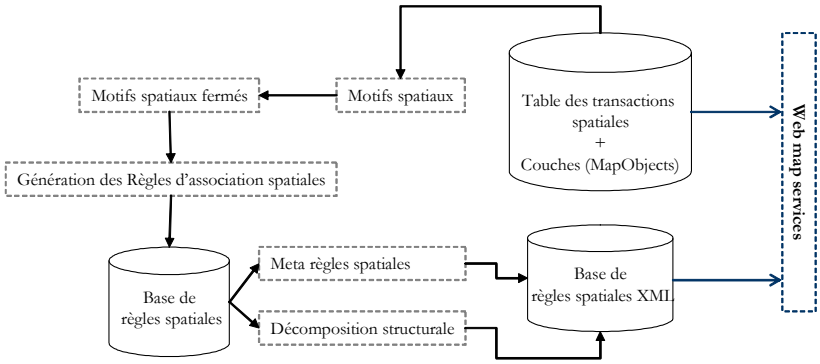


FIG. 9 – *Sous système de visualisation des règles.*

La solution offre plusieurs fonctionnalités. En effet, elle permet de manipuler et de visualiser des données spatiales. Elle génère aussi des fichiers de transactions spatiales grâce à diverses requêtes d'analyse. Elle procède également aux fouilles de données via des algorithmes ap-

propriés. Dans cette partie nous proposons quelques interfaces fournis par le prototype actuel.

FIG. 10 – *Interface de l'extracteur des transactions spatiales.*

1. *Pour créer un nouveau projet.*
2. *Pour ouvrir un projet.*
3. *Pour ajouter des couches (layers) à la carte depuis le projet déjà ouvert.*
4. *Pour fermer un projet ouvert.*
5. *Pour quitter l'application.*
6. *Pour Mettre le curseur manipulant la carte en mode normal.*
7. *Zoom avant.*
8. *Zoom arrière.*
9. *Pour faire bouger le contenu de la carte.*
10. *Pour ajuster la carte au cadre.*
11. *Pour visualiser une correspondance entre deux couches de la carte.*
12. *Pour extraire les fichiers de transactions depuis la carte.*
13. *Pour effectuer des fouilles de données sur la carte ou sur les fichiers de transactions.*
14. *La légende de la carte, elle permet de sélectionner les couches afin de les manipuler.*
15. *La carte est constituée de couches placées les unes sur les autres selon l'ordre perçu dans la légende.*

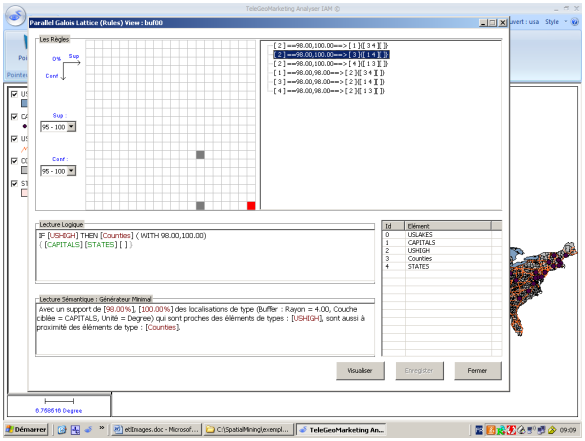


FIG. 11 – *Interface de visualisation des règles d'associations spatiales.*

Exemple de règles générées par l'algorithme de découverte:

IF [USHIGH] THEN [CAPITALS] (WITH 98.00,100.00) { [Counties] [STATES] [ ] }
IF [USHIGH] THEN [Counties] (WITH 98.00,100.00) { [CAPITALS] [STATES] [ ] }
IF [USHIGH] THEN [STATES] (WITH 98.00,100.00) { [CAPITALS] [Counties] [ ] }
IF [CAPITALS] THEN [USHIGH] (WITH 98.00,98.00) { [Counties] [STATES] [ ] }
IF [Counties] THEN [USHIGH] (WITH 98.00,98.00) { [CAPITALS] [STATES] [ ] }
IF [STATES] THEN [USHIGH] (WITH 98.00,98.00) { [CAPITALS] [Counties] [ ] }

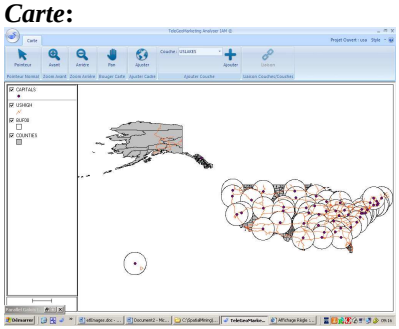
5.1 Détails sur les règles obtenues

Ci-dessous, nous donnons les éléments générés pour une règle d'association spatiale. Ces éléments sont explicités par des TDT XML. Les règles spatiales sont ensuite exportées vers un module de structuration et de visualisation *Webmapping*.

Description sémantique : Avec un support de [98.00%], [100.00%] des localisations de type Buffer qui sont proches des éléments de types : [USHIGH], sont aussi à proximité des éléments de types : [Counties].

Description logique : IF [USHIGH] THEN [Counties] WITH ([98.00%], [100.00%]) { [CAPITALS] [STATES] }{ }

***Carte*:**



Détails sur la méthode d'analyse utilisée	
Type	Buffer
Rayon	4.00
Couche ciblée	CAPITALS
Unité	Degree
Couches traitées	USLAKES CAPITALS USHIGH Counties STATES
Détails sur l'algorithme utilisé	
Type	Treillis de Galois parallèle (génération des règles)
Support	0.00
Confiance	0.00

TAB. 4 – *Méta données de la méthode d'analyse déployée.*

6 Conclusions et perspectives

Nous disposons actuellement d'un prototype avancé en phase de pré-industrialisation. Son déploiement chez l'opérateur de télécommunication *Maroc Telecom*, a montré la force du produit pour répondre à des interrogations en géomarketing. L'architecture proposée est validée ainsi que l'ensemble des composants élaborés. Les tests de sacalabilité du système ainsi que la performance des algorithmes sont très encourageants. La distribution de l'algorithme de construction de treillis de Galois nous a permis de dissocier ses tâches principales : la génération des agents et la gestion de l'arbre des concepts. La première est gourmande en capacité du processeur et la deuxième en mémoire. Ce ci 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. L'autre perspective de cette approche est de généraliser cette distribution sur plusieurs algorithmes afin de pouvoir combiner les Agents et les Managers les plus performants d'entre eux. Ceci générera des algorithmes hybrides mais sûrement plus robustes que les originaux.

Références

Agrawal R., Imielinski T. and Swami A. (1993) *Mining association rules between sets of items in large databases*, proceedings of ACM-SIGMOD International Conference Management of Data, pages 207-016, 1993

Barbut M. et Montjardet B. (1970) *Ordre et Classification.* Algèbre et Combinatoire. Hachette.

Berry A. et Sigayret A. (2004) *Discrete Applied Mathematics*, volume 144, Issue 1-2, Discrete Mathematics & Data mining (DM & DM), pp. 27-42.

Boulmakoul, A. Idri, R. Marghoubi (2007) *Closed frequent itemsets mining and structuring association rules based on Q-analysis,* The 7th IEEE International Symposium on Signal Processing and Information Technology, De-

cember 15-18, 2007, Cairo, Egypt. ISBN: 978-1-4244-1834-3, Digital Object Identifier: 10.1109/ISSPIT.2007.4458017, pp. 519-524.

Bordat J. P. (1986) *Calcul pratique du treillis de Galois d'une correspondance*, Math. Sci. Hum. 96 pp. 31-47.

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.

Estivill-Castro V. and Lee I. (2001) *Data Mining Techniques for Autonomous Exploration of Large Volumes of Geo-referenced Crime Data*, proceedings 6th International Conference on Geocomputaion, GeoComputation, CD-ROM, ISBN 1864995637, 2001

Ganter B. and Reuter K. (1991) *Finding all closed sets: a general approach.* Order, 8:283-290.

Ganter B. and Wille R. (1999) *Formal Concept Analysis*: Mathematical Foundations. Springer Verlag.

Idri A., Boulmakoul A. (2008) *Une approche parallèle distribuée pour la génération des motifs fermés fréquents basée sur une infrastructure corba*.197-210, in Les systèmes décisionnels : applications et perspectives (Noulmakoul et al . eds), ISBN 978-9981-1-3000-1, ASD 10-11 octobre 2008.

Koperski K. and Han J. (1995) *Discovery of Spatial Association Rules in Geographic Information Databases*, proceedings of 4th International Symposium on Large Spatial Databases, pages 47-66, 1995

Malerba D. (2008) *A relational perspective on spatial data mining*, IJDMMM 1(1): 103-118 (2008))

Njiwoua P. et Nguifo E. M.(1997) A *Parallel Algorithm to build Concept Lattice*, In proceedings of 4 Groningen Intl. Information Tech. Conf. for Students, pp. 103-107.

ORBACUS (2008) www.orbacus.com

Pasquier N., Bastide Y., Taouil R. et Lakhal L (1999) *Efficient mining of association rules using closed itemset lattices.* Information systems, 24 (1), pp. 25-46.

Rinzivillo S. and Turini F. (2007) *Knowledge discovery from spatial transactions*, Journal of Intelligent Information Systems, 28:1-22.

Shekhar S. and Chawla S. (2003) *Introduction to Spatial Data Mining, in Spatial Databases: A tour*, Prentice Hall, ISBN 013-017480-7, 2003

Stumme G. (1999), *Conceptual knowledge discovery with frequent concept lattices.* FB4-Preprint 2043, TU Darmstadt.

Zaki M. J. and Ogihara M. (1998) *Theoretical foundations of association rules.* Proc. 3rd ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pp. 1-7.

Summary

In this work we present the architecture and functionality of a spatial data mining system. The proposed system implements techniques for discovering spatial association rules based on the Galois closure. This implementation is parallel, distributed and deployed on a CORBA bus. The system architecture specification and the various software components are developed. The spatial transaction extractor module is described. It makes use of structural elements of neighborhood-type grid, buffer and Voronoi polygon.

Une solution web mapping pour la visualisation, la navigation, la structuration des règles d'association spatiales

Azedine Boulmakoul, Abdelfatah Idri, Mohamed Bendaoud
FST Mohammedia, Département informatique, B.P. 146 Mohammedia MAROC

azedine.boulmakoul@yahoo.fr

Résumé. Ce papier aborde la visualisation des règles d'association spatiales issues d'un processus de découverte de connaissances spatiales. La solution logicielle développée est fondée sur la technologie *web mapping*, et permet de déployer la visualisation pour plusieurs acteurs décideurs. La navigation est rendue flexible grâce à un mécanisme de structuration basée sur la décomposition d'une similarité floue. Le prototype actuel est validé auprès d'un opérateur de télécommunication pour une application en géomarketing.

1 Introduction

Ce travail porte sur la visualisation des règles d'association spatiales issues d'un processus de data mining spatial. L'un des problèmes récurrents est le nombre de règles produites encombrant le décideur du fait d'une saturation cognitive. L'infrastructure logicielle de cette visualisation sera portée par le web mapping. Cette solution visuelle permet la diffusion de l'information spatiale en ligne sur le web. Une règle d'association spatiale est une règle d'association classique, telles que la prémisse et la conclusion de la règle, pourront contenir des prédicats spatiaux. Les approches classiques de génération des règles associatives génèrent un nombre exorbitant de règles, rendant leur visualisation une tâche difficile. Il existe très peu de travaux sur l'aspect exploitation et visualisation de ces règles comparativement au nombre de travaux dédiés à l'extraction de ces règles. Les approches récentes proposées dans la littérature se segmentent en trois grandes familles: l'exploitation du treillis de Galois pour les *itemsets fermés fréquents* (Wille, 1982), (Zaki et al., 2003), (Boulmakoul et al., 2007), la hiérarchisation ou la décomposition (partition/recouvrement) des règles d'association extraites, et enfin la visualisation de bases génériques de règles associatives basée sur une méta-connaissance (Mephu et al., 2005). En plus de l'affichage à la demande explicite des règles associatives dérivables, est affichée une connaissance additionnelle matérialisant des connexions sémantiques entre ces règles. Ainsi, cette connaissance additionnelle permet d'améliorer l'interaction homme machine (Buono and Costabile, 2004), (Blanchard et al. 2003), (Buttenfield, 2003), (Han and Cercone, 2000). Le nombre élevé de règles associatives ayant une faible précision nuit à l'efficacité de l'utilisation de cette connaissance et ne fait qu'accroître la frustration de l'utilisateur. Ainsi, les connaissances extraites par les techniques de la fouille de données ont intérêt à présenter les connaissances extraites d'une manière graphique et interactive permettant à l'utilisateur d'inférer une connaissance avec une plus-value. Les techniques d'analyse de données visuelles ont besoin aussi d'être couplées avec la gestion des méta-connaissances utilisées pour gérer le nombre important de connaissances. Aussi, il est

important que cette méta-connaissance soit exprimée d'une manière adéquate, sans la dissoudre avec des représentations internes de connaissances. Le travail de Zaki et al (2003), introduisant l'outil MIRAGE, qui utilise la base générique des règles associatives. Cet outil présente une approche permettant d'extraire et d'explorer visuellement les règles minimales. MIRAGE utilise une approche de visualisation de règles interactive basée sur le treillis de Galois, permettant d'afficher les règles dans une forme très compacte.

Dans ce travail, nous proposons une approche structurale de décomposition pour la visualisation des règles d'association spatiales. Cette approche fait référence à un algorithme de partitionnement des relations floues. Nous définissons le concept de règle structurante pour faciliter la navigation dans l'espace des règles. Une règle structurante est un élément attracteur qui constitue le générateur d'une classe de règles. Les aspects liés à l'extraction des règles d'association spatiales ne sont pas considérés dans ce papier. La section deux aborde la structure des règles d'association spatiales. La troisième section détaille la méthode de décomposition proposée. La quatrième section décrit la solution webmapping pour la visualisation des règles d'association spatiales. Elle incorpore le prototype réalisé. La conclusion résume notre contribution et aborde quelques investigations projetées pour la valorisation industrielle de ce travail.

2 Règles d'association spatiales

Le Data Mining Spatial (DMS) est né du besoin d'exploitation dans un but décisionnel de données à référence spatiale produites, importées ou accumulées, susceptibles de délivrer des informations ou des connaissances par le moyen d'outils d'exploration, d'analyse et de fouille de données. Il constitue un domaine à part entière car il considère les couplages des objets dans l'espace. Ce domaine intègre des techniques provenant à la fois des bases de données spatiales et des SIG (Zeitouni, 2000), (Ester et al., 1998) du Data Mining et des statistiques spatiales (Chelghoum et al., 2002). La formalisation du problème d'extraction de règles associatives a été introduite par Agrawal et al. (1993). La découverte des règles d'association consiste à déterminer l'ensemble des règles dont le support et la confiance sont au moins égaux, respectivement, à un seuil minimal de support *minsup* et à un seuil minimal de confiance *minconf* prédéfinis par l'utilisateur. La littérature a largement abordé ce problème.

Dans ce travail, nous n'abordons pas les technqiues d'extraction des règles d'association spatiales. Ci-dessous nous présentons le modèle de donnée qualifiant une règle d'association spatiale (figure 1). Une règle d'association spatiale est caracterisée par un ensemble de couches thématiques spatiales qui apparaissent en prémisse ou en conclusion de la dite règle (Boulmakoul et al., 2002), (Marghoubi et al., 2006). Les couches thématiques sont données par des shapefiles. A une règle sont attachés des documents de divers format (XML, PDF, IMAGE, etc.). A une règle est associé un motif fermé fréquent qui a permis la génération de la règle. La spécialisation d'une règle d'association spatiale donne naissance à d'autre type : règle structurante qui correspond à une règle d'attraction d'une classe de règles, une méta-règle qui est une règle extraite à partir d'un ensemble de règles (règle sur les règles), etc.

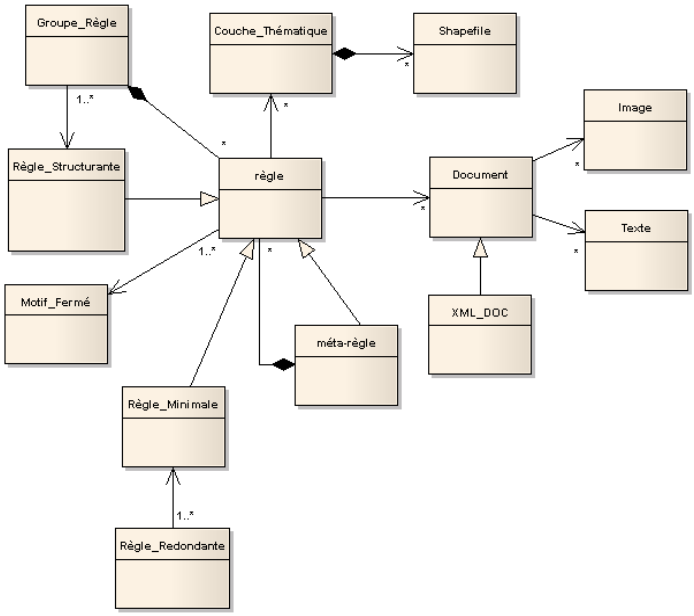


FIG. 1 — *Diagramme de classes du domaine «visualisation des règles d'association spatiales»*

Commençons par donner une illustration pour les règles d'association spatiales (Idri and Boulmakoul, 2008). Dans l'exemple schématisé dans la figure 2, le contexte d'extraction correspond à des données spatiales des Etat Unis, fournies par ESRI et qui concernent : l'infrastructure autoroutière, les états, les countries, les capitales, et les lacs. L'extraction des règles d'association spatiales est faite à partir de la couche de référence « capital ». La figure 3, propose l'histogramme des règles extraites. Une sélection possible de ces règles est donnée ci-dessous :

IF [STATES] THEN [USHIGH] (WITH 98.0, 98.0)

Une interprétation sémantique de cette règle d'association spatiale de support égal à 0,98 et de confiance égale à 0,98, pouvant être associée à ce type de connaissance est la suivante : la probabilité de trouver les items "*states*", "*ushigh*" ensemble est égale à 0,98. Ceci peut exprimer la connaissance spatiale « tous les états sont traversés par des autoroutes »

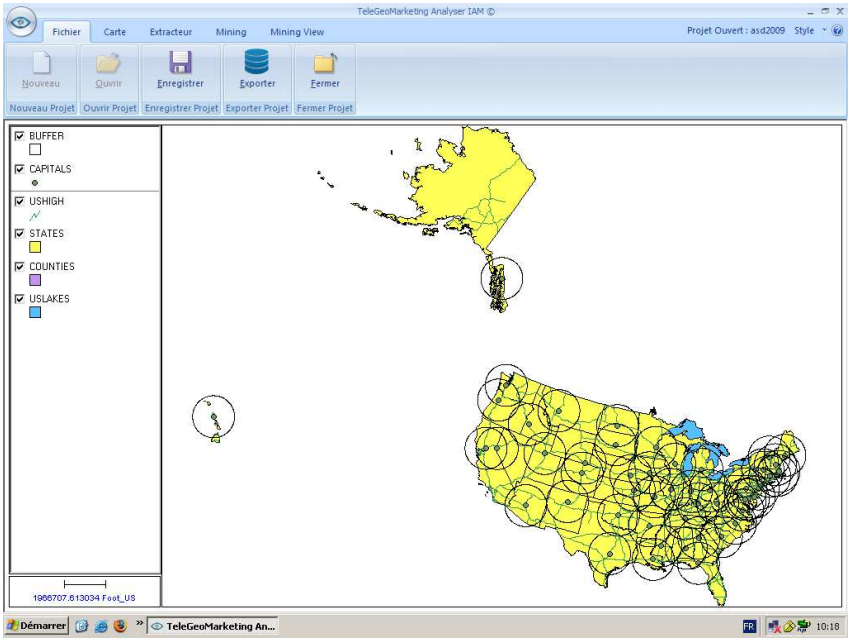


FIG. 2 — *Contexte spatial d'extraction des règles d'association spatiales.*

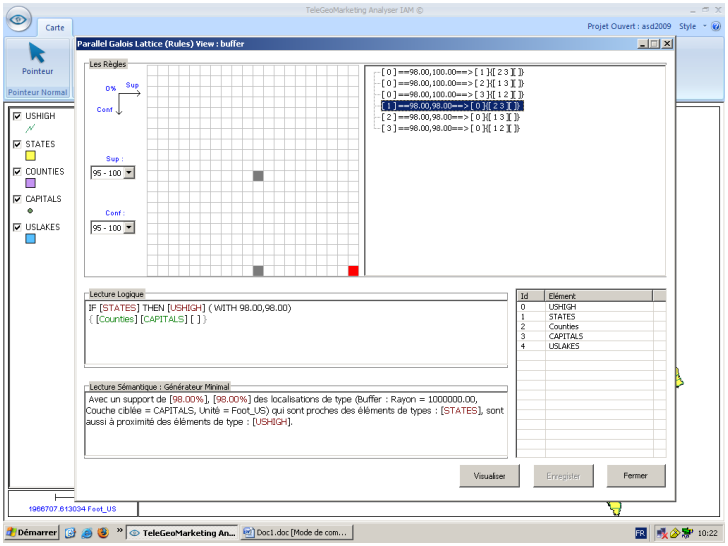


FIG. 3 — *Histogramme des règles d'association spatiales extraites.*

La figure 4, schématise les couches thématiques associées à la règle extraite. La table 1, propose sa description XML.

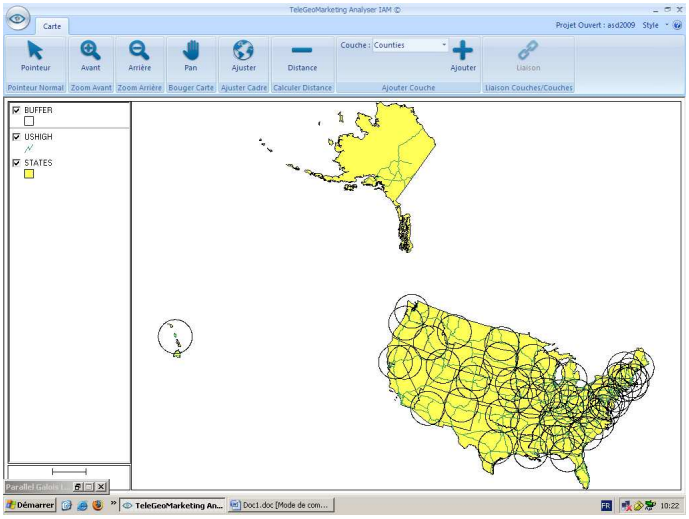


FIG. 4 — *Représentation spatiale de la règle d'association spatiale extraite.*

TAB. 1 — *Description XML d'une règle d'association spatiale*

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<?xml version="1.0" encoding="ISO-8859-1"?>
<?xml-stylesheet href ="buffer2009-06-12-102254\buffer2009-06-12-102254.xsl"
type="text/xsl"?><REGLE>
<METHODE NOM="Buffer">
<PARAMETRE NOM="Rayon" VALEUR="1000000.00"/>
<PARAMETRE NOM="Couche ciblée" VALEUR="CAPITALS"/>
<PARAMETRE NOM="Unité" VALEUR="Foot_US"/>
<CIBLE NOM="USHIGH"/>
<CIBLE NOM="STATES"/>
<CIBLE NOM="Counties"/>
<CIBLE NOM="CAPITALS"/>
<CIBLE NOM="USLAKES"/>
</METHODE>
<ALGORITHME NOM="Parallel Galois Lattice (Rules)">
<PARAMETRE NOM="Support" VALEUR="0.00"/>
<PARAMETRE NOM="Confiance" VALEUR="0.00"/>
</ALGORITHME>
<SUPPORT VALEUR="98.00"/>
<CONFIANCE VALEUR="98.00"/>
<PREMISSE>
<ITEM NOM="STATES"/>
</PREMISSE>
<CONSEQUENCE>
<ITEM NOM="USHIGH"/>
</CONSEQUENCE>
```

```
<PREMISSE_PLUS>
<ITEM NOM="Counties"/>
<ITEM NOM="CAPITALS"/>
</PREMISSE_PLUS>
<CONSEQUENCE_PLUS>
</CONSEQUENCE_PLUS>
<CARTE CHEMIN="buffer2009-06-12-102254/map.bmp"/>
<LEGENDE>
<ITEM CHEMIN="buffer2009-06-12-102254/legend0.bmp"/>
<ITEM CHEMIN="buffer2009-06-12-102254/legend1.bmp"/>
<ITEM CHEMIN="buffer2009-06-12-102254/legend2.bmp"/>
</LEGENDE>
</REGLE>
```

3 Décomposition hiérarchique des règles d'association spatiales

Des méthodes ont été proposées pour grouper et synthétiser de grands ensembles de règles d'association selon les items contenus dans chaque règle. Les techniques de classification hiérarchique sont utilisées pour partitionner les règles initiales en sous-ensembles thématiques cohérents (Gupta et al., 1999). Ceci permet de résumer un ensemble de règles, en choisissant la proportion de règles représentatives pour chaque sous-ensemble, et aide aussi dans l'exploration interactive des règles par l'utilisateur. La figure suivante, donne le processus permettant de mettre en œuvre une telle approche.

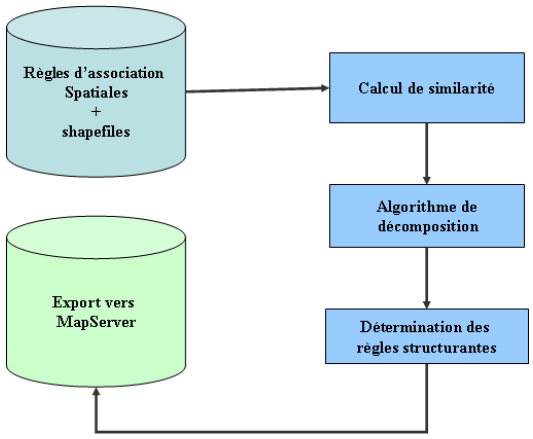


FIG. 5 — *Schéma de regroupement et thématisation des règles d'association.*

Dans un contexte plus général, nous proposons dans la suite les développements récents dans le domaine de l'analyse hiérarchique. L'algorithme général décrit dans le paragraphe suivant offre une possibilité de choix de similarité pour mieux apprécier le couplage entre règles d'association. Les méthodes de classification fondées sur les relations de similarité floues sont simples et s'appliquent facilement aux diverses applications (Backer, 1978).

Zadeh (Goguen, 1973) a proposé la transitivité max-min pour les relations de similarité floues. Cette opération associée à la décomposition convexe (analyse de similitude) d'une relation réflexive, symétrique et transitive max-min permet de déterminer une hiérarchie de partitions. Ce résultat a été exploité sur des relations réflexives et symétriques par l'application de la fermeture transitive max-min et de la décomposition convexe de la relation obtenue. Boulmakoul (2002) a proposé l'extension des procédures dues à Tamura à d'autres type de fermeture transitive (max-Δ, max-prod, etc.). La composition max-Δ reste fidèle et n'altère pas sévèrement la relation d'origine, et par conséquent, elle permet d'obtenir de meilleurs résultats (la composition max-Δ est plus « soft » que la composition max-min). La décomposition de la fermeture transitive max-Δ d'une relation de similarité floue ne permet pas de générer une hiérarchie de partitions : les relations binaires obtenues après décomposition ne correspondent pas à des relations d'équivalence. Pour tirer profit de la composition max-Δ, propose un algorithme qui permet d'extraire des partitions pour les relations de similarité transitives max-Δ. Cet algorithme reste valable pour les relations de similarité transitives max-min. Dans cet article nous proposons un algorithme général simple qui permet de générer des partitions pour toute relation floue réflexive et symétrique. En particulier pour celles qui sont transitives max-min ou transitives max-Δ. Cet algorithme généralise celui proposé par Yang et al. (2001), d'une part pour la simplicité de son implémentation, d'autre part, il s'inscrit dans les fondements des méthodes de classification prétopologiques (Belmandt, 1993) (Tremolières, 1979).

La classification «clustering » est un processus de groupement d'objets en classes. Plusieurs techniques ont été développées, elles se distinguent selon la typologie suivante : les méthodes de partitionnement, les méthodes de hiérarchisation, les méthodes de densité et les méthodes de grilles. Dans ce contexte, nous proposons un modèle d'analyse de classification fondé sur les graphes flous. Ces graphes sont construits à partir de relations floues.

Dans cette approche, les propriétés et les opérations des relations floues de similarité sont sollicitées (l'analyse de similitude hiérarchique, et la composition des relations). Plusieurs cas sont à envisager :

- si le graphe ne possède pas de structure connue, des opérateurs de composition appropriés seront appliqués (composition max-min, max-Δ, etc.),
- pour les autres graphes flous, selon leurs propriétés des caractérisations sont à définir.

Sur les graphes flous transformés, il est possible d'appliquer l'algorithme général, pour extraire des arbres de partitions.

3.1 Algorithme de décomposition

Dans cette partie nous proposons un algorithme général simple qui permet de générer des partitions pour toute relation floue réflexive et symétrique. En particulier pour celles qui sont transitives max-min ou transitives max-Δ.

3.1.1 Relations de similarité floues

Une relation de similarité floue est une généralisation de la notion de relation d'équivalence dans le cadre classique. Soit Ω un ensemble, une relation floue R dans Ω est une relation de similarité si elle vérifie les propriétés suivantes :

- (réflexivité) $\mu_R(x, x)=1 \ \forall x \in \Omega$;
- (symétrie) $\mu_R(x,y)=\mu_R(y,x)= \forall (x,y) \in \Omega^2$;
- (transitivité max-T) $\mu_R(x,z) \geq \max_{y\in\Omega}(T(\mu_R(x,y),\mu_R(y,z)))$

$\forall (x,z) \in \Omega^2$; où T est une T-norme.

- $T(\mu_R(x,y),\mu_R(y,z)) = \min(\mu_R(x,y),\mu_R(y,z))$ R est dite transitive max-min,
- $T(\mu_R(x,y),\mu_R(y,z)) = \max(0,\mu_R(x,y)+\mu_R(y,z)-1)$ R est dite transitive max-Δ,
- $T(\mu_R(x,y),\mu_R(y,z)) = \mu_R(x,y)\bullet\mu_R(y,z)$, R es dite transitive max-prod.

Si R est une relation floue, sa décomposition convexe est donnée par $R = \max_{\alpha} (\alpha R^{\alpha})$, où R^{α} est la coupe α de la relation R, $\alpha \in [0,1]$. Si R est une relation de similarité transitive max-min alors R^{α} est une relation d'équivalence.

3.1.2 Algorithme général de recherche des partitions

Données :

- Ω : ensemble des règles à structurer.
- μ_R : l'indicatrice d'une relation floue réflexive et symétrique définie sur Ω.
- Θ liste des classes obtenues.
- $\alpha \in [0,1]$
- $\sigma(e,\wp) = \sum_{x\in\wp} \mu_R(e,x)$ fonction de similarité objet ensemble.
- $\delta(e,\wp) = \prod_{x\in\wp} \mu_R(e,x)$
- $\chi(e) = \{\wp \in \Theta,\ \delta(x,\wp) \neq 0\}$
- $\hat{R} = \min(R, R^{\alpha})$, avec R^{α} est la coupe de niveau α de la relation R.

Recherche_partitions$(\mu_R, \Omega, \Theta, \alpha)$

{- Calculer $\hat{R} = \min(R, R^{\alpha})$

- Calculer la fonction φ,

$\varphi(\omega) = \max_{x\in\Omega}(\mu_{\hat{R}}(\omega, x))$, $\forall \omega \in \Omega$

- Trier les éléments de Ω *dans une pile T, selon les valeurs décroissantes de* φ.

- $\Theta \leftarrow \varnothing$

Tant que $T \neq \varnothing$

{ e←T.tête

si $\Theta == \varnothing$

alors

Création(e,T,Θ*)*

Sinon

{// calculer la similarité objet–ensemble

- $\forall \wp \in \Theta$, *calculer* $\sigma(e,\wp)$, $\delta(e,\wp)$ *; calculer* $\chi(e)$

si $\chi(e) \neq \varnothing$ *Alors Attraction(e,*Θ*)*

Sinon Création(e,T,Θ*)*

}

La procédure création permet de créer une nouvelle classe et de supprimer de la pile T tous les objets classés.

Création*(e,T,Θ)*
{
- créer une nouvelle classe C ;
- $\Gamma(e) = \left\{ y \in \Omega , y \neq e / \mu_R(y,e) = \max_{x \in \Omega}(\mu_R(e,x)) \right\}$;
y←Γ(e).tête*
C←{e,y }
Θ←{C}
T.supprimer(e)
T.supprimer(y)*
}

La procédure attraction permet d'affecter les objets aux classes existantes « les plus similaires ». Elle supprime de la pile T tous les objets affectés aux classes.

Attraction *(e,Θ)*
{- Soit la classe C telle que* $\sigma(e, C^*) = \max_{\wp \in \chi(e)} \sigma(e, \wp)$;

$C^* \leftarrow C^* \cup \{e\}$
T.supprimer(e)
}

3.1.3 Concept de similarité pour les règles d'association spatiales

Nous adoptons la similarité transitive max-min suivante :

$$S(R_i, R_j) = \frac{\{items\ R_i\} \cap \{items\ R_j\}}{\{items\ R_i\} \cup \{items\ R_j\}} \times \frac{\min(\psi(R_i), \psi(R_j))}{\max(\psi(R_i), \psi(R_j))}$$

Où ψ correspond à la confiance de la règle. La similarité entre deux règles est proportionnelle au nombre d'occurrence d'items dans les deux règles.

3.1.4 Règles structurantes

Nous appelons règle structurante toute «génératrice» des classes. Elles représentent les règles exprimant le plus fort couplage avec les autres règles.

Chaque règle structurante représente un ensemble de règles; ceci permet une meilleure navigation dans l'espace des règles. La visualisation sera guidée par le seuil de perception α. Pour chaque seuil seront présentés : la règle structurante, son support et sa confiance, le cardinal de la classe de règles associée, les thèmes relatifs à la règle structurante, son fichier XML à visualiser, et le taux de couverture de la règle. Après la sélection d'une règle représentative, une liste de règles triées selon la valeur de couplage φ sera affichée (figures 6-8).

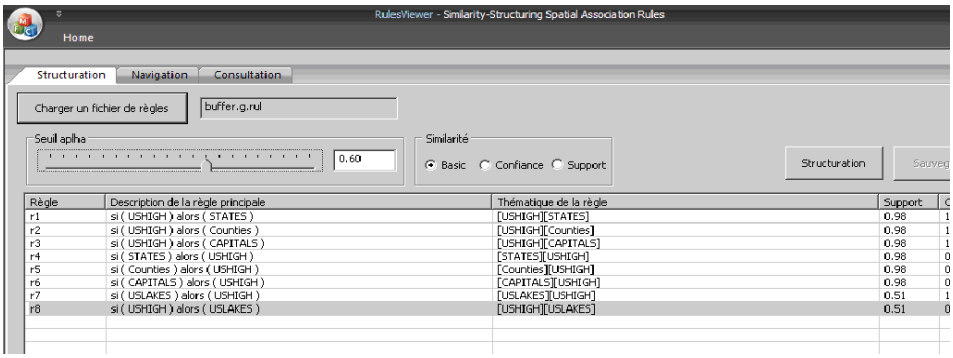


FIG. 6 — *Règles d'association à structurer.*

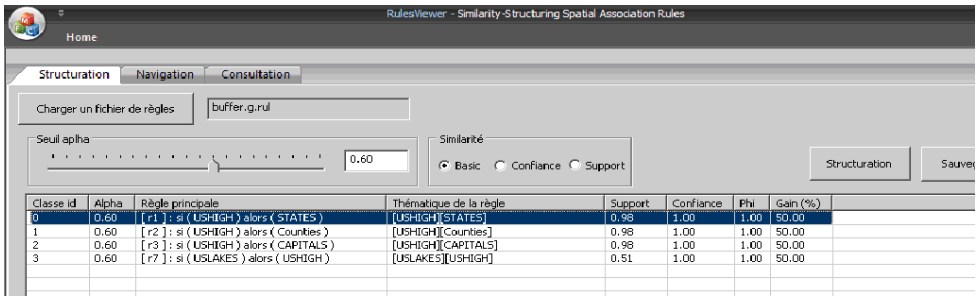


FIG. 7 — *Règles d'association structurées en 4 classes (seuil = 0.6).*

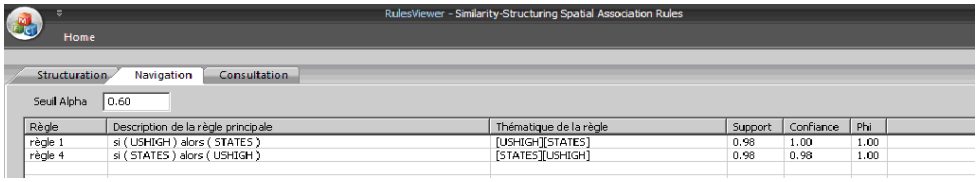


FIG. 8 — *navigation dans les règles d'association attachées à la classe 0.*

4 Rule web viewer basé sur le Web mapping

Le terme *web mapping* (Kooistra et al., 2009), (Mitchell, 2005) désigne la diffusion de cartes dynamiques ou statiques ainsi que des données attributaires pouvant être associées sur un réseau (intranet/extranet/internet). Il s'agit d'un domaine en pleine expansion grâce au développement de solutions open source arrivées à maturité. Les informations cartographiques brutes ou les données géoréférencées sont ainsi consultables à partir de postes clients. Elles sont en général stockées dans un système de gestion de base de données (SGBD) sur un ou plusieurs serveurs et administrables de façon centralisée. Les SIG «en ligne » se distinguent donc des SIG bureautiques classiques, nécessitant une installation

logicielle sur chaque poste nécessaire (ou au minimum un viewer) ainsi que parfois, une copie des données si celles-ci ne sont pas accessibles par le réseau local. Evoluant rapidement, le web mapping est souvent présenté comme étant l'avenir des SIG. La plupart des solutions ne nécessitent pas d'installation lourde côté client. L'échange d'information se fait par le navigateur internet via les requêtes HTTP envoyées par le poste client et les pages HTML que le serveur lui envoie en réponse. La mise en place d'une application spatiale en ligne est facilitée, par l'existence de solutions libres éprouvées. Les solutions basées sur les technologies Open Source permettent justement de faire l'économie d'une offre éditeur tout en permettant d'envisager des développements ultérieurs afin de répondre aux besoins spécifiques et constatés des utilisateurs. Le caractère open source d'une application implique non seulement la libre diffusion en totalité ou en partie de son code source mais aussi, sa libre redistribution et la possibilité de développer librement des applications dérivées ou des fonctionnalités complémentaires. L'OGC (2008) (Open Geospatial Consortium - anciennement OpenGIS Consortium) est l'organisation à but non lucratif consacrée au développement des solutions libres en géomatique et à l'interopérabilité des systèmes (OSGeo, 2008).

4.1 Technologie et distribution choisies

Le choix de la technologie *web mapping* s'est porté sur le serveur spatial, trois possibilités sont offertes : MapGuide d'Autodesk (2008), GeoServer (2008) et MapServer (2008). C'est MapServer qui a été retenu compte tenu de sa grande adaptabilité qui permet le développement d'applications personnalisées, de la richesse de la documentation disponible et du dynamisme de sa communauté de développeurs. MapServer n'est ni compilé ni associé avec aucun client web prédéfini. A ce serveur cartographique s'ajoute *p.mapper* (Configurable Web Mapping Client Component - CWC2) (Maptools, 2008).

Enfin, le serveur cartographique devait être couplé à un SGBD relationnel supportant les données géographiques. Nous avons ôté pour la version 4.1 de MySQL (2008) qui supporte les données géométriques et géoréférencées. L'architecture de l'application est modulaire et nécessite l'intégration et le paramétrage de l'ensemble des composantes logicielles. Les architectures disponibles du web mapping sont données sommairement dans les figures 9-12.

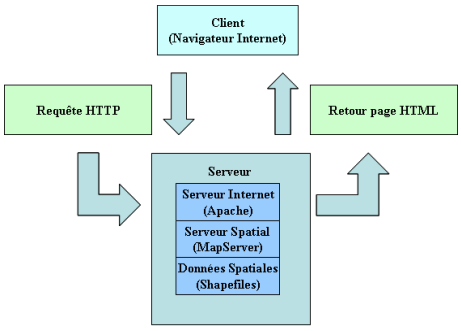


FIG. 9 — *Architecture générale d'une solution webmapping.*

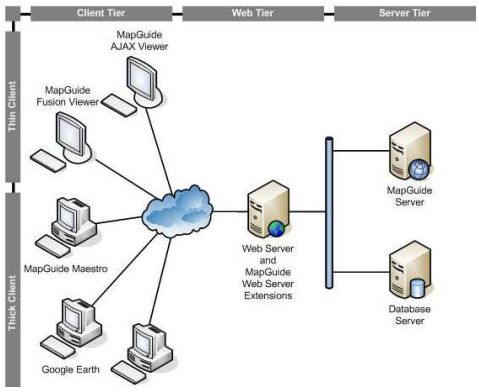


FIG. 10 — *MapGuide 3-tier architecture*

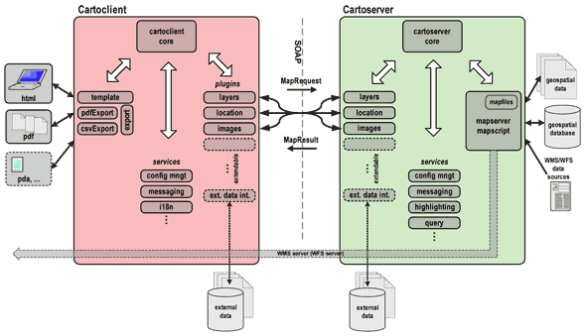


FIG. 11 — *Web-Service Architecture – SOAP, CartoWeb (2008).*

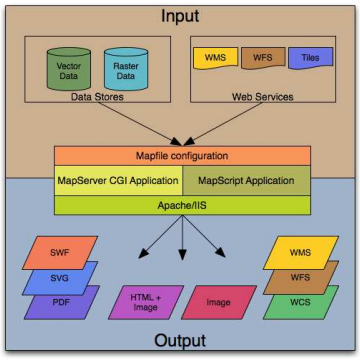


FIG. 12 — *http:MapServer architecture*

4.2 Intégration des composants

L'environnement de travail étant installé et fonctionnel, il s'agissait à partir de ce moment de préparer et intégrer les données. Deux bases de données ont été créées, l'une contenant les règles d'associations spatiales et les couches (shapefiles), l'autre destinée à l'administration du serveur.

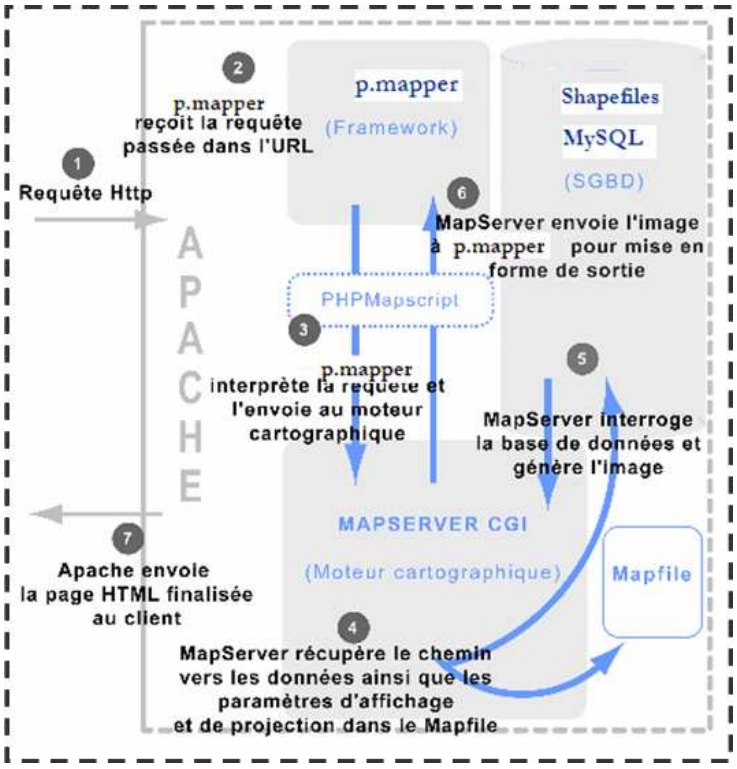


FIG. 13 — Intégration des composants Web mapping.

Le framework *p.mapper* permet des configurations multiples afin de faciliter l'installation d'une application de MapServer basée sur PHP/MapScript et offre une panoplie de fonctionnalités : DHTML (DOM) zoom/pan, Zoom/pan, requêtes (identify, select, search), etc. La figure 13, trace les étapes d'exécution d'une requête de visualisation d'une règle. Les interfaces obtenues pour la visualisation des règles d'association spatiales sont données ci-après (figures 14-15).

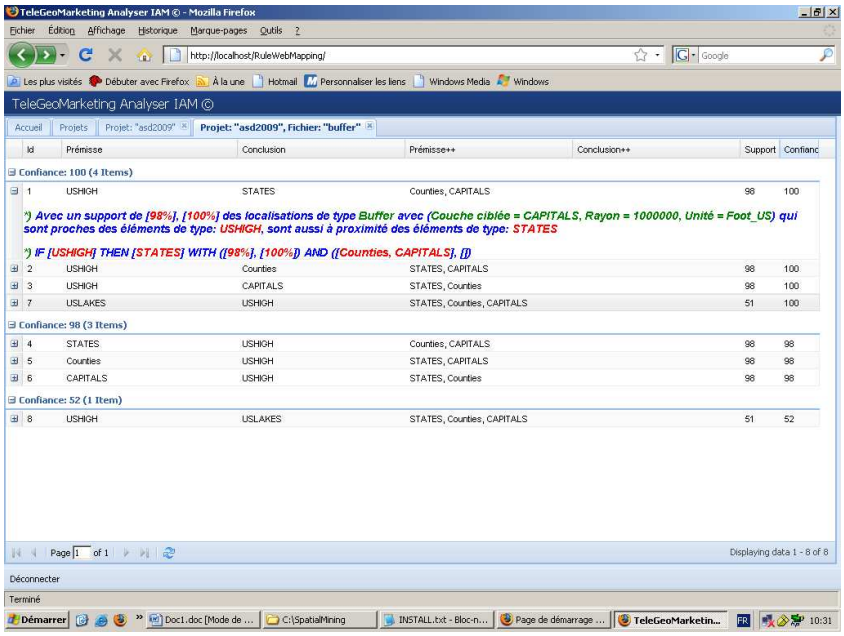


FIG. 14 — *Interface Webmapping du prototype.*

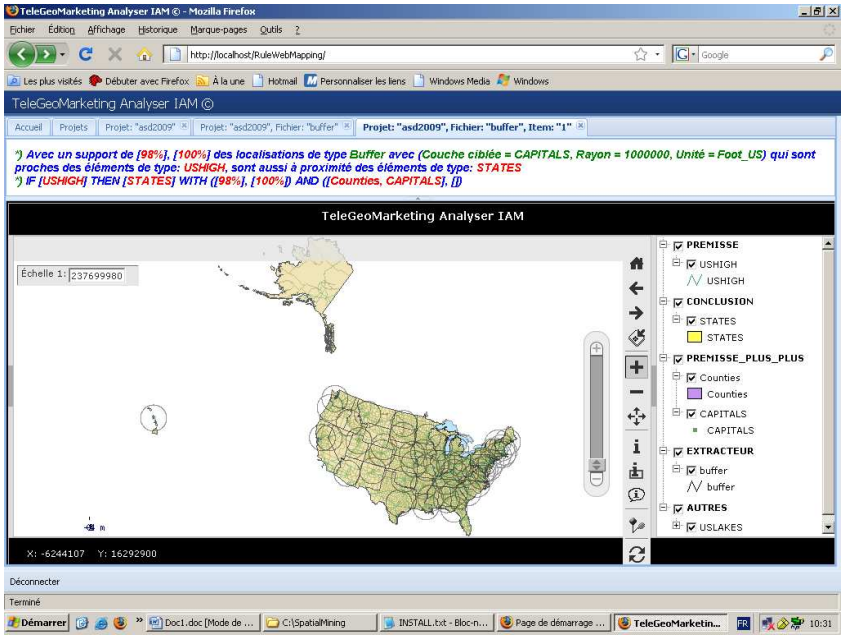


FIG. 15 — *Interface Web mapping du prototype (suite).*

5 Conclusion

Les résultats issus de cette recherche ont été testés et s'avèrent concluants. Nous disposons actuellement d'un prototype *TeleGeoMarketing Analyser*, avec une capacité d'export des règles d'association spatiales vers le web. Les composants de structuration et de visualisation augmentent la flexibilité et la maniabilité pour naviguer dans l'espace des règles. L'architecture proposée est validée ainsi que l'ensemble des composants élaborés. Pour répondre à la qualité de l'extraction des connaissances spatiales, l'analyse de la pertinence des règles d'association spatiales est une tâche nécessaire. De nombreuses mesures de qualité sont disponibles dans la littérature (Lenca et al., 2004), (Tan et al., 2004). Le concept de fonction structurante avec certains critères définis par le décideur sera prochainement étudié pour répondre à cette problématique.

Références

Agrawal R., Imielinski T. and Swami A. (1993) *Mining association rules between sets of items in large databases*. In Proceedings of the ACM SIGMOD Intl. Conference on Management of Data, Washington, USA, pages 207–216, June 1993

Backer E. (1978) *Cluster analysis by optimal decomposition of induced fuzzy sets*, PhD Thesis, Delftse Universitaire Pers, 1978.

Belmandt Z. (1993), Manuel de prétopologie et ses applications (coll. Interdisciplinarité et nouveaux outils), Lavoisier, 1993.

Blanchard J., Guillet F., Briand H. (2003) *A User-driven and Quality-oriented Visualization for Mining Association Rules*, Proceedings of Third IEEE International Conference on Data Mining, Melbourne, Florida, November 2003, p. 493.496.

Boulmakoul A., Karine Zeitouni, Nadjim Chelghoum, Rabia Marghoubi (2002) *Fuzzy structural primitives for spatial data mining*, 2nd IEEE International Symposium on Signal Processing and Information Technology, Marrakech, Morocco, pp. 266-27, December 18-21, 2002.

Boulmakoul, A. Idri, R. Marghoubi (2007) *Closed frequent itemsets mining and structuring association rules based on Q-analysis,* The 7th IEEE International Symposium on Signal Processing and Information Technology, December 15-18, 2007, Cairo, Egypt. ISBN: 978-1-4244-1834-3, Digital Object Identifier: 10.1109/ISSPIT.2007.4458017, pp. 519-524.

Buono, P., Costabile, M. F. (2004) *Visualizing Association Rules in a Framework for Visual Data Mining*, From Integrated Publication and Informations Systems to Virtual Information and Knowledge Environments, LNCS, Springer-verlag, february 2004, p. 221.231.

Buttenfield B. (2003) *Representing information for knowledge discovery: pattern detection and database structure*, Proceedings of the UCGIS workshop on Knowledge Discovery and Visualization, Boulder, Colorado, USA, 2003.

CartoWeb (2008), *http://www.cartoweb.org/*

Chelghoum N., K. Zeitouni, A. Boulmakoul (2002) *A decision Tree for multi-layered spatial data, in Advances in Spatial Data Handling*, pp. 1-10, (D. Richardson & P. Van Oosterom ed.) Springer-Verlag 2002, ISBN 3-540-43802-5.

Goguen, J. A. *Review: L. A. Zadeh, Fuzzy Sets; L. A. Zadeh, Similarity Relations and Fuzzy Orderings,* J. Symbolic Logic Volume 38, Issue 4 (1973), 656-657.

Ester M., Frommelt A., Kriegel H-P., and Sander J.(1998) *Algorithms for Characterisation and Trends Detection in Spatial Databases*, Proc. 4th Int. Conf. On Knowledge Discovery and Data Mining, New York City, 1998, pp. 44-50.

GeoServer (2008), *http://geoserver.org/*

Gupta G. and Strehl A. and Ghosh J. (1999) *Distance Based Clustering of Association Rules,In Intelligent Engineering Systems Through Artificial Neural Networks*, 1999, pp. 759--764,ASME Press.

Han J., Cercone N. (2000) *Ruleviz : A model for visualizing knowledge discovery process*, Proceedings of 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,Boston, MA, USA, August 2000, p. 244.253.

Idri A., Boulmakoul A. (2008) *Une approche parallèle distribuée pour la génération des motifs fermés fréquents basée sur une infrastructure corba*.197-210, in Les systèmes décisionnels : applications et perspectives (Boulmakoul et al . eds), ISBN 978-9981-1-3000-1, ASD 10-11 octobre 2008.

Kooistra, L., Bergsma, A.; Chuma, B.; de Bruin, S. (2009) *Development of a Dynamic Web Mapping Service for Vegetation Productivity Using Earth Observation and in situ Sensors in a Sensor Web Based Approach*. Sensors 2009, 9, 2371-2388.

Lenca P., Meyer P., Picouet P., Vaillant B. et Lallich S. (2004) *Evaluation et analyse multi-critères des mesures de qualité des règles d'association*, Mesures de Qualité pour la Fouille de Données, RNTI-E-1, pp 219-246, 2004

MapGuide OS (2008), *http://mapguide.osgeo.org/*

Maptools (2008), *http://www.maptools.org/*

Marghoubi A. Boulmakoul, K. Zeitouni (2006) *Utilisation des treillis de Galois pour l'extraction et la visualisation des règles d'association spatiale,* INFORSID 2006, VOL 2., pp. 703-718., 01-03 juin 2006 Hammat Tunisie, ISBN 2-906855-22-7

Mephu N. E., Njiwoua P. (2005) *Treillis de concepts et classification Supervisée*, Technique et Science Informatiques, RSTI, Hermès - Lavoisier, Paris, France, 2005

Mitchell Tyler (2005) *Web Mapping Illustrated*. Sebastopol (USA) : O'Reilly Media, 2005 .

MySQL(2008) *http://www-fr.mysql.com/*

OGC (2008) *http://www.opengeospatial.org/*

OSGeo (2008) *http://www.osgeo.org/*

Tan P.N., Kumar V. et Srivastava J. (2004), *Selecting the right objective measure for association analysis*, Information Systems 29(4), pp 293-313, 2004.

Tremolières R. (1979) *The percolation method for an efficient grouping of data*, Pattern recognition, vol. 11, n°4, 1979.

UMN MapServer (2008) *http://mapserver.gis.umn.edu/*

Wille R. (1982) *Restructuring lattice theory: an approach based on hierarchies of concepts*. Ordered Sets (I. Rival, ed.), pp. 445–470, Reidel, Dordrecht-boston, 1982.

Yang M. S., Shih H. M. (2001) *Cluster analysis based on fuzzy relations*, Fuzzy Sets and Systems, Volume 120, Number 2, 1 June 2001 , pp. 197-212(16)

Zaki M., and Phoophakdee B. (2003) *MIRAGE: A framework for mining, exploring and visualizing minimal association rules*. RPI CS Dept Technical Report 03-04, July 2003.

Zeitouni K. (2000) *Data mining spatial*, Revue internationale de géomatique Vol.9 N° 4/1999.

Summary

This paper approaches the visualization of a set of spatial association rules resulting from a spatial knowledge discovery process. The developed software solution is founded on webmapping technology, and makes it possible to deploy visualization for several actors and decision makers. Navigation is made flexible thanks to a structuring mechanism based on fuzzy similarity decomposition. The current prototype is validated with a telecommunication operator in GeoMarketing area.

Capitalisation des connaissances pour le diagnostic industriel : approche hybride Datamining-RàPC

Noureddine Mekroud[1], Abdelouahab Moussaoui[2]

Département de l'informatique, Faculté des Sciences de l'Ingénieur, Université Ferhat Abbas, Sétif
{Mekroud_n, Moussaoui_abdel}@yahoo.fr

Résumé. Le Raisonnement à Partir de Cas (RàPC) est un réflexe puissant et très naturel, qui vise la réutilisation des expériences passées dans la résolution des nouveaux problèmes ; ceci est confirmé par des expériences en psychologie et en sciences cognitives. Le RàPC, comme méthodologie d'ingénierie des connaissances, peut être renforcé dans les différentes étapes de son processus par la richesse des techniques du Datamining. Dans cet article, on propose une solution hybride, RàPC et Datamining, appliquée au domaine du diagnostic industriel. Le processus proposé commence par une fragmentation de la base des cas en deux espaces : Symptomes-Pannes & Symptomes-Solutions ; suivie d'un clustering des deux espaces, et d'un mappage entre leurs fragments; on appliquera enfin un cycle RàPC pour chaque espace. La plateforme JCOLIBRI 2.1 sera utilisée pour l'implémentation de notre démarche. Les avantages de cette approche seront présentés, ainsi que des réflexes et perspectives diverses.

1 Introduction

L'application du Retour d'Expérience est fréquente dans la vie quotidienne de chacun, puisqu'il est bien évident et légitime que les problèmes similaires auront des solutions similaires, et qu'on se retrouve souvent face à un problème que l'on a déjà rencontré.

De nos jours, le savoir faire est le patrimoine principal des individus et établissements ; la valorisation des entreprises repose de plus en plus sur des facteurs immatériels. Selon une étude sur des centaines de sociétés industrielles américaines, l'estimation de la valeur de leur patrimoine de connaissances (Knowledge Capital) correspond à 217% de leur capital financier net (Ermine, 2005) . La préservation et la capitalisation du savoir, du savoir faire et des meilleures pratiques existantes dans une entreprise, nécessite la mise en place d'un système permettant de fournir à une personne, pas nécessairement hautement qualifiée, l'information utile au moment où elle en a besoin, de façon exploitable pour superviser les décisions à prendre.

Les connaissances d'une entreprise, considérées comme un patrimoine fragile, circulent et s'enrichissent et sont exploitées plus ou moins avec fiabilité, mais peuvent également disparaître par un départ en retraite, une mutation ou un licenciement d'un expert. Cette richesse d'expertise, qui doit être capturée, capitalisée, protégée et distribuée, forme la partie principale d'une mémoire d'entreprise, qui rapproche le niveau de connaissances individuel au niveau d'expérience collective de l'organisme.

Le RàPC, comme une méthodologie de capitalisation des connaissances, propose des solutions aux problèmes actuels à résoudre, en utilisant les connaissances acquises des expériences passées, et en enrichissant en continu la base de connaissances. Cette méthodologie, qui forme un point de rencontre entre l'intelligence artificielle et les sciences

cognitives ; repose sur des notions beaucoup plus théoriques que techniques, ce qui lui donne l'aspect d'une méthodologie et pas une technologie. Le Datamining, utilisé dans la découverte et la modélisation des informations utiles, cachées dans une masse de données grande et complexe, offre des solutions techniques incontournables dans la découverte et la capitalisation des connaissances de l'entreprise.

L'objectif de ce travail est d'étudier l'utilisation des techniques du Datamining, qui émergent actuellement, dans le cycle du processus RàPC, en éclairant l'appui qu'elles peuvent fournir pour améliorer la fiabilité de cette méthodologie. Visant le domaine du diagnostic industriel, notre approche est basée sur la mise en cause de la liaison directe et stable entre les pannes de leurs solutions possibles. On propose dans cet article la fragmentation verticale de la base des connaissances du domaine étudié en deux partitions : l'espaces des symptômes et leurs pannes correspondantes, et l'espace des symptômes et leurs solutions possibles ; suivi d'un clustering du contenu de chaque espace, et d'un mappage entre les fragments des pannes et de leurs solutions possibles ; et on finira par l'application du cycle RàPC sur les deux partitions conjointement. L'approche proposée offre non seulement la possibilité de réutilisation des expériences passées en limitant la recherche aux segments les plus pertinents, mais aussi de traiter le cas où les mêmes symptômes donneront plusieurs pannes possibles, ce qui indique l'insuffisance de ces symptômes pour diagnostiquer les pannes. Aussi, si les mêmes symptômes auront plusieurs solutions possibles, on pourra choisir entre ces solutions suivant la stratégie de maintenance, corrective ou préventive. D'autres avantages et arguments pertinents seront fournis en ce qui suit.

Dans la section 2, on propose une présentation générale de la gestion des connaissances dans les entreprises. La section 3 étudie la méthodologie du raisonnement à partir de cas. Dans la section 4, on présente notre idée de fragmentation et mappage de la base de cas. La section 5 expliquera notre démarche de prétraitement de données. Dans la section 6, on donnera les étapes de notre processus, suivis de leurs avantages et inconvénients. Dans la section 8, la plateforme JCOLIBI 2.1 sera utilisée dans l'implémentation et la validation de la solution proposée. À la fin, une conclusion et des perspectives seront proposées.

2 La gestion des connaissances dans l'entreprise

L'un des aspects de développement de l'efficacité d'une entreprise est la capitalisation de ses expériences développées au cours du temps, vu le risque de centralisation du savoir faire dans les experts humains, qui peut causer la non disponibilité de cette richesse suite à une sur-occupation de l'expert, mutation, départ à la retraite, démission, licenciement … (Dieng-Kuntz et al., 2001).

Les connaissances sont une représentation réduite du monde réel (Duizabo et Guillaume, 1997). Dans une entreprise, deux catégories de connaissances sont à distinguer, à savoir : les connaissances *Tacites* (implicites, non formalisables) : qui sont difficiles à décrire, comme les compétences, les habilités, la connaissance historique de l'organisation ...etc, c'est le savoir faire de l'entreprise ; et les connaissances *Explicites* (formalisables) : qui sont plus facilement codifiables (manuels, plans, modèles, documents d'analyse, données, ...), c'est le savoir de l'entreprise. Par exemple, avant l'invention du système de notation (solfège) au XII^{éme} siècle, on ne pouvait pas représenter les connaissances musicales ; elles étaient apprises uniquement par l'expérience directe, l'écoute (tacites) ; mais grâce au solfège, elles sont devenues codifiables, donc explicites (Cortes Robles, 2006) .

Dans la littérature de la gestion moderne des entreprises, de nouvelles notions sont évoquées, comme : les travailleurs de la connaissance, la société et l'économie de la

connaissance, la mémoire d'entreprise, les entreprises apprenantes, la gestion des richesses immatérielles, la gestion de l'innovation...etc., où la richesse de l'entreprise est désormais basée sur ses activités intellectuelles, donnant naissance à une économie basée sur la connaissance (Duizabo et Guillaume, 1997). La gestion des connaissances, ou Knowledge Management en anglais (baptisée KM), désigne l'ensemble de concepts et outils permettant la production des connaissances et le développement des compétences individuelles, collectives et organisationnelles. Le KM vise à rassembler le savoir et le savoir faire sur des supports accessibles, et faciliter leur transmission en temps réel à l'intérieur de l'établissement, ou les différer à nos successeurs (Rasovska, 2006). C'est une application pratique des sciences cognitives, de l'intelligence artificielle et des sciences de l'organisation (Rakoto, 2004).

Il existe une diversité de méthodologies de capitalisation des connaissances d'une entreprise, on peut distinguer celles spécifiques à la construction des mémoires d'entreprises, à savoir : REX, MEREX, CYGMA, atelier FX et Componential Framework ... etc ; et d'autres empruntées du domaine de l'ingénierie des connaissances, comme : KADS, CommonKADS, MKSM, MASK, et KOD (Rasovska, 2006).

3 Le Raisonnement à Partir de Cas

Le RàPC est un processus qui vise la réutilisation des expériences passées. Cette méthodologie, provenant du domaine de l'Intelligence Artificielle, a été utilisée dans les systèmes experts et les sciences cognitives. Dans cette approche, l'utilisateur essaie de résoudre un nouveau problème en reconnaissant les similarités avec des problèmes préalablement résolus, appelés : cas. Un cas est communément un problème spécifique qui a été identifié, résolu, stocké et indexé dans une mémoire avec sa solution, et éventuellement le processus d'obtention de celle-ci (Cortes Robles, 2006). Les systèmes de RàPC sont appliqués dans de nombreux domaines comme : la médecine, le commerce, le diagnostic industriel, le contrôle et l'analyse financière (Rasovska, 2006).

3.1 Le cycle du RàPC

Le cycle du RàPC se décompose habituellement en cinq phases (Rasovska, 2006) :

1. L'élaboration d'un nouveau problème (cas cible) : représente l'acquisition et la modélisation des informations connues sur le nouveau problème, pour lui donner une description initiale, d'une manière similaire aux cas existants dans la base des cas.
2. La remémoration des cas (cas sources) : rechercher les cas les plus similaires, cela signifie la recherche des correspondances entre les descripteurs des cas de la base et ceux du cas actuel à résoudre.
3. L'adaptation des cas (cas source) : réutiliser totalement ou partiellement la solution du cas trouvé le plus similaire, pour résoudre le nouveau problème.
4. La révision de la solution proposée (solution cible) : signifie l'évaluation de la solution proposée, par son application dans le monde réel.
5. La mémorisation d'un nouveau cas (cas cible) : représente l'ajout éventuel du cas cible dans la base des cas, c'est l'apprentissage d'un nouveau cas, qui pourra ainsi être utilisé pour la résolution des problèmes futurs.

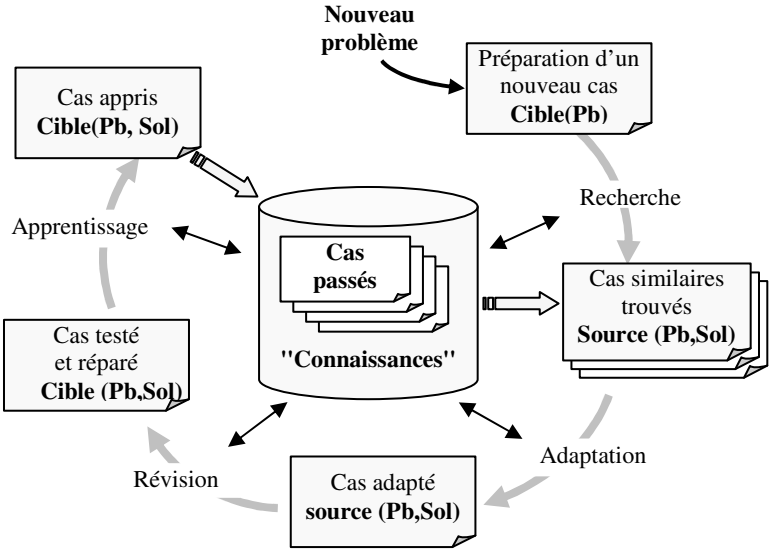


Fig 1 : Les étapes du cycle de RàPC (Rakoto, 2004)

3.2 Avantages et inconvénients du RàPC

Puisque cette approche n'a besoin que d'un ensemble de problèmes résolus pour commencer, le RàPC diffère des autres approches de l'IA, qui visent la représentation globale des connaissances utilisées dans un domaine, sous forme de système expert.

Ainsi, l'approche RàPC a l'avantage d'être une démarche plus simple à mettre en oeuvre que celles basées sur un modèle de domaine, puisque elle permet d'éviter les difficultés de modélisation du savoir-faire des experts (complexité des ontologies et des représentions logique…). Aussi, c'est un bon choix pour les domaines n'exigeant pas de solutions optimales, ou dont les principes sont mal formalisés ou peu éprouvés (Rasovska, 2006). Le RàPC vise l'utilisation des connaissances spécifiques et pragmatiques, qui concernent les problèmes précédemment expérimentés, en les capitalisant d'une façon progressive avec le temps ; l'apprentissage sera ainsi incrémental et basé sur les expériences vécues (Cortes Robles, 2006).

Mais, par contre, le RàPC ne trouve pas nécessairement la solution concrète à un problème ; et parfois, juste proposer un ensemble de solutions possibles (Devèze et Fouquin, 2004). Aussi, vu la nécessité d'une intervention et d'une mobilisation en continu des experts, lors de la capitalisation progressive des connaissances, l'individualisme constitue un frein redoutable ; les experts hésitent de partager leurs connaissances, acquises après des années de travail, par méfiance d'une restructuration ou compression dans l'entreprise, ou ils estiment leurs savoir-faire comme leur plus grande assurance (Dieng-Kuntz et al., 2001).

Un aspect, qu'on juge très intéressant, est que le RàPC est une méthodologie et pas une technologie de résolution des problèmes, et la majorité de ses définitions se basent sur la présentation du « Quoi » et pas le « Comment ». Ainsi, présenter le RàPC comme une méthodologie est important pour son développement. Si le RàPC est considéré comme une technologie, il pourrait sembler que les recherches dans ce domaine sont en grande partie achevées. Mais, considéré comme une méthodologie, les chercheurs auront toujours le défi de perfectionner le cycle de ce processus (Watson, 1999).

4 Notre approche : Hybridation RàPC-Datamining

Le RàPC est basé sur deux hypothèses concernant la nature du monde réel (Abasolo, 2004) : la *régularité* des situations, ou problèmes, similaires qui doit impliquer des solutions similaires ; et la *récurrence* qui assume que c'est fortement probable que les situations futures seront des variantes des situations actuelles. Ainsi, les approches standard du RàPC proposent un mappage entre le problème cible (actuel) et le problème source (déjà résolu), basé sur la similarité ; et la solution du cas similaire trouvé sera, suivant le raisonnement par analogie, la solution proposée au problème cible actuel. Mais, les contraintes internes et l'environnement extérieur changent au fil du temps. Ainsi, la liaison entre les descripteurs des cas et leurs solutions peut ne pas être stable, ce qui donne la possibilité d'avoir des expériences erronées (Leake et Wilson, 1999). Avec des systèmes RàPC utilisés à longs termes, on pourra être face à des situations dans lesquelles la réutilisation des expériences deviendra une erreur. Démarrant de ces hypothèses de base, notre approche propose un cycle de RàPC fondé sur des similarités appliquées dans deux espaces différents : espace des pannes et espace des solutions (Fig 2).

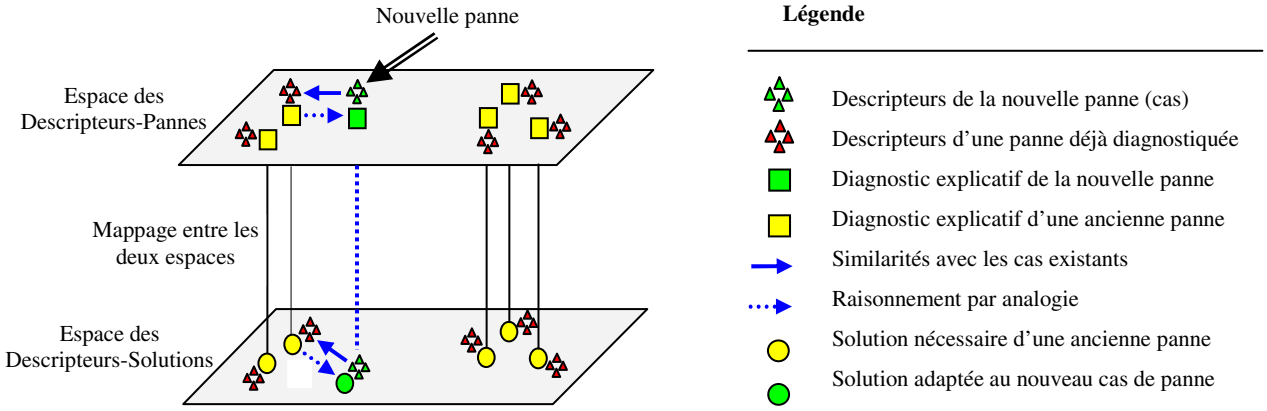


Fig 2 : Tracé de résolution des problèmes proposé par notre approche

Notre démarche commence par la fragmentation verticale de la base des cas en deux espaces : Descripteurs-Pannes & Descripteurs-Solutions (Fig 3), suivie d'un clustering de chacun des deux espaces, et d'un mappage (via des mesures de similarité globales) entre les clusters (fragments) de ces deux espaces. Ensuite, et lors d'une nouvelle panne, on prépare les descripteurs de ce nouveau cas à résoudre ; et on applique dans l'espace des Descripteurs-Pannes, dans le but de trouver un diagnostic adéquat à cette panne, un cycle RàPC sur le cluster le plus similaire au cas actuel, en utilisant des mesures de similarité locales adéquates à chaque descripteur de panne, ainsi que des similarités basées sur les ontologies du domaine. Après, on mappe vers le cluster pertinent dans l'espace des solutions, où on recherche, suivant un autre cycle RàPC, la solution adéquate au cas de panne actuel à résoudre.

Descripteurs						Panne	Solution	
Texte erreur	Température moteur	Catégorie fonctionnelle	Alarme	Pression moteur	Indicateurs lumineux	Explication de la panne	Composant défaillant	Solution nécessaire
					ID1 … ID6			

Fragmentation de la base des cas

Descripteurs						Panne
Texte erreur	Température moteur	Catégorie fonctionnelle	Alarme	Pression moteur	Indicateurs lumineux	Explication de la panne
					ID1 … ID6	

Descripteurs						Solution	
Texte erreur	Température moteur	Catégorie fonctionnelle	Alarme	Pression moteur	Indicateurs lumineux	Composant défaillant	Solution nécessaire
					ID1 … ID6		

Fig 3 : Les attributs proposés de la base des cas, réparties en deux espaces

5 Le Datamining et le prétraitement des données

Notre base de cas est composée de variables quantitatives et d'autres qualitatives. Mais, seules les attributs à valeurs discrètes peuvent être utilisés dans la construction des clusters par l'application de la SHA (Segmentation Hiérarchique Ascendante) et la génération des K-Means (segmentation par les K centres mobiles). Cela nous a obligé à travailler en deux étapes : (1) Réaliser une Analyse en Composantes Principales (ACP) pour réduisant du nombre de variables quantitatives à traiter ; suivie d'une Analyse des Correspondances Multiples (ACM), pour la discrétisation des variables qualitatives ; (2) Lancer les K-Means et les SHA sur les x premiers axes factoriels issus des analyses par ACP et ACM.

Il existe des indicateurs lumineux (ID1, …, ID6) qui peuvent prendre des valeurs soit rouge (cas de problème) ou vert (bon fonctionnement). Ces indicateurs sont sémantiquement proches ; et vu leur nombre, on a opté pour leur réduction par une ACP, en les projetant sur des axes factoriels résumant leur contenu. Ainsi, nous ne retenons que l'information essentielle, nous évacuons celles assimilables à du bruit. Suivant les résultats illustrés dans la figure suivante, nous avons constaté que les trois premiers axes factoriels restituent 87,15% de l'information disponible.

La même ACP sera appliquée sur l'espace des Descripteurs-Pannes et des Descripteurs-Solutions, puisque les champs à réduire (les indicateurs lumineux) sont les mêmes pour ces deux espaces. Cette analyse sera suivie par une ACM pour les attributs qualitatifs.

Axis	Eigen value	% explained	Histogram	% cumulated
1	2,536281	42,27%		42,27%
2	1,747969	29,13%		71,40%
3	0,944998	15,75%		87,15%
4	0,526897	8,78%		95,94%
5	0,203969	3,40%		99,34%
6	0,039886	0,66%		100,00%
Tot.	6,000000	-	-	-

Fig 4 : Les résultats d'une analyse ACP

5.1 Choix du nombre de clusters par l'algorithme SHA

Nous avons réalisé une Segmentation Ascendante Hiérarchique basée sur la méthode du saut minimum, en utilisant les axes factoriels générés de l'ACP et l'ACM. La figure ci-dessous affiche les résultats de construction des groupements SHA sur l'espace des Descripteurs-Solutions de notre base des cas. Tandis que l'a SHA appliquée sur l'espace des Descripteurs-Pannes a généré une segmentation en quatre clusters ; dans l'espace des solutions, la SHA a généré une segmentation optimale en six clusters. Le dendrogramme correspondant confirme que le partitionnement en 6 segments est judicieux. Les données d'entrée pour la SHA de l'espace des Descripteurs-Solutions sont les mêmes pour l'espace Descripteurs-Pannes, mais en ajoutant l'attribut Composant qui spécifie l'élément de la machine concerné par la panne déjà diagnostiquée.

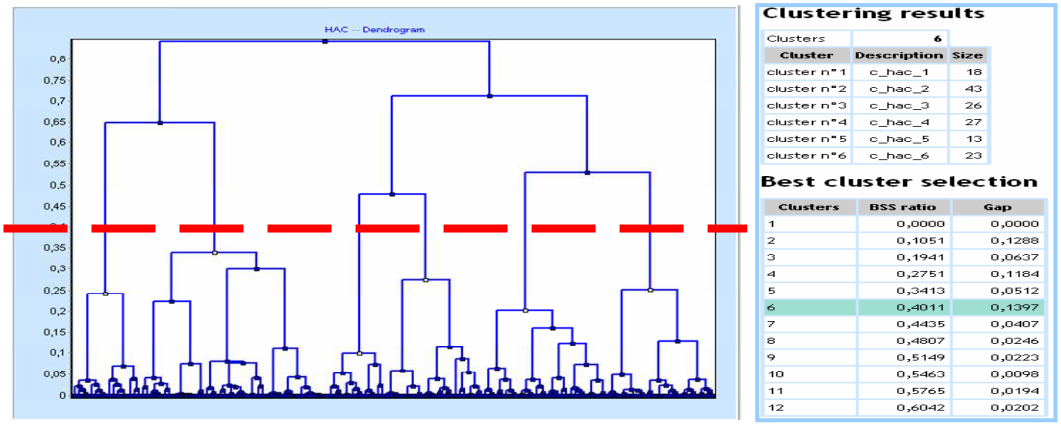


Fig 5 : Les résultats de la SHA dans l'espace Descripteurs-Solutions

5.2 Le K-Means sur les axes factoriels et les données continues

Suivant les résultats de la SHA sur les deux espaces, nous effectuons une segmentation par les K-Means sur quatre clusters pour l'espace des Descripteurs-Pannes, et six clusters pour l'espace des Descripteurs-Solutions. Le calcul des centres lors des itérations est basé sur l'algorithme de McQueen. Aucune normalisation n'est effectuée c.-à-d. la distance euclidienne simple est utilisée. Le graphique des nuages de points est un outil de visualisation très intéressant pour comprendre et bien interpréter les groupes. La figure ci-dessous présente les clusters générés, dans chaque espace, ainsi que le mappage entre ces derniers selon des mesures de similarité globales. En revenant sur les clusters représentés graphiquement, nous constatons que les groupes sont assez discernables. Nous avons réalisé presque 20 essais d'optimisation, avec un nombre d'itération maximum à 40.

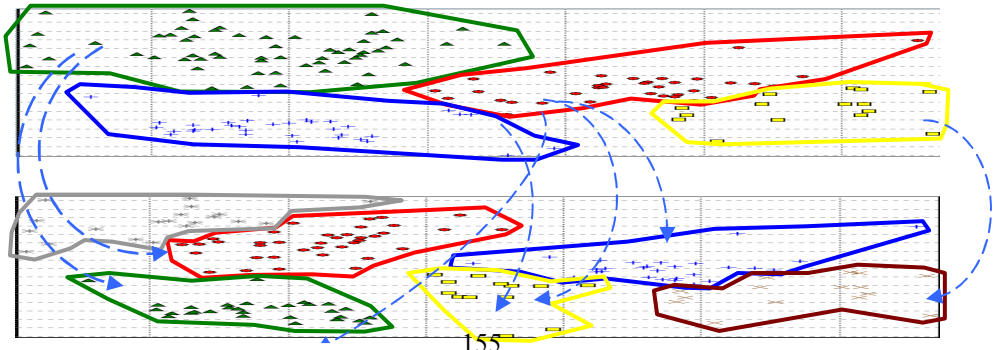


Fig 6 : Mappage entre les clusters générés dans les deux espaces (Descripteurs-Pannes & Descripteurs-Solutions)

6 Les étapes du processus proposé

La figure suivante illustre la succession des étapes de notre approche :

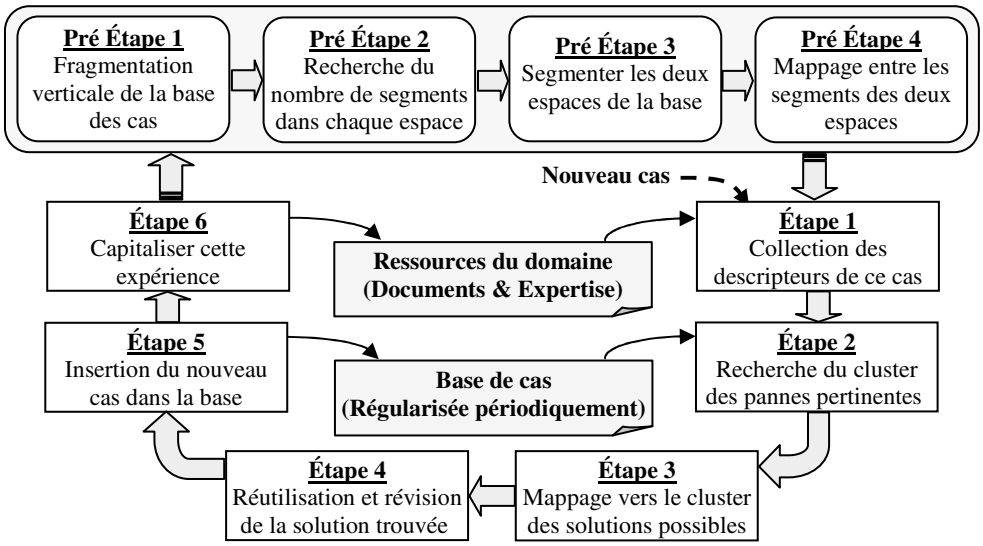


Fig 7 : Processus d'intégration des techniques du Datamining dans le cycle RàPC

Pré Étape 1 : Fragmenter la base des cas verticalement, en Descripteurs-Pannes et Descripteurs-Solutions, et isoler les causes des pannes de leurs solutions possibles.

Pré Étape 2 : prétraitement des données de la base, par les analyses ACP et ACM, suivi d'une SHA pour fixer le nombre optimal de clusters dans chaque espace.

Pré Étape 3 : Segmenter les deux partitions de la base pas un algorithme de K-Means. Des significations et utilités seront fournies lors d'une valeur différente du nombre de clusters dans les deux espaces.

Pré Étape 4 : Faire un tracé (mappage) entre les clusters Symptomes-Pannes et Symptomes-Solutions, suivant des mesures de similarités globales, pour lier les pannes à leurs solutions les plus pertinentes.

Étape 1 : Collection des descripteurs du nouveau cas de panne.

Étape 2 : Recherche du cluster des pannes pertinentes, suivant les Descripteurs-Pannes ayant la plus grande similarité avec les caractéristiques du nouveau cas. La recherche est basée sur l'algorithme des K Plus Proches Voisins (KPPV).

Étape 3 : Faire un mappage vers le (les) cluster(s) Descripteurs-Solutions correspondant(s) au cluster des pannes trouvé dans l'étape précédente.

Étape 4 : Réutiliser la solution, ayant les descripteurs les plus adéquats au cas actuel, dans le contexte du nouveau cas. Réviser et commenter les résultats trouvés.

Étape 5 : Maintenance de la base par l'insertion du nouveau cas résolu.

Étape 6 : Capitaliser cette expérience par la conservation des commentaires et des évaluations fournis, suivie d'une nouvelle segmentation des deux espaces de la base.

7 Avantages et inconvénients de notre approche

1. La mise en cause de la liaison directe entre les pannes et leurs solutions possibles, par l'isolation des Descripteurs-Pannes et Descripteurs-Solutions, aidera à détecter les cas d'insuffisance des descripteurs ; puisque si les mêmes descripteurs donneront plusieurs pannes possibles et/ou plusieurs solutions possibles, alors ces descripteurs ne sont pas suffisants et discriminants.
2. Si on considère le mappage entre les segments des deux espaces comme une relation par cardinalités dans les deux sens, cela est interprétable comme suite :
   - ☞ Les cardinalités dans le sens Discripteurs-Pannes vers Discripteurs-Solutions décrivent le nombre de solutions possibles pour les mêmes symptômes de panne, et cela nous donnera les interventions possibles sur cette panne dans les différentes stratégies de maintenance (préventive, corrective …).
   - ☞ Les cardinalités dans le sens Discripteurs-Solutions vers Discripteurs-Pannes donneront le nombre de pannes qu'on peut résoudre par la même intervention, et cela nous aidera à planifier les interventions par priorité, selon le nombre de problèmes possibles à résoudre par chaque intervention.
   - ☞ Il est possible d'avoir une cardinalité nulle entre les segments des deux espaces ; c'est le cas d'un ensemble de symptômes qui prédisent une panne avant son arrivé. Ainsi, on aura à faire à une maintenance préventive.
3. Les algorithmes des K-Means et KPPV ont une fiabilité reconnue, et aussi une complexité faible (polynomiale). Cela facilitera la refragmentation de la base par les K-Means, qui est nécessaire après chaque modification de son contenu.
4. L'utilisation de divers mesures de similarité, suivant le type et les valeurs de chaque champs, ainsi que la similarité basée sur les ontologies du domaine, renforcera l'efficacité de la recherche des cas similaires au problème à résoudre.

Mais, comme toute travail, notre approche présente quelques inconvénients :

1. Si la panne actuelle est malle classifiée, lors de la recherche du segment des Descripteurs-Pannes le plus pertinent, alors on aura un mappage vers un cluster de solutions non adéquates au problème à résoudre.
2. cette approche nécessite une bonne connaissance du domaine étudié, interprétée par une base de cas suffisamment grande, pour assurer la performance des résultats issus de l'application des algorithmes de segmentation et de mappage.

8 Implémentation

8.1 Plateforme d'expérimentation

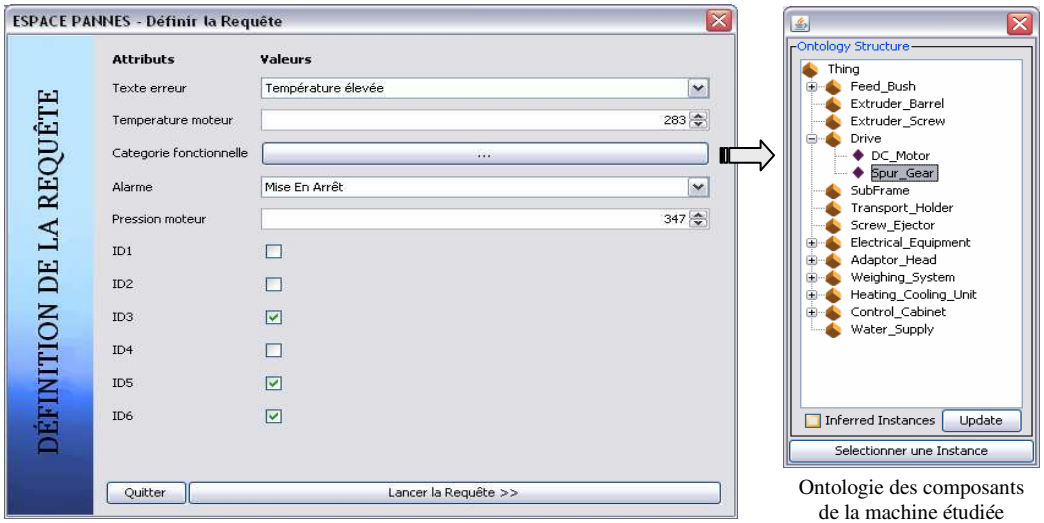
Pour étudier la faisabilité de notre démarche, système de diagnostic industriel est fondé sur plusieurs solutions logicielles, composée de trois axes conjoints : (1) L'éditeur d'ontologies Protégé 2000 pour la création de l'ontologie du domaine d'étude, et la représentation de ces concepts et taxonomies, dans une hiérarchie exprimant les relations entre ces éléments (Rasovska, 2006); (2) Le logiciel de Datamining Tanagra 1.4.29 pour réaliser, visualiser et analyser les algorithmes du Datamining nécessaires à l'implémentation

de notre approche ; (3) La personnalisation de la plateforme RàPC JCOLIBRI 2.1, une solution modulaire Open Source écrite en Java.

Comme équipement étudié, on a choisi une machine extrudeuse des tubes de plastique (*Extrudeuse Monovis BEX 1-90-30B* du fabricant allemand *BATTENFELD*). Un soin particulier, lors du développement de notre système, a été apporté à l'élaboration des mesures de similarité pour chaque attribut de la base, selon le type et la modalité des descripteurs de pannes. Les similarités basées sur les ontologies, appliquées sur l'attribut représentant les composants concernés par la panne, sont les plus importantes ; puisqu'elles sont les plus pertinentes et vont assurer une meilleure exactitude des résultats de recherche des cas similaires. Le cycle RàPC sera exécuté dans deux itérations, l'une sur l'espace des Descripteurs-Pannes, et l'autre sur l'espace des Descripteurs-Solutions.

Notre recherche par les K plus proches voisins est basée sur des mesures de similarité diverses : similarité basée sur les ontologies pour le champs Composant ; similarité basée sur la distance simple ou cyclique entre deux valeurs d'un attribut quantitatif ; égalité directe (1 si les deux valeurs sont identiques, 0 sinon) ; similarité par intervalle entre deux valeurs lui appartenant ; similarité basée sur les plus grandes sous-chaîne égaux, ou sur une table de similarité entre deux attributs qualitatives, et enfin une similarité basée sur un seuil (1 si la similarité ≥ seuil, 0 sinon). Ces mesures de similarité aideront à optimiser la pertinence des résultats de recherche.

La figure suivante, projetée de notre implémentation, illustre la première phase du cycle de RàPC (élaboration d'un nouveau cas), en utilisant les descripteurs de panne et l'ontologie des composants de la machine à étudier, réalisée par l'éditeur Protégé, pour préparer une requête de recherche des cas similaires les plus proches.



Ontologie des composants de la machine étudiée

Fig 8 : Préparer une requête recherchant les cas similaires au problème à résoudre

Après, le cycle RàPC recherche dans l'espace des Descripteurs-Pannes les pannes les plus similaires aux descripteurs de panne actuels. Et après avoir adapter et réviser le cas trouvé le plus pertinent, le nouveau cas résolu sera ajouté à la base des cas, bien sur si il n'existé par auparavant (ie : similarité moins de 100% par rapport au cas trouvé). À la fin, le processus continua par un deuxième cycle RàPC sur l'espace des Descripteurs-Solutions. Un nouveau clustering de la base des cas peut être nécessaire.

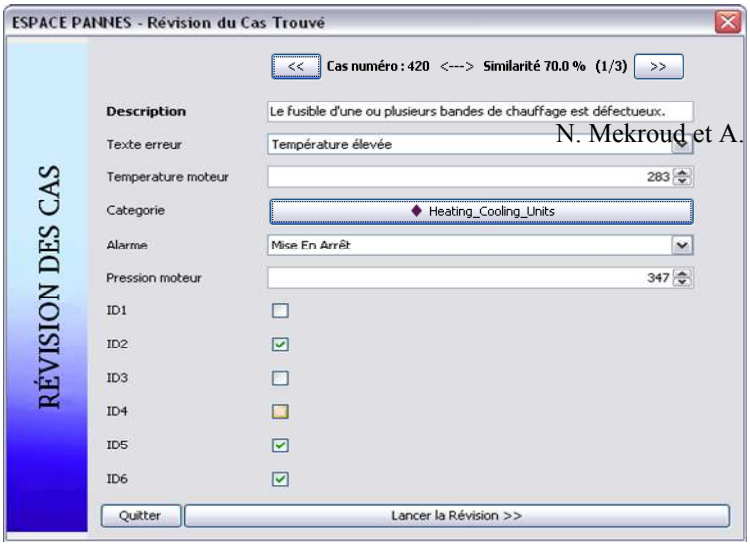


Fig 9 : Révision et adaptation du cas similaire trouvé

8.2 Analyse des performances du cycle RàPC implémenté

Le système de gestion de connaissances développé a proposé des solutions de plus en plus pertinentes et utiles aux pannes survenant au fil du temps ; ce qui prouve un apprentissage progressif de notre solution. La figure suivante montre l'évolution de la moyenne de similarité pour les 3 plus proches voisins trouvés pour les cas de panne à résoudre, suivant le nombre de cas figurant dans la base. On a commencé par une base contenant 100 cas de pannes résolues, et les tests ont allé jusqu'à 500 cas capitalisés. Les premiers tests ont donnés une similarité moyenne des 3 cas retrouvés de 37 % pour 100 cas initiaux. À la fin des tests, cette similarité a augmenté à 83%.

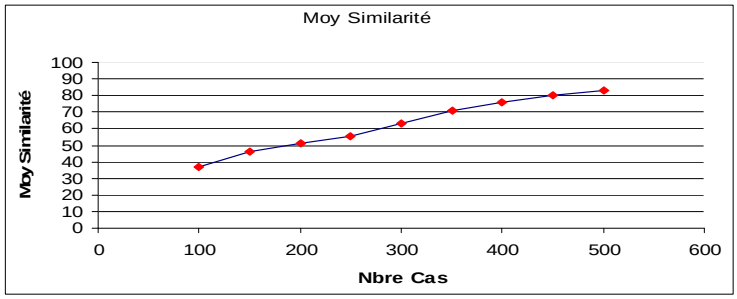


Fig 10 : Augmentation de la pertinence des solutions proposées suivant le nombre de cas stockés

10 Conclusion

Le RàPC a apporté un courant d'air frais, et un certain degré d'optimisme au secteur de l'intelligence artificielle, d'une manière générale, et aux systèmes à base de connaissances en particulier. Poussé par l'intérêt de capitalisation du savoir-faire des entreprises, le renforcement du RàPC par les techniques du Datamining améliorera la fiabilité de cette méthodologie. Après avoir présenter un état de l'art sur la gestion des connaissances dans l'entreprise, on a proposé dans cet article une démarche pour l'intégration de plusieurs techniques, empruntées du Datamining, dans le processus du RàPC, surtout dans l'étape de recherche des cas similaires, par l'utilisation de diverses mesures de similarité locales, adaptées au type et à la modalité de chaque champs. Aussi, exploiter les ontologies du

domaine d'étude dans la recherche, en utilisant des mesures de similarité basées sur les concepts, influa sur la pertinence des résultats de la recherche. On a étudié les connaissances qu'on peut extraire par les différentes formes de mappage entre les clusters de l'espace des pannes et ceux de l'espace des solutions.

Les opérations de prétraitement nous ont donné la possibilité de bien comprendre et visualiser le contenu de la base des cas utilisée. Aussi, la transformation de quelques champs, en leur appliquant des analyses factorielles, nous a aidé à appliquer les algorithmes de Datamining nécessaires.

Notre démarche a montrée des prérogatives claires, elle aidera à mieux comprendre le comportement fonctionnel des équipements, et facilite la planification des interventions de maintenance, préventives et correctives ; ainsi qu'une exploitation optimale de la base des connaissances, avec plus d'interactivité et de visibilité.

Les perspectives du couplage de la puissance et l'efficacité des solutions du Datamining avec la maturité du RàPC nous laisse estimer, dans l'avenir, des systèmes de retour d'expérience bien fondés, soit du coté méthodologique, ou de la qualité des outils d'implémentation utilisés. Dans notre domaine de diagnostic industriel, et parmi plusieurs perspectives possibles, les règles d'associations pourront aider à extraire les dépendances entre les symptômes des pannes ; où un symptôme pourra être le résultat d'un autre. Cela nous aidera à mieux comprendre le comportement fonctionnel des équipements étudiés, et à analyser l'enchaînement des défaillances, ce qui donnera la possibilité d'éviter la panne, en intervenant par une maintenance préventive.

Références

Abasolo, J.M. (2004), T*owards a Component-based Platform for Developing Case-based Reasoning Systems*. Thèse doctorale. Université polytechnique de Catalunya, Espagne.

Cortes Robles, G. (2006), *Management de l'innovation technologique et des connaissances : Synergie entre la théorie TRIZ et le Raisonnement à Partir de Cas, Application en génie des procédés et systèmes industriels.* Thèse doctorale, Institut Polytechnique, Toulouse.

Devèze, B., Fouquin, M. (2004), *Case-Based Reasoning*. Rapport des études en spécialité SCIA, école de l'ingénieur EPTA, France.

Dieng-Kuntz, R., Corby, O., Gandon, F., Giboin, A., Golebiowska, J., matta, N., Ribière, M. (2001), *Méthodes et outils pour la gestion des connaissances : une approche pluridisciplinaire du Knowledge management.* Dunod, 2ème édition, Paris.

Duizabo, S., Guillaume, N. (1997), *Les problématiques de gestion des connaissances dans les entreprises*. Centre de recherches DMSP, Université Paris Dauphine. France.

Ermine, J.L. (2005), *Enjeux, démarches et processus de la gestion des connaissances*. Support de cours, université de trier, France.

Leake, D.B., Wilson, D.C. (1999), *When Experience is Wrong : Examining CBR for Changing Tasks and Environments?*. Computer Science Department, Indiana University, U.S.A.

Rakoto, H. (2004), *Intégration du Retour d'Expérience dans les processus industriels, Application à Alstom Transport.* Thèse doctorale. Institut National Polytechnique de Toulouse.

Rasovska, I. (2006), *Contribution à une méthodologie de capitalisation des connaissances basée sur le raisonnement à partir de cas, Application au diagnostic dans une plateforme d'e-maintenance.* Thèse doctorale, UFRST, Université Franche-Comté.

Watson, I. (1999), *Case-based reasoning is a methodology not a technology.* In: AI-CBR, University of Salford, United KingDoom.

Vers une indexation cellulaire dans les approches à raisonnement à base de cas : Application à la régulation d'un réseau de transport urbain collectif

Fouzia Amrani*,
Karim Bouamrane**, Atmani Baghdad**
* ENSET Oran, 1523 EL Mnaouer Oran,
amranibf@yahoo.fr
** Laboratoire LIO, Université d'Oran, BP 1524 EL Mnaouer Oran
Bouamrane.karim@univ-oran.dz, bbatmani@yahoo.fr

Résumé. Cet article propose l'intégration d'une nouvelle technique d'indexation des cas dite cellulaire par rapport à l'indexation usuellement utilisée dans les approches de raisonnement à base de cas.
Afin d'illustrer cela, nous exploitons cette technique dans un module de raisonnement à base de cas dédié à la régulation d'un réseau de transport urbain collectif.
Le résultat du calcul de similarité qui fait partie du processus de recherche du raisonnement à base de cas entre les perturbations affectant le réseau et les perturbations déjà enregistrées avec leurs solutions dans la base de cas est affiné par ce processus d'indexation cellulaire. Ce raffinement se fait par un langage de modélisation booléen (BML) adopté par le moteur d'inférence cellulaire (CIE).

Mots-clés : Automate cellulaire, Régulation, Réseau de Transport Urbain (RTU), Raisonnement à Base de Cas (RBC), indexation classique, Indexation cellulaire, CASI, Système d'aide à la régulation (SAR).

1 Introduction

La régulation du trafic est une tâche complexe, où les décisions sont prises en fonction de l'état courant du réseau de transport urbain (RTU). Les exploitants du réseau rencontrent de nombreuses difficultés pour maintenir un trafic conforme à la planification prévisionnelle (tableau de marche théorique) et respecter les règles en usage (règles de régulation, règles de sécurité, rôle commercial de l'entreprise…). Les difficultés rencontrées sont dues par exemple aux mauvaises conditions de circulation, au manque de personnel, aux pannes de matériel, aux désynchronisations entre les différents modes de transport, Bouamrane *et* al. (2005).

En cas de perturbations, les régulateurs ne peuvent prendre en compte l'ensemble des informations fournies par le SAE (Système d'Aide à l'Exploitation). C'est le rôle du Système d'Aide à la Régulation (SAR) qui doit identifier les perturbations qui apparaissent, proposer et évaluer des actions correctives. Par ailleurs, le SAR est nécessairement interactif, et doit laisser au régulateur le contrôle de la gestion des perturbations et le choix des actions correctives.

Les travaux menés dans le domaine des systèmes d'aide à la décision en régulation de transport urbain s'intéressent principalement au développement d'algorithmes de régulation automatique, une synthèse des travaux a été proposée dans Bouamrane et Amrani (2007). De même que l'opportunité de réfléchir sur un raisonnement à base de cas pour la régulation d'un RTU.

L'objectif de ce papier est de proposer une amélioration de l'indexation des cas pour un RBC. Le résultat du calcul de similarité entre les perturbations est affiné par un processus d'indexation cellulaire. Ce raffinement se fait par un langage de modélisation booléenne (BML) adopté par un moteur d'inférence cellulaire (CIE) cœur de la machine CASI (Cellular Automaton for Symbolic Induction) proposée par Atmani et Beldjilali (2007), Atmani (2007). La finalité de cette démarche booléenne est double à savoir réduire la complexité de stockage et le temps de réponse. Le deuxième objectif a été d'intégrer ce nouveau module de Raisonnement à base de cas cellulaire dans la plateforme de régulation SARRT, Bouamrane (2006) au niveau du Poste de Commande Centralisé d'un réseau de transport urbain (PCC) afin de comparer et valider le choix d'une telle technique par rapport à celle déjà utilisée.

2 Exploitation d'un RBC avec une indexation classique : Exemple illustratif

La plateforme SARRT intègre actuellement un module de régulation via un raisonnement à base de cas (RBC), Bouamrane et Amrani (2007). Ce module RBC intègre actuellement au niveau de l'étape recherche et classification, une indexation par situation comportementale. Nous proposons, dans ce travail de modifier cette méthode par une indexation cellulaire en exploitant la machine cellulaire présentée en section 5. Nous proposons dans ce qui suit de vous montrer le fonctionnement du RBC avec une indexation classique, ensuite une indexation cellulaire avec le même exemple illustratif.

La construction d'un système à base de cas passe par cinq étapes, Lamontagne et Lapalme (2000). A cet effet, nous proposons l'exemple illustratif ci-dessous qui reprend en détail ces étapes.

Soit une perturbation qui affecte le bus N° 02 sur la ligne N°11, les conséquences de cette perturbation est un retard du véhicule à quelques mètres du prochain arrêt appelé « Valéro».

2.1 Construction des cas (perturbation)

La construction des cas correspond à la localisation de la perturbation à traiter ainsi que ces différentes caractéristiques obtenues via le SAE. Nous pouvons citer parmi ces caractéristiques : la même ligne, l'arrêt suivant, la durée, la période de la journée, etc.

.

N° d'ordre	ligne	N° véhicule	Arrêt suivant	perturbation	durée	Horaire
06	11	02	valéro	retard	10	13 : 53 : 04

FIG. 1 - *Identification de la perturbation*

2.2 La recherche ou remémoration

Il s'agit de rechercher les cas similaires. Cela se traduit comme suit : Etant donné *Pb1*, une nouvelle perturbation acquise (*cible*) via le SAE ou un appel radio et *Pb2*, la perturbation présente dans la base de cas (*source*).

Chaque cas-source (perturbation) enregistré devient une solution source pour la base de cas. A ce stade, on lui associe un index *indx(source)*. Une solution source est ajoutée à la hiérarchie en considérant un seuil à partir duquel elle peut être retenue dans la hiérarchie. Ce seuil est établi sur la base d'un certain nombre de paramètres, tableau (1).

Paramètres	Retard	Localisation	Période	Mode de transport
Seuil	± 2 min	± soit un arrêt précédent, soit un arrêt suivant sauf si c'est le terminus	même période (horaire de la perturbation)	même mode de transport

TAB. 1 - *Paramètres et seuils de fm1*

Le calcul de la similarité dans le cas de la régulation d'un RTU est basé sur six données, les quatre proposées dans le tableau (1), plus deux paramètres concernant le calcul des distances par rapport au prochain arrêt et au prochain terminus.

Dans le cas de notre exemple, la recherche des perturbations similaires en exploitant les paramètres du tableau (1) a permis :

- de trouver trois cas similaires dont les numéros d'ordre dans la liste des perturbations sont (34, 36, 37), figure (2).
- Les autres cas ne sont pas retenus, car ils ne correspondent pas aux critères comme par exemple : le cas N°38 ou la durée n'appartient pas à l'intervalle [8min, 12min].

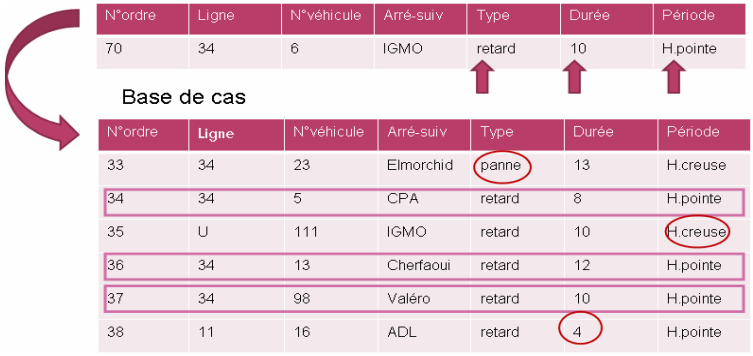


FIG. 2 - *Identifier les perturbations similaires*

Ensuite, nous procédons à la recherche des *solutions* des cas similaires trouvés, en fonction du N° ordre (N° index), dans notre exemple, pour la ligne N° 34, nous avons deux solutions pertinentes (échange conducteurs et véhicules, demi-tour en ligne), figure (3).

2.3 Adaptation

Dans la phase d'adaptation, le système RBC aide l'usager à modifier et à réutiliser les solutions de ces cas pour résoudre son problème courant.

Le RBC choisit le cas similaire le plus proche : par exemple, dans notre cas le véhicule N° 34 est le plus proche. Enfin, le RBC propose sur la base d'une approche multicritère, la solution la mieux adaptée au cas le plus proche du module RBC implémenté sur la plateforme SARRT. Nous schématisons l'indexation des solutions par la structure proposée dans la figure (3).

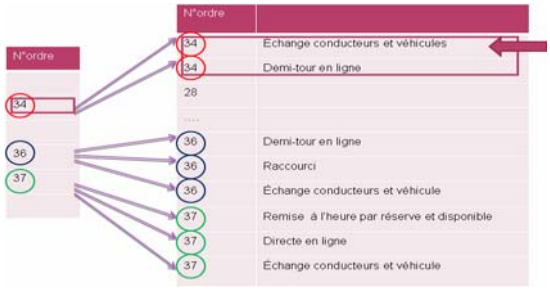


FIG. 3 - *Fonctionnement du RBC*

Dans la figure (4), les Pi représentent les perturbations et les Ai représentent les solutions (les actions de régulation).

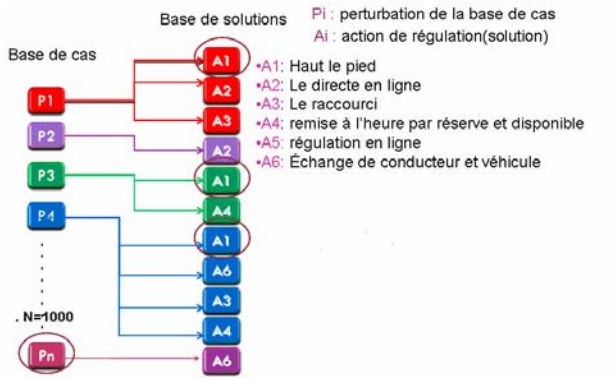


FIG. 4 - *Indexation classique dans un RBC*

Deux problèmes se posent dans l'indexation classique :

- La duplication : les solutions (actions de régulations) se répètent plusieurs fois dans la base de solution comme c'est le cas pour A1.
- La saturation de la base de solution : puisque chaque cas (perturbation) peut posséder plusieurs solutions (actions de régulation). Ceci peu induire une explosion combinatoire.

Pour pallier à ces deux problèmes, nous recourons aux automates cellulaires.

2.4 Révision ou Vérification

Poursuit deux buts complémentaires Lamontagne et Lapalme (2000) à savoir : évaluer la pertinence de la solution proposée selon des critères de réussite spécifiques au problème et vérifier que la solution ne contredit pas des règles générales qui doivent toujours être vérifiées par exemple. Dans le cadre du RBC classique proposé pour la régulation, la révision consiste à évaluer de nouveaux les trois indicateurs permettant d'assurer la qualité de service à savoir le gain en ponctualité, le gain en régularité et le gain en correspondance.

2.5 Maintenance ou Apprentissage

Les concepteurs doivent préconiser certaines stratégies pour intégrer de nouvelles solutions dans la base de cas et pour modifier les structures du système RBC pour en optimiser les performances. Dans le cas actuel du travail, l'ensemble des solutions adaptées et révisées sont incluses dans la base de cas en respectant l'index de la perturbation.

3 Passage d'une indexation par situation comportementale à une indexation cellulaire

Tout en gardant le même exemple illustratif, nous proposons ci-dessous comment assurer le passage d'une indexation classique vers une indexation cellulaire. Pour cela, il faut d'abord construire des règles (en spécifiant la prémisse et la conclusion de chaque règle).

Règle : prémisse→conclusion

La prémisse représente la perturbation et la conclusion représente les solutions de cette perturbation.

La solution la plus logique est d'affecter à chaque cas une règle, soit un cas P1 ses solutions sont par exemple A1, A2, A3 (exemple pour illustrer notre proposition). Avec Ai, les actions de régulation, par conséquent, la règle devient P1→A1, A2, A3.

FIG. 5 - *Principe de passage vers l'Indexation cellulaire dans un RBC*

Dans ce qui suit, nous allons décrire le déroulement du processus d'indexation sur un exemple. Le processus d'indexation ne concerne que la phase « recherche ou remémoration » du cycle d'un RBC (section 3). Supposons que notre base d'indexation se compose de 5 perturbations et 6 actions qui se répartissent sur 5 règles. La figure (6) récapitule l'exemple de la base de connaissances.

Base perturbations/solutions		
P1→	A1, A2, A3	I1
P2→	A2	I2
P3→	A1, A4	I3
P4→	A1, A3, A4, A6	I4
P5→	A6, A5	I5

FIG. 6 - *Un exemple de base de connaissances*

3.1 Le Moteur d'Inférence Cellulaire : Architecture et principe de fonctionnement

La problématique du moteur d'inférence cellulaire CIE (Cellular Inference Engine) est : "Comment synchroniser une ligne de faits (respectivement de règles) de façon à ce que toutes les cellules de la ligne participent (respectivement se déclenchent) ensemble ? ". La modélisation naturelle de ce problème fut concrétisée par la conception d'un Automate Cellulaire pour des Systèmes d'Inférence à base de Règles appelé ACSIR, Atmani (2007).

3.1.1 Architecture du CIE

Le module CIE simule le fonctionnement du cycle de base d'un moteur d'inférence en utilisant deux couches finies d'automates finis. La première couche, CELFACT, pour la base des faits et, la deuxième couche, CELRULE, pour la base de règles. Chaque cellule au temps t+1 ne dépend que de l'état des ses voisines et du sien au temps t. Dans chaque couche, le contenu d'une cellule détermine si et comment elle participe à chaque étape d'inférence : à chaque étape, une cellule peut être active (1) ou passive (0), c'est-à-dire participe ou non à l'inférence.

Le principe est simple :

- Toute cellule i de la première couche CELFACT est considérée comme fait établi si sa valeur est 1, sinon, elle est considérée comme un fait à établir.
- Toute cellule j de la deuxième couche CELRULE est considérée comme une règle candidate si sa valeur est 1, sinon, elle est considérée comme une règle qui ne doit pas participer à l'inférence.

Les états des cellules se composent de trois parties : EF, IF et SF, respectivement ER, IR et SR, étant l'entrée, l'état interne et la sortie d'une cellule de CELFACT, respectivement d'une cellule de CELRULE. L'état interne, IF d'une cellule de CELFACT indique le rôle du fait : IF = 0 correspond à un fait incertain avec une probabilité = 0 ; IF = 1 correspond à un fait certain avec une probabilité = 1. Pour une cellule de CELRULE, l'état interne IR peut être utilisé comme coefficient de vraisemblance que nous n'utiliserons pas dans cette étude, Atmani et Beldjilali (2007).

Pour illustrer l'architecture du module CIE, nous considérons les 5 règles, extraites de la figure (6), obtenues en utilisant les index I1, I2, I3, I4 et I5, figure (7). De même, nous utilisons l'appellation Celindex au lieu de Celrule compte tenu de l'objectif du travail concernant l'indexation cellulaire.

FIG. 7- *Configuration initiale du CIE*

La figure (7), montre comment la base de connaissance extraite à partir de la base d'indexation est représentée par les couches CELFACT et CELINDEX. Initialement, toutes les entrées des cellules dans la couche CELFACT sont passives (EF = 0), exceptées celles qui représentent la base des faits initiale (EF(1) = 1).

3.1.2 Voisinage du CIE

Nous supposons qu'il y a *l* cellules dans la couche *CELFACT*, et *r* cellules dans la couche *CELINDEX*. Le voisinage d'ACSIR est représenté par deux relations d'entrée (&E), et la relation de sortie (&S) formulées comme suit :

- la relation d'entrée, notée iR_Ej, est formulée comme suit : $\forall i \in [1, l]$, $\forall j \in [1, r]$, si (le Fait i ∈ à la Prémisse de la règle j) alors RE (i, j) ← 1.
- a relation de sortie, notée iR_Sj, est formulée comme suit : $\forall i \in [1, l]$, $\forall j \in [1, r]$, si (le Fait i ∈ à la Conclusion de la règle j) alors RS (i, j) ← 1.

Dans la figure (8), respectivement la figure (9), est représentée la matrice d'incidence d'entrée RE, respectivement de sortie RS, de l'automate.

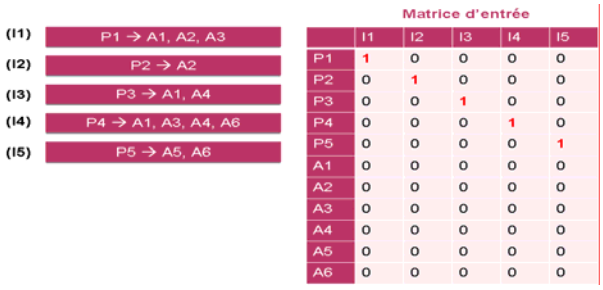
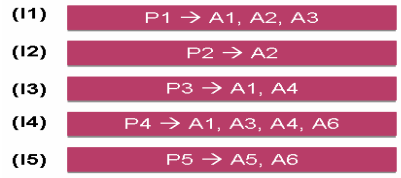


FIG. 8 - *Matrice d'entrée R_E du CIE*



Matrice de sortie

	I1	I2	I3	I4	I5
P1	0	0	0	0	0
P2	0	0	0	0	0
P3	0	0	0	0	0
P4	0	0	0	0	0
P5	0	0	0	0	0
A1	1	0	1	1	0
A2	1	1	0	0	0
A3	1	0	0	1	0
A4	0	0	0	1	0
A5	0	0	0	0	1
A6	0	0	0	1	1

FIG. 9 - *Matrice de sortie R_S du CIE*

3.1.3 Dynamique du CIE

Le cycle de base d'un moteur d'inférence, pour établir un fait F en chaînage avant, fonctionne traditionnellement comme suit :

1. Recherche des règles applicables (évaluation et sélection) ;
2. Choisir une parmi ces règles, par exemple R (filtrage) ;
3. Appliquer et ajouter la partie conclusion de R à la base des faits (exécution).

Le cycle est répété jusqu'à ce que le fait F soit ajouté à la base des faits, ou s'arrête quand aucune règle n'est applicable. La dynamique de l'automate cellulaire CIE, pour simuler le fonctionnement d'un Moteur d'Inférence, utilise deux fonctions booléennes de transitions δFact et δIndex, Atmani et Beldjilali (2007), où δFact correspond à la phase d'évaluation, de sélection et de filtrage, et δIndex correspond à la phase d'exécution.

- La fonction de transition δFact :

$$(EF, IF, SF, EI, II, SI) \rightarrow \delta Fact\ (EF, IF, EF, EI + (M^{T}_{E} \cdot EF), II, SI)$$

- La fonction de transition δIndex :

$$(EF, IF, SF, EI, II, SI) \rightarrow \delta Index\ (EF + (M_S \cdot EI), IF, SF, EI, II, \overline{EI})$$

Où la matrice R^{T}_{E} désigne la transposée de $\boldsymbol{R_E}$

Nous considérons G0 la configuration initiale de notre automate cellulaire, figure (7) et, $\Delta = \boldsymbol{\delta Index} \circ \boldsymbol{\delta Fact}$ la fonction de transition globale : Δ (G0) = G1 si G'0= $\boldsymbol{\delta Fact}$(G0) et G1= $\boldsymbol{\delta Index}$(G'0). Supposons que G = {G0, G1,..., Gq} est l'ensemble des configurations de notre automate cellulaire. L'évolution discrète de l'automate, d'une génération à une autre, est définie par la séquence G0, G1,..., Gq, où Gt+1 = Δ (Gt).

3.2 Indexation cellulaire

Supposons qu'après avoir exécuté le RBC, nous ayons trouvé que P4 est le cas le plus similaire de la perturbation, alors P4 est dans la base de faits, figure (10). Pour trouver les solutions de P4, nous devons exécuter $\boldsymbol{\delta Fact}$ et $\boldsymbol{\delta Index}$.

	EF	IF	SF
P1	0	1	0
P2	0	1	0
P3	0	1	0
P4	1	1	0
P5	0	1	0
A1	0	1	0
A2	0	1	0
A3	0	1	0
A4	0	1	0
A5	0	1	0
A6	0	1	0

	EI	II	SI
I1	0	1	1
I2	0	1	1
I3	0	1	1
I4	0	1	1
I5	0	1	1

FIG. 10 - *Recherche du cas similaire, le plus proche*

En utilisant maintenant notre principe cellulaire, la figure (11) présente l'état global des deux couches, CELFACT et CELINDEX, après évaluation, sélection et filtrage en mode synchrone : application de la première loi de transition ***δFact***.

	EF	IF	SF
P1	0	1	0
P2	0	1	0
P3	0	1	0
P4	1	1	1
P5	0	1	0
A1	0	1	0
A2	0	1	0
A3	0	1	0
A4	0	1	0
A5	0	1	0
A6	0	1	0

	EI	II	SI
I1	0	1	1
I2	0	1	1
I3	0	1	1
I4	1	1	1
I5	0	1	1

FIG. 11 - *Configuration G'0 du CIE obtenue avec δFact*

De même, après l'application de la seconde loi de transition, δIndex, nous obtenons la configuration G1 qui illustrée par la figure (12). La fonction Δ constitue une loi de transition globale en chaînage avant qui transforme itérativement notre automate cellulaire d'une configuration initiale G0 en une configuration finale G1 présentée dans la figure (12).

	EF	IF	SF
P1	0	1	0
P2	0	1	0
P3	0	1	0
P4	1	1	1
P5	0	1	0
A1	1	1	0
A2	0	1	0
A3	1	1	0
A4	1	1	0
A5	0	1	0
A6	1	1	0

	EI	II	SI
I1	0	1	1
I2	0	1	1
I3	0	1	1
I4	1	1	0
I5	0	1	1

FIG. 12 - *Configuration G1 du CIE obtenue avec δIndex*

Nous avons : A1, A3, A4, A6 qui s'ajoutent dans la base de fait et qui sont les solutions de notre cas similaire P4. Par conséquent, en adoptant cette méthode d'indexation, nous avons pu éliminer les problèmes de l'indexation classique (duplication, saturation).

4 Interprétation des résultats

Suite à la mise en œuvre de la nouvelle approche d'indexation cellulaire au niveau du RBC, et son intégration dans la plateforme SARRT pour la régulation d'un RTU, nous avons constaté les conséquences suivantes sur les deux paramètres temps de réponse et espace de stockage:

1. Le temps de réponse

Nous avons introduit une minuterie pour comparer les résultats réels. En effet, les résultats obtenus avec le RBC cellulaire avec un codage binaire montre une légère amélioration par rapport au RBC classique compte tenu que la recherche n'est plus effectuée dans une base de données. Pour l'exemple illustratif utilisé, les résultats étaient comme suit:
RBC classique: 0,01
RBC cellulaire: 0009

2. L'espace mémoire de stockage

Dans le RBC classique où les solutions sont stockées sous forme d'enregistrements, le fichier occupe près de 36 Ko. Par contre, pour le RBC cellulaire, le fichier d'initialisation utilisé pour la sauvegarde n'occupe que 4 Ko.

Il y a lieu de noter que ces résultats sont encore partielles et que des expérimentations, principalement, l'augmentation du volume de la base est en cours afin d'affiner ces résultats. De plus, au niveau adaptation, c'est encore le régulateur qui visuellement choisisse la solution à adopter.

5 Conclusion

Deux motivations concurrentes nous ont amenés à proposer ce principe cellulaire pour la génération, la représentation et l'utilisation d'une indexation booléenne dans le module de régulation à base de cas intégrant la plateforme SARRT. Nous avons souhaité avoir une base de cas de perturbations indexé par automate cellulaire reduisant le plus possible le nombre de duplication afin d'alléger la base de cas. Les avantages de notre approche basée sur le principe de CASI peuvent être récapitulés comme suit :

- La représentation de la base de cas ainsi que son contrôle sont simples, sous forme de stockage binaire exigeant un prétraitement minimal.
- La facilité de l'implémentation des fonctions de transition qui sont de basse complexité, efficaces et robustes concernant des valeurs extrêmes. D'ailleurs, elles sont bien adaptées aux situations ayant beaucoup de cas.
- Les résultats de l'indexation sont simples, ils peuvent être réorganisé et réutilisé par le régulateur.

Références

B. Atmani & B. Beldjilali (2007). Knowledge Discovery in Database: Induction Graph and Cellular Automaton. Computing and Informatics Journal, Vol.26, N°2, 171-197.

Atmani B. (2007). CNNS : Un système Neuro-Symbolique cellulaire pour l'apprentissage automatique a partir de connaissance empiriques. Thèse de doctorat, Université d'Oran.

Bouamrane K & T. Bonte & C. Tahon. (2005). SART : un système d'aide à la décision pour la régulation d'un réseau de transport bimodal Méthodologies et Heuristiques pour l'Optimisation des Systèmes Industriels MHOSI'2005. Hamamet (Tunisie).

Bouamrane K. (2006). Un système Interactive d'aide à la décision pour la régulation d'un réseau de Transport urbain Bimodal : Approche multi agent et raisonnement à base de cas. Thèse de Doctorat, Université d'Oran.

Bouamrane K. & F. Amrani (2007). Un système d'aide à la régulation pour un réseau de transport urbain collectif : Vers une approche à base de cas. Journal of Decision Systems (JDS), Hermès Science Publications, Vol 16 N°4, pp 469-504.

Kacem I. & Dridi M. (2004). Un problème de régulation dans les réseaux de transport urbains. CIFA 2004 – Conference Internationale Francophone d'Automatique, Douz, November.

Laichour H. (2002). Modélisation multi agent et aide à la décision : application à la régulation des correspondances dans les réseaux de transport urbain, thèse de doctorat, Université de Lille.

Lamontagne L. & Lapalme L. (2000). Raisonnement à base de cas textuels. Etat de l'art et perspectives. RSTI – RIA. Volume 15 – n°3/2002, pp. 339-366.

Soulhi A. (2000). Contribution de l'intelligence artificielle à l'aide à la décision dans la gestion des systèmes de transport urbain collectif, thèse de doctorat, Université de Lille 1.

Zidi S. (2007). SARR : Système d'aide à la régulation et la reconfiguration des réseaux de transport multimodal. Thèse Université de Lille 1.

L'utilisation d'un système à base de cas dans le cadre d'une mémoire d'entreprise juridique

* Karima DHOUIB ** Faïez GARGOURI

* Institut Supérieur des Etudes Technologiques de Sfax
B.P. 88A Elbustan ; 3099 Sfax TUNISIE
dk_karima@yahoo.fr
** Institut Supérieur d'Informatique et du Multimédia de Sfax
B.P. 242- 3021 Sfax TUNISIE
faiez.gargouri@fsegs.rnu.tn
Laboratoire MIRACL

Résumé. Passé l'effet de mode qui a entouré son développement, la gestion de connaissance ou knowledge management (KM) est désormais reconnu comme une discipline à part entière. Il puise ses origines dans les principes de l'organisation apprenante et de la systémique, se nourrissant des progrès technologiques qui rendent possibles la mise en œuvre de concepts de travail collaboratif, jusque-là relativement théoriques. La mise en œuvre d'un processus de KM repose souvent sur l'élaboration d'une mémoire d'entreprise. Nous décrivons, dans ce papier, nos travaux actuels, portant sur la construction d'une mémoire d'entreprise juridique en mettant l'accent sur les problématiques liées à la maintenance de cette mémoire. Nous proposons également une justification de l'utilisation d'un système de raisonnement à partir de cas dans le cadre de cette mémoire.

1. Introduction

Dans un environnement en constante mutation où la concurrence se fait de plus en plus féroce, où le cycle de vie des produits devient de plus en plus court, où les clients sont en recherche perpétuelle de nouveautés, de produits personnalisés et de services individuels, les entreprises doivent sans cesse innover et se démarquer de la concurrence. De ce fait, la capacité à produire des entreprises modernes ne repose plus uniquement sur leurs ressources industrielles, mais de plus en plus sur leur capital intellectuel, leurs connaissances.
Moteur de la pérennité et de la croissance des entreprises, le capital de connaissances est une richesse qu'il est devenu impératif de gérer et de valoriser sous peine de céder du terrain sur un marché de plus en plus concurrentiel.

La gestion des connaissances ou Knowledge Management (KM) est une réponse efficace à la nécessité de préserver ce capital de savoirs souvent lié à l'expérience pratique des personnes, ou figurant dans des textes ou des procédures dispersés manquant d'accessibilité. Elle consiste à décrire, capitaliser, valoriser et transmettre les savoirs internes de l'entreprise, ainsi que les savoirs externes acquis par la veille concurrentielle et stratégique.

La mise en œuvre d'un processus de KM repose souvent sur l'élaboration d'une **mémoire d'entreprise** qui sera le réservoir des connaissances de l'organisation.
La construction d'une mémoire d'entreprise repose sur les étapes suivantes: détection des besoins, construction, diffusion, utilisation, évaluation et finalement maintenance et évolution de la mémoire.

C'est à cette dimension que nous voulons aborder le concept du Knowledge Management, notamment en application dans le domaine juridique et nous allons situer notre problématique au niveau de l'étude des problèmes liés à la maintenance d'une mémoire d'entreprise.

Dans cet article, nous commençons par définir le contexte de notre problématique en présentant le contenu de la mémoire d'entreprise juridique que nous allons créer ainsi qu'exposer les problèmes liés à la maintenance de cette mémoire. Puis nous présenterons dans la troisième section une justification de l'utilisation du système à base de cas dans notre mémoire. Dans la section suivante, nous présenterons l'architecture, le fonctionnement et la maintenance du système à base de cas. Nous conclurons cet article en donnant les perspectives de notre travail de recherche.

2. Définition de la problématique

2.1 Contexte

La gestion de connaissances peut être vu comme étant un processus menant à administrer le patrimoine des connaissances d'une organisation, et de les rendre accessibles et utilisables par tous les membres (Tixier, 2001). Ce processus est souvent matérialisé par la construction d'une mémoire d'entreprise.

Nous retenons ici la définition de la mémoire d'entreprise comme étant la matérialisation explicite désincarnée et persistante des connaissances et informations cruciales dans une organisation, afin de faciliter leur accès, partage et réutilisation par des membres de l'organisation dans le cadre de leurs différentes tâches individuelles et collectives (Dieng et al, 2001).

Nous décrirons dans ce qui suit les étapes du processus de gestion de la mémoire d'une entreprise (Dieng et al, 2001):

Détecter les besoins: dans cette première phase, il va falloir chercher les utilisateurs potentiels de la mémoire et délimiter leurs profils, leurs localités, leur environnement de travail ainsi que leur niveau d'expertise. Nous devons définir également quel le est l'utilisation prévue de la mémoire après sa construction et quand sera-t-elle utilisée: à cours terme, moyen terme ou long terme.

Construire: à ce niveau, Il s'agit de dégager les sources de connaissances disponibles dans l'entreprise (documents papier, documents semi-structurés ou structurés, spécialistes humains, base de données). Et puis identifier quels types de connaissances doivent être pris en compte dans la construction de la mémoire. Le type de la matérialisation préférée de la mémoire doit être aussi précisé selon l'environnement informatique des développeurs et des futurs utilisateurs.

Diffuser et utiliser: dans cette étape, le scénario d'interaction souhaité entre les utilisateurs futurs et la mémoire d'entreprise doit être défini. Il va falloir également définir le moyen ainsi que l'organisation qui sera mise en place pour la diffusion.

Evaluer: à ce stade, des critères d'évaluation ainsi que les responsabilités de la conduite de cette évaluation doivent être identifiés.

Maintenir et évoluer: dans cette dernière étape du processus, il faut tenir en compte du résultat de l'évaluation et trouver les moyens pour ajouter de nouvelles connaissances ou détecter et éliminer (ou contextualiser) les connaissances obsolètes ou Incohérentes. Il faut se décider également de la centralisation ou non de la démarche de la maintenance et de l'évolution.

Après une étude de ces différentes phases ainsi que des outils utilisés dans chacune, une constatation a été faite ; c'est que la majorité des travaux se focalisent sur les étapes d'identification, de diffusion et d'utilisation, cependant on accorde souvent moins de priorité pour l'étape de maintenance (Antonova, 2006).

C'est pour cela que nous allons situer notre problématique au niveau de l'étude des problèmes liés à la maintenance et l'évolution d'une mémoire d'entreprise.
Il serait essentiel alors de prendre en compte les résultats de l'évaluation de l'existant. Les problèmes liés à l'intégration de nouvelles connaissances et à la suppression ou modification des connaissances obsolètes; doivent être abordés. De même, il existe des problèmes organisationnels et des problèmes techniques sous-jacents à l'évolution possible de la memoire (Dieng et al, 2001).

Il est à noter également que les techniques utilisées pour maintenir et faire évoluer la mémoire dépendent également du type de mémoire (révision de bases de connaissances, révision d'une base de cas, etc.).

Dans la même optique, nous nous sommes intéressés à choisir un domaine d'application dont les connaissances sont très évolutives. Le domaine que nous proposons alors est le ***domaine juridique*** vue que le Droit est une science en constante évolution**.** En effet, L'accroissement considérable du nombre de textes législatifs et réglementaires applicables et l'extension simultanée des prérogatives d'appréciation confiées aux juridictions rendent de plus en plus délicate la maîtrise de l'information juridique.

L'idée sera alors de créer une mémoire d'entreprise juridique et de suivre son évolution et sa maintenance. Nous présentons brièvement, dans ce qui suit, les besoins de la construction de cette mémoire (Baudot et al, 2006):

- Conservation des savoirs et préservation d'un capital intellectuel. Cet aspect est essentiel pour pallier la perte de données due au départ d'avocats.
- Accès au "savoir-faire" interne avec une grande facilité de recherche dans le corpus des documents.
- Création d'une « doctrine interne » constituant une valeur ajoutée par rapport à la concurrence.
- Identification des compétences de chacun, dans un but de mise en relation d'avocats ou de juristes dans un domaine particulier (langue, pratique juridique spécifique).
- Partage des connaissances et de l'information à l'occasion d'évolutions règlementaires.

2.2 Contenu de la mémoire

Nous proposons dans ce qui suit une architecture pour la mémoire permettant de mesurer l'impact sur la mémoire après chaque changement législatif (Dhouib, 2008).

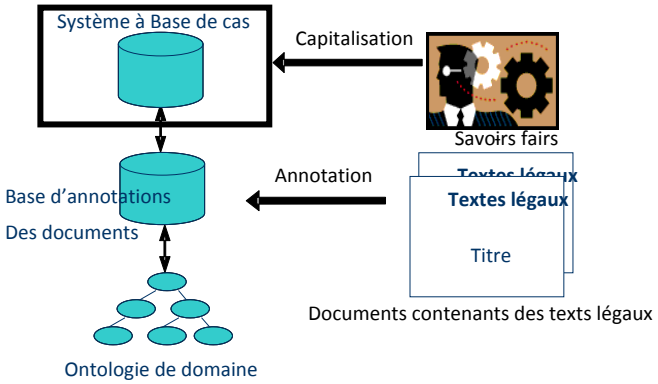


Figure.1. Contenu de la mémoire d'entreprise juridique

Suivant l'architecture que nous proposons (Figure.1.), la mémoire sera constituée de:

Système à Base de cas: Une entreprise dispose généralement d'une collection d'expériences passées ayant des solutions bien définies et qui peuvent être facilement transformées en cas. De plus, une mémoire évolue toujours et par conséquent, ne peut pas être construite en même temps. La représentation en termes de cas permet une construction incrémentale de la mémoire par ajout progressif de nouveaux cas (Simon, 1996).

La base de cas, dans notre contexte, sera alors une collection des savoir-faire des experts juristes et regroupant différentes expériences vécues lors de la résolution des problèmes juridiques. Nous justifierons davantage l'utilisation de la base de cas dans la section suivante.

Base d'annotation: dans le cadre d'une mémoire d'entreprise matérialisée dans des documents, il est intéressant d'associer à de tels documents une connaissance formelle sur laquelle pourra être effectuée un raisonnement afin de rechercher les documents adéquats ou les parties adéquates du document. Cette connaissance formelle peut soit représenter une partie du document, soit consister en des méta-informations sémantiques sur le document, avec parfois des informations supplémentaires non explicites dans le document lui-même.

De ce fait, nous avons pensé à utiliser une base d'annotation dans laquelle nous allons annoter les documents contenant les textes légaux qui pourront être utilisés éventuellement dans la description des solutions des cas juridiques.

Ontologie: pour une mémoire d'entreprise, les ontologies constituent en elles-mêmes une connaissance intéressante pour les utilisateurs, en établissant notamment des référentiels terminologiques.

Notre ontologie servira alors pour garder les concepts juridiques ainsi que les éventuelles relations qui peuvent exister entre eux.

2.3 Maintenance de la mémoire

L'évolution croissante des textes légaux et la présence du risque d'obsolescence d'utiliser un document traitant d'un sujet ayant connu des évolutions réglementaires ou légales récentes, nous ont amenés à dégager quelques problématiques récurrentes liées à la maintenance de la mémoire juridique (Dhouib, Gargouri, 2009):

- Il faut étudier l'interface entre la base de cas et la base d'annotations des documents, car tout cas juridique se réfère sur des textes légaux. Il va falloir tenir en

compte alors que tout changement qui va se faire au niveau de la base d'annotation des documents doit entraîner une mise à jour des cas contenus dans la base de cas, sinon nous risquons la réutilisation de cas obsolète.

- De même, il faut penser à mettre à jour l'ontologie de domaine si de nouveaux textes légaux ; annotés dans la base de documents ; introduisent de nouveaux concepts juridiques.

Ainsi, pour pouvoir maintenir la mémoire d'entreprise juridique, il faut non seulement étudier les problèmes liés à la maintenance de la base de cas, la base d'annotations de documents et de l'ontologie, mais également l'impact de la mise à jour de l'un de ces composants sur les autres composants de la mémoire.
Dans la section suivante, nous nous intéressons à l'étude de de l'utilité de l'intégration d'un système de raisonnement à partir de cas dans le cadre d'une mémoire d'entreprise notamment en application au domaine juridique.

3. Justification de l'intégration d'un système à base de cas dans la mémoire d'entreprise juridique

Dans cette section, nous allons justifier d'une part, l'utilisation d'un système à base de cas dans les systèmes de gestion de connaissance, puis d'autre part, son application au domaine juridique.

3.1 Le système à base de cas, base méthodologique pour les systèmes de KM

Les solutions typiques de KM sont décrites en termes d'un cycle de connaissance incluant des tâches comme la capture, la distribution et la réutilisation. Les cycles de connaissance sont fortement corrélés avec le cycle de raisonnement à base de cas (CBR) qui inclut des étapes de récupération, réutilisation, révision et de conservation. (Aamodt et Plaza, 1994). L'association forte entre le cycle de raisonnement à partir de cas et les cycles de connaissance du KM justifie l'utilisation consistante du CBR pour guider la conception des systèmes de KM (Kitano et al, 1993) (Weber et Aha, 2003) (Althoff et all, 1998) (Aamodt et Nyagaard, 1995) (Aha, 1999).

L'affinité entre le KM et CBR va au-delà de leurs cycles. En effet, la littérature du KM recommande que les solutions efficaces du KM doivent cibler les utilisateurs, les processus et la technologie (Abecker et all, 2000). D'une perspective CBR, Aamodt et Nygaard ont depuis longtemps suggéré que les recherches en CBR doivent se concentrer sur l'optimisation, non seulement du système CBR, mais la combinaison d'un système CBR et son utilisateur. Cela a représenté un point de départ important pour considérer CBR comme une approche contribuant au KM.

Par conséquent, il y a eu beaucoup d'activités de recherche sur le CBR et le KM. La relation entre ces deux domaines est illustrée, par exemple, par un certain nombre d'événements Concernant le KM et le CBR. En 1999, l'Atelier AAAI Explorant les Synergies de Gestion des connaissances et Raisonnement à base de cas (Aha, 1999) s'est concentré sur les exigences pour la contribution efficace de CBR au KM.

En 2001, le comité de programme de l'Atelier allemand traditionnel sur CBR a décidé de changer le nom de leur événement CBR annuel à l'Atelier allemand sur la Gestion d'Expériences.

Les liens proches entre le KM et le CBR sont aussi mis en évidence dans les ouvrages. Tautz (Tautz, 2000) décrit comment personnaliser des systèmes de gestion d'expériences aux besoins organisationnels particulièrement d'un point de vue d'ingénierie logicielle. Bergmann (Bergmann,2000) représente un manuel sur la gestion d'expériences, décrivant tous les aspects des applications du CBR. Watson (Watson, 2003) présente les mémoires d'entreprise d'une perspective CBR.

3.2 Le raisonnement à partir du cas pour le domaine juridique

Dans le raisonnement à partir de cas, l'expérience (basée sur les cas déjà résolus) guide la compréhension des nouvelles situations. L'interprétation et la mémorisation des cas se fait grâce à une indexation basée essentiellement sur des traits de surface. Si une situation nouvelle ne s'apparie pas avec un cas déjà référencé, la base de cas est complète. Le raisonnement par cas est bien adapté aux domaines où il n'existe ni théorie ni modèle formalisé, et où le rôle de l'expérience est prédominant. Un des avantages souvent cité en faveur du raisonnement par cas est qu'il opère à un haut niveau pragmatique d'expertise sans s'occuper des modèles théoriques. Tel est le cas pour le domaine juridique du fait que les experts autorisés professent des visions parfois divergentes. Le raisonnement à base de cas est aussi une approche intéressante lorsque différents points de vue sont en compétition, et où les experts eux-mêmes utilisent des cas concrets dans leur argumentation, dans l'enseignement, dans l'explication, et dans la planification. C'est également le cas dans le domaine juridique. Les problèmes juridiques ont souvent pour objectif de régler un litige où les partenaires invoquent des connaissances juridiques conflictuelles. Le raisonnement par cas y est utilisé pour l'argumentation(www.droit.univ-paris5.fr).

D'autre part, les juristes emploient souvent le raisonnement analogique en comparant une situation réelle donnée avec des décisions passées. Dans le processus de rappel d'une situation semblable face à un nouveau problème, les systèmes de raisonnement à base de cas peuvent être utilisés pour simuler le raisonnement analogique. Un autre argument en faveur de l'utilisation du RPC, c'est que les professionnels juridiques ont deux sources de recherche de jurisprudence: les livres et les systèmes de base de données. La recherche dans les livres est consommatrice de temps et souvent imprécise. Les systèmes de bases de données textuels disponibles, quant à eux, ne garantissent pas la récupération de documents utiles (Weber et all, 1999).

4. Architecture, fonctionnement et maintenance du système à base de cas

4.1 Architecture du système

Notre mémoire d'entreprise juridique sera composée, comme s'est déjà avancé, d'une base de cas, d'une base d'annotation et d'une ontologie.
Nous nous intéressons, en premier lieu à la construction du système à base de cas.

L'objectif opérationnel de ce système est de fournir une aide pour le juriste pour résoudre une situation juridique donnée et ce en mettant à sa disposition une sélection de cas représentant des situations similaires, ce qui améliorera son raisonnement futur.
Un cas est donc l'association d'un problème d'une situation juridique et de la solution de ce problème : ***cas=(pb,Sol(pb))***.
Un **cas source** est un cas dont on va s'inspirer pour résoudre un nouveau cas que l'on appelera un **cas cible**. Un cas source s'écrit : ***cas-source=(source,Sol(source))*** et un cas cible s'écrit donc ***cas-cible=(cible,Sol(cible))***.

Un cas, son problème et sa solution sont décrits par un ensemble de de scripteurs. Un descripteur de est défini par une paire d=(a,v) où a est un attribut et v la valeur qui lui est associée dans ce cas. Conformément à ce vocabulaire, source et cible sont définis de la manière suivante :

- source=$\{d^s_1..d^s_n\}$ où d^s_i est un descripteur du problème source.
- Sol(source)=$\{D^s_1..D^s m\}$ où D^s_i est un descripteur de la solution source.
- cible=$\{d^c_1..d^c_n\}$ où d^c_i est un descripteur du problème cible.
- Sol(cible)=$\{D^c_1..D^c_n\}$ où D^c_i est un descripteur du problème cible.

La base de cas comprend la collection de cas ainsi que les mécanismes qui sont utilisés pour relier les cas. La structure organisationnelle dans la théorie CBR se réfère à la manière dont les cas sont organisés dans la base de cas. Il s'agit alors de définir un modèle pour organiser nos cas.

D'un autre coté, il faut dresser la liste des descripteurs qui vont être associés à un cas. Ceci ne peut être réalisé qu'avec la collaboration avec les experts juristes.
Pour permettre de comparer les cas les uns avec les autres, il faut pouvoir comparer leurs valeurs d'attributs de façon à établir à quels points ces valeurs sont proches. Chaque attribut doit donc être typé. C'est la connaissance du type qui permet de connaitre les opérations de comparaison licites et par là d'établir des similarités. Une ontologie est un ensemble de termes reliés entre eux par des relations vérifiées. Les relations les plus classiques sont les relations d'héritage (sorte-de, est-un, a-pour-spécialité, a-pour-instance), les relations de composition (est-composé-de, composant-de). Les autres relations associant les termes n'ont pas de sémantique (règle permettant de dire ce qui est vrai si la relation est en place) implicite évidente. Dans notre cas, il s'agit d'utiliser une ontologie pour décrire les relations qui sont vraies entre les termes utilisés pour les valeurs de descripteurs.
Notre système aura l'architecture suivante (Figure.2.):

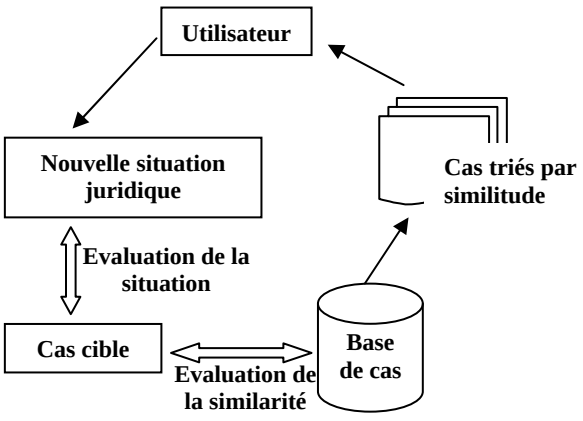


Figure2. *Architecture du système à base de cas*

4.2 Fonctionnement du système

L'inférence commence par l'identification d'une nouvelle situation légale. Cela arrive quand un professionnel juridique exécutant des activités légales habituelles rencontre une nouvelle situation légale qui exige la recherche de jurisprudence.

Le professionnel légal commence par une interprétation de cette nouvelle situation légale dans sa mémoire. Le système essaye de mettre à jour la nouvelle situation légale de l'utilisateur à travers un processus d'évaluation de situation.

Les méthodes d'évaluation de situation déduisent des valeurs pour assigner les attributs dans la représentation semblable à celle des cas du système, puis procède à la modélisation de la nouvelle situation légale de la même façon que les cas dans la base de cas.
Le système compare alors la nouvelle situation légale, dorénavant nommée le cas cible, à chaque cas candidat dans la base de cas. Une métrique de similitude mesure la valeur de chaque similitude pour pouvoir trier les cas candidat.

4.3 Maintenance du système

4.3.1 Etat de l'art de la maintenance d'une base de cas

Les connaissances évoluent, les modes de raisonnement aussi, ce qui rend les travaux de maintenance pour les bases de cas essentielles pour garantir la cohérence et la compatibilité des connaissances, anciennes et nouvelles.
La **M**aintenance de **B**ase de **C**as (MBC) est la mise en œuvre des politiques permettant de réviser l'organisation et/ou le contenu de la BC afin d'améliorer le raisonnement futur (Leak, 1998)]. La MBC (Smyth , 1 998a) est un ensemble de réalités différentes, telles que la suppression des cas incohérents, la sélection de groupes de cas pour l'élimination de la redondance, et la récriture des cas afin de réparer les problèmes d'incohérences.
Deux approches sont proposées pour la MBC (Haouchine et al, 2006), la première traite le problème d'optimisation de la BC et la deuxième traite le problème du partitionnement de la BC.

Une BC est de bonne qualité si elle permet au système de raisonnement à partir de cas de résoudre le plus de problèmes possibles de manière correcte en un temps raisonnable (Zehraoui et al, 2004).
Pour évaluer la qualité de la BC, il existe deux critères:

- La Compétence, est mesurée par le nombre de problèmes différents pour lesquels le système apporte une bonne solution (Smyth, 1995).
- La Performance, d'un système est mesurée par le temps de réponse qui lui est nécessaire pour proposer une solution à un cas cible (Smyth, 1995). Cette mesure est liée directement aux coûts d'adaptation et aux coûts de recherche.

4.3.2 Ebauche pour une méthode de suppression de cas

A partir d'une BC, la stratégie de suppression de cas, évalue les cas suivant un critère afin de pouvoir les supprimer et ramener la BC à un nombre de cas donné. Les critères d'évaluation tels que la compétence, la redondance et l'inconsistance ont été utilisés dans différentes méthodes.

Dans notre contexte, étant donné qu'un cas juridique a une durée de validité qui est exprimée suivant la durée de validité des textes légaux auxquels il se réfère, donc nous proposons un nouveau critère d'évaluation, que nous appellons critère de **« validité »** à travers lequel nous introduisons l'aspect temporel des connaissances.

De ce fait, nous devons attribuer une durée de vie et de fiabilité à la connaissance juridique. Ainsi que déterminer des critères objectifs et subjectifs de validité de la connaissance.

Cette méthode gagne à être enrichie et détaillée dans nos travaux futurs.

5. Conclusion

Dans ce papier, nous avons présenté la notion de mémoire d'entreprise, ainsi que les différentes étapes du processus de sa gestion. Par la suite, nous avons situé notre problématique au carrefour de deux domaines: L'Ingénierie législative, pour la détermination des méta-données, des notions à appréhender et des liens entre ces notions; et l'Ingénierie des Connaissances, permettant de gérer des savoirs juridiques au moyen d'une mémoire d'entreprise.

Dans une autre étape, on a défini quelques problèmes récurrents de la maintenance d'une mémoire d'entreprise qui sera composée, dans notre contexte, d'une base de cas, d'une base d'annotation et d'une ontologie. Nos travaux actuels de recherche, portent principalement sur la construction de la base de cas et à l'étude des problèmes liés à sa maintenance.

6. Références

Aamodt, A and Plaza, E.(1994). Case-based reasoning: foundational issues, methodological variations, andsystem approaches. AI Communication 7(1), 39–59.

Aamodt, A and Nygaard, M. (1995). Different roles and mutual dependencies of data, information, and knowledge—an AI perspective on their integration. Data and Knowledge Engineering , 191–222.

Abecker, A , Decker, S and Maurer, F .(2000). Organizational memory and knowledge management. Information Systems Frontiers 2(3–4), 251–252.

Aha, D W, Becerra-Fernandez, I, Maurer, F and Muñoz-Avila, H (eds).(1999). Exploring Synergies of Knowledge Management and Case-Based Reasoning: Papers from the AAAI Workshop (Technical Report no. WS-99-10). Menlo Park, CA: AAAI Press.

Antonova, A.(2006) Insight into Practical Utilization of Knowledge Management Technologies. *IEEE John Vincent Atanasoff International Symposium on Modern Computing*.

Althoff, K-D, Bomarius, F and Tautz, C. (1998). Using case-based reasoning for building learning software organizations. In Proceedings European Conference on Artificial Intelligence 1998 Workshop on Building, Maintaining, and Using Organizational Memories (OM'98), Brighton, UK.

Baudot, B ., C hrissement, A ., Gi de L oyrette, N. (2006). N ote de sy nthèse: k nowledge management & droit. *Juriconnexion.*

Bergmann, R . (20 02). Ex perience M anagement: Foun dations, Dev elopment M ethodology, and Internet-Based Applications. Berlin: Springer.

Dhouib K. (2008). Construction et maintenance d'une mémoire d'entreprise juridiques, 8ème journées scientifiques des jeunes chercheurs en génie électrique et informatique.

Dhouib K ., Gar gouri F.(2009) Problématiques d e maintenance d es co mposantes d'une mémoire d'entreprise juridique, 2ème Conférence Internat ionale: Systèmes d'Information et Intelligence Economique.

Dieng, R., Corby, O., Gandon, F., Giboin, A., Golebiowska, J., Matta, N., Ribière, M.(2001). *Méthodes et outils pour la gestion des connaissances.* Dunod, Paris, 2ème édition 2001.

Haouchine, K., Ch ebel-Morello, B., Zer houni, N. (2006) Méth ode d e suppression de cas pour une maintenance de base de cas. *14e Atelier de Raisonnement à Partir de ca.*

http://www.droit.univ-paris5.fr/HTMLpages/recherche/griad/RaisCas.htmlMM

Kitano, H, S himazu, H an d Shi bata, A.(1993). C ase-method: a methodology fo r b uilding large-scale cas e-based syste ms. In Procee dings of the National Conference on Art ificial Intelligence (AAAI-93). Menlo Park, CA: AAAI Press.

Leake, D.B., Wilson, D.C. (1998). Catego rizing case- base m aintenance: d imensions and directions. A dvances i n C ase-Based R easoning. *4th European Workshop, EWCBR 98. Proceedings*8: 196-207. Springer- Verlag, Berlin, Germany.

Tautz, C . (2000). C ustomizing so ftware engineering e xperience m anagement sy stems to organizational nee ds, P hD Thesi s, Department of C omputer Sci ence, University of Kaiserslautern, Germany.

Tixier, B.(2001). La p roblématique d e la g estion d es con naissances. *Rapport de recherche N°01.9 IRIN*, Septembre.

Simon, G.(199 6). Knowledge Acqu isition an d Mod eling for Corp orate Me mory: Lesso ns learnt fr om Expe rience. In B. Gai nes, M . M usen eds , Pro c. o f K AW'96, B anff, Canada, November, pp. 41-1 41-18.

Smyth, B . (1998a). C ase-base m aintenance. Task s and Meth ods i n Applied Artificial Intelligence. *11th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, IEA-98-AIE. Pr oceedings. 199 8: 507-516 vo l.2. Springer-Verlag, Berlin, Germany.

Smyth, B., Keane, M .T.(1995).. Rem embering T o Forget : A com petence Preserving Deletion P olicy for C ase- B ased R easoning Sy stems. In : *Proceeding of the 14th International Joint Conference on Artificial Intelligent*. Morgan-Kaufmann. 377-382.

Watson, I. (2 003). Ap plying Kn owledge Management: Techniques for B uilding Corporate Memories. San Francisco, CA: Morgan Kaufmann.

Weber, R an d Ah a, DW.(200 3). Intelligent d elivery of military lesso ns learn ed. Decision Support Systems 34(3), 287–304.

Weber, R. et a l ,(1 997). A Large Case-Based Reasoner for Le gal Cases. Lecture Notes In Computer Science; Vol. 1266. *Proceedings of the Second International Conference on Case-Based Reasoning Research and Development* : 190-199.

Zehraoui, F. (2004). Sy stèmes d'a pprentissage co nnexionnistes et rai sonnement à par tir de cas pour la classification et le classem ent de séquences. Thèse de doctorat, université Paris-Nord.

Summary

Crossed the effect of mode which surrounded its development, the knowledge management (KM) is he nceforth recognized as a sepa rate d iscipline. It draws its o rigins fro m th e principles o f t he ap prenante orga nization, feeding on t echnological pr ogress which m ake possible t he i mplementation o f c oncepts of c ollaborative w ork.. T he i mplementation o f a process of KM often rests on the elaboration of a m emory of company. We describe, in this paper, o ur cu rrent works, co ncerning t he c onstruction of a m emory of l egal c ompany by emphasizing p roblems conne cted t o t he maintenance of this memory. We al so p ropose a justification of the use of a syste m case ba sed reason ing with in th e fram ework of th is memory.

Intégration de la logique floue dans le raisonnement à base de cas : application dans le domaine du bâtiment

Abed Hafidha*, Rezoug Nachida**

* Département d'informatique,
Faculté des sciences, Université de Blida(Algérie)
hafidabouarfa@hotmail.com
** Département d'informatique,
Faculté des sciences, Université de Blida(Algérie)
rnac1972@yahoo.fr

Résumé : L'objectif principal de notre travail est le développement et la validation d'un système dénoté *Vulnérabilité Floue.* Le système permet l'estimation de la vulnérabilité sismique d'une construction. Pour la circonstance, *Vulnérabilité Floue* a utilisé le raisonnement à base de cas flou (RBCF). Notre système transite par trois principales phases : la première étape consiste à décrire les paramètres jugés nécessaires à l'estimation de la vulnérabilité sismique d'une construction. L'influence de ces paramètres sur la réponse sismique de la construction est évaluée par des valeurs linguistiques. *Vulnérabilité Floue* utilise une représentation floue afin de tolérer les imprécisions au niveau de la description de la construction. Lors de la deuxième phase, nous avons évalué la similarité entre une nouvelle construction et l'ensemble des cas historiques. Cette similarité est évaluée sur deux niveaux : 1) similarité individuelle : se base sur les techniques d'agrégation floues (max-min); 2) similarité globale : utilisent les quantificateurs linguistiques monotones croissants (RIM) pour combiner les différentes similarités individuelles entre deux constructions. La troisième phase du processus d'estimation de *Vulnérabilité Floue* consiste à utiliser des vulnérabilités des constructions historiques étroitement similaires à la construction courante pour en déduire une estimation à sa vulnérabilité. Nous avons validé notre système en utilisant 50 cas du CTC de Tlemcen et Blida. Pour cela nous avons évalué les performances de *Vulnérabilité Floue* sur la base de deux critères de base : la précision des estimations et la tolérance des imprécisions tout le long du processus d'estimation. La comparaison s'est faite avec des estimations faites par des modèles d'évaluation fastidieux et longs. Les résultats sont satisfaisants.

1. Introduction

Dans toute organisation ou métier, les gens sont confrontés, quotidiennement, à des problèmes ou des situations auxquels ils doivent apporter des solutions, les meilleurs possibles. Pour ce faire, ils doivent avoir le maximum d'informations sur le problème ou la situation, et puiser, surtout, dans leurs expériences, savoir et savoir-faire. Plusieurs approches ont été proposées pour implémenter cet état de fait dans une machine et permettre à celle-ci de trouver des solutions à des problèmes, en lui fournissant les connaissances nécessaires. Parmi ces approches, le raisonnement à base de cas (RBC) semble le plus proche du raisonnement humain et le plus utilisé dans la vie courante. En effet, Il peut être utilisé

lorsqu'on a peu de connaissances et d'informations sur le problème à résoudre et pour lequel une solution optimale est a priori inconnue.

Dans la plupart des cas on trouve que les connaissances représentées dans un RBC sont linguistiques, imprécises, incertaines et vagues. La tolérance de l'imprécision et de l'incertitude avec ce type de connaissances s'avère nécessaire. Cette fonctionnalité est assurée en utilisant la puissance des outils de la logique floue. L'intérêt de la logique floue réside dans son aptitude à manipuler des grandeurs imprécises utilisées notamment dans le langage humain.

L'intégration de ces outils dans le processus du raisonnement à base de cas, pourrait remédier aux limites observées dans les différents types de RBC classiques.

Nous préconisons dans ce travail d'intégrer la logique floue dans le RBC et afin de valider cela, nous l'avons appliqué dans le domaine de la vulnérabilité des constructions.

2. Raisonnement à base de cas

Le raisonnement à base de cas (RBC), est une technique d'apprentissage automatique issue de l'intelligence artificielle.

Dans un RBC, les connaissances sont représentées sous forme de cas. Un cas est une partie conceptuelle de la représentation de la connaissance expérimentale J. L. Kolodner et al., (1993)**.** Un cas est constitué de la description du problème et de la solution correspondante. Un ensemble représentatif de cas comporte une bibliothèque de cas pour un domaine de problème. Le processus du raisonnement à base de cas transite généralement par quatre étapes : (1) recherche des cas les plus similaires, (2) réutilisation de la solution des cas retrouvés, (3) révision de la solution proposé si nécessaire, et (4) retenir la nouvelle solution comme une partie d'un nouveau cas Aamodt et al., (1994). Le système RBC recherche un ou plusieurs cas similaire à partir d'une bibliothèque de cas quand un nouveau se présente. La solution proposé par le(s) cas plus similaire(s) est réutilisée ou adaptée pour résoudre le nouveau problème. Le problème doit être retenu comme nouveau cas dans la base de cas pour mettre à jour les connaissances du système RBC. Retrouver le(s) cas le(s) plus similaire(s) est la première étape dans le processus du RBC. Cette étape (retrouver) est considérée comme étant la plus importante. Car sans elle les séquences du processus qui suivent ne pourraient avoir lieu. Retrouver le(s) cas le(s) plus similaires revient à évaluer les degrés de similarité entre n'importe quels deux cas à comparer. L'approche la plus utilisée pour cela est une fonction de distance. Cependant, l'imprécision et les incertitudes sont omni présente dans cette fonction. Donc, l'intégration de la logique floue dans le processus du RBC permet de traiter en premier lieu l'imprécision dans les mesures de similarité. La logique floue fut introduite par la suite pour la représentation de la connaissance. En effet comme les résultats du processus du RBC est le produit des retours d'expérience des experts. La représentation pour la plus part des attributs est imprécise et incertaine. La logique floue en prend charge ce type de donnée. Le RBC flou (RBCF) a été d'un grand apport dans plusieurs domaines d'application.

3. Raisonnement à base de cas flou

3.1 Théorie des ensembles flous

La théorie des ensembles flous a été introduite par L.A Zadeh (1965), il a introduit premièrement le concept flou à la place des valeurs précises et binaires pour décrire les phénomènes se produisant autours de nous. Ce concept est très connu et utilisé dans plusieurs domaines de recherche. L'idée de la logique floue est de " capturer" l'imprécision de la pensée humaine et de l'exprimer avec des outils mathématiques appropriés .J. Godjevac (1999). La logique floue, dont les variables peuvent prendre n'importe quelles valeurs entre 0 et 1, permet de tenir compte de cette réalité. Les limites ne varient pas soudainement, mais progressivement. La représentation de ce type de valeur peut être sous plusieurs formes comme indiqué dans la figure 1.

Figure 1. Différentes formes de représentation d'une valeur floue

3.2 Le raisonnement à base de cas flou

Le RBC flou est une approche différente de celle du RBC conventionnel. Dans ce contexte la théorie des ensembles flous a été utilisée dans les différentes phases du processus du CBR mais essentiellement et principalement au niveau de la phase de recherche. En d'autre terme elle a contribué à évaluer les similarités entre un nouveau cas et les cas existants dans la base de cas. En général, le mécanisme de correspondance dans un RBC classique est une affaire du tous ou rien (correspondance totale, ou nulle). Mais cette approche pose des difficultés car les intervalles doivent être petits ou des résultats spécifiques n'existent pas dans la correspondance pour l'ensemble des cas observés ou sinon nécessite une large bibliothèque de cas pour couvrir l'espace en entré.

La similarité floue est proposée pour assurer une indexation efficace et une correspondance exacte R.J. Ku et al., (2005), les données jugées pertinentes dans un cas sont transformées en termes floues basées sur les types de données « textuelle, simple, intervalle ou linguistique » Dvir et al., (1999), le RBC flou utilise les similarités évaluées de sorte que les similarités entre les attributs entre un ancien et un nouveau cas peuvent être des nombres floues comme 0.8, 0.4, 0.2 au lieu d'une correspondance booléenne exacte dans le RBC traditionnel , dans ce contexte divers travaux ont vu le jour entre autre M. Q. Xu et al., (1999) ont calculé la distance du centre de gravité entre deux fonctions d'appartenance floues comme similarité, et a été appliqué dans la conclusion légale pour l'interprétation des jugements de la cour. Zwick, Carlstein et Budescu ont proposé une fonction de similarité avec un paramètre de distance pour combiner les similarités de n-dimensions, des données floues R.J. Ku et al., (2005). Les recherches qui ont incorporé la logique floue dans l'adaptation et maintenance ont été réalisées par Corchado et Torres en 2002, Portinale et

Montari 2002 respectivement Ya-jun Jiang et al., (2006). Liang et Shi (2003) ont proposé diverses mesures numériques pour différencier les ensembles floues intuitifs et Wang a tous simplement proposé des mesures simples utilisant le maximum, minimum et les paramètres de distance pour définir la similarité entre deux ensembles flous ou deux nombre flous, Dvir et al par exemple ont dénoté un graphe simple de mesure de correspondance entre les fonctions d'appartenances des attributs du cas courant et le cas historique. L'espace de chevauchement résultant est le degré de similarité entre les attributs Ya-jun Jiang et al., (2006).

4. Approche Adoptée

L'originalité de notre travail consiste à appliquer un modèle d'estimation pour le cas d'estimation de la vulnérabilité sismique des constructions.

Ce domaine d'application est intéressant pour les raisons suivantes:

1. Estimation difficile car: Quantité des ouvrages et la variabilité des types de constructions sont généralement importantes et dont on a peu de connaissances.
2. Méthodes d'estimation basées sur les retours d'expériences (subjectives).

4.1 Notions de base

La définition donnée lors de la décennie internationale pour la prévention des catastrophes naturelles AFPS (2008) est :

- ✓ **Vulnérabilité d'un élément** est définit comme étant le degré de perte qu'il subit lors d'une catastrophe naturelle. Sa nature et son estimation varient selon que l'élément représente une population, des structures sociales, des structures physiques, ou des actifs économiques.

4.2 Méthodes d'estimation de la vulnérabilité sismique

Plusieurs méthodes d'estimation de la vulnérabilité ont été développées, en particulier la méthode de l'index de vulnérabilité.

Principe de la méthode de l'index de vulnérabilité

L'index de vulnérabilité « Iv » est un indicateur de l'état de la structure qui peut être estimé avant comme après l'occurrence de l'événement sismique. Il nous permet de connaître l'état des constructions d'une région et de les classer selon leur vulnérabilité. Cette méthode offre la possibilité d'une mise à jour pratiquement continue de la qualité sismique des bâtiments d'une région, F.I. Belheouane (2006).

Les différents paramètres, pris en compte par cette méthode, sont classés en trois catégories. Chaque paramètre a une valeur numérique exprimant la qualité sismique des éléments structuraux et non structuraux influant sur le comportement. La somme des valeurs numérique de ces paramètres représente l'index de vulnérabilité « Iv » de la construction étudiée, F.I. Belheouane (2006).

Un paramètre ne peut prendre qu'une seule valeur et représente ainsi la classe (A,B,C) à

la quelle il appartient, F.I. Belheouane (2006).Ou la classe A représente une construction pas du tous vulnérable ; B : moyennement vulnérable ; C : construction vulnérable.

Dans la littérature, des méthodes d'évaluation de l'index de vulnérabilité ont utilisé la théorie de la logique floue pour définir le degré d'appartenance de la construction à une classe de vulnérabilité et assurer un passage progressif et non brusque entre les classes de vulnérabilité. L'une des premières méthodes à avoir utilisé la logique floue est la méthode Echelle Macrosismique Européenne (EMS-98), M.J. Nollet (2004).

.Mais cette dernière ne l'utilise que dans la représentation de l'Iv final, ce qui rend la précision de l'estimation partielle et non finale.

4.3 Approche Adoptée pour l'estimation de la vulnérabilité

Nous proposons un système dénoté ***Vulnérabilité Floue,*** son objectif est de permettre la tolérance des imprécisions tout au long du processus d'estimation par analogie ainsi que la gestion des incertitudes au niveau de la vulnérabilité estimée. Pour être plus précis, nous avons adopté la méthode de IDRI (A.Idri et al., 2004) développée pour l'estimation des coûts de développement de logiciels que nous avons adapté à la méthode de l'index de vulnérabilité EMS-98, M.J. Nollet (2004).

Pour cela, et comme dans un processus d'estimation dans un RBC classique, ce processus transitera par les principales étapes suivantes:

1) Identification de la vulnérabilité sismique par un ensemble d'attributs ;

2) Similarité et Adaptation :

2.1) Évaluation de la similarité entre la vulnérabilité d'une nouvelle construction et celles existantes dans la base de cas.

2.2) Utilisation des vulnérabilités réelles des constructions les plus similaires à la nouvelle construction pour en déduire une estimation à sa vulnérabilité (adaptation compositionnelle).

4.3.1 Phase d'identification

Nous avons pris un seul type qui est les bâtiments en béton armé pour notre étude, la liste des paramètres sont au nombre de quatorze, et chacun d'eux est jugé nécessaire à l'estimation de la vulnérabilité, F.I. Belheouane (2006).. Les facteurs sont recueillis par les experts du CTC, F.I. Belheouane (2006) dans une fiche technique qui contient tous les facteurs nécessaires à l'estimation de la vulnérabilité. Ces paramètres sont tirés, en partie, des méthodes d'évaluation de la vulnérabilité, complétés par d'autres paramètres dont l'influence sur le comportement global de la structure, donc sur sa réponse sismique, a été mise en évidence lors des retours d'expérience sismiques en Algérie, F.I. Belheouane (2006).

Chaque paramètre a un indice de vulnérabilité estimé pour chacun d'eux à partir d'un ensemble de facteurs, cet indice sera mesuré par une valeur linguistique (A_k^j). Chaque valeur linguistique, (A_k^j), est représentée par un ensemble flou ayant une fonction d'appartenance μA_k^j . Dans notre cas les valeurs linguistiques utilisé explicitement sont (bas, moyen, élevé), bas représente une vulnérabilité faible formellement décrite par la classe A, moyen représente une vulnérabilité moyenne formellement décrite par la classe B, élevé

représente une forte vulnérabilité au séisme décrite formellement par la classe C.

Pour la représentation des fonctions d'appartenance on a utilisé la forme trapézoïde.

Les scores obtenus pour chaque paramètre sont obtenus à partir de données observables (scores affectés manuellement par l'expert) ou calculables (scores obtenus en appliquant des formules mathématiques utilisés à cet effet) on donnera pour chaque type un exemple :

1. ***Système de contreventement (Donnée observable):***
   - ✓ Domaine de valeurs : 0 - 2.25
   - ✓ Termes linguistiques

N° d'ordre	**Terme linguistique**	***a***	***b***	***c***	***d***
1	Bas	0.10	0.15	0.25	0.35
2	Moyen	0.25	0.75	1.25	1.75
3	Elevé	1.25	1.85	2.25	0

   - ✓ Représentation graphique

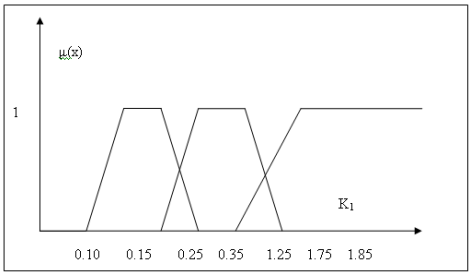


Figure 2 : Représentation graphique des classes de vulnérabilité du paramètre système de contreventement.

La figure 2 permet d'identifier les trois classes possibles du paramètre système de contreventement, les valeurs allant de 0,10 à 1,85 sont données par l'expert selon le type de système de contreventement.

2.Capacité sismique (donnée calculable) : La capacité sismique est obtenue en déterminant un facteur α, ce facteur est calculé par une formule de calcul , cette valeur a trois intervalles si $\alpha > 1.2$ alors le paramètre correspondant est classé dans la classe A, si $0.7 < \alpha \leq 1.2$, le paramètre capacité sera classé dans la classe B, si maintenant $\alpha \leq 0.7$ alors le paramètre sera classé dans la classe C, mais ce classement se fera par un degré d'appartenance comme suit :

Capacité sismique α :

- ✓ Domaine de valeurs : 0- 2
- ✓ Termes linguistiques :

N° d'ordre	**Terme linguistique**	***a***	***b***	***c***	***d***
1	Bas	1.0	1.2	1.30	0
2	Moyen	0.7	0.90	1.00	1.2
3	Elevé	0.20	0.50	0.70	0.80

✓ Représentation graphique

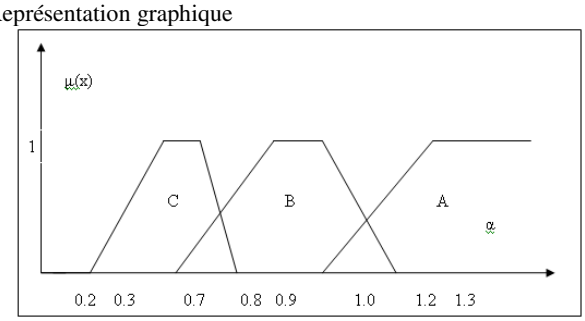


Figure 3 Représentation graphique des classes de vulnérabilité de α

Après avoir calculé α , on calcule son degré d'appartenance aux différentes classes, on prend le degré d'appartenance le plus élevé qui nous indique alors la classe à qui il appartient, cette valeur serait alors le degré d'appartenance du K_3 à la classe correspondante. Exemple $\alpha = 0.44$, μ_C=0.8, μ_B=0, μ_A=0, alors le degré d'appartenance de K_3 est de 0.8 dans la classe C. et donc sa classe la plus probable est la classe C.

Afin de prendre en considération l'importance de chaque attribut dans le processus d'estimation, nous affectons des poids de pondération aux attributs sélectionnés, uj , F.Yépez (1994).

N°	Nom de l'attribut	Poids de l'attribut
1	Système de contreventement	4
2	Qualité du système de contreventement	1
3	La capacité sismique	1
4	Type de sol	1
5	Diaphragme horizontal	1
6	Régularité en plan	1
7	Régularité en élévation	2
8	Qualité des nœuds	1
9	Phénomène du poteau court	1
10	Détails	1
11	Maintenance	2
12	Modifications	1
13	Entrechoquement	1
14	Implantation de l'ouvrage	1

Tableau 1 Poids des paramètres

Une fois la phase d'identification terminée il est temps de passer au calcul de la similarité.

4.3.2 Evaluation de la similarité

La similarité dans notre système sera calculée sur deux niveaux :

1. Similarité locale

Cette étape consiste à évaluer la similarité entre deux constructions C1 et C2 selon chaque attribut V_j, $SV_j(C_1,C_2)$. $SV_j(C_1,C_2)$ sera calculé par la formule d'agrégation flou :

$$SVj(C_1,C_2) = \max_{3} \min(\mu A_k^j(C_1), \mu A_k^j(C_2)) \quad (1)$$

Ou k est le nombre d'ensembles flous pour chaque attribut j, μA_k^j est la fonction d'appartenance de chaque ensemble flou d'un attribut Vj et Ci est la construction i. Les ensembles flous sont (bas, moyen, élevé)

2. Similarité globale

La similarité entre deux constructions C1 et C2 S(C1,C2) est évalué en combinant les similarités individuelles $SVj(C_1,C_2)$, par un quantificateur linguistique Q tel que all, most, many, at-most α et there exists.. Ce genre de quantificateur est appelé RIM (*Regular Increasing Monotone Quantifier)* R. Yager (2001). Donc, la similarité globale entre deux constructions C_1 et C_2 est définie par l'expression informelle suivante :

$$S(C_1,C_2) = mostof\,(SV_j(C_1,C_2)) \quad (3)$$

Ou $SVj(C_1,C_2)$ est la $j^{ième}$ similarité individuelle selon un ordre croissant .
Ou mostof veut dire prendre en considération la plupart des attributs dans le calcul des similarités

L'implémentation du quantificateur RIM de l'équation 3 est assurée par un opérateur OWA. Donc, la similarité globale entre deux constructions C_1 et C_2 est calculée par :

$$S(C_1,C_2) = \sum_{j=1}^{M} w_j(C_1,C_2) SV_j(C_1,C_2) \quad (4)$$

Ou $SVj(C_1,C_2)$ est la $j^{ième}$ similarité individuelle selon un ordre croissant .
$w_j(C_1,C_2)$ est le poids de l'attribut j selon un ordre croissant.

La procédure utilisée pour retrouver le vecteur W associé à un quantificateur RIM, Q, est composée de deux étapes R. Yager (2001), A. Idri et al., (2001a). La fonction d'appartenance associée à Q est monotone croissante Q(0)=0 et Q(1)=1. Deuxièmement les poids $w_j(C_1,C_2)$ sont calculés par l'équation 5.

$$w_j = Q\left(\frac{\sum_{k=1}^{j} p_k}{T}\right) - Q\left(\frac{\sum_{k=1}^{j-1} p_k}{T}\right) \quad (5)$$

où p_k est le poids associé au $k^{ème}$ critère selon un ordre croissant et T est la somme des p_k.
La dernière phase dans l'estimation de la vulnérabilité sismique est la phase adaptation.

4.3.3 Phase d'adaptation

Le but de cette étape est de déduire une estimation de la vulnérabilité sismique d'une construction C, en utilisant les vulnérabilités réelles des constructions les plus similaires à C.

La vulnérabilité d'une construction Ci est étroitement similaire à C si son degré de similarité à C est approximativement égal à 1.

$$\text{Vulnérabilité(C)} = \frac{\sum_{i=1}^{N} \mu_{Voisinage}(S(C,Ci)) \times Vulnérabilité(Ci)}{\sum_{i=1}^{N} \mu_{Voisinage}(S(C,Ci))} \quad (6)$$

où N est le nombre de construction qui vérifie la qualification étroitement similaire (dans notre cas les constructions ayant des degrés d'appartenance supérieur à 0,8).

Une fois cette vulnérabilité estimée (Ve), on classe une construction selon les intervalles suivants (ces intervalles ont été déterminés par F.I. –Belheouane (2006) et validés par les experts du CTC):

- La classe verte : $3{,}25 \leq Ve \prec 6{,}25$: cette classe exprime que la structure étudiée n'est pas vulnérable au séisme. Donc la construction présente une bonne résistance sismique.
- La classe orange : $6{,}25 \leq Ve < 10{,}25$: cette classe exprime que les constructions ont une résistance sismique moyenne et donc la structure est moyennement vulnérable.
- La classe rouge : $10{,}25 \leq Ve < 13{,}25$: cette classe exprime que les constructions ont une résistance sismique faible, et donc présentent une vulnérabilité sismique élevée.

5. Discussion et Résultats

Une validation d'un nouveau système peut être effectué sur deux niveau :1) axiomatique ;2) empirique. Le niveau axiomatique a été assuré par [11].

La validation empirique de notre système *vulnérabilité floue* consiste en l'évaluation de la précision de ses estimations de la vulnérabilité sur une base de constructions. Dans notre cas, nous utilisons les cas d'un canevas d'expertise de Tlemcen et de Blida pour évaluer la précision des estimations de notre approche *Vulnérabilité floue*. Cette évaluation utilise le prototype logiciel *Vulnérabilité floue* qui implémente notre approche. Les résultats de cette validation empirique sont comparés aux estimations faites par les experts du CTC de Tlemcen et Blida. Avant de décrire les résultats obtenus par notre système *Vulnérabilité* ***floue,*** nous présentons un échantillon des écrans de traitement de notre prototype.

5.1 Présentation du prototype *Vulnérabilité floue*

Cette présentation permettra de mettre en valeur les différentes phases de l'estimation de la vulnérabilité décrite dans la section précédente.

Mais avant tous, nous présentons l'architecture globale de notre système :

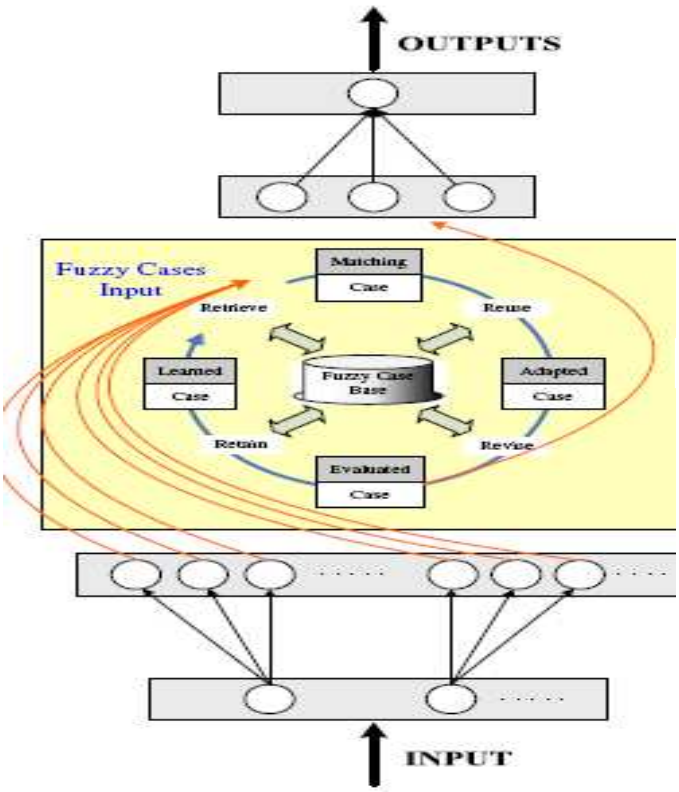


Figure 4 : Architecture de Vulnérabilité floue

a) Identification des structures

Notre prototype permet en premier lieu la saisie d'un nouveau cas .

Une fois le cas saisi, on passe à l'identification du cas courant (Fuzzification), cette phase permettra d'affecter comme précisé dans la section précédente un indice de vulnérabilité à chaque paramètre. L'appartenance à une classe est évaluée avec un degré d'appartenance défini par la forme trapèze (Figure 6)

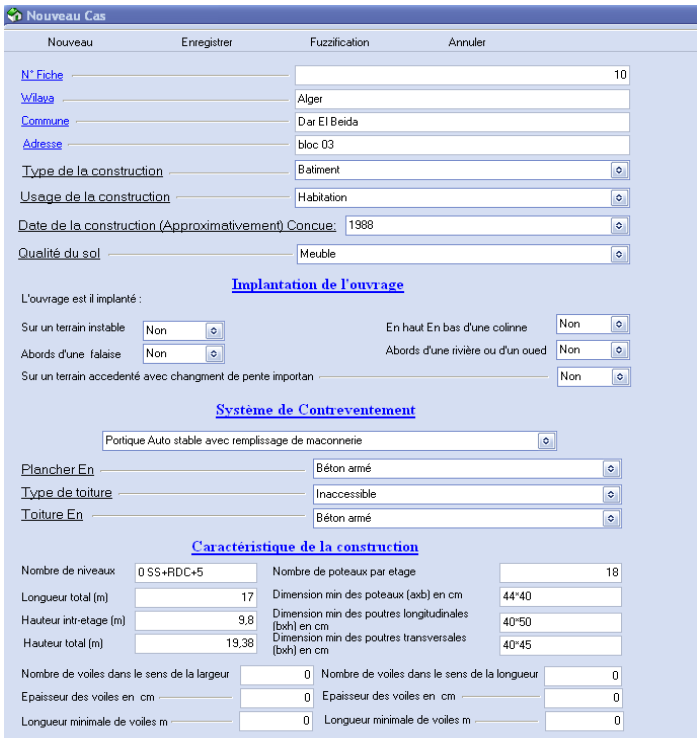


Figure (5) Ecran de saisi d'un nouveau cas

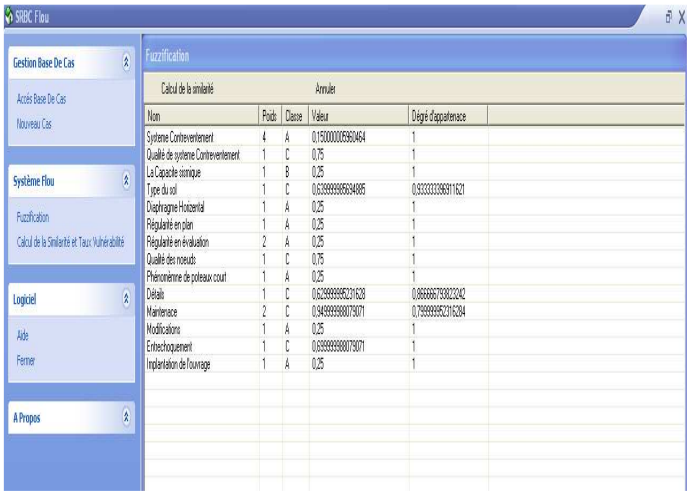


Figure 6: phase d'identification (précision de chaque paramètre avec son indice de vulnérabilité, sa classe, le degré d'appartenance à cette classe et son poids)

On confirme par la Figure 7 la phase de fuzzification :

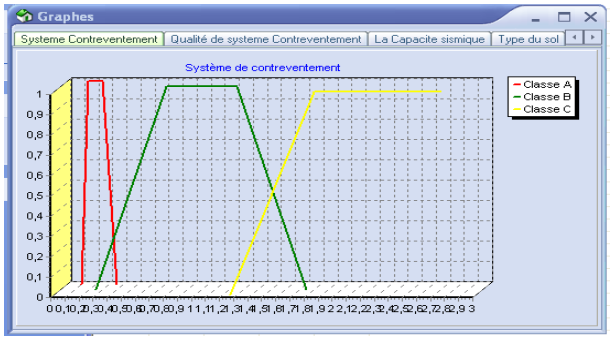


Figure 7 Fuzzification du paramètre système de contreventement

b) Calcul de similarité et adaptation

Dans cette étape, on a introduit le calcul de similarité et l'estimation de l'indice de vulnérabilité. La Figure 8 nous présente les similarités globales entre un cas et la bibliothèque de cas existante et l'estimation de sa vulnérabilité sismique.

Figure 8 : Similarité globale et Estimation de la vulnérabilité d'un cas

La figure 8 nous permet d'identifier les différentes similarités d'un nouveau cas avec les cas historiques ces derniers sont classés de la plus grande similarité à la plus petite, on prend celles qui ont une similarité >=0.8, on applique la formule d'adaptation (équation 6), et le résultat n'est en autre que l'estimation la vulnérabilité sismique du nouveau cas. Cette valeur

se rapproche considérablement de la vulnérabilité calculée par les experts du domaine.

5.2 Validation Empirique

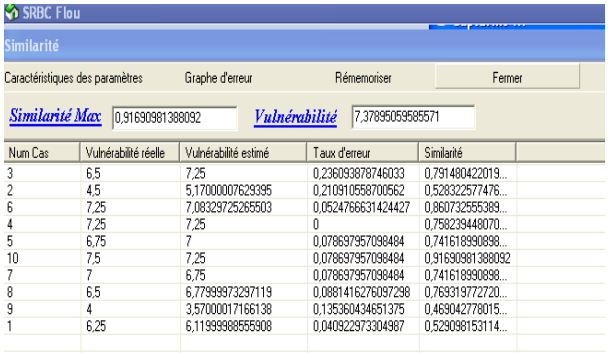
La validation empirique de notre système est évaluée selon deux critères la précision des estimations et la prise en charge de l'imprécision : l'imprécision est prise en charge par la flexibilité et la puissance des outils de la logique floue. En effet on évalue l'appartenance d'un paramètre à une classe par un degré d'appartenance et non par une valeur booléenne vraie ou fausse. Cette fonctionnalité est requise pour prendre en charge les valeurs incertaines et imprécises.

La Figure (9) montre que les estimations effectuées se rapprochent considérablement des vulnérabilités réelles des constructions. Ce qui avantage notre système en termes de précision des résultats.

Et la précision est évaluée à partir d'un taux d'erreur dont la formule est calculée par l'équation 7.

$$\text{Taux d'erreurs Cas i} = \frac{\left|(Ve - Vr)_i\right|}{\sum_{i=1}^{i=n}\left|(Ve - Vr)_i\right|} \quad (7)$$

Où Ve est la vulnérabilité estimée, Vr est la vulnérabilité réelle, et n est le nombre de cas de la base de cas. i est le i^{ème} cas



Num Cas	Vulnérabilité réelle	Vulnérabilité estimé	Taux d'erreur	Similarité
3	6,5	7,25	0,236093878746033	0,791480422019...
2	4,5	5,17000007629395	0,210910558700562	0,528322577476...
6	7,25	7,08329725265503	0,0524766631424427	0,860732555389...
4	7,25	7,25	0	0,758239448070...
5	6,75	7	0,078697957098484	0,741618990898...
10	7,5	7,25	0,078697957098484	0,916909813880 92
7	7	6,75	0,078697957098484	0,741618990898...
8	6,5	6,77999973297119	0,0881416276097298	0,769319772720...
9	4	3,57000017166138	0,135360434651375	0,469042778015...
1	6,25	6,11999988555908	0,040922973304987	0,529098153114...

Figure 9: taux d'erreur des estimations.

Pour confirmer ces résultats nous avons élaboré un graphe qui illustre les différences entre les valeurs réelles de la vulnérabilité sismique et les valeurs estimées.

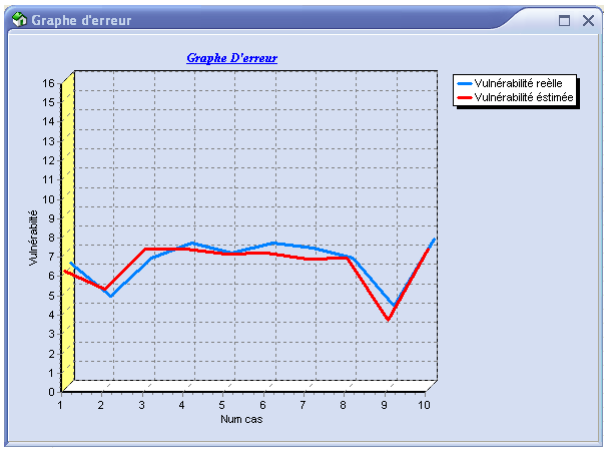


Figure 10 : Graphe d'erreur.

On confirme par les figures précédentes que notre système a pu en utilisant le RBC flou de prendre en charge l'estimation de la vulnérabilité sismique d'une construction, le retour d'expérience dans le domaine a pu nous faire gagner du temps pour estimer sans grand effort (utilisation des méthodes techniques fastidieuses) la vulnérabilité sismique. Dans les 10 cas présentés dans la figure 10 on remarque que pratiquement les estimations (en rouge) se rapprochent des valeurs réelles (bleu).

Vulnérabilité floue prend en charge aussi la gestion des incertitudes, l'incertitude a été vérifiée par rapport aux erreurs de mesurage des valeurs d'attributs.

Vulnérabilité floue a utilisé des valeurs linguistiques pour les quatorze paramètres, ceci permet de réduire les effets des erreurs commises lors du mesurage des attributs de vulnérabilité. Les erreurs de mesurage des valeurs linguistiques sont évaluées par la différence de degrés d'appartenance relatifs aux valeurs linguistiques. Prenons l'exemple de la capacité sismique : ce dernier est estimé avec trois ensembles flous bas, moyen et élevé :

La figure 11 présente plusieurs situations ou les erreurs de mesurage de l'attribut capacité sismique n'ont aucun effet sur les estimations des vulnérabilités résultantes de *Vulnérabilité floue*

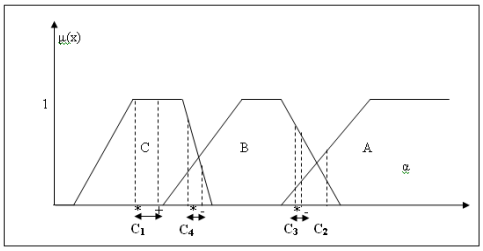


Figure 11: Exemples de situations où les erreurs de mesurage des attributs affectent (ou non) les estimations fournies par *Vulnérabilité floue.*

L'utilisation de quantificateurs linguistiques RIM pour l'évaluation de la similarité globale entre deux constructions permet aussi, dans certaines situations, d'éviter les effets des erreurs de mesurage des attributs sur l'estimation de la vulnérabilité sismique d'une construction.

En conclusion, les erreurs de mesurage sur un attribut peuvent être masquées (n'influent par sur l'estimation de la vulnérabilité) au cours des trois étapes de *Vulnérabilité floue*.

6. Conclusion

Dans ce modeste travail, nous avons conçu et réalisé un système dénoté *vulnérabilité floue*, permettant l'estimation de la vulnérabilité sismique d'une construction. Vu la variété des types de constructions en Algérie, on s'est limité dans notre étude à un seul type de construction, celles qui sont en béton armé. Ce système a été développé en utilisant le RBC flou. Le processus d'estimation de vulnérabilité passe par trois étapes :

- ✓ Phase d'identification ou de fuzzification : l'indice de vulnérabilité d'une construction n'est autre que la somme des indices de vulnérabilité de l'ensemble des paramètres influant sur la réponse sismique de la structure. Cet indice est estimé soit par des données observables ou calculables. Une fois obtenu il doit être apparenté à une classe de vulnérabilité (A, B, C). ***Vulnérabilité floue*** a représenté les valeurs linguistiques de cet indice par des ensembles flous « bas, moyen, élevé ». L'appartenance à une classe est évaluée par un degré d'appartenance. Cette fonctionnalité a permis la tolérance de l'imprécision.
- ✓ Evaluation de la similarité : Dans cette étape, nous avons utilisé des mesures de similarité spécifiques qui ont été développées spécialement pour un processus d'estimation . La similarité entre deux constructions est évaluée en deux étapes :
  - o Similarité individuelle : évalué au niveau de chaque attribut par $SV_j(C_1,C_2)$, en utilisant les techniques d'agrégation floues (max-min, som-produit).
  - o Similarité globale : évalué au niveau de l'ensemble des attributs constituant une construction, pour cela on a utilisé les quantificateurs linguistiques monotones croissants RIM
- ✓ Adaptation : Pour une bonne estimation de la vulnérabilité sismique d'une construction, il faudrait définir le seuil de similarité (dans notre cas (0,80)) sur le quel se basera le choix des constructions semblables à la construction courante. Pour cela on a utilisé la qualification étroitement similaire pour définir l'ensemble des constructions à inclure dans l'estimation de vulnérabilité de la construction courante. La qualification a été définie par un ensemble flou. L'estimation de la vulnérabilité a été évaluée par la moyenne arithmétique des indices de vulnérabilité des différentes constructions choisis par la qualification étroitement similaire.

La validation s'est faite en comparant les résultats obtenus par les experts du CTC et les résultats de notre système.

Cette comparaison de performance considère deux critères :

1. la tolérance des imprécisions lors de l'affectation des scores.
2. la précision des estimations

Les résultats obtenus ont montré qu' une précision de + 90% est constatée dans les estimations obtenues et que la tolérance des imprécisions est prise en charge lors des erreurs de mesurage des attributs influant sur la réponse sismique.

Notre système *vulnérabilité floue* prévoit de prendre en considération les différents types de constructions (maçonnerie, portique, mixte). Et ceci en organisant notre base de cas en utilisant comme critère de base de classification le type de construction. La méthode prévisible à être exploitée à cet effet est un des algorithmes de la classification supervisée floue par exemple l'ID3 flou qui a montré son efficacité dans les différents domaines d'application du RBC flou. Afin de permettre à notre modèle de s'adapter facilement à son environnement, nous prévoyons aussi d'intégrer des mécanismes d'apprentissage. A ce niveau il faudrait donner la possibilité à l'estimateur de changer dans les valeurs linguistiques, ce qui permettra d'être continuellement en cohérence avec leur environnement. Changer la pondération des paramètres, afin de permettre à notre système de prendre en charge n'importe quel type de construction et enfin changer la valeur linguistique étroitement similaire pour permettre de prendre en considération les nouvelles exigences sur la sélection des constructions qui vont contribuer à l'estimation de la vulnérabilité sismique d'une nouvelle construction.

Dans notre travail, on a assuré la gestion des incertitudes juste au niveau des erreurs de mesurage, il est intéressant d'introduire la théorie des possibilités de ZADEH pour prendre en charge les incertitudes au niveau des erreurs d'estimation.

Les résultats obtenus ont montré qu' une précision de + 90% est constatée dans les estimations obtenues et que la tolérance des imprécisions est prise en charge lors des erreurs de mesurage des attributs influant sur la réponse sismique.

Références

J. L. Kolodner. Case-Based Reasoning. Morgan Kaufmann Publishers, San Mateo, CA, 1993.
Aamodt, Agnar, et E.Plaza. Case-Based Reasoning: Foundational Issues, Methodological Variations and System Approaches. AI Communications, IOS Press, vol. 7, no 1, p. 39-59. 1994
.J. Godjevac, " Idées nettes sur la logique floue ", Presses polytechniques et universitaires romandes, Lausanne,1999.
R.J. Ku, Y.P. Kuo, Kai-Ying Chen Developing a diagnostic system through integration of fuzzy case-based reasoning and fuzzy ant colony system .Science direct 2005
Dvir, G.,Langholz, G.,&Schneider,M.(Matching attributes in a fuzzy case based reasoning. Fuzzy Information Processing Society , pp33–36, 1999.
M. Q. Xu, K. Hirota, and H. Yoshino, A fuzzy theoretical approach to representation and inference of case in CISG, International Journal of Artificial Intelligence and Law, vol. 7, no. 2–3, pp. 259–272, 1999.
Ya-jun Jiang, Jun Chen, Xue-yu Rua Fuzzy similarity-based rough set method for case-based reasoning and its application in tool selection. Science Direct 2006.
AFPS, Vulnérabilité Sismique du Bâtis Existant: approche d'ensemble, Cahier Technique n°25, France, 2005
F-I-Belheouane, Détermination de l'indice de vulnérabilité pour les bâtiments en béton armé, Thèse de Magister, Université Saad Dahleb de Blida, Algérie, 2006.
MARIE-JOSÉ NOLLET, Évaluation de la vulnérabilité sismique des bâtiments existants, État des connaissances -2004
A.Idri, A. Abran,." La logique floue appliquée aux modèles d'estimation d'effort de développement de logiciels-cas du modèle COCOMO'81" , IEEE Computer Society ,2004.
R.Yager « Induced OWA Aggregation in Case Based Reasoning" disponible depuis

09/07/2007 sur http://www.aic.nrl.navy.mil/papers/2001/AIC-01-003/ws5/ws5toc5.pdf
A.Idri, et A.Abran. 2001a. «A Fuzzy Logic Based Measures For Software Project Similarity: Validation and Possible Improvements», *Proceedings of the 7th International Symposium on Software Metrics*, avril, Londres, IEEE Computer Society, p. 85-96, 2001.
F.Yépez, A.H.Barbat, J.A.Canas,A Method to Perform Computer Simulations of Damage in Buildings for Seismic Risk Evaluation, Faculty of Civil Engineering, Technical University of Catalonia, Barcelona, Spain,1996.
L.A Zadeh, Fuzzy sets, information control, vol.8,pp. 338-353 1965.

Summary

The main object of our work is the development and the validation of a system indicated *Fuzzy Vulnerability.* This system estimates a construction seismic vulnerability. *Fuzzy Vulnerability* uses a fuzzy representation in order to tolerate the imprecision during the description of a construction. For the circumstance, *Fuzzy Vulnerability* used the fuzzy case based reasoning (FCBR). Our system forwards by three principals' phases: the first stapes consists in describing the parameters considered to be necessary to the construction seismic vulnerability estimation. The influence of these parameters on the seismic answer of construction is evaluated by linguistic values. *Fuzzy vulnerability* uses a fuzzy representation in order to tolerate the imprecision on the level of the construction description. At the second phase, we evaluated the similarity between new construction and the whole of the historical cases. This similarity is evaluated on two levels: 1) individual similarity: bases on the fuzzy techniques of aggregation (max-min); 2) global similarity: uses the increasing monotonous linguistic quantifiers (RIM) to combine the various individual similarities between two constructions. The third phase of the estimation process of *Fuzzy Vulnerability* consists in using vulnerabilities of historical constructions narrowly similar to current construction to deduce its estimate vulnerability. We validated our system by using 50 cases of the CTC of Tlemcen and Blida. For that we evaluated the performances of *Fuzzy Vulnerability* on the basis of two basic criteria: the precision of estimations and the tolerance of the imprecision all along the process of estimation. The comparison was done with estimations made by tiresome and long models. The results are satisfactory.

Identification du type de diabète par une approche cellulo-floue

Abdelkader Beldjilali, Baghdad Atmani

Equipe de recherche « Simulation, Intégration et Fouille de données »
Département Informatique, Faculté des Sciences, Université d'Oran
BP 1524, El M'Naouer, Es Senia, 31 000 Oran, Algérie
beldjilali.abdelkader@gmail.com
atmani.baghdad@univ-oran.dz

Résumé. La classification par apprentissage inductif trouve son originalité dans le fait que souvent les humains l'utilisent pour résoudre et pour manipuler des situations très complexes dans leurs vies quotidiennes. Cependant, l'induction chez les humains est souvent approximative plutôt qu'exacte. En effet, le cerveau humain est capable de manipuler des informations imprécises, vagues, incertaines et partielles. Le cerveau humain est, également, capable d'apprendre et d'opérer dans un contexte où la gestion de l'incertitude est indispensable. Dans cet article, nous proposons la conception et l'expérimentation d'un Moteur d'Inférence Cellulaire de gestion de données floues, baptisé FLOU-CIE, qui utilise les caractéristiques de la classification par apprentissage inductif artificiel. Le système FLOU-CIE doit tolérer les imprécisions tout au long de son processus d'Extraction des Règles de classification à partir des données, gérer les incertitudes et permettre le traitement des situations complexes rencontrées dans sa tâche de classification.

Mots-clés : Arbre de décision, Apprentissage automatique, Extraction de règles, Fouille de données, Logique floue, Modélisation booléenne.

1 Introduction

L'extraction des connaissances à partir de données (ECD) est une nouvelle préoccupation qui est apparue dans la recherche informatique (Kodratoff, 1997). Outre, la découverte de nouvelles règles de classification dans différents domaines, une application des techniques d'extraction de règles peut être une contribution dans la réalisation des Systèmes Experts (Atmani et Beldjilali, 2007a). En effet, le coût de réalisation d'un Système Expert en utilisant des techniques traditionnelles d'acquisition des connaissances, à partir d'experts humains, constitue un obstacle à son utilisation et à sa large diffusion. Une solution à ce problème est de concevoir des programmes informatiques capables d'apprendre et de découvrir leurs propres connaissances à partir de cas pratiques (exemples). Dans ce type de système, l'expertise n'est plus fournie par l'expert humain, mais doit être construite à partir de données dont

on dispose sur le domaine. L'ECD est un processus complexe qui se déroule suivant une série d'opérations (Fayyad et al, 1996). Nous pouvons regrouper ces opérations en trois étapes majeures: 1) préparation des données, 2) construction des règles de classification par fouille de données (étape centrale de l'ECD) et enfin 3) la validation du modèle ainsi élaboré.

La problématique de la classification consiste à affecter les différentes observations à des catégories ou classes prédéfinies (Duhamel et al., 2001) . En général les méthodes de classification sont constituées de plusieurs étapes. L'étape la plus importante consiste à élaborer les règles de classification à partir des connaissances disponibles à priori; il s'agit de la phase d'apprentissage.

La classification par apprentissage inductif trouve son originalité dans le fait que souvent les humains l'utilisent pour résoudre et pour manipuler des situations très complexes dans leurs vies quotidiennes (Zadeh, 1994). Cependant, l'induction chez les humains est souvent approximative plutôt qu'exacte. Le cerveau humain étant capable de manipuler des informations imprécises, incertaines est apte à apprendre et à opérer dans un contexte où la gestion de l'incertitude est indispensable. Dans cet article, nous proposons la conception d'un modèle cellulaire de classification de données floues, qui s'inspire des caractéristiques de la classification par apprentissage inductif chez les humains (Zighed et Rakotomalala, 2000). Il doit donc tolérer les imprécisions tout au long de son processus d'ECD, gérer les incertitudes et permettre l'apprentissage inductif afin de pouvoir traiter les situations complexes rencontrées dans sa tâche de classification.

Cet article est structuré selon deux sections. La section 2 est consacrée à la classification par apprentissage inductif à partir de données et en particulier à l'induction des règles par graphe d'induction. La section 3 aborde l'élaboration du modèle de classification par automate cellulaire. La logique floue est explicitée dans la section 4. Enfin, nous présenterons notre contribution dans la section 5.

2 Classification par apprentissage inductif

Etablir un diagnostic d'identification du diabète dans le domaine médical, signifie être capable d'associer le type du diabète à un certain nombre de symptômes présentés par des malades diabétiques. On repère, dans ce type de problème, trois objets essentiels : les patients, les types du diabète et les symptômes. Les patients forment la population, les symptômes sont les descriptions des patients et les types du diabète sont les classes. Notre contribution consiste en la classification des patients diabétiques selon leurs symptômes. On suppose qu'il existe un classement correct, c'est-à-dire qu'il existe une application qui associe à tout patient un type de diabète. Apprendre à établir un diagnostic, c'est associer un type de diabète à une liste de symptômes de telle manière que cette association corresponde au classement défini ci-dessous. Pour formaliser ces propos, nous utiliserons ces notations : $\Omega=\{\omega_1,\omega_2,\ldots,\omega_n\}$ pour désigner une population de n patients diabétiques. $\Gamma=\{\gamma_1,\gamma_2,\ldots,\gamma_d\}$ pour l'ensemble des d descriptions et $\Theta=\{\theta_1,\theta_2,\ldots,\theta_m\}$ pour l'ensemble des m classes.

Soit Ω une population d'individus concernés par le problème de classification. A cette population est associé un attribut particulier appelé attribut classe noté Y. La variable Y est appelée dans le domaine de la statistique variable endogène ou simplement classe. A chaque individu ω peut être associée sa classe $Y(\omega)$. On dit que la fonction Y prend ses valeurs dans l'ensemble des étiquettes Θ, appelé également ensemble des classes. Dans notre cas la population Ω est celle des patients diabétiques et Y le résultat de l'identification du diabète type 1

noté θ_1, et type 2 noté θ_2; alors $Y(\omega)$ sera le résultat de l'identification du type de diabète du patient ω (Atmani et Beldjilali, 2007a et b).

La détermination du modèle de classification φ est liée à l'hypothèse selon laquelle les valeurs prises par la variable Y ne relèvent pas du hasard, mais de certaines situations particulières que l'on peut caractériser (Zighed et Rakotomalala, 2000). Pour cela le Diabétologue, l'expert du domaine concerné, établit une liste à priori de p variables statistiques appelées variables exogènes et notées $X=\{X_1,X_2,\ldots,X_p\}$. Ces variables sont également appelées attributs prédictifs ou attributs explicatifs. La valeur prise par une variable exogène X_j est appelée modalité ou encore valeur de l'attribut X_j du patient ω. Nous désignons par l_j le nombre de modalité qu'une variable X_j peut recevoir. Pour illustrer cette notation, considérons la description par quatre variables exogènes d'un patient diabétique :

X_1 : Antécédent, qui permet de préciser les antécédents du diabète et qui peut prendre trois modalités X_{11}=famille, X_{12}=personnel ou X_{13}=aucun.

X_2 : Poids, qui indique le poids du patient et qui peut prendre quatre modalités X_{21}=maigre, X_{22}=normal, X_{23}=obèse ou X_{24}=surchargé.

X_3 : Age, qui indique l'âge du patient et qui peut prendre des valeurs continues.

X_4 : Infection virale, qui peut prendre deux valeurs X_{41}=oui ou X_{42}=non. « oui » correspond à l'existence d'une infection virale (grippe, oreillon, ...).

L'objectif de l'apprentissage inductif est de rechercher un modèle de classification φ, permettant, pour un nouveau patient ω, pour lequel nous ne connaissons pas la classe $Y(\omega)$ mais dont nous connaissons l'état de toutes ses variables exogènes, de prédire cette valeur grâce à φ. La mise au point de φ nécessite de prélever dans la population Ω deux échantillons notés Ω_A et Ω_T. Le premier dit d'apprentissage servira à la construction de φ et le second dit de test servira à tester la validité de φ. Ainsi, pour tout patient ω, nous supposons connues à la fois ses valeurs $X(\omega)$ dans l'espace de représentation et sa classe $Y(\omega)$ dans l'espace des étiquettes Θ.

La population Ω_A des patients diabétiques, prise en compte pour une classification n'est autre qu'une suite de n patients ω_i (exemples) avec leur classe correspondante $Y(\omega_i)$. Supposons que l'échantillon Ω_A est composé de 14 observations et qu'il s'agit d'expliquer le comportement des patients, âgés de plus de 60 ans, par rapport au type du diabète (θ_1=*oui*, pour type 1 et θ_2=*non*, pour type 2) à partir des observations listées dans la table 1.

Ω	$Y(\omega)$	$X_1(\omega)$	$X_2(\omega)$	$X_3(\omega)$	$X_4(\omega)$	Ω	$Y(\omega)$	$X_1(\omega)$	$X_2(\omega)$	$X_3(\omega)$	$X_4(\omega)$
ω_1	oui	famille	75	70	oui	ω_8	oui	perso	64	65	Oui
ω_2	non	famille	80	90	oui	ω_9	oui	perso	81	75	Non
ω_3	non	famille	85	85	non	ω_{10}	non	aucun	71	80	Oui
ω_4	non	famille	72	95	non	ω_{11}	non	aucun	65	70	Oui
ω_5	oui	famille	69	70	non	ω_{12}	oui	aucun	75	80	Non
ω_6	oui	perso	72	90	oui	ω_{13}	oui	aucun	68	80	Non
ω_7	oui	perso	83	78	non	ω_{14}	oui	aucun	70	96	Oui

TAB. 1 – *Exemple d'un échantillon d'apprentissage*

L'apprentissage inductif supervisé se propose donc de fournir des outils permettant d'extraire, à partir de l'information dont on dispose sur l'échantillon d'apprentissage, le modèle de classification φ. Ce modèle peut prendre la forme d'un réseau de neurones (φ^{RN}) (Atmani et Beldjilali, 2007b), d'un graphe d'induction (φ^{GI}) (Zighed, 1996) ou d'un automate

cellulaire (φ^{AC}) (Atmani et Beldjilali, 2007a). Cela constitue l'essentiel de notre contribution. Le processus général d'apprentissage inductif comporte généralement trois étapes que nous récapitulons ci-dessous :

1. **Elaboration du modèle :** C'est l'étape qui fait appel à un échantillon d'apprentissage noté Ω_A, dont tous les individus ω_i sont décrits dans un espace de représentation et appartiennent à l'une des *m* classes notées θ_j, j=1,…,m. Il s'agit de construire l'application φ qui permet de déterminer la classe à partir de la description.
2. **Validation du modèle :** Il s'agit de vérifier, sur un échantillon test Ω_T pour lequel nous connaissons pour chacun de ses individus, la représentation et la classe, sous réserve que le modèle de classification φ issue de l'étape précédente donne bien la classe attendue.
3. **Généralisation du modèle :** c'est l'étape qui consiste à étendre l'application du modèle à tous les individus de la population Ω.

3 Apprentissage supervisé par induction booléenne

Dans cette section, nous présentons les principes de construction, par modélisation booléenne, des graphes d'induction dans les problèmes de discrimination et de classification (Atmani et Beldjilali, 2007a) : on veut expliquer la classe prise par une variable à prédire catégorielle ***Y***, dite attribut classe ou variable endogène; à partir d'une série de variables X_1, X_2, ..., X_p, dites variables prédictives (descripteurs) ou exogènes, discrètes ou continues. Selon la terminologie de l'apprentissage automatique, nous nous situons donc dans le cadre de l'apprentissage supervisé. Nous n'aborderons pas les autres types d'utilisation que sont les arbres de régression où il s'agit d'un problème de prédiction mais la variable à prédire est continue Lefebure et Venturi (2001) ; et les arbres de classification (Breiman et al., 1984), où l'objectif est de construire des groupes homogènes dans l'espace de descripteurs.

Le processus général d'apprentissage que le système cellulaire *WCSS* (Benamina et Atmani, 2008) applique à une population est organisé sur quatre étapes (voir figure 1) :

1. Acquisition et préparation des données par Weka qui consiste à utiliser les différentes techniques de prétraitement des données déjà intégrées dans l'environnement Weka ;
2. Elaboration du Modèle φ^{AC} par ACSIPINA qui se résume sur 4 étapes :
   a. Initialisation du graphe d'induction par automate cellulaire (coopération *COG* et *CIE*) ;
   b. Génération des règles de production (coopération *COG* et *CIE*) ;
   c. Validation des règles cellulaires (coopération *CV* et *CIE*) ;
3. Validation du Modèle φ^{AC} par Weka qui consiste à exploiter toutes les méthodes de visualisation et d'analyse déjà intégrées dans la plate forme Weka.

3.1 Construction du graphe d'induction

A partir de l'échantillon Ω_A nous commençons le traitement symbolique pour la construction du graphe d'induction (méthode *SIPINA* (Zighed, 1996) (Zighed et al, 2000)).

1) Choisir la mesure d'incertitude (Shannon ou quadratique) ;
2) Initialiser les paramètres λ, μ et la partition initiale S_0 ;

3) Appliquer la méthode *SIPINA* pour passer de la partition S_t à S_{t+1} et générer le graphe d'induction.
4) Enfin, génération des règles de prédiction.

Les paramètres λ, μ, les partitions et toutes les autres notions utilisées dans ce processus, sont présentés à l'aide d'exemples dans la référence (Zighed et al, 2000).

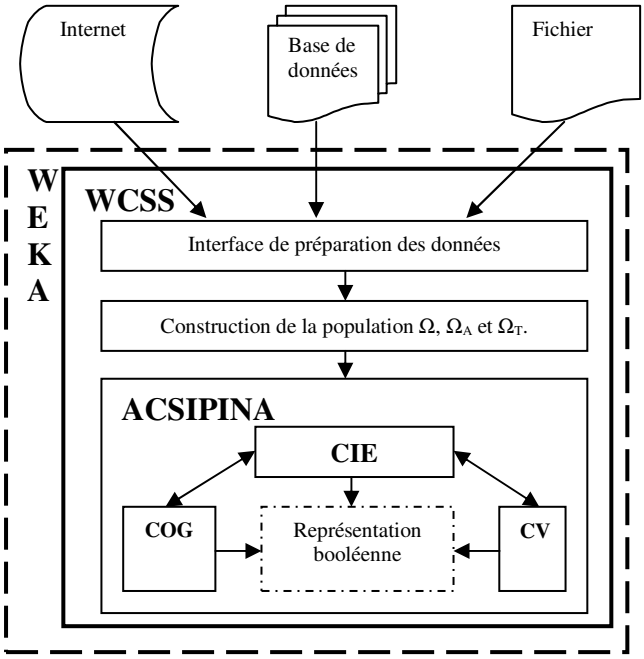


FIG. 1 – *Diagramme général du système cellulaire WCSS*

3.1.1 Définition d'une partition

L'algorithme de la méthode *SIPINA* (Zighed, 1996) est une heuristique non arborescente pour la construction d'un graphe d'induction. Son principe consiste à générer une succession de partitions par fusion et/ou éclatement des nœuds du graphe. L'objectif est d'optimiser un critère τ_λ. Dans ce qui suit nous allons décrire le déroulement du processus sur l'exemple fictif de la table 1. Supposons que notre échantillon d'apprentissage Ω_A se compose de 14 patients diabétiques qui se répartissent en deux classes *oui* et *non* (voir la table 1).

La partition initiale S_0 comporte un seul élément noté s_0, qui regroupe tout l'échantillon d'apprentissage avec 9 patients appartenant à la classe *oui* et 5 appartenant à la classe *non*. La partition suivante S_1 est engendrée par la variable X_1 et les individus dans chaque nœud s_i sont définis comme suit : $s_1=\{\omega\in\Omega_A|X_1(\omega)=\text{famille}\}$, $s_2=\{\omega\in\Omega_A|X_1(\omega)=\text{perso}\}$ et $s_3=\{\omega\in\Omega_A|X_1(\omega)=\text{aucun}\}$. De même dans le nœud s_0, on distingue dans s_1, s_2 et s_3, les individus des classes *oui* et *non*. La figure 2 résume les étapes de construction de s_0, s_1, s_2 et s_3.A partir de la partition S_1, le processus est réitéré à la recherche d'une partition S_2 qui serait meilleure selon certains critères qu'on peut consulter dans les références (Zighed et Rakotomalala, 2000).

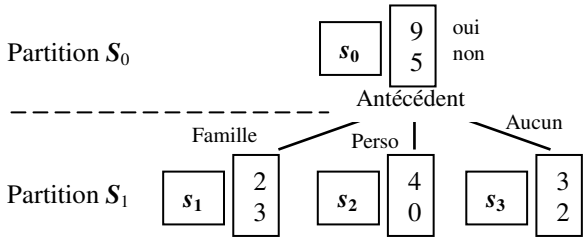


FIG. 2 – *Construction des nœuds s_0, s_1, s_2 et s_3*

3.1.2 Génération des règles

Regardons le graphique de la figure 3 comme s'il s'agissait d'un résultat final sans se préoccuper de vérifier dans le détail tous les calculs qui ont conduit à ce graphe. Sous réserve que notre échantillon Ω_A soit représentatif de la population originelle nous pouvons donc, pour alimenter la base de connaissance, extraire cinq règles R_1, R_2, R_3, R_4 et R_5 de prédiction qui sont de la forme :
Si *Condition* Alors *Conclusion*.

- R_1 : Si (X_1=famille) et ($X_2 < 77{,}5$) Alors *oui*.
- R_2 : Si (X_1=famille) et ($X_2 \geq 77{,}5$) Alors *non*.
- R_3 : Si (X_1=perso) Alors *oui*.
- R_4 : Si (X_1=aucun) et (X_4=*oui*) Alors *non*.
- R_5 : Si (X_1=aucun) et (X_4=*non*) Alors *oui*.

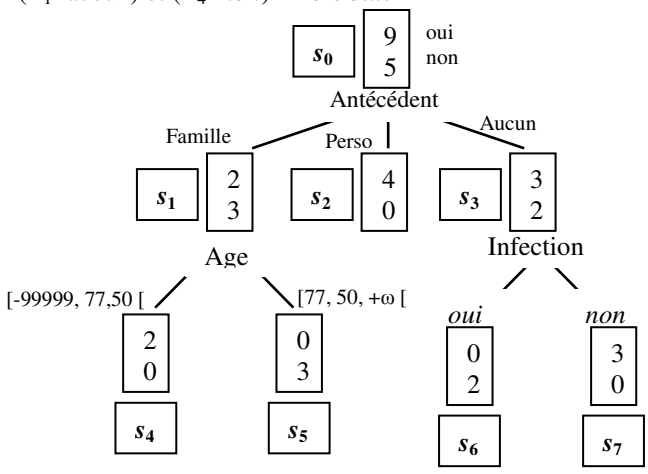


FIG. 3 – *Graphe d'induction final généré par SIPINA*

3.2 Moteur d'inférence cellulaire « CIE »

Les automates cellulaires (Wolfram, 2002) ont été introduits dans les années 1950 pour modéliser la vie artificielle. Depuis lors, ils ont été utilisés pour modéliser divers phénomènes en physique, biologie, ... mais aussi en informatique où ils apparaissent comme un modèle bien formalisé du parallélisme massif. Un automate cellulaire est une grille compo-

sée de cellules qui changent d'état dans des étapes discrètes. Après chaque étape, l'état de chaque cellule est modifié selon les états de ses voisines calculés dans l'étape précédente. Les cellules sont mises à jour d'une manière synchrone, et les transitions sont effectuées dans la théorie simultanément. En appliquant des règles simples et des transitions spécifiques, un automate cellulaire peut effectuer d'une manière globale, des opérations complexes.

Le module *CIE* (*Cellular Inference Engine*) simule le fonctionnement du cycle de base d'un moteur d'inférence en utilisant deux couches finies d'automates finis. La première couche, *CELFACT*, pour la base des faits et, la deuxième couche, *CELRULE*, pour la base de règles (Atmani et Beldjilali, 2007a). Chaque cellule au temps *t+1* ne dépend que de l'état des ses voisines et du sien au temps *t*. Dans chaque couche, le contenu d'une cellule détermine si et comment elle participe à chaque étape d'inférence: à chaque étape, une cellule peut être active (1) ou passive (0), c'est-à-dire participe ou non à l'inférence. Le principe est simple:

- Toute cellule *i* de la première couche *CELFACT* est considérée comme fait établi si sa valeur est 1, sinon, elle est considérée comme fait à établir.
- Toute cellule *j* de la deuxième couche *CELRULE* est considérée comme une règle candidate si sa valeur est 1, sinon, elle est considérée comme une règle qui ne doit pas participer à l'inférence.

Nous supposons qu'il y a *l* cellules dans la couche *CELFACT*, et *r* cellules dans la couche *CELRULE*. Les états des cellules se composent de trois parties : *CELFACT_E*, _I *et* _*S*, respectivement *CELRULE_E*, *_I* et *_S*, sont l'entrée, l'état interne et la sortie d'une cellule de *CELFACT*, respectivement d'une cellule de *CELRULE*. L'état interne, I d'une cellule de *CELFACT* indique le coefficient d'incertitude: *CELFACT*_I=0 correspond à un fait incertain, *CELFACT*_I=1 correspond à un fait certain. Pour une cellule de *CELRULE*, l'état interne _I peut être utilisé comme coefficient de probabilité que nous pouvons affecter à une règle. En effet et comme vous allez le voir, notre contribution concernera la représentation et la gestion des coefficients d'incertitudes.

Pour illustrer l'architecture et le principe de fonctionnement du module *CIE*, nous considérons la partie du graphe, extraite de la figure 3, obtenue en utilisant les partitions $S_0=(s_0)$, $S_1=(s_1, s_2, s_3)$ et $S_2=(s_4, s_5)$. La figure 4 montre comment la base de connaissance extraite à partir de ce graphe est représentée par les couches *CELFACT* et *CELRULE*. Initialement, toutes les entrées des cellules dans la couche *CELFACT* sont passives (E=0), excepté celles qui représentent la base des faits initiale (E=1).

Dans la figure 5 sont respectivement représentées les matrices d'incidence d'entrée R_E et de sortie R_S de l'automate.

- la relation d'entrée, notée i R_E j, est formulée comme suit : $\forall$ i $\in$ [1,l], $\forall$ j$\in$ [1,r], si (le Fait i $\in$ à la Prémisse de la règle j) alors $R_E(i, j) \leftarrow 1$.
- la relation de sortie, notée i R_S j, est formulée comme suit : $\forall$i $\in$ [1,l], $\forall$ j$\in$ [1,r], si (le Fait i $\in$ à la Conclusion de la règle j) alors $R_S(i, j) \leftarrow 1$.

Les matrices d'incidence R_E et R_S représentent la relation *entrée/sortie* des faits et sont utilisées en *chaînage avant*. On peut également utiliser R_S comme relation d'entrée et R_E comme relation de sortie pour lancer une inférence en chaînage arrière. Notez qu'aucune cellule du voisinage d'une cellule qui appartient à *CELFACT* (respectivement à *CELRULE*) n'appartient à la couche *CELFACT* (respectivement à *CELRULE*).

ARC_1 : Si s_0 Alors (X_1=famille) et s_1.
ARC_2 : Si s_0 Alors (X_1=perso) et s_2.
ARC_3 : Si s_0 Alors (X_1=aucun) et s_3.
ARC_4 : Si s_1 Alors ($X_2 < 77{,}5$) et s_4.
ARC_5 : Si s_1 Alors ($X_2 \geq 77{,}5$) et s_5.

CELFACT	*E*	*I*	*S*
s_0	1	1	0
X_1=famille	0	1	0
s_1	0	1	0
X_1=pesro	0	1	0
s_2	0	1	0
X_1=aucun	0	1	0
s_3	0	1	0
$X_2 < 77{,}5$	0	1	0
s_4	0	1	0
$X_2 \geq 77{,}5$	0	1	0
s_5	0	1	0

CELRULE	*E*	*I*	*S*
ARC_1	0	1	0
ARC_2	0	1	0
ARC_3	0	1	0
ARC_4	0	1	0
ARC_5	0	1	0

FIG. 4 – *Configuration initiale de l'automate cellulaire*

R_E	ARC_1	ARC_2	ARC_3	ARC_4	ARC_5
s_0	1	1	1	0	0
X_1=famille	0	0	0	0	0
s_1	0	0	0	1	1
X_1=pesro	0	0	0	0	0
s_2	0	0	0	0	0
X_1=aucun	0	0	0	0	0
s_3	0	0	0	0	0
$X_2 < 77{,}5$	0	0	0	0	0
s_4	0	0	0	0	0
$X_2 \geq 77{,}5$	0	0	0	0	0
s_5	0	0	0	0	0

R_S	ARC_1	ARC_2	ARC_3	ARC_4	ARC_5
s_0	0	0	0	0	0
X_1=famille	1	0	0	0	0
s_1	1	0	0	0	0
X_1=pesro	0	1	0	0	0
s_2	0	1	0	0	0
X_1=aucun	0	0	1	0	0
s_3	0	0	1	0	0
$X_2 < 77{,}5$	0	0	0	1	0
s_4	0	0	0	1	0
$X_2 \geq 77{,}5$	0	0	0	0	1
s_5	0	0	0	0	1

FIG. 5 – *Les matrices d'incidences d'Entrée/Sorties*

La dynamique de l'automate cellulaire *CIE*, pour simuler le fonctionnement du *Moteur* d'*Inférence Cellulaire*, utilise deux fonctions de transitions δ_{fact} et δ_{rule}, où δ_{fact} correspond à la phase d'*évaluation*, de *sélection* et de *filtrage*, et δ_{rule} correspond à la phase d'*exécution* (Atmani et al, 2007a). Pour définir les deux fonctions de transition nous allons adopter la notation suivante : EF, IF et SF pour désigner *CELFACT*_E, _I et _S ; Respectivement ER, IR et SR pour désigner *CELRULE*_E, _I et _S.
La fonction de transition δ_{fact}:

$$(EF, IF, SF, ER, IR, SR) \xrightarrow{\delta_{fact}} (EF, IF, EF, ER + (R_E^T \cdot EF), IR, SR)$$

La fonction de transition δ_{rule}:

$$(EF, IF, SF, ER, IR, SR) \xrightarrow{\delta_{rule}} (EF + (R_S \cdot ER), IF, SF, ER, IR, \overline{ER})$$

Où la matrice R_E^T désigne la transposé de R_E et où $\overline{ER}$ désigne la négation logique du vecteur *ER*. Les opérateurs + et · utilisés sont respectivement le OU et le ET logiques.

Nous considérons G_0 la configuration initiale de l'automate cellulaire (voir la figure 5) et, la fonction $\Delta=\delta_{rule}o\delta_{fact}$ la fonction de transition globale : $\Delta(G_0)=G_1$ si $\delta_{fact}(G_0)= G'_0$ et $\delta_{rule}(G'_0)=G_1$. Supposons que $G=\{G_0, G_1, ...,G_q\}$ est l'ensemble des configurations de l'automate cellulaire. L'évolution discrète de l'automate, d'une génération à une autre, est définie par la séquence $G_0, G_1, ..., G_q$, où $G_{i+1}=\Delta(G_i)$.

4 Pourquoi la logique floue ?

4.1 Définition

La logique floue est née de la constatation que la plupart des phénomènes ne peuvent pas être représentés à l'aide de variables booléennes qui ne peuvent prendre que deux valeurs (0 ou 1). Peut-on considérer une eau à 18° comme étant chaude ou froide ? N'est elle pas ni vraiment chaude, ni vraiment froide mais tous simplement tiède ? Pour répondre à ce type de question, la logique floue considère la notion d'appartenance d'un objet à un ensemble non plus comme une fonction booléenne mais une fonction qui peut prendre toutes les valeurs entre 0 et 1**.**

Selon Zadeh, la logique floue est la théorie des ensembles flous. La théorie des ensembles flous est une théorie mathématique dont l'objectif principal est la modélisation des notions vagues et incertaines du langage naturel. Ainsi, elle évite les inadéquations de la théorie des ensembles classiques quant au traitement de ce genre de connaissances. La caractéristique fondamentale d'un ensemble classique est la frontière rigide entre deux catégories d'éléments: ceux qui appartiennent à l'ensemble et ceux qui n'appartiennent pas à cet ensemble; ils appartiennent plutôt à son complémentaire. La relation d'appartenance est représentée dans ce cas par une fonction μ qui prend des valeurs de vérité dans {0,1}. Ainsi, la fonction d'appartenance d'un ensemble classique A est définie par :

$$\mu_A(x)=\begin{cases}1 & si\ x\in A\\ 0 & si\ x\notin A\end{cases}$$

Cela signifie qu'un élément *x* est soit dans A ($\mu_A(x)=1$) ou non ($\mu_A(x)=0$). Or dans plusieurs situations, il est parfois ambigu que *x* appartienne ou non à A.

4.2 Exemple

Pour mettre en évidence le principe fondamental de la logique floue, nous allons présenter l'exemple qui permet de classer les patients diabétiques en trois ensembles «jeune», «entre-deux-âges», et enfin «âgé». Selon la logique classique (logique de Boole), qui n'admet pour les variables que les deux valeurs 0 et 1, une telle classification pourrait se faire comme la figure 6. Toutes les personnes âgées de mois de 25 ans sont alors considérées jeunes et toutes les personnes âgées de plus de 50 ans comme vieux.

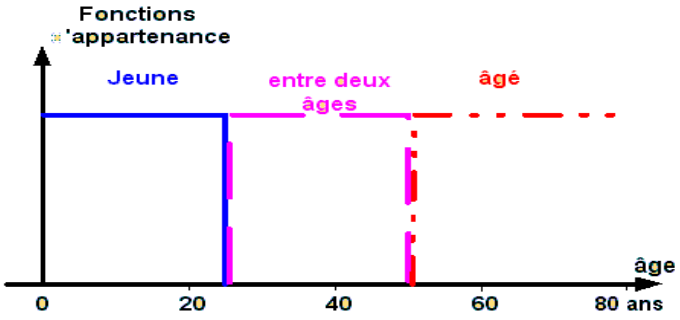


FIG. 6 – *Classification d'âges selon la logique classique.*

Cependant, une telle logique de classification n'est même pas logique. Pourquoi une personne, lorsqu'elle a eu 50 ans, doit-elle être considérée comme appartenant à l'ensemble âgé? En réalité, un tel passage se fait progressivement et individuellement. La logique floue, dont les variables peuvent prendre n'importe quelles valeurs entre 0 et 1, permet de tenir compte de cette réalité. Les limites ne varient pas soudainement, mais progressivement. La figure 7 montre une classification possible pour l'exemple précédent, cette fois-ci à l'aide de la logique floue. Ainsi une personne de 25 ans appartient à l'ensemble «jeune» avec une valeur μ=0.75 de la fonction d'appartenance et à l'ensemble «entre deux âges» avec μ=0.25. Par contre une personne âgée de 65 ans appartient avec une valeur μ=1 de la fonction d'appartenance à l'ensemble «âgé».

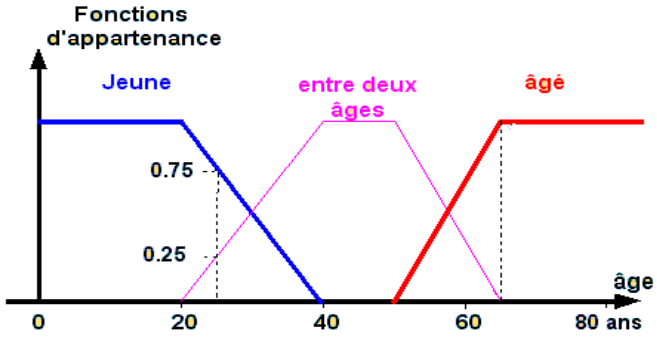


FIG. 7 – *Classification d'âges selon la logique floue.*

5 Architecture générale du système Flou-CIE

Comme tout système expert, Flou_CIE est composé de quatre principaux modules. Chaque module de Flou_CIE se distingue avec ses caractéristiques structurelles et fonctionnelles. L'architecture de notre Flou_CIE est illustrée par la figure 8 ci-dessous : 1) Une base de connaissance ; 2) Une interface de fuzzification cellulaire ; 3) Un moteur d'inférence cellulaire « CIE » et 4) Une interface de défuzzification cellulaire.

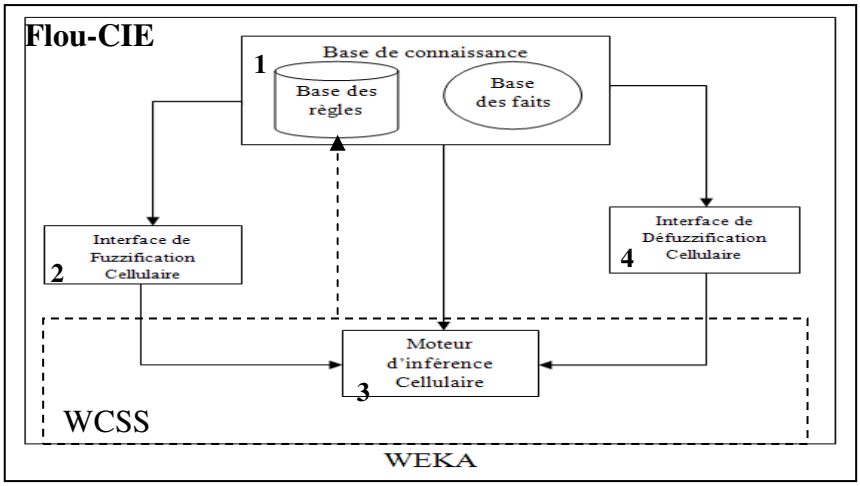


FIG. 8 – *Architecture du système Flou_CIE.*

5.1 Base de connaissance

La BC est l'élément capital d'un SE qui contient la représentation des connaissances de l'expert. Elle est constituée d'une base des faits et d'une base des règles. Pour illustrer nous considérons la base des faits suivante :

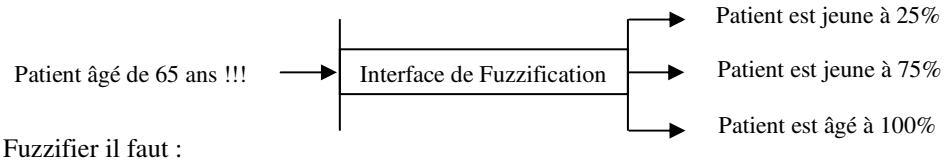
F1 : X_1=famille (0.9).
F2 : X_1=perso (0.18).
F3 : X_1=aucun (0.675).
F4 : $X_2 < 77{,}5$ (0.045).
F5 : $X_2 \geq 77{,}5$ (0.6).
F6 : X_4=*oui* (0.5).
F7 : X_4=*non* (0.5).

Dans cette BF, le fait X_1=famille est certain à 90 % (avec un CF=0,9) et ainsi de suite pour les autres faits. De même, la base des règles contient les connaissances expertes sous forme de règles qui symbolisent les raisonnements effectuées par les experts. Les règles sont exécutées les unes à la suite des autres afin de créer des enchainements de raisonnements. Ci-dessous un exemple de base des règles utilisées par Flou_CIE :

R_1 : Si (X_1=famille) et ($X_2 < 77{,}5$) Alors *oui*. (0.2)
R_2 : Si (X_1=famille) et ($X_2 \geq 77{,}5$) Alors *non*. (0.75)
R_3 : Si (X_1=perso) Alors *oui*. (0.05)
R_4 : Si (X_1=aucun) et (X_4=*oui*) Alors *non*. (0.5)
R_5 : Si (X_1=aucun) et (X_4=*non*) Alors *oui*. (0.5)

5.2 Interface de Fuzzification cellulaire

Le système Flou_CIE traite des variables d'entrées floues et fournit des résultats sur des variables de sorties elles-mêmes floues. La Fuzzification, illustrée par l'exemple suivant, est l'étape qui consiste à la quantification floue des valeurs réelles d'une variable.



Pour Fuzzifier il faut :

L'univers du discours, c'est-à-dire une Plage de variations possibles de l'entrée considérée.

Une partition en Intervalle floue de cet univers, pour l'identification du diabète nous avons partitionné l'espace de X_3 en 7 avec une modélisation booléenne sur 3 bits de 000 à 110.

Enfin, les fonctions d'appartenances des classes.

	E	I	S
F1	1	*100*	0
F2	0	*000*	0
F3	0	*011*	0
F4	0	*000*	0
F5	1	*010*	0
F6	1	*010*	0
F7	0	*010*	0
CELFACT			

	E	I	S
R1	0	*000*	1
R2	0	*011*	1
R3	0	*000*	1
R4	0	*010*	1
R5	0	*010*	1
CELRULE			

5.3 Interface de Défuzzification cellulaire

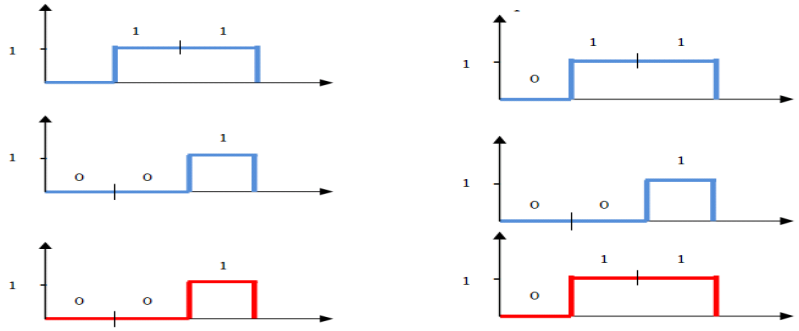
En sortie le système Flou_CIE ne peut pas communiquer à l'utilisateur des valeurs floues. Le rôle de la défuzzification est donc de fournir des valeurs précises. Durant cette étape, notre système Flou_CIE va effectuer des tests pour définir l'intervalle du but prouvé. Ce test dépend du nombre de règles candidates et du nombre de fait de chaque règle qui a participé à l'inférence selon le principe suivant :

- Cas d'une seule règle et un seul fait : « **Si** fait **alors** conclusion »
  *CELFACT*_I (conclusion) = minimum (*CELFACT*_I (fait), *CELRULE*_I (règle)).
- Cas d'une seule règle avec plusieurs faits : « **Si** *fait1 et fait2 et fait3* **alors** *conclu*
- *sion »* : *CELFACT*_I (conclusion) = minimum (*CELFACT*_I (fait1), *CELFACT*_I (fait2), *CELFACT*_I (fait3)).

L'opérateur 'minimum' représente en logique booléenne le « ET logique ».

- Cas de plusieurs règles :
  *CELFACT*_I (but) = maximum (*CELRULE*_I (règle1), *CELRULE*_I (règle2), …).

L'opérateur 'maximum' représente en logique booléenne le « OU logique ». La figure 9 illustre le principe booléen adopté par notre système Flou_CIE.



(a) L'opérateur minimum (b) L'opérateur maximum

FIG. 9 – *Principe booléen pour la déffuzification*

5.4 Moteur d'inférence cellulaire Flou « Flou-CIE »

Pour définir la dynamique du *FCIE*, nous allons rappeler que le cycle de base d'un moteur d'inférence, pour établir un fait *F* en chaînage avant, fonctionne traditionnellement comme suit :

Recherche des règles applicables (évaluation et sélection) ;
Choisir une parmi ces règles, par exemple *R* (filtrage) ;
Appliquer et ajouter la partie conclusion de *R* à la base des faits (exécution).

Le cycle est répété jusqu'à ce que le fait *F* soit ajouté à la base des faits, ou s'arrête lorsque aucune règle n'est applicable.

La dynamique de l'automate cellulaire *FCIE*, pour simuler le fonctionnement d'un *Moteur* d'*Inférence cellulaire Flou*, utilise les mêmes fonctions de transitions δ_{fact} et δ_{rule}, avec une modification au niveau de δ_{rule} que nous avons baptisé ***F2M*** acronyme de Flou-Min-Max.

La fonction de transition δ_{fact}:

$$(EF, IF, SF, ER, IR, SR) \xrightarrow{\delta_{fact}} (EF, IF, \boldsymbol{EF}, \boldsymbol{ER} + (\boldsymbol{R_E^T} \cdot \boldsymbol{EF}), IR, SR)$$

La fonction de transition δ_{rule}:

$$(EF, IF, SF, ER, IR, SR) \xrightarrow{\delta_{rule}} (\boldsymbol{EF} + (\boldsymbol{R_S} \cdot \boldsymbol{ER}), \boldsymbol{F2M}(IF), SF, ER, IR, \overline{\boldsymbol{ER}})$$

Où la matrice R_E^T désigne toujours la transposé de la matrice R_E et ***F2M*** désigne l'opération de défuzzification booléenne.

6 Les résultats expérimentaux obtenus avec Flou-CIE

Une manière classique d'évaluer la qualité de l'apprentissage est de confronter la prédiction du modèle avec les valeurs observées sur un échantillon de la population. Cette confron-

tation est résumée dans un tableau croisé appelé matrice de confusion (voir table 2). Il est possible d'en extraire des indicateurs synthétiques, le plus connu étant le taux d'erreur ou taux de mauvais classement qu'on note ξ.

$\xi = 0.0889$	Type 2	Type 1	Total
Type 2	835	39	874
Type 1	91	496	587
Total	926	535	1461

TAB. 2 – *Matrice de confusion obtenue par SIPINA*

Nous pouvons donc dire qu'en classant un individu pris au hasard dans la population, nous avons 8.89 chances sur 100 de réaliser une mauvaise affectation. Le principe intérêt du taux d'erreur est qu'il est objectif ; il sert généralement à comparer les méthodes d'apprentissage sur un problème donné. Pour obtenir un indicateur non biaisé, il est impératif, en pratique, de ne pas le mesurer sur l'échantillon qui a servi à élaborer le modèle. A cet effet, le praticien met souvent de coté un échantillon, dit de test, qui servira à évaluer et à comparer les modèles.

La validation est la phase qui consiste à calculer le taux ξ, sur un échantillon test Ω_T, en utilisant les règles de prédiction produites par *Flou-CIE*. La généralisation est la dernière phase qui consiste à calculer de nouveau la valeur de ξ en appliquant le modèle à tous les individus de la population $\Omega = \Omega_A + \Omega_T$.

Pour évaluer notre *Flou-CIE*, nous avons expérimenté la base diabétique avec plusieurs méthodes déjà implémenté dans la plateforme Weka, à savoir ID3, C4.5, CART et kppv. Nous avons obtenu les résultats expérimentaux que résume la table 3.

WCSS		Weka			
ξ (Flou-CIE)	ξ (CIE)	ξ (ID3)	ξ (C4.5)	ξ (CART)	ξ (kppv)
0.0033	0.0082	0.0130	0.0034	0.0034	0.0185

TAB. 3 – *Résultats expérimentaux*

7 Conclusion

Deux motivations concurrentes nous ont amenés à adopté le principe des automates cellulaire pour les systèmes à base de règles floues. En effet, nous avons non seulement souhaité avoir une base de règles optimale, mais aussi, nous avons également souhaité améliorer la gestion des connaissances incertaines par le moteur d'inférence cellulaire *CIE* intégrant une nouvelle technique cellulaire dans la plate forme Weka.

Quand il s'agit de l'inférence cellulaire, nous devons impérativement passer par les étapes suivantes :

1) Importer la base de connaissances dans la plate forme Weka ;
2) Définir le problème à résoudre, c'est-à-dire sélectionner les faits à prédire ou à déduire ;
3) Lancer la fuzzification cellulaire pour définir les intervalles et initialiser les deux champs *CELFACT*_I et *CELRULE*_I.
4) Lancer l'inférence floue par le *Flou-CIE* : soit en chaînage avant ou en chaînage arrière.

5) Lancer la défuzzification cellulaire qui permet de convertir les valeurs binaires en valeurs réels et donner un résultat précis pour que l'utilisateur puisse le comprendre.

Les inférences du moteur *Flou-CIE* dans la plateforme Weka se résume comme suit :

- **Initialisation de la BC par automate cellulaire :** Cela revient à initialiser les couches *CELFACT* et *CELRULE* et générer les matrices d'incidences R_E et R_S
- **Inférence des règles en avant :** C'est le rôle du *Flou-CIE* ; Pour déduire les faits buts le module d'inférence utilise les fonctions de transition δ_{fact} et δ_{rule} ;
- **Inférence des règles en arrière :** C'est le rôle du même *Flou-CIE* ; Pour induire les faits hypothèses le module d'inférence utilise les mêmes fonctions de transition δ_{fact} et δ_{rule} mais en permutant les deux matrices d'incidences R_E d'entrée et R_S de sortie ;

Les avantages de ce principe booléen basé sur l'automate cellulaire peuvent être récapitulés selon Benamina et Atmani (2008) comme suit :

- La représentation de la connaissance ainsi que son contrôle sont simples, sous forme de matrices binaires exigeant un prétraitement minimal.
- La facilité de l'implémentation des fonctions de transition δ_{fact} et δ_{rule} qui sont de basse complexité, efficaces et robustes pour des valeurs extrêmes.
- Les résultats sont simples pour être insérés et utiliser par un système expert.
- Le système de prédiction obtenu est un modèle cellulaire composé d'un ensemble simple de fonctions de transition et de règles de production, qui permettent non seulement de décrire le problème actuel mais d'établir également une fonction de classification pour la prévision.
- La matrice d'incidence, R_E, facilite la transformation de règles dans des expressions équivalentes booléennes, qui nous permet d'utiliser l'algèbre de Boole élémentaire pour examiner d'autres simplifications.

Références

Atmani, B., Beldjilali, B (2007a) : Knowledge Discovery in Database: Induction Graph and Cellular Automaton, Computing and Informatics Journal, V.26, N°2 (2007) 171-197.

Atmani, B., Beldjilali, B (2007b) : Neuro-IG: A Hybrid System for Selection and Elimination of Predictor Variables and non Relevant Individuals, Informatica, Journal International, Vol. 18, N°2 (2007) 163-186.

Beldjilali, A, Atmani, B : Traitement des coefficients d'incertitude dans les Arbres de Décision: Application sur la machine Cellulaire 'CASI ', Journée des jeunes chercheurs en Informatique 2008 – Guelma, Algérie.

Benamina, B., Atmani, B., (2008). WCSS: un système cellulaire d'extraction et de gestion des connaissances, Troisième atelier sur les systèmes décisionnels, 10 et 11 octobre 2008, Mohammadia – Maroc, pp. 223-234.

Brahami, M., Atmani, B. (2009). Vers une fouille visuelle des données par automate cellulaire : application à la cartographie des connaissances critiques. Workshop on machine learning and visualization (cap'2009), plateforme AFIA, 25 mai 2009, Hammamet, Tunisie.

Brahami, M., Atmani, B. (2009). Vers une cartographie des connaissances guidée par fouille de données : 1ère Etape - Modélisation booléenne. Second colloque Gestion des connaissances, société et organisations (GeCSO'2009), 14-15 mai 2009, Bordeaux, France.

Breiman, L.- Friedman, J. H.- Olshen, R. A. –Stone, C. J.(1984) : Classification and regression and trees, Technical report, Wadsworth International, Monterey, CA, 1984.
Duhamel, A., M. Picavet, P. Devos et R. Beuscart (2001) : From Data Collection to Knowledge Data Discovery - A Medical Application of Data Mining, Studies in Health Technology and informatics, Vol 84, 2001, 1329-1333.
Fayyad, U., Shapiro, G.P., Smyth, P.(1996): The KDD process for extraction useful knowledge from volumes data, Communication of the ACM, 1996.
Kodratoff, Y (1997) : The extraction of knowledge from data, a new topic for the scientific research, Magazine electronic READ, 1997.
Lefebure, R. G. Venturi (2001) : Data Mining, Paris, EYROLLES, 2001.
Wolfram, S (2002) : Cellular Automata and Complexity, Perseus Books Group, 2002.
Zadeh, Lotfi A (1968) : Probability measures of fuzzy events. Journal of Mathematical Analysis and Applications, 23 :421–427, 1968.
Zadeh, Lotfi A (1994) : «Fuzzy Logic. Neural Networks and Soft Computing». Communications of ACM, vol. 37, no. 3, p.77-84, 1994.
Zadeh, Lotfi A (1997) : «Some Reflections on the Relationship Between AI and Fuzzy Logic: A Heretical View». IJCAI, Springer, p.1-8, 1997.
Zadeh, Lotfi A (2001) : «From Computing with Numbers to Computing with Words-from Manipulation of Measurements to Manipulation of Perceptions». Computing with words, Éditeur: Paul P. Wang, John Wiley et Sons, p. 35-68, 2001.
Zighed, D. A.(1996) : SIPINA for Windows, ver 2.5 Laboratory ERIC, University of Lyon 2, 1996.
Zighed, D.A., Rakotomalala, R. (2000): Graphs of induction, Training and Data Mining, Hermes Science Publication, 2000, 21-23.

Summary

The main objective of our work is the design and validation of a cell model for extracting fuzzy knowledge, by developing a diagnostic for identifying different types of diabetes. Our cell model adopts the principle of learning by induction that we have incorporated the fuzzy logic to meet three criteria of intelligence: 1) tolerance to inaccuracies, 2) the management of uncertainties in the data provided, and 3) learning. We propose in this paper an experimental study on the design of a cellular data mining fuzzy.

Keywords: Machine learning, cellular automata, extraction rules, data search, decision tree, fuzzy logic.

L'utilisation des Chemins hiérarchiques des lieux pour la Désambiguïsation des Toponymes

Imene Bensalem*, Mohamed-Kireddine Kholladi*

* Faculté de sciences de l'ingénieur, Département Informatique
Université Mentouri Constantine, Algérie
bens.imene@gmail.com
kholladi@yahoo.fr

Résumé. La collecte et l'intégration de données depuis plusieurs sources est une opération de préparation de données pratiquement présente dans tout projet du data mining. Avec l'avènement du Web et des bibliothèques numériques, le texte en langue naturelle est devenu une source importante d'informations pour le data mining. L'utilisation du texte comme source de données pâtit d'un grand obstacle à l'intégration et à la précision de données. Cet obstacle est l'ambiguïté des sens des mots y compris l'ambiguïté des toponymes (les noms des lieux) : un seul toponyme peut avoir plusieurs sens c.-à-d. peut se référer à plusieurs lieux dans la Terre. La Désambiguïsation des Toponyme est la tâche d'associer à un toponyme le lieu à lequel il se réfère. Cet article présente une nouvelle heuristique de désambiguïsation des toponymes dans le texte basée sur la mesure de corrélation entre les chemins hiérarchiques des référents candidats des toponymes du même contexte.

1 Introduction

En raison de la grande quantité de données, il est coûteux et souvent irréaliste de les examiner en détail. Le data mining vise à automatiser un tel processus de découverte de connaissances. Les données du data mining peuvent être collectées depuis plusieurs sources. Les données brutes qu'offrent ces différentes sources peuvent être structurées, comme les bases de données relationnelles, ou non structurées comme les images et le texte. Avec l'avènement du Web et des bibliothèques numériques, et le développement des techniques du traitement automatique des langues naturelles (TALN), le texte libre est devenu une source importante d'informations pour le data mining. Les dates, les numéros de téléphones, les noms de personnes, de médicaments, de maladies,…etc. sont tous des informations qui peuvent être extraites du texte, puis intégrées dans une base de données susceptible à subir une analyse tel que le data mining.

L'utilisation du texte comme source de données pâtit d'un grand problème qui est l'ambiguïté des sens des noms propres. Généralement, cette ambiguïté consiste à l'utilisation d'un seul nom pour représenter des entités différentes. L'ambiguïté est un problème pour le data mining pour deux raisons, d'un coté, elle réduit la qualité des données, qui est un facteur

important pour la réussite du data mining, et d'un autre côté c'est un obstacle à l'intégration de données de plusieurs sources.

Les toponymes c.-à-d. les noms des lieux sont parmi les noms qui peuvent être extraits du texte ; en fait, il a été estimé qu'au moins 70% des documents textuels contiennent des références aux lieux géographiques sous forme de toponymes (Hill 2006).

À l'instar des autres types de noms, les toponymes sont très ambigus. En effet, il y a deux types d'ambiguïté des toponymes : l'ambiguïté géo/géo et l'ambiguïté géo/non-géo (Amitay, et al. 2004). L'ambiguïté géo/géo se pose lorsqu'un toponyme représente plusieurs lieux, par exemple, selon le gazetteer[1] Getty[2], « Alger » est le nom de 26 lieux géographiques dans le monde qui représentent des lieux habités, des reliefs, ou des étendus d'eau. L'ambiguïté géo/non-géo apparaît lorsqu'un toponyme se réfère à d'autres types d'entités (ex. Arafat est le nom d'un lieu à coté de La Mecque et aussi le nom de l'ex-président de Palestine) ou possède d'autres sens (ex. java est un langage de programmation et Java est une ile indonésienne).

Nous nous intéressons dans cet article à la désambiguïsation des toponymes (DT) (aussi appelée la résolution des toponymes) qui est une technique qui adresse l'ambiguïté de type géo/géo, et elle est définie comme la tâche d'attribuer un emplacement géographiques à un nom de lieu ambigu dans un contexte textuel donné. La désambiguïsation des toponymes est une tâche indépendante en elle-même mais elle peut être considérée comme une étape primordiale dans une multitude d'applications comme l'indexation géographique des documents textuels, l'extraction d'informations et le data mining.

Cet article a deux buts, d'un coté montrer la relation du data mining et la désambiguïsation des toponymes, et d'un autre coté proposer une nouvelle heuristique de désambiguïsation des toponymes. Dans ce qui suit, nous commençons par présenter des exemples de travaux qui utilisent la DT comme prétraitement des données du data mining. Nous présentons en suite un bref aperçu de l'état de l'art des méthodes de désambiguïsation des toponymes. Après nous classifions les différents type relations qui peuvent exister entre les lieux mentionnés dans le même texte, puis nous présentons notre méthode en introduisant la mesure de la *Densité Géographique* que son calcul se base principalement sur les chemins hiérarchiques des lieux. Nous fournissons dans la section 5 les résultats d'évaluation de notre heuristique en la comparant avec une autre. Enfin nous terminons par une conclusion qui résume les différents points discutés dans cet article.

2 Quelque travaux sur l'utilisation de la DT dans le processus du data mining

Dans (Morimoto, et al. 2003) les auteurs ont présenté un système d'extraction de connaissances géographiques à partir des collections de pages web. La désambiguïsation des

[1] Un gazetteer est un terme anglais qui désigne traditionnellement un dictionnaire de toponymes. Maintenant, le gazetteer est considéré comme un type de Systèmes d'Organisation des Connaissances (SOC), qui organise des informations sur les lieux géographiques nommés (Hill 2006). Nous avons choisi d'utiliser cette appellation anglaise dans cet article car il n'y a pas une traduction unique et précise en français.

[2] Getty Thesaurus of Geographic names online http://www.getty.edu/research/conducting_research/vocabularies/tgn (dernière consultation le 30 juin 2009).

toponymes est parmi les techniques qu'ils ont utilisés pour associer à chaque page web les coordonnées géographiques des lieux qu'elle contient. Ensuite, ils ont extrait les concepts-clés des pages Web, puis formé une table d'associations géographiques dont chaque tuple contient les concepts-clés d'une page web et les coordonnées géographiques des lieux qu'elle renferme. Finalement, des techniques du data mining spatial ont été appliquées sur cette table pour découvrir des patterns spatiaux tel que les collocations spatiales.

Li, Srihari, Niu, et Li (2003) ont utilisé la désambiguïsation des toponymes comme un module dans un moteur d'extraction d'information depuis des articles d'actualités et des guides de touristes. Le résultat du traitement est ensuite stocké dans un entrepôt dynamique de connaissances qui sert à supporter plusieurs applications comme le texte mining, la visualisation et l'analyse des évènements.

3 État de l'art des méthodes de désambiguïsation des toponymes

Malgré le fait que les méthodes de désambigüisation des toponymes sont très différentes dans l'esprit, mais ils ont des facteurs en commun (Leidner 2007). La plupart des méthodes de DT comprennent 2 phases principales qui sont : (1) l'extraction des référents candidats et (2) le choix du référent correct. La première phase consiste à déterminer les référents possibles de chaque toponyme dans le texte à main en utilisant des ressources de connaissances géographiques, comme, les gazetteers et les ontologies. La deuxième phase consiste à l'application d'un ensemble d'heuristiques en vue de déterminer parmi l'ensemble des candidats le référent le plus susceptible d'être le sens voulu par le toponyme ambigu. Ces heuristiques utilisent principalement les connaissances fournies par le contexte et des ressources externes comme sources d'évidence.

Nous classifions les heuristiques existantes de la désambiguïsation des toponymes en deux catégories principales: les *heuristiques de désambiguïsation par les règles de préférence* et *les heuristiques de désambiguïsation par le contexte*.

Les heuristiques de la première catégorie désambiguïsent les toponymes en se basant sur des préférences et des intuitions de l'être humain. Par exemple, les méthodes de Pouliquen, et al.(2004), Amitay, et al. (2004), et Rauch, et al.(2003) utilisent des heuristiques qui résolvent les toponymes ambigus par les référents à plus grande population. La méthode de (Stokes, et al. 2008) utilisent une heuristique qui attribue aux toponymes ambigus les référents les plus fréquents, par exemple si le toponyme à résoudre est Gaza, cette heuristique lui associe le référent Gaza>Palestine au lieu de Gaza>États-Unis car le premier est le plus connu.

Les heuristiques de la deuxième catégorie dépendent principalement des toponymes qui existent dans le même contexte dans lequel le toponyme à désambiguïser apparaît. Cela rend la tâche de désambiguïsation des toponymes similaire à la désambiguïsation des sens des mots (DSM) (Navigli 2009) qui est parmi les tâches connues du traitement automatique des langues naturelles (TALN). Le contexte est le texte en langue naturelle qui contient le(s) toponyme(s) à désambiguïser. La taille du contexte dans les méthodes de DT varie de quelques toponymes autour du toponyme ambigu jusqu'à tous les toponymes du texte du document.

Parmi les travaux dans cette catégorie d'heuristiques, Leidner et al. (2003) attribuent aux toponymes ambigus du même contexte les référents qui diminuent le plus les distances bilatérales, et par conséquent, ils occupent ensemble l'espace géométrique le plus réduit.

Cette heuristique prend en compte toutes les coordonnées spatiales possibles pour chaque toponyme et fait des traitements d'optimisation dont le critère est la proximité spatiale.

Une heuristique proposée par Clough (2005) est basée sur le calcul du score de chevauchement entre les chemins hiérarchiques des référents et le contexte (c.-à-d. calculer le nombre de toponymes en commun). Plus le score est grand plus le référent aura une chance d'être le référent correct.

Il existe aussi des heuristiques qui cherchent la mention du nom du lieu père dans le texte, par exemple, chercher Liban ou Lybie si le toponyme ambigu est Tripoli, cette heuristique est utilisée par Pouliquen et al.(2004) et Li et al. (2006).

Smith et Crane (2001) ont proposé une heuristique qui consiste à calculer le centroïde (barycentre) géographique des référents candidats des toponymes mentionnés dans le même document, puis éliminer tous les référents situés à plus de 2 écarts-types loin du centre. Une méthode similaire est proposée par Rauch et al. (2003).

La méthode de Buscaldi et Rosso (2008) est basée sur le calcul de la *Densité Conceptuelle* pour chaque référent candidat d'un toponyme ambigu. Le référent qui maximise cette valeur (c.-à-d. la densité conceptuelle) est celui qui sera attribué au toponyme ambigu. La Densité Conceptuelle (DC) est une mesure de la corrélation entre le sens d'un mot et son contexte. Elle a été présentée dans le domaine de DSM par Agirre et Rigau (1996) puis reformulée par Rosso et al. (2003). Cette dernière est ensuite adaptée à la désambiguïsation des toponymes par Buscaldi et Rosso (Buscaldi et Rosso 2008). La densité conceptuelle est calculée en se basant sur les chemins hiérarchiques des référents candidats du toponyme ambigu et des lieux qui apparaissent avec lui dans le même contexte. Ces chemins hiérarchiques sont obtenus de WordNet. Un chemin hiérarchique d'un lieu est composé d'un ensemble de lieux reliés avec une relation de holonymie c.-à-d. « tout-partie ». Par exemple, le chemin hiérarchique de Constantine est Afrique>Algérie>Constantine. On dit alors que le toponyme Algérie est holonyme du toponyme Constantine.

L'heuristique que nous allons proposer désambiguïse les toponymes en se basant sur les chemins hiérarchiques des lieux du contexte. Donc, elle se situe dans la deuxième catégorie d'heuristiques de DT citée ci-dessus, et en plus, elle a des principes en commun avec celle de Buscaldi et Rosso (2008). Les sections suivantes sont consacrées à la description de notre contribution.

4 Les types de relations entre les toponymes du même contexte

En observant les heuristiques de la désambiguïsation des toponymes par le contexte, nous remarquons que derrière la plus part des heuristiques de cette classe se cache une intuition qui consiste à supposer l'existence d'une certaine proximité géographique entre les toponymes du même contexte. Dans les travaux présentés ci-dessus, Leidner et al. (2003), Smith et Crane (2001), et Rauch et al. (2003) désambiguïsent les toponymes par les référents les plus proches en termes de distance, ce qui implique à faire des calculs géométriques en utilisant les coordonnées spatiales des référents. Cependant, les heuristiques de Clough (2005), Pouliquen et al.(2004), Li et al. (2006) et Buscaldi et Rosso (2008) désambiguïsent les toponymes par les référents les plus proches dans l'arbre hiérarchique des lieux du monde. Dans ce cas, les référents doivent être représentés par leurs chemins hiérarchiques.

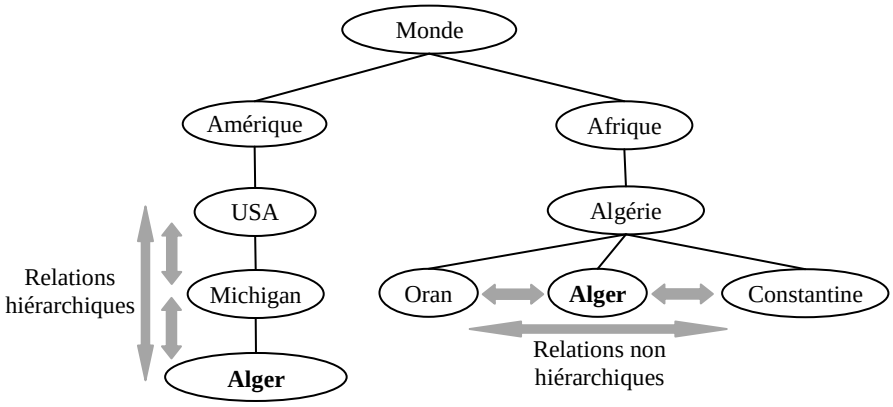


Fig. 1– *Une partie de l'arbre hiérarchique du monde : Alger est un toponyme ambigu*.

Nous appelons « relation spatiale » entre les référents toute relation résultante de la proximité des distances, et « relation arborescente » toute relation résultante de proximités dans l'arbre hiérarchique des lieux du monde (voir Fig. 1).

En outre, nous distinguons deux types de relations arborescentes: les relations hiérarchiques, et les relations non hiérarchiques.

Les relations hiérarchiques existent entre les lieux de la même branche dans l'arbre. Par exemple entre un pays et une de ses villes ; comme entre l'Algérie et Constantine dans la Fig. 1. Les relations non hiérarchiques sont celles qui existent entre les nœuds qui se trouvent dans des branches différentes mais qui ont une (ou plusieurs) racine commune (ex. Constantine et Oran dans la Fig. 1). Une racine (un holonyme) d'un lieu peut être directe (ex. Michigan par rapport à Alger) ou indirecte (ex. Afrique par rapport à Constantine).

La FIG. 2 résume les différents types des relations qui peuvent exister entre les lieux du même contexte.

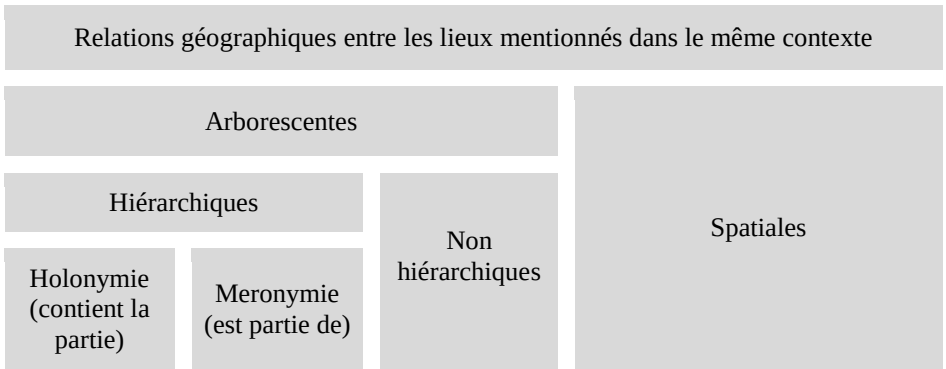


FIG. 2– *Les différents types de relations géographiques qui peuvent exister entre les lieux mentionnés dans le même contexte*.

Les relations arborescentes les plus exploitées dans les méthodes de désambiguïsation des toponymes sont les relations hiérarchiques. C'est le cas des heuristiques de (Clough 2005), (Pouliquen, et al. 2004) et (Stokes, et al. 2008). Cependant, au meilleur de nos connaissances, la seule méthode qui peut détecter des relations non hiérarchiques sans la mention de l'holonyme dans le texte est celle proposée par Buscaldi et Rosso (2008). Néanmoins, cela reste notre point de vue sur leurs méthodes et ce n'est pas déclaré explicitement par les auteurs.

Nous proposons dans le reste de cet article une nouvelle heuristique de désambiguïsation des toponymes. À la différence des autres heuristiques de DT, notre heuristique est conçue explicitement sur l'idée de chercher des relations arborescentes de tous types entre les toponymes du même contexte. Pour cela nous introduisons une nouvelle mesure de corrélations arborescentes que nous appelons la *Densité Géographique.*

5 Une nouvelle heuristique de désambiguïsation des toponymes

5.1 Notation

Soit:

- T l'ensemble des toponymes qui apparaissent dans un document D. $T = \{t_i \in D, i = 1...n\}$, tel que n est le nombre de toponymes.
- G: un gazetteer, $G = \{ r_{id}, r_{id}$ est un lieu géographique dans la Terre$\}$. Chaque r_{id} est représenté par un ensemble de caractéristiques qui diffèrent selon le gazetteer utilisé, qui sont : l'identifiant, le nom et le chemin hiérarchique. On dit que le lieu r_{id} est un référent de t_i si t_i est le nom de r_{id}.
- h_{id} est le chemin hiérarchique de r_{id} dans l'arbre d'hiérarchie de G. Chaque nœud de h_{id} est un référent $r_{id.k}$, tel que le premier nœud $r_{id.0}$ est le monde et le dernier nœud $r_{id.1}$ est r_{id}.
- R_i l'ensemble des référents du toponyme t_i. $R_i = \{r_{ij} \in G, j = 1..m\}$.
- R est l'ensemble des ensembles R_i des toponymes qui apparaissent dans le document D. $R = \{Ri, i = 1..n\}$

5.2 Principe et méthode

Notre heuristique est basée sur l'hypothèse que les toponymes qui apparaissent ensembles dans le même document sont reliés géographiquement. Cette relation peut être hiérarchique ou non hiérarchique.

L'heuristique proposée résout un toponyme par le référent qui est :

- Le plus relié géographiquement aux référents des autres toponymes, c.-à-d. celui qui possède relativement beaucoup de relations arborescentes avec les référents des autres toponymes (on peut dire que c'est une relation indirecte avec le contexte), et ;
- Le plus relié au contexte, c.-à-d. son chemin hiérarchique et le contexte contiennent relativement beaucoup de toponymes en commun.

Ces deux caractéristiques sont quantifiées par le calcul de ce que nous appelons la Densité Géographique. Nous définissons donc la Densité Géographique (DG) comme une

mesure de corrélation (directe ou indirecte) entre un référent d'un toponyme et le contexte de ce dernier.

La désambiguïsation des toponymes par le calcul de la densité géographique suit les étapes suivantes :

1. Extraire tous les toponymes *T* du contexte
2. Déterminer la liste des référents candidats R_i pour chaque toponyme t_i. Chaque référent candidat r_{id} doit être représenté par son chemin hiérarchique h_{id}.
3. Calculer la densité géographique pour chaque référent candidat dans R_i de chaque toponyme t_i.
4. Attribuer à chaque toponyme t_i le référent r_{id} qui possède la plus grande densité géographique $DG(r_{id})$ parmi l'ensemble de ses référents candidats.

5.3 La densité géographique

Les connaissances principales sur lesquelles se base le calcul de la densité géographique sont les chemins hiérarchiques des référents candidats de tous les toponymes du contexte (les chemins hiérarchiques des référents de l'ensemble *R*). Le chemin hiérarchique d'un référent est composé du référent lui-même, est ces holonymes c.-à-d. sa racine directe, et ces racines indirectes. Il représente une branche dans l'arbre hiérarchique des lieux du monde.

La DG d'un référent r_{id} d'un toponyme ambigu t_i augmente lorsque :

(a) ce référent apparaît parmi les holonymes (les racines) des autres référents dans *R–Ri*, et /ou,
(b) ses holonymes sont parmi les référents candidats des autres toponymes (c.-à-d. dans *R–Ri*), et /ou,
(c) ses holonymes sont aussi des holonymes pour d'autres référents, et
(d) les toponymes qui composent son chemin hiérarchique existent partiellement ou totalement dans le contexte.

Les caractéristiques (a), (b) et (d) signifient la présence d'une relation hiérarchique entre le référent cible r_{id} et certains référents des autres toponymes, et (c) signifie la présence d'une relation non hiérarchique.

Les caractéristiques (a), (b) et (c) sont quantifiées par le calcul des fréquences du référent r_{id} et ses holonymes (c.-à-d. de $r_{id.1}$,..$r_{id.2}$, …$r_{id.l}$) dans les chemins hiérarchiques des référents de l'ensemble *R*. La fréquence d'un référent $r_{id.k}$ est la somme de ses poids dans chaque R_i (l'équation (2)).

Le poids P est une fonction booléenne qui indique l'existence ou l'absence d'un référent $r_{id.k}$ dans les chemins hiérarchiques d'un ensemble R_i (l'équation (3)). Par conséquent, La plus grande valeur que peut prendre une fréquence est égale à *n* : le nombre des ensembles R_i dans *R*, et ce qui représente aussi le nombre de toponymes dans le texte.

La caractéristique (d) est quantifiée par le calcul du score du chevauchement du chemin hiérarchique du référent r_{id} avec le contexte *D*, cela est représenté par la valeur $SC(h_{id}, D)$.

La densité géographique *DG* (r_{id}, R) d'un référent candidat r_{id} est la somme de ces deux valeurs décrites ci-dessus (la fréquence des référents qui compose son chemin hiérarchique h_{id} et le score du chevauchement de ce dernier avec le contexte) (l'équation (1)).

$$\mathrm{DG}\,(r_{id}, R) = \sum_{k=1}^{l} (\text{Fréquence}\,(r_{id.k}, R)) + \mathrm{SC}(h_{id}, D) \qquad (1)$$

$$\text{Fréquence}\ (r_{id.k}, R) = \sum_{i=1}^{n} \mathrm{P}(r_{id.k}, R_i) \quad (2)$$

$$\mathrm{P}(r_{id.k}, R_i) = \begin{cases} 0, \text{si le nombre de } r_{id.k} \text{ dans les } h_{id} \text{ de } R_i = 0 \\ 1, \text{si le nombre de } r_{id.k} \text{ dans les } h_{id} \text{ de } R_i \neq 0 \end{cases} \quad (3)$$

5.4 Évaluation

L'évaluation des méthodes de la désambiguïsation des toponymes nécessite l'utilisation de deux ressources principales qui sont les corpus textuels et les gazetteers. L'évaluation est encore problématique dans ce domaine dû au manque de ressources standards qui permettent la comparaison entre les performances des différentes méthodes. Leidner (2004, 2006) a adressé ce problème mais malheureusement ses données ne sont pas disponible gratuitement[3].

Buscaldi et Rosso (2008) ont évalué leur méthode basée sur la densité conceptuelle en utilisant l'ontologie WordNet comme gazetteer, et le corpus GeoSemCor qui est une version de SemCor où chaque toponyme peut être relié avec son référent correct dans WordNet.

WordNet[4] est une large base de données lexicale qui contient plusieurs types de mots avec leurs sens ; donc, ce n'est pas une source de connaissances purement géographiques, et par conséquent, elle n'est pas aussi riche de toponymes et de référents pour chaque toponyme que les gazetteers. Le tableau

TAB. **1** fournit des toponymes pris du corpus GeoSemCor et des toponymes de quelques wilayas d'Algérie et compare leur nombre de référents récupérés du WordNet (version 2.1) et du Gazetteer Getty.

Toponyme	Nombre de référents dans WordNet	Nombre de référent dans le gazetteer Getty
China	2	264
Georgia	3	74
New York	3	104
Paris	2	102
Palestine	2	44
Russia	4	14
Annaba	1	3
Constantine	1	17
Mila	0	4
Oran	1	14

TAB. 1 – *Comparaison du nombre de référents pour certains toponymes dans WordNet et le Gaztteer Getty.*

[3] D'après une communication personnelle avec Jochen Leidner.
[4] http://wordnet.princeton.edu

WordNet et GeoSemCor ne sont pas vraiment adaptées à la tâche de désambiguïsation des toponymes ; ce sont plutôt utilisées dans le domaine de la désambiguïsation des sens des mots. Toutefois, Nous avons choisi d'évaluer notre heuristique en utilisant ces ressources. Cela est pour deux raisons, d'un coté, ce sont les seules ressource de DT gratuitement disponible[5], et de l'autre coté cela nous permet de comparer notre méthode à celle de Buscaldi et Rosso (2008) qui est la seule méthode qui ressemble à la notre dans le fait qu'elle puisse détecter des relations non hiérarchiques entre les toponymes même si leurs holonymes ne sont pas présents dans le contexte. Le TAB. 2 fourni les résultats de l'évaluation.

	Précision	Recall
DG	88,2%	**87,4%**
DC	**89,9%**	77,5%

TAB. 2 – *Résultats d'évaluation en utilisant WordNet et GeoSemCor. La ligne DG représente les performances de notre heuristique basée sur la densité géographique. La ligne DC représente les performances de l'heuristique de (Buscaldi et Rosso 2008) basée sur la densité conceptuelle.*

L'estimation des performances des méthodes de désambiguïsation des toponymes se fait par les métriques utilisées dans les domaines de la recherche d'information et le traitement automatique des langues naturelles. Ces métriques sont : la précision et le recall. Ils se calculent dans le domaine de la DT comme montré dans les équations (4) et (5) respectivement.

$$\text{Précision} = \frac{\text{nombre de toponymes résous correctement}}{\text{nombre de toponymes résous}} \quad (4)$$

$$\text{Recall} = \frac{\text{nombre de toponymes résous correctement}}{\text{nombre total de toponymes}} \quad (5)$$

Le recall de notre heuristique dépasse avec une valeur significative (+9.9%) celui de (Buscaldi et Rosso 2008). Or il n'y a pas une grande différence (-1.9%) entre les deux heuristiques en termes de précision.

La précision ainsi que le recall de notre heuristique ont dépassé la valeur 80%, ce qui indique de bonnes performances.

6 Conclusion

Nous avons montré dans cet article que la désambiguïsation des toponymes est l'une des technique utilisées dans la phase de préparation des données du data mining dans le cas où le texte en langue naturelle est utilisé comme une source de données.

[5] GeoSemCor est disponible dans l'adresse http://users.dsic.upv.es/grupos/nle/downloads.html

Il convient de noter que la désambiguïsation des toponymes est une tâche indépendante en elle-même, et le data mining n'est pas le seul champ de son application.

La contribution principale de ce papier est la proposition d'une nouvelle heuristique de désambiguïsation des toponymes. Notre heuristique est basée sur l'hypothèse de l'existence d'une relation géographique arborescente entre les toponymes du même contexte. Et donc elle résout les toponymes ambigus par les référents les plus reliés entre eux dans l'arbre hiérarchique des lieux du monde. Pour quantifier le degré de cette relation nous avons introduit une mesure de corrélation géographique que nous avons appelé la *densité géographique* (DG), cela est par analogie à la densité conceptuelle (DC) utilisée pour la désambiguïsation des sens des mots, et appliquée par Buscaldi et Rosso (2008) pour la DT.

L'évaluation de notre heuristique en utilisant WordNet et GeoSemCor a montré la validité de notre hypothèse et la performance de notre heuristique.

7 Remerciements

Nous somme très reconnaissons a Simon Overell qui nous a proposé d'évaluer notre méthode en utilisant le corpus GeoSemCor. Nous tenons à remercier également Davide Buscaldi de nous avoir envoyé une version originale de son article (Buscaldi et Rosso 2008) et aussi d'avoir partagé ses données gratuitement sur le Web. Nous remercions aussi les reviewers anonymes de leurs remarques pertinentes qui nous ont permis d'améliorer la qualité de cet article.

Références

Agirre, E., & Rigau, G. (1996). Word sense disambiguation using conceptual density. *Proceedings of the 16th conference on computational linguistics (COLING '96)* (pp. 16–22). Copenhaghen: Association for Computational Linguistics.

Amitay, E., Har'El, N., Sivan, R., & Soffer, A. (2004). Web-a-where: geotagging web content. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 273 - 280). New York: ACM.

Buscaldi, D., & Rosso, P. (2008). A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Science , 22* (3), 301-313.

Clough, P. (2005). Extracting metadata for spatially-aware information retrieval on the Internet. Dans C. Jones, & R. Purves (Éd.), *Proceedings of the ACM Workshop on Geographic Information Retrieval (GIR) held at the Conference on Information and Knowledge Management (CIKM)* (pp. 25-30). ACM Press.

Hill, L. L. (2006). *Georeferencing: The geographic associations of information.* Cambridge, MA, USA: The MIT Press.

Leidner, J. L. (2006). An evaluation dataset for the toponym resolution task. *Computers,Environment and Urban Systems , 30* (4), 400–417.

Leidner, J. L. (2007). *Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names.* PhD dissertation, University of Edinburgh, Institute for Communicating and Collaborative Systems, School of Informatics.

Leidner, J. L. (2004). Towards a reference corpus for automatic toponym resolution evaluation (Extended abstract). *Proceedings of the Workshop on Geographic Information Retrieval held at the 27th Annual International ACM SIGIR Conference (SIGIR)*, (p. pages unnumbered). Sheffield, England, UK.

Leidner, J. L., Sinclair, G., & Webber, B. (2003). Grounding spatial named entities for information extraction and question answering. *Proceedings of the Workshop on the Analysis of Geographic References held at the Joint Conference for Human Language Technology and the Annual Meeting of the North American Chapter of the Association for Computational Linguistics 2003 (HLT/NAACL 2003)*, (pp. 31-38). Edmonton, Alberta, Canada.

Li, H., Srihari, R. K., Niu, C., & Li, W. (2003). InfoXtract location normalization: a hybrid approach to geographic references in information extraction. *Proceedings of the HLT-NAACL 2003 Workshop: Analysis of Geographic References* (pp. 39–44). Edmonton, Alberta, Canada: Association for Computational Linguistics.

Li, Y., Moffat, A., Stokes, N., & Cavedon, L. (2006). Exploring Probabilistic Toponym Resolution for Geographical Information Retrieval. *Proceedings of SIGIR Workshop on Geographical Information Retrieval*, (pp. 17-22). Seattle, Washington.

Morimoto, Y., Aono, M., Houle, M. E., & McCurley, K. S. (2003). Extracting spatial knowledge from the web. *Proceedings of the 2003 Symposium on Applications and the Internet (SAINT'03)* (p. 326). Los Alamitos, CA, USA: IEEE Computer Society.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys , 41* (2), 69 pages (10:1-10:69).

Pouliquen, B., Steinberger, R., Ignat, C., & Groeve, T. D. (2004). Geographical information recognition and visualization in texts written in various languages. *Proceedings of the 2004 ACM Symposium on Applied Computing* (pp. 1051–1058). ACM Press.

Rauch, E., Bukatin, M., & Baker, K. (2003). A confidence-based framework for disambiguating geographic terms. *HLTNAACL 2003 Workshop: Analysis of Geographic References* (pp. 50-54). Edmonton, Alberta, Canada: Association for Computational Linguistics.

Rosso, P., Masulli, F., Buscaldi, D., Pla, F., & Molina, A. (2003). Automatic noun sense disambiguation. Dans A. Gelbukh (Éd.), *Computational linguistics and intelligent text processing: 4th International Conference, CICLing 2003 Mexico City, Mexico, February 16–22, 2003 Proceedings , 2588 of Lecture Notes in Computer Science* (pp. 273–276). Berlin: Springer.

Smith, D. A., & Crane, G. (2001). Disambiguating geographic names in a historical digital library. *Research and Advanced Technology for Digital Libraries: Fifth European Conference (ECDL 2001)*, (pp. 127–136).

Stokes, N., Li, Y., Moffat, A., & Rong, J. (2008). An empirical study of the effects of NLP components on Geographic IR performance. *International Journal of Geographical Information Science , 22* (3), 247-264.

Summary

Collecting and integrating data from multiple sources is a process of data preparation which must be present in any data mining project. With the advent of Web and digital libraries, natural language text has become an important source of information for data mining. Using text as data source suffers a major obstacle to integration and accuracy of data. This obstacle is the ambiguity of words’ senses including the ambiguity of toponyms’ (place names) senses: a toponym may have several senses i.e. it may refer to several places in the Earth. Toponym disambiguation is the task of associating a toponym to the place to which it refers. This paper presents a new toponym disambiguation heuristic based on the quantification of the arborescent relationships between toponyms of the same context.

Un Modèle Multi-Agents pour l'Aide à la Décision Coopérative

Bakhta Nachet*, Abdelkader Adla***

***Département Informatique, Université d'Oran, Algérie**
nachdal@yahoo.fr
****IRIT, Université Paul Sabatier, Toulouse**
adla@irit.fr

Résumé. Nous proposons dans ce papier d'utiliser les systèmes multi-agents (SMA) pour modéliser les Systèmes Intelligents d'Aide à la Décision (SIAD) Coopératifs. Ces derniers prennent en charge la coopération de deux agents : le décideur (l'utilisateur) et la machine pour résoudre conjointement un problème de prise de décision. Le but est de tirer profit des capacités du décideur et de la machine. L'idée présentée dans ce travail est la modélisation des SIAD Coopératifs par les SMA en couplant deux SMA, l'un réactif et l'autre cognitif. Le SMA réactif prend en charge la production des différents plans d'actions possibles pour la résolution du problème posé. Ces plans d'actions sont ensuite validés et épurés par un agent observateur extérieur au SMA réactif. L'agent observateur collabore avec le décideur pour choisir le meilleur plan à présenter en entrée du SMA cognitif. Ce dernier considère le plan d'actions en entrée et prend en charge la coopération homme/machine pour résoudre le problème d'aide à la décision.

1 Introduction

La prise de décision et l'exécution des décisions sont les buts fondamentaux de toute organisation, de tout management. Toute organisation dépend, structurellement de la nature des décisions qui sont prises en son sein par des décideurs qu'ils soient individuels ou collectifs.

Les décisions sont souvent prises sur la base d'intuitions et d'expériences passées. Elles sont issues d'heuristiques observables au travers de biais systématiques (Ferber, 1995). Comme l'a observé Simon (1977), ce type de stratégies ne peut s'appliquer qu'à des problèmes familiers. Lorsque nous sommes confrontés à des situations nouvelles, la tâche de prise de décision devient beaucoup plus difficile et l'environnement des décideurs est de plus en plus complexe et évolue rapidement. L'un des principaux problèmes est de déterminer les informations pertinentes pour la prise de décision (Holtzman, 1989).

Des systèmes d'aide à la décision, notés SIAD, sont alors utilisés, ils permettent d'évaluer la situation, les diverses alternatives et leurs impacts. Cependant, ces systèmes sont réduits à un état insulaire et très technique, les seules opportunités d'action offertes par les SIAD traditionnels se limitent à pouvoir arrêter et relancer le processus, inspecter certains paramètres, …etc. Cela prend une importance d'autant plus grande qu'en situation complexe la décision n'est pas structurée.

Les SIAD sont complexes par le fait qu'ils manipulent une quantité importante et variée d'informations ainsi que différents mécanismes de son exploitation. Afin de rendre ces systèmes plus efficaces, c'est à dire plus simples à concevoir et plus simple dans leur fonctionnement, il devient primordial de concevoir des systèmes intelligents et coopératifs permettant une résolution conjointe du problème et une répartition dynamique des tâches de résolution entre l'utilisateur et le système en fonction des problèmes à résoudre et d'un mode de coopération approprié. Ceci favorisera une distribution des connaissances et des compétences sur différents modules indépendant (agents humains et agents machines) qui doivent se coordonner, coopérer et donc s'organiser afin de constituer une unité capable de reproduire le processus du système global.

Dans cette perspective nous nous sommes intéressés à la technologie Multi-Agents. Ces derniers sont issus de l'intelligence artificielle distribuée (IAD). L'IAD s'intéresse entre autres à la résolution distribuée des problèmes qui recherche la meilleure manière de diviser un problème en un ensemble d'entités distribuées et coopérantes et à la façon de partager la connaissance du problème afin d'en obtenir la solution (Ferrand, 2003).

Comme les Systèmes Multi-Agents modélisent un système en termes d'agents, d'environnement, d'interaction, d'organisation, de coopération, et d'émergence de comportements (Ferber, 1995) (Jarras et al., 2002), nous proposons de les appliquer aux SIAD Coopératifs et de faire coopérer l'ensemble des agents (humain et artificiels) à la résolution du problème de prise de décision.

Ce papier est organisé comme suit : nous introduisons d'abord les systèmes d'aide à la décision coopératifs. Nous présentons, dans la section 3, le paradigme tâches-méthodes et, dans la section quatre, l'approche SMA pour les SIAD coopératifs. Dans la section 5, nous décrivons notre proposition de modèle multi-agents suivi de la présentation d'un exemple d'application avant de conclure.

2 Les Systèmes d'aide à la décision Coopératifs

Les SIADs ont été conçus pour résoudre des problèmes de décision peu ou mal structurés (Marakas, 2003). Des problèmes où les préférences, jugements, intuitions et l'expérience du décideur sont essentiels, où la séquence des opérations telles que la recherche d'une solution, la formalisation et la structuration du problème n'est pas connue à l'avance, où les critères pour la prise de décision sont nombreux, en conflit ou fortement dépendant de la perception de l'utilisateur et où la solution doit être obtenue en un temps limité.

De nombreuses définitions ont été proposées. Scott-Morton (1978) qui virtuellement inventait la discipline au début des années 70 offrait cette définition des systèmes d'aide à la décision : « Les systèmes d'aide à la décision font coupler les ressources intellectuelles des individus avec les capacités de l'ordinateur pour améliorer la qualité des décisions. C'est un système d'aide informatique aux décideurs qui traitent des problèmes semi-structurés ».

Turban (1993) donne une définition qui porte à la fois sur les fonctions et la structure du système : « Un SIAD est un système d'information interactif, flexible, adaptable et spécifiquement développé pour aider à la résolution d'un problème de décision, en améliorant la prise de décision. Il utilise des données, fournit une interface utilisateur simple et autorise l'utilisateur à développer ses propres idées ou points de vue. Il peut utiliser des modèles standards ou spécifiques, supporter les différentes phases de la prise de décision et inclure une base de connaissances ».

De même, il n'existe pas une architecture standard pour un SIAD. Une architecture de base a été présentée par Marakas dans (Marrakas, 2003). Cette architecture considère 5 composants : le système de gestion des bases de données (SGBD), le système de gestion des modèles (MBMS), le moteur (KE), l'interface utilisateur et l'utilisateur.

La notion d'interactivité dans un SIAD renvoie au rôle indispensable de l'utilisateur dans son fonctionnement, rôle non passif qui sous-tend le terme aide à la décision, mais aussi à la qualité de l'intégration des différents composants du système et à la nature de l'interface homme/machine (Lévine, 1989). On reproche aux SIAD traditionnels une mauvaise prise en compte de l'utilisateur. En effet, ces systèmes prédéfinissent les rôles des agents en donnant au système le rôle de pure résolution alors que l'utilisateur est confiné dans des tâches d'entrées de données, voire de résolution de conflits.

Résoudre un problème décisionnel nécessite de faire appel à l'intuition et au savoir faire du décideur qui devient l'élément prépondérant du couple Homme /Machine, c'est rendre le système coopératif, à le doter capacités supplémentaires afin de coopérer avec son utilisateur et de le guider dans son processus de résolution de problèmes. Coopérer signifie en particulier distribuer les tâches à réaliser entre l'utilisateur et le système. Le partage des tâches est une condition de la mise en œuvre de la coopération entre les deux agents. La tâche qui fait l'objet de la coopération doit faire l'objet d'un découpage en sous-ensembles cohérents.

Nous considérons dans ce travail, le modèle d'architecture de SIAD coopératif proposé dans (Adla, 2007). Ces systèmes supportent la coopération homme/machine. Le système est doté de capacités afin d'assurer une collaboration avec le décideur qui doit toujours avoir l'avantage sur la machine. En effet, le processus de prise de décision repose sur le savoir faire et les compétences du décideur. Pour cela, le système tout en étant capable de faire le choix d'une stratégie de décision et parfois même de sa mise en œuvre, doit permettre au décideur d'intervenir à tout moment afin de modifier les caractéristiques du processus de prise de décision.

La figure 1 présente l'architecture à base de modèles proposée dans (Adla, 2007), celle ci nous servira alors de base pour notre proposition de modélisation par les SMA.

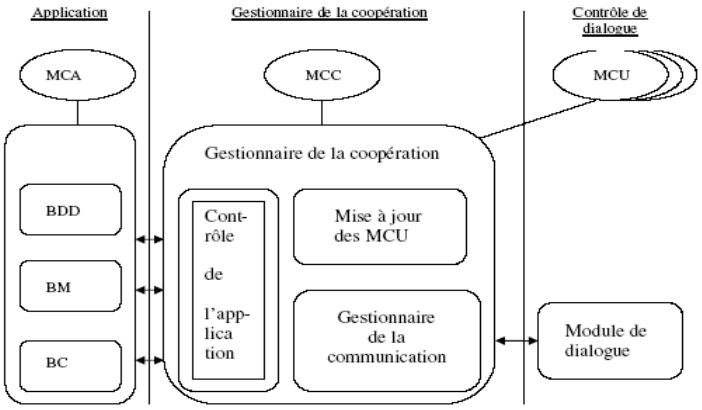


Figure 1: Architecture d'un SIAD coopératif selon Adla [26].

Cette architecture présente les modèles conceptuels nécessaires à la conception d'un système coopératif d'aide à la décision. Trois types de modèles sont requis : un Modèle Conceptuel de l'Application (MCA) représentant la connaissance du système sur le domaine d'expertise (ce modèle s'articule autour de trois bases à savoir : Base de données, Base de modèles et Base de connaissance), un Modèle Conceptuel de la Coopération (MCC) décrivant les modes de coopération entre les différents agents (système/utilisateurs) et un Modèle Conceptuel de mise à jour des Utilisateurs (MCU) spécifiant les connaissances, buts, etc. que le système possède sur l'utilisateur.

Pour la mise en œuvre de la coopération homme/système, il faudra donc décomposer le problème à résoudre en un ensemble de tâches à distribuer entre le système et le décideur. Chacun de ces deux derniers possède des compétences qui lui permettent de se voir affecter des tâches. Les compétences de l'utilisateur et du système sont parfois complémentaires, parfois « redondantes ». Dans ce dernier cas, utilisateur et système sont souvent capables de jouer un même rôle. Se pose donc le problème du choix de l'agent (utilisateur ou système) qui devra tenir chaque rôle. Suivant le contexte, on pourra donner différentes indications sur la façon de réaliser ce choix.

Nous avons considéré la modélisation de la décision dans (Adla, 2007) qui, se basant sur le paradigme tâches – méthodes, modélise le problème à résoudre sous forme d'arbre de tâches et de sous-tâches ainsi que les méthodes associées pour les réaliser. On obtient de ce fait une relation d'ordre entre les différentes tâches à réaliser.

3 Le paradigme tâches – méthodes

Le problème à résoudre est modélisé en une hiérarchie de tâches et de méthodes (Fig. 2). Le principe est de décomposer les tâches complexes en sous tâches. A chaque sous-tâche, au moins une méthode est associée pour la réaliser.

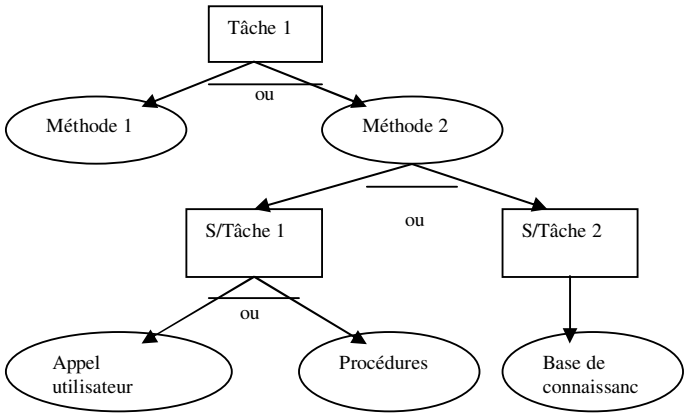


Fig. 2 *: Hiérarchie de tâches et de méthodes*

Une tâche représente l'ensemble des problèmes et des sous-problèmes à résoudre sans connaissance a priori de la façon de les résoudre. Elle est caractérisée par un but, un type (décision, exécution, …), des pré-conditions, des post-conditions et les méthodes qui lui sont

-

associées. A chaque tâche sont associées des méthodes qui sont déclarées a priori comme les mieux adaptées pour la traiter. Une décomposition récursive en sous-tâches de plus en plus élémentaires avec l'ordre de leur exécution, conduit à un enchaînement de méthodes dont l'exécution mène à la résolution du problème correspondant.

Une tâche est définie par les composants suivants :

Nom : le nom de la tâche

Par : la liste de paramètres manipulés par la tâche

Objectif : le but de la tâche

Méthodes : la liste des méthodes réalisant la tâche

Une méthode représente l'ensemble des mécanismes de résolution ou des savoir-faire qu'il est possible de mettre en œuvre pour traiter les différentes tâches. Une méthode peut consister en un plan d'actions (composition de sous-tâches), un ensemble d'heuristiques (base de connaissances), un code de calcul ou une procédure (base de modèles), une requête (base de données) ou encore un appel à l'utilisateur. Elle est caractérisée par un but, des événements, des résultats, des contraintes et le processus de traitement. Les méthodes peuvent être de deux types : méthodes de décomposition ou méthodes terminales. Les méthodes de décomposition servent à diminuer la complexité des tâches. Les méthodes terminales expriment le fait que la tâche en entrée est non décomposable et ne fera pas l'objet de coopération, elle est donc mono agent.

Une méthode est définit par les caractéristiques principales suivantes :

Nom : le nom de la méthode

En-tête : la tâche réalisée par la méthode

Contrôle : contrôle de l'ordre d'exécution des sous-tâches

Sous-tâche : l'ensemble des sous-tâches issues de la décomposition. Dans le cas ou la méthode est terminale, elle n'a pas de sous tâches. Ce champ est remplacé par la nature de l'action qui réalise la tâche.

4 L'approche Multi-Agents dans les SIAD

Notre étude des SIAD coopératifs nous a permis de relever certaines ressemblances avec les SMA. En effet, les SIAD coopératifs utilisent les concepts d'agent (homme-machine), de mode de coopération entre agents et aussi de rôles attribués aux agents afin de résoudre un problème décisionnel de façon coopérative. Il nous a semblé naturel de faire l'analogie et de rechercher les correspondances entre ces concepts et ceux des SMA.

Dans cette perspective, nous avons étudié plusieurs approches de conception de SMA en vue de les appliquer dans notre démarche de conception de SIAD Coopératifs.

L'approche voyelle de Demazeau (1995) présentait une structure organisationnelle où un agent principal (gestionnaire de la coopération homme/machine) supervisait les autres agents qui représentaient le décideur et le système. La distribution des connaissances est bien réalisée. Cependant, nous reprochons à ce système une pauvreté en termes de dynamique interactionnelle et d'autonomie des agents.

L'application du modèle AGR (Gutknecht, 2001) permet d'organiser les agents dans des groupes. Au sein d'un groupe, l'organisation est assurée par des rôles que peuvent jouer les agents. Le rôle est lié au comportement de l'agent, il sert à contrôler ses interactions au sein du groupe selon les services qu'il pourra fournir. Nous avons tenté d'exploiter le concept de rôle d'AGR afin de le mettre en correspondance avec les rôles des deux acteurs du SIAD

Coopératif. Cette approche, bien qu'elle apporte un avantage relativement aux contributions des concepts de rôles et de groupes auxquels peuvent appartenir les agents, présente des insuffisances. En effet, plus le SIAD Coopératif considéré est complexe, plus l'agent responsable de l'élaboration du graphe de tâches – méthodes devient complexe. En outre, de même que pour l'approche voyelle, nous relevons un manque de dynamique interactionnelle dans ce cas aussi.

5 Proposition d'un modèle multi-agent pour la conception de SIAD Coopératif

5.1 Architecture générale du système

Après analyse des résultats obtenus suite à l'application des deux approches de conception précédentes, nous proposons une nouvelle approche qui fait coupler un SMA réactif (Muller, 1998) avec un SMA cognitif (Jarras et al., 2001). La figure 4 présente l'architecture globale de ce système.

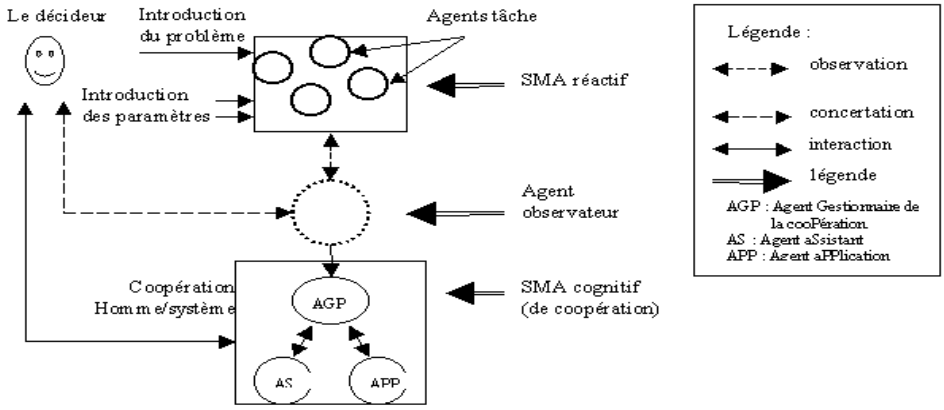


Figure 4 : architecture globale du SMA

5.2 Le SMA réactif

Le SMA réactif consiste en une population d'agents réactifs, ce sont les agents tâche avec des comportements assez simplistes. L'entrée du SMA c'est le problème à résoudre avec les différents paramètres liés à ce problème. Et, la sortie c'est le ou les graphes de tâches et sous tâches.

Un graphe de tâches et de sous tâches représente l'ensemble des plans d'actions – des solutions possibles – au problème posé en entrée du SMA réactif. Cet ensemble de solutions

émerge du comportement réactif des agents tâches. Comme ces agents n'ont pas de connaissances ni de compétences pour analyser les différentes solutions, nous utiliserons un agent observateur, extérieur au système réactif, qui lui, aura la tâche de l'analyse de l'ensemble des solutions en épurant celles qui sont valides de celles qui ne le sont pas. Aussi, il recherchera la (ou les) meilleure(s) solution(s). Une seule d'entre elles sera fournie comme entrée au SMA cognitif qui devra l'exécuter dans un contexte coopératif entre le système et décideur.

L'entrée du SMA réactif c'est le problème à résoudre. Le SMA est constitué d'une population d'agents réactifs et d'un environnement que les agents peuvent percevoir. Dans notre contexte, nous avons considéré un seul type d'agents, ce sont les agents *tâche décomposable.* Les tâches terminales sont des tâches qui ne possèdent pas de méthodes de décomposition. A ces tâches, il ne correspond aucun plan et donc aucune sous tâche. Ces tâches ne vont pas s'exécuter, pour cela on ne leur fait pas correspondre d'agents. Ces tâches apparaissent dans le graphe de tâches et de sous tâches au niveau des feuilles de l'arbre. Les maillons relatifs à ces tâches sont crées par l'agent qui correspond au nœud père. La sortie ou la finalité du SMA est le ou les graphes de tâches et de sous tâches qui représentent l'entrée du SMA cognitif.

Au départ il n'existe aucune organisation prédéfinie et en sortie c'est une auto-organisation des agents qui émerge des interactions, ainsi que de l'environnement qui représente les paramètres relatifs au problème à résoudre. La figure 4 présente l'architecture générale de ce système.

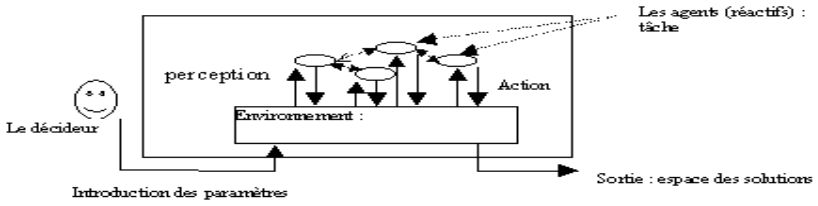


Figure 5 : architecture générale du SMA réactif

5.2.1 Les agents

Les agents du SMA représentent les tâches décomposables du graphe tâche-méthode. Nous avons donc un seul type d'agent : les agents « *tâche décomposable* ». Un agent tâche possède les informations suivantes :

Nom : le nom de la tâche
Par : la liste de paramètres manipulés par la tâche
Objectif : le but de la tâche
Méthodes : la liste des méthodes réalisant la tâche

Une méthode et un ensemble de relations possibles avec d'autres agents. Pour un agent *tâche*, l'application d'une méthode crée des liens avec les sous-tâches de la tâche active. Ces relations sont orientées de l'agent tâche source vers les agents tâches spécifiques relativement à la méthode appliquée. Le critère de satisfaction de ces agents c'est de s'organiser avec d'autres agents pour former un arbre de tâches et de sous tâches. Cet arbre représente un plan d'actions possible du problème à résoudre.

Un agent ne connaît que les agents avec lesquels il est lié par des relations de spécificité, grâce à ses relations méthodes. Par conséquent, il n'a pas de vision globale des autres agents du système. Un agent tâche ne peut créer de liens de spécificités qu'avec d'autres agents relativement à l'application d'une seule méthode. Lorsqu'un agent est activé, cela veut dire qu'il existe un autre agent *tâche* « générique » qui l'a activé suite à l'application d'une de ses méthodes. Cette dernière contient l'agent spécifique parmi ses relations. A ce moment là, l'agent *tâche* (spécifique), doit chercher parmi ses méthodes celle qu'il faudra appliquer. Un Agent *tâche* activé peut avoir plusieurs méthodes candidates. Une méthode candidate est une méthode dont les paramètres sont vérifiés. Si pour un agent *tâche* donné, une seule de ses méthodes est candidate alors elle est appliquée. Dans le cas où pour un agent *tâche* donné, plusieurs de ses méthodes sont candidates, cela voudra dire que le problème en cours de résolution admet plusieurs solutions et donc génère un graphe contenant plusieurs sous graphes du graphe tâches – méthodes. Chacun de ces sous graphes est un arbre de tâches et de sous tâches. Chacun de ces arbres représente une solution possible au problème posé. Les agents doivent avoir dans leurs comportements réactifs la possibilité de construire cet espace de solutions. Ils vont créer dynamiquement les différents maillons qui vont former le graphe de tâches et de sous tâches.

5.2.2 L'environnement

Il est constitué de deux parties : La première partie sert d'environnement de perception pour les agents ; c'est là où les paramètres du problème en entrée sont spécifiés. La deuxième partie représente l'environnement d'action ; c'est là où les agents agissent individuellement et collectivement pour construire l'espace de solutions.

L'environnement d'action est constitué de 2 espaces :

a) Le premier espace est un tableau noir constitué d'une matrice de tâches et de méthodes relatives à ces tâches. Ce tableau noir est perceptible et accessible par tous les agents. C'est là où se déroulera la propagation des perturbations données en entrée, ainsi que l'application du processus de régulation de ces perturbations ;

b) Le deuxième espace est l'espace des solutions qui est construit progressivement à base de composants maillons pour former le graphe de tâches et de sous tâches.

Conceptuellement le tableau noir est une matrice M, ses lignes sont les identifiants de tâches (décomposables et terminales) et les colonnes représentent les identifiants de toutes les méthodes. Une case M(i,j) marquée représente donc l'application de la méthode j de la tâche i. Lorsqu'un agent *tâche* est activé, il cherche parmi ses méthodes celles qui sont candidates.

Tous les agents perçoivent l'environnement à partir duquel les agents sous-tâches sauront construire l'espace des solutions. La construction de l'espace des solutions (c'est-à-dire le graphe de tâches et de sous-tâches) se fait dynamiquement et progressivement en greffant au fur et à mesure de nouveaux maillons jusqu'à ce que toutes les tâches marquées soient des tâches non décomposables :

. Lors de la spécification du problème il y'a obligatoirement désignation d'une tâche initiale. L'agent correspondant à cette tâche est activé, il évalue ses différentes méthodes, il met à jour la matrice tableau noir, puis, il initialise l'espace des solutions en créant : le maillon racine, les maillons correspondant aux méthodes candidates de la tâche, le (les) maillon(s) (méthode appliquée, première sous tâche et la méthode suivante s'il existe d'autres méthodes à appliquer), les maillons correspondant aux sous tâches. Les agents correspondant aux maillons des sous-tâches crées au niveau de l'espace des solutions seront activés et procéderont de la même façon que l'agent initial.

. A la fin de l'exécution, l'espace des solutions est construit. Cet espace est un arbre qui contient l'ensemble des sous arbres tel que chacun d'eux représente une solution au problème posé. Il en ressort une auto-organisation émergente des agents relativement au problème posé. Cet espace présente toutes les solutions possibles au problème de prise de décision. Toute solution est une structure d'arbre tel que les profondeurs paires sont les tâches et les profondeurs impaires sont les méthodes. Il reste à les parcourir et à les valider. Ceci est pris en charge par un agent extérieur au SMA réactif : c'est l'agent observateur.

5.2.3 Processus de SMA réactif

Le processus du SMA réactif consiste en trois étapes :

Enoncé du problème : la spécification du problème implique la désignation d'une tâche. C'est la tâche initiale, elle doit être obligatoirement une tâche décomposable. Pour cette tâche initiale, l'agent *tâche décomposable* correspondant est activé. Ce dernier créera au niveau de l'espace des solutions le maillon correspondant à cette tâche racine ainsi que ceux correspondant aux méthodes applicables et sous tâches correspondantes.

Introduction des différents paramètres liés au problème : L'introduction des paramètres initialise toutes les tâches concernées par ces paramètres.

Exécution du SMA : elle démarre à partir de la tâche racine donnée par la spécification du problème.

5.3 L'agent observateur

Son rôle est d'exploiter les sorties du SMA réactif. Il doit ainsi analyser l'espace des solutions. Cet agent doit :

1. Parcourir l'espace à partir de la tâche d'entrée ou la tâche racine (c'est la première tâche décomposable qui a servi à la spécification du problème) ;
2. Epurer l'ensemble des solutions : supprimer de l'espace des solutions celles qui n'aboutissent pas. Ces dernières sont les arbres de tâches et sous-tâches dont les feuilles ne sont pas toutes des tâches terminales ;
3. Parmi les solutions valides, choisir la meilleure. Les critères de choix doivent être définis. Cela peut être la solution la plus rapide (en terme de temps), la solution avec le moins de tâches, ou encore impliquer le décideur dans le choix de la meilleure solution.

5.4 Le système cognitif

Notre étude et analyse de l'architecture du SIAD coopératif que nous proposons nous a permis de recenser 7 rôles classés dans 2 groupes (Figure 6). Le critère de formation du

groupe que nous avons utilisé est l'existence d'un mécanisme de communication commun au sein d'un groupe.

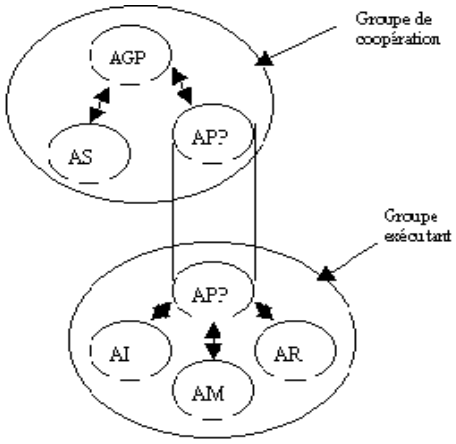


Figure 6 : architecture du SMA cognitif

5.4.1 Le groupe de coopération

Le premier groupe est le groupe de coopération, il est composé de 3 agents :

L'Agent Assistant (AS) : il assiste l'utilisateur dans ses différentes tâches, il renferme entre autres le MCU (le modèle de l'utilisateur dans l'architecture du SIAD coopératif considéré dans ce travail). l'AS : (1) Gère l'interface utilisateur selon le profile de l'utilisateur, (2) Assiste le décideur lors de l'identification du problème, (3) Introduit dans le SMA réactif les paramètres relatifs au problème à résoudre, (4) Permet de fixer, selon les choix du décideur, le mode de coopération et par conséquent les rôles attribués au décideur et au système, (5) Lance et initialise l'agent AGP, et (6) Il joue le rôle qui lui est dicté par le mode de coopération. Lorsque le mode de coopération change, l'AS change de rôle, et par conséquent change de compétences et de comportement au sein du même groupe. Il participe aussi dans le modèle de coopération du SIAD.

L'Agent APPlication (APP) : représente le savoir faire automatisable du décideur. L'APP coopère avec l'acteur participant pour l'exécution d'une tâche donnée, en utilisant un mode de coopération donné. Son rôle est de veiller à l'exécution des tâches automatisées. Selon le mode de coopération choisi ; l'APP est initialisé par L'AGP avec un rôle adéquat. Lorsque le mode de coopération change, l'APP change de rôle, et par conséquent, de compétence et de comportement. L'APP peut assurer plusieurs rôles au sein de ce groupe.

L'Agent Gestionnaire de la Coopération (AGP) : son but est de gérer la coopération entre l'utilisateur et le système pendant la résolution conjointe du problème de prise de décision. Il assure toujours le même rôle dans le groupe. Il commence par lancer l'agent APP. Ce dernier représente le système dans le groupe. L'agent AGP initialise l'agent APP avec le rôle qui lui correspond dans le mode de coopération. Il précise aussi le rôle de l'agent AS, en faisant la

correspondance entre le rôle joué par l'acteur « décideur » dans la coopération homme-machine et celui que doit assurer l'AS au niveau du groupe du modèle AGR. Ensuite, l'AGP commence l'exécution à partir de la tâche racine de l'arbre des tâches et de sous-tâches communiqué par l'agent observateur. Cet arbre de tâches et de sous-tâches représente une vue globale du plan qui spécifie l'ensemble des actions que doivent exécuter les agents.

Ainsi, le problème de prise de décision est décomposé en tâches qu'il faudra distribuer entre le système et l'utilisateur selon les compétences de chacun. Les tâches distribuées ne sont pas forcément atomiques. En effet, l'agent APP peut se voir attribuer une tâche dont il a la compétence de gérer sans pour cela qu'elle soit atomique. Il reçoit donc la branche de l'arbre qu'il sait exécuter en coordonnant entre les différentes actions automatisables et en gérant la solution partielle du problème. Ceci permet d'augmenter l'efficacité du système en diminuant le nombre de messages envoyés entre l'AGP et l'APP. De plus, cela permet de réduire la complexité du problème puisque la coordination est partagée entre les deux agents l'AGP et l'APP et n'incombe pas totalement à l'AGP.

Il existe certaines tâches terminales qui peuvent être réalisées par les deux agents coopératifs. Dans ce cas, l'attribution de ces tâches aux agents artificiels l'APP et l'AS prend en compte le rôle de chacun des 2 agents dans la coopération. Cela veut dire que le mode de coopération en cours d'exécution lève l'ambiguïté sur l'affectation des tâches aux deux agents acteurs dans la coopération. Donc, les conflits qui peuvent survenir lors de l'affectation des tâches aux agents sont résolus par la fonction du rôle que peuvent jouer les agents coopératifs.

Aussi, lorsqu'un mode de coopération n'aboutit pas à la résolution du problème, l'AGP propose au décideur un nouveau mode de coopération. Et, par conséquent une réaffectation de rôles aux deux agents. L'AGP représente le MCC (le modèle de coopération dans l'architecture du SIAD Coopératif considéré dans ce travail).

5.4.2 Le groupe exécutant

Le deuxième groupe du modèle AGR est le groupe exécutant, il est composé des agents suivants : l'APP, l'Agent de recherche d'Information (AI), l'Agent Modèle (AM), et l'Agent de Raisonnement (AR). Ce deuxième groupe matérialise le MCA (le modèle de l'application dans l'architecture du SIAD Coopératif considéré dans ce travail).

L'agent APP : cet agent participe dans ce deuxième groupe avec un autre rôle, c'est le coordinateur des actions automatisables. Ces dernières sont classées en 3 catégories (selon l'approche par les modèles du SIAD Coopératif considéré) : les actions de recherche d'informations, les actions de calcul, statistique etc. et les actions de raisonnement. L'APP à pour rôle de coordonner entre les 3 autres agents présents dans ce groupe afin de résoudre la partie du problème qui lui a été affectée par l'AGP. Au sein de ce groupe l'APP ne change pas de rôle.

L'Agent de Recherche d'Information (AI) : est dédié à la recherche d'informations dans la base de données.

L'Agent Modèle (AM) : cet agent exploite la base de modèles.

L'Agent Raisonnement (AR) : a pour rôle l'exploitation de la base de connaissances lors de l'exécution des tâches de raisonnement.

Les agents APP et AS communiquent les solutions partielles à l'AGP. Ce dernier les intègre d'une façon incrémentale et forme progressivement la solution globale. Il est le seul agent qui a une vision globale de la solution.

6 Exemple de scénario de résolution

6.1 Présentation de l'exemple

La mise en service d'une turbine à gaz est réalisée en six étapes ; chaque étape est divisée à son tour en séquences liées par un certain nombre de conditions devant être vérifiées pour permettre le passage d'une séquence à la suivante. Si une certaine condition pour une séquence donnée n'est pas remplie, la mise en service s'arrête, la séquence ne passe pas. Le processus de mise en service continue automatiquement si la condition est remplie. Les séquences sont répertoriées par étape de mise en service : Pré-démarrage, démarrage, arrêt normal, arrêt d'urgence et refroidissement. Les défauts peuvent se présenter aux différentes séquences. Les défauts de mise en service de la machine sont étroitement liés à des paramètres (pression, température, vitesse, …etc.) ; ces derniers sont mesurés par des instruments spécifiques (capteurs). Différents capteurs sont mis en place pour détecter des anomalies aux différentes étapes. S'il y a un défaut, une alarme sera déclenchée. Les différentes alarmes sont interceptées par l'opérateur humain qui a la responsabilité de prendre une décision en fonction de l'alarme déclenchée et ce, selon un mode de coopération approprié. Dans le cas où la panne est directement relevée par l'opérateur (alarme non déclenchée), ce dernier peut prendre les mesures nécessaires. Ce cas de figure se passe lorsque le problème est au niveau du capteur (panne non automatiquement signalée)

Le système devrait assister alors l'opérateur dans la détection des défauts lors de la mise en service de la machine. Il est basé sur un modèle du domaine orienté objet et un modèle de raisonnement fondé sur le paradigme tâches-méthodes (Figure 8). Le système que nous avons conçu prend en charge les différents cas de figures (alarmes signalés à l'opérateur, alarme non déclenchée - panne directement détectée par l'opérateur).

6.2 Déroulement d'une session

L'interface opérateur est constituée d'une suite de fenêtres. Celles-ci présentent une vision hiérarchique pour toutes les pannes existantes (soit au niveau d'une étape, d'une séquence ou directement signalée par une alarme). Dans le cas d'une alarme signalée à l'opérateur : Le drapeau (la référence donnée à chaque alarme) est indiqué sur le tableau de l'opérateur (dans la salle de contrôle). L'opérateur prend connaissance de l'alarme et situe le défaut.

Pour la résolution de ce problème : (1) l'opérateur énonce au système la tâche à résoudre (le problème situé) ; (2) le système identifie la tâche et la localise dans le graphe tâches-méthodes, calcul un chemin des actions à entreprendre (un sous-graphe de tâches-méthodes).Le plan d'action est élaboré en fonction des paramètres relevés directement via les capteurs ou demandés à et introduits par l'opérateur.

Deux cas se présentent pour le système: (i) la méthode terminale est « appel à l'utilisateur », la tâche est donc affectée à l'utilisateur pour la résoudre manuellement. Une interaction s'établit entre le système et l'utilisateur pour l'introduction et la présentation des données et résultats. (ii) la méthode terminale est « procédure », « requête » ou « heuristique », le système vérifie, à partir du modèle utilisateur, si elle est « redondante » (pouvant être exécutée aussi bien par le système que par l'utilisateur) : En fonction du type

de la tâche et du mode de coopération, la tâche est affectée à l'agent suggéré (système ou l'opérateur), sinon elle est prise en charge par le système.

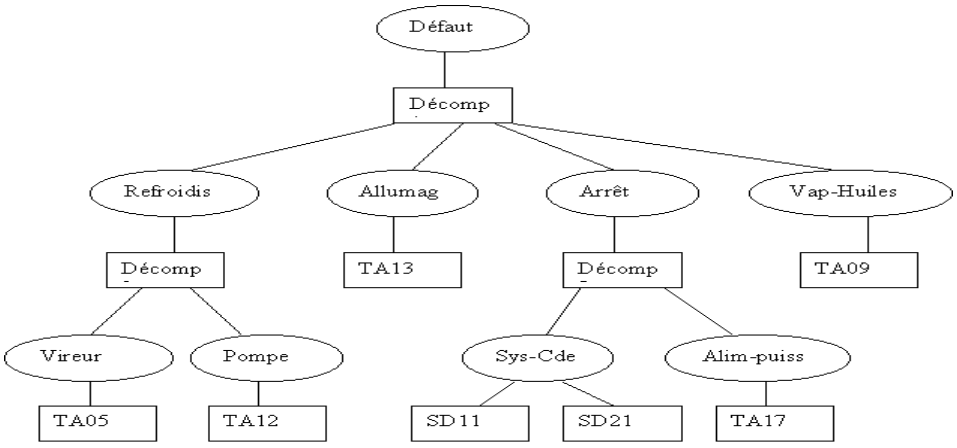


Figure 8 : Extrait du modèle tâches méthodes du traitement de défauts de mise en service

7 Conclusion

Un SIAD coopératif est doté de capacités supplémentaires afin de coopérer avec son utilisateur et de le guider dans son processus de résolution de problèmes. A ce titre, le système doit joue le rôle de collaborateur avec le décideur.

Nous nous sommes fixés dans le cadre de ce papier de définir un univers multi-agent pour la modélisation des SIAD coopératif fondé sur un modèle de raisonnement en terme de tâches-méthodes qui permet de prendre en compte les compétences de l'utilisateur et ainsi l'intégrer dans le processus de résolution. Pour réaliser une dynamique interactionnelle et une autonomie des agents, nous avons proposé un modèle hybride articulé autour d'un SMA réactif et un SMA cognitif.

Références

Holtzman, S. (1989). *Intelligent decision systems*. Addison Wesley.

Lévine, P. et J. Pomerol (1989). *Systèmes interactifs d'aide à la décision et systèmes experts*. Editions Hermès.

Marakas, G. M. (2003). *Decision support systems in the 21st century*. Prentice Hall.

Simon, H. A. (1977). *The new science of management decision*, Prentice Hall,

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

Ferrand N. (2003). *Modèles Multi-Agents pour l'Aide à la Décision et la Négociation en Aménagement du Territoire*. Thèse de doctorat, Université Joseph Fourier, France.

Ferber J. (1995). *Les systèmes multi-agents vers une intelligence collective*. InterEditions.

Jarras I et B. Chaib-draa (2002). Aperçu sur les systèmes multi-agents. Série Scientifique. Montréal.

Scott-Morton (1978).

Toussain H. (2004). Systèmes Informatiques d'aide à la décision distribués : Le modèle fédéraliste. *Mini-Workshop Systèmes Coopératifs Matière Approfondie.*

Adla A. (2007). *Architecture Coopérative pour L'Aide à la Décision de Groupe Distribuée.* Thèse de Doctorat d'Etat. Université d'Oran – Es-Sénia.

Demazeau Y. (1995). From Interactions to Collective Behaviour in Agent-Based Systems. *Proceedings of the 1st European Conference on Cognitive Science. Saint-Malo, France.*

Gutknecht O. (2001). *Proposition d'un modèle organisationnel générique de systèmes multi-agents Examen de ses conséquences formelles, implémentatoires et méthodologiques*. Thèse de doctorat. Université de Montpellier II.

Muller J. (1998). *Vers une méthodologie de conception de systèmes multi-agents de résolution de problèmes par émergence*. JFIADSMA 98. p. 355-371.

Marion N. (2006). Étude de modèles d'organisation sociale pour les environnements virtuels de formation ». Mémoire de Master 2 Recherche Informatique. IFSIC.

Summary

We propose to use the Multi-Agent Systems (MAS) technology to model the Cooperative Intelligent Decision Support Systems. These systems carry out cooperation between two agents: the decision-maker (user) and the system to resolve collaboratively a decision problem. The main idea of our work is to model the Cooperative DSS with a hybrid system which consists of reactive and cognitive MAS. The reactive one provides the potential action plans for the problem solving. These action plans are then validated before being considered by the cognitive MAS to solve the decision problem at hand. The proposed system is under development and experimentation.

Une approche d'intégration d'agents dans l'ERP

Hadia HOADJLI *, Okba KAZAR**

*Département d'informatique
université de Biskra
hadiamail2007@gmail.com
** Département d'informatique
université de Biskra
07000, Algérie
kazarokba@yahoo.fr

Résumé. Les travaux dans le domaine des ERP «Enterprise Ressource Planning » ont offert aux sociétés une maîtrise renforcée de ses activités de production grâce à une optimisation de l'utilisation des ressources. Mais ils rencontrent un problème lié à leur implantation, pour ce fait nous avons fait appel à une branche de l'intelligence artificielle qui est les systèmes multi agent afin de manier aux insuffisances des ERP, le résultat de notre recherche est une modélisation d'un système ERP composé d'agents, notre approche qui est une approche cognitive se base sur la connaissance pour assurer le bon fonctionnement du système et sa fiabilité.

1 Introduction

Les systèmes des entreprises sont un ensemble d'outils de système d'information permettant la circulation de l'information à l'intérieur de l'organisation. L'accès des entreprises aux nouvelles technologies et donc l'intégration d'outils basés sur les technologies de l'information et la communication au sein de l'entreprise, ont conduit vers l'apparition des Enterprise Ressource Planning (ERP), qui présente un outil permettant une gestion homogène et cohérente du système d'information de l'entreprise, en particulier pour la gestion commerciale de la chaîne de production jusqu'à la vente d'un produit. Ils sont conçus pour contrôler potentiellement les centaines de fonctions dans une grande organisation, Ils sont des outils utiles pour la centralisation des opérations et la prise de décision, et ils facilitent la communication entre les différents modules.

N'empêche qu'ils soufrent de certaines difficultés dans leur mise en œuvre qui nécessite un énorme volume de travail, aussi la plupart exigent un certain niveau de reconfiguration ce qui représente un risque pour la stabilité du système de l'entreprise, tous cela est due à la complexité de l'organisation de l'entreprise et au taux gigantesque d'informations.

Une branche de l'Intelligence Artificielle Distribuée consiste à ce que les composants possédant une certaine autonomie, doivent être dotés de capacités de perception et d'action sur leur environnement, on parle alors d'agents et par conséquent de systèmes multi agents. Ces systèmes deviennent indispensables dans plusieurs domaines d'applications dans le but de résoudre les problèmes de complexité surtout quand il s'agit de grand système tel que ceux des entreprises.

L'objectif de ce travail est d'intégrer la technologie multi agents dans les systèmes ERP pour manier à ces problèmes. Pour ce fait nous commençons par introduire la notion de l'ERP.

2 L'ERP et les systèmes multi agent

Dans un contexte international de plus en plus compétitif, les entreprises ont constamment besoin d'adapter et d'optimiser leurs outils industriels en vue d'augmenter leur productivité. En particulier, le pilotage et l'ordonnancement des lignes de production sont évalués en fonction de paramètres de coût et de délais, qui contribuent à définir le rendement et donc la compétitivité d'une entreprise, c'est pour cela que les système ERP ont paru, mais il reste encore le besoin accru de flexibilité, d'agilité et d'efficacité des systèmes de production qui se traduit par une complexité grandissante qu'il faut savoir maîtriser. Les système multi agents de leur coté permettent d'acquérir les caractéristiques voulues et en même temps résoudre le problème de complexité, grâce aux propriétés suivantes :

1. **La modularité :** un système ERP est composé de différents modules partageant une base de donnée unique, d'un autre coté un système multi agent est destiné pour l'implémentation de solution modulaire, ce qui fait de lui un bon outil pour les ERP.
2. **L'autonomie :** est la propriété principale d'un agent logiciel, donc l'utilisation d'un système multi agent dans un ERP lui garantie cette caractéristique qui devient de plus en plus nécessaire, pour l'optimisation des applications actuelles qui sont destinées à résoudre des problèmes nécessitant une solution distribuée, ces applications sont composés d'unités capable d'agir sans intervention externe dans le but de manier à la complexité due à ce type de problème.
3. **L'adaptabilité :** un système multi agent est adaptable ce qui veut dire qu'il a une capacité de répondre aux changements de l'environnement, nous parlons de l'auto organisation des agents. Cette caractéristique permet à l'entreprise de s'adapter aux changements du marché.
4. **L'intégration d'expertise:** dans le cas de l'ERP il a besoin de connaissances de domaines permettant la gestion de l'entreprise, les systèmes multi agent lui facilitent la tâche.

Dans le but de profiter des bénéfices que peut apporter les caractéristiques déjà citées, les systèmes ERP se dirigent vers une approche multi agent. Dans la partie qui suit nous allons présenter quelques travaux dans ce sens.

3 Synthèse de quelques travaux dans le domaine

Les projets que nous présentons, sont une intégration des systèmes multi agents dans les ERP dans le but d'apporter à ces systèmes de la flexibilité et pour les rendre plus performants vu qu'ils sont des systèmes complexes, mais chacun d'eux a utilisé une manière différente.

Le projet PABADIS a adopté une approche qui lui a permit de passer de la solution fonctionnelle vers l'intégration des systèmes multi agent, il aborde le problème de la flexibilité et la reconfiguration de systèmes de fabrication. L'approche suivie est la répartition de certaines fonctions du système d'exécution Manufacturier dans un environnement multi agent pour augmenter la flexibilité de la planification de la production. Pour atteindre ces objectifs, le projet PABADIS utilise :

- des agents mobiles pour décrire les différentes pièces, où l'agent oriente le produit associé à travers le processus de fabrication.
- des agents stationnaires représentant les installations de fabrication de l'entreprise nous parlons des machines, systèmes de transport, et bases de données.

Ainsi les agents du PABADIS ne sont implémenter qu'au niveau du MES qui sert de support à l'ERP dans le domaine de gestion de ressources et qui est bien sure un outil propre au système PABADIS.

Pendant que le projet ExPlanTech a renforcé le système ERP avec des agents qui répondent à des besoins supplémentaires spécifiques, car il intègre le système Proplant avec ses agents, qui sont destinés à la planification, dans un système ERP déjà existant, cela dans le but de profiter des bénéfices en terme de performance et de la faciliter de mise à niveau qualitative de la méthode de planification de la production. Alors les agents du système ExPlanTech ne sont pas intégrés à l'intérieur de l'ERP mais par contre il lui sert de support.

De sont coté l'approche MAERP, conserve l'ancien système d'information en plus du nouveaux système ERP, tout en ajoutant quatre agents qui ont pour rôle le transfert des données entre les deux systèmes dans le but d'éviter la perte d'information sans avoir à changer la conception de l'architecture de données utilisée.

Donc les projet PABADIS, ExPlanTech, et l'MAERP ainsi que les projets que nous avons rencontrés lors de notre recherche et qui sont des projets destinés à l'intégration de l'approche agents dans les systèmes ERP repose sur l'ajout d'agents pour des taches précise afin d'améliorer le rendement du système, et non pas sur la distributions des services du système ERP lui-même et c'est l'objet de notre recherche.

Vu que les systèmes ERP sont des systèmes complexes composés de plusieurs service partageant une base de données centralisée il est nécessaire de leurs garantir une certaine indépendance et donc une autonomie, mais une entreprise est vue comme un ensemble de sous systèmes inter opérables et complexes qui doivent coordonner leurs taches pour satisfaire les besoins des utilisateurs en temps et en qualité de réponse, c'est pour cela que nous devons garantir la communication entre les services de notre entreprise, ce qui nous mène à être sure que la solution multi agents est meilleur pour l'implantation des services de l'entreprise.

Rappelons qu'un système ERP est composé d'un ensemble de modules tel que le module de gestion commerciale, celui de gestion comptable et financière, de ressources humaines, ou de gestion client…etc. Dans le modèle que nous proposons nous allons remplacer chaque module du système ERP par un agent, aussi l'approche que nous proposons est une approche plutôt cognitive alors nos agents seront des agents cognitifs, pour plus de précision nos agents seront des systèmes experts, ils possèdent un savoir est peuvent raisonner. Aussi parce qu'un système expert regroupes les connaissances du domaine et raisonne afin d'aboutir à des résultat plus fiable, et puisque les systèmes experts sont des programmes conçus pour raisonner habilement à propos des tâches dont on pense qu'elles requièrent une

expertise humaine considérable, et c'est le cas des modules de l'ERP chacun représente tout un domaine d'expertise, nous pensons que c'est un bon choix.

4 Présentation générale du modèle proposé

L'ERP comme nous l'avons déjà présentée dans le premier chapitre est un ensemble de modules fonctionnels spécialisé qui tourne autour d'une base de donnée unique, pour ce fait et en se basant sur l'approche S.M.A qui fait distribuer l'expertise sur un ensemble d'agents qui modélisent les différents modules de l'ERP.

Nous proposons l'architecture de notre système qui est illustré dans la figure 1, et qui montre que notre système comprend un ensemble d'agents cognitifs chacun expert dans un domaine utile au fonctionnement de l'entreprise, ces agents appartiennent à cinq classe distinct es qui sont :

- La classe d'agent GS : c'est l'agent responsable de la gestion du stock.
- La classe d'agent GRH : responsable de la gestion des ressources humaines.
- La classe d'agent planification : prend en charge la fonction de planification de l'entreprise.
- La classe d'agent vente : c'est l'agent responsable du processus de vente de produits.
- La classe d'agent accueil : qui contrôle l'accès aux différents agents.

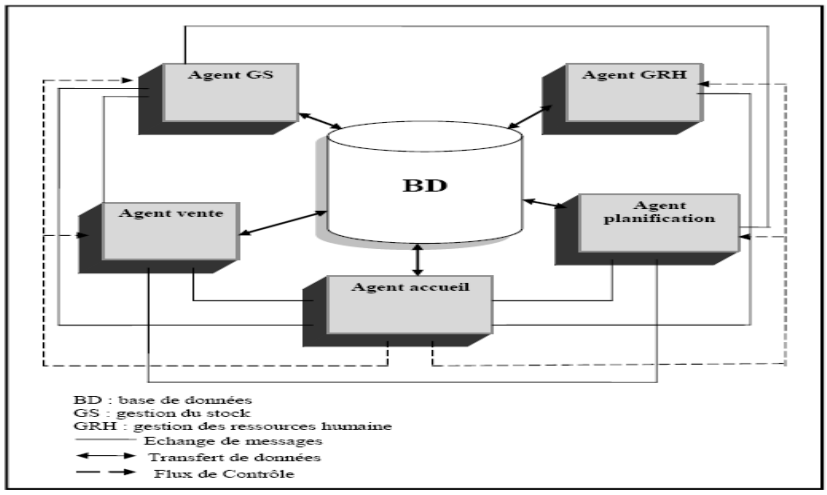


FIG. 1– *schéma représentant la structure du modèle*

Cette architecture garantie l'indépendances des module de l'entreprise et en même temps les maintient en liaison puisque les fonctions d'une entreprise ne doivent pas être isolées mais plutôt complémentaires, cette liaison est assuré via la communication des agents entre eux dans le cas de notre système il s'agit d'une communication par envoi de messages.

Le flux de contrôle indiqué dans le schéma, représente le contrôle d'accès qu'effectue l'agent administrateur sur les autre agents et qui garantie la sécurité du système.

De plus la base de données unique et centralisé qui garantie l'unicité et l'intégrité de l'information est géré bien sure par un système de gestion de base de données qui permet le transfert de ces dernière depuis et vers les agent du système.

5 Présentation détaillée des agents de notre modèle

L'architecture des agents de notre système est composée d'une base de connaissance notée «BC » qui à sont tour contient une base de règle «BR » et une base de fait «BF », de plus le deuxième composant de notre système expert est un moteur d'inférence qui est nommé « raisonneur », le raisonneur utilise la base de connaissances pour son raisonnement dans le but de déterminer le plan et contrôler l'exécution de ses actions, qui sont réparti sur un nombre de modules qui varient selon l'agent, prenons comme exemple le cas de l'agent accueil qui est illustré dans la figure suivante .

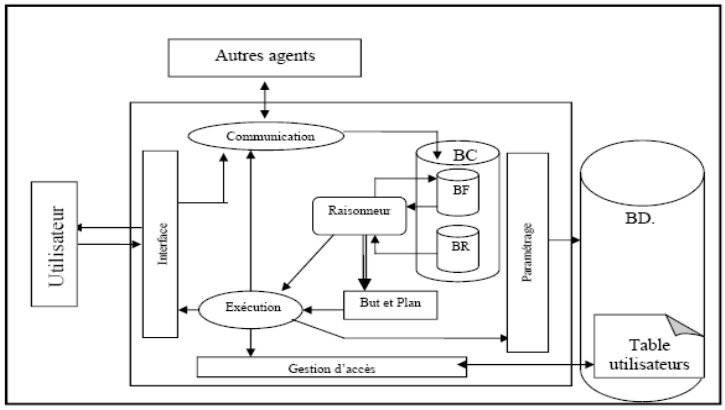


FIG. 2 – *Architecture de l'agent accueil*

5.1 Agent accueil (AC)

Il joue le rôle de l'interface qui relie l'utilisateur aux différents agents de l'ERP, il contrôle l'accès à ces derniers et assure la sécurité du système, en protégeant les comptes utilisateurs.

Dans le cas de notre système il existe deux types d'utilisateurs, le type employé et celui de l'administrateur. Les actions propres à l'agent accueil, sont réparties sur trois modules qui sont :

- **Un module interface :** responsable de la liaison de l'utilisateur avec l'agent, il collecte les données entrée par l'utilisateur, et saisie la demande d'activation de l'un des agents du système selon le choix de l'utilisateur bien sure.
- ***Un module gestion d'accès :*** qui gère les droit d'accès, cet agent effectue l'identification des utilisateurs, vu que l'accès à chaque agent est protégé par un

nom d'utilisateur et un mot de passe pour cela le module de gestion d'accès fait appel à la base de donnée unique et centralisé du système et plus précisément la table utilisateurs.

- ***Un module de paramétrage :*** ce module est accessible seulement par l'utilisateur du type administrateur, il permet d'initialiser la base de données par les informations de base nécessaires au fonctionnement, tel que les salaires de base pour chaque grade d'employé, et création de comptes utilisateur du type employé.

5.2 L'agent vente

Cet agent se charge du coté commercial de l'entreprise il prend en charge la gestion des clients car ils garde leurs coordonnées dans la table des client qui réside dans la base de données du système aussi les commande des client sont enregistrer dans une table spécifique, encore l'agent vente établit ces commandes et effectue le suivi du processus de vente.

L'agent vente comporte de plus de ses composants de bases trois modules qui lui font distingué, et qui assurent son fonctionnement. Ces modules sont :

- **Un module interface :** fournit une interface entre l'utilisateur et l'agent, il saisie les données entrées par l'utilisateur, ces données concerne soit les client ou bien leurs demandes. Aussi il peut saisir la demande de consultation de l'état commercial de l'entreprise ; par exemple ce qui concerne le budget et le taux de vente, et suite à cette demande il affiche la table associée.
- ***Un module gestion des clients :*** ce module à pour rôle l'inscription des clients ce qui veut dire la collecte d'information sur les clients, mais aussi l'enregistrement de leurs demandes.
- ***Un module de suivi :*** c'est au niveau de ce module que le processus de vente est effectué, selon la commande cet agent vérifie la disponibilité des produit requis par le client en accédant à la table de produit existant en stock, ensuite il gère la vente en établissant la facture et la mise à jour du montant de la caisse. Dans le cas ou le produit demander n'existe pas cet agent envoi une requête à l'agent de planification.

5.3 Agent GS

L'agent de gestion du stock a pour charge le suivi du magasin, càd les produits qui entrent et qui sortent du magasins aussi ce qui concerne les matières premières nécessaires à la fabrication des produits.

L'agent GS réparti son fonctionnement sur deux modules qui sont :

- **Un module interface :** cet interface permet, la saisie des informations sur les produits qui sont entrée par l'utilisateur, il permet aussi de consulter les tables contenant ces informations aux choix de l'utilisateur.
- ***Un module gestion:*** possède trois tâches à accomplir, l'enregistrement des nouveaux produits venant en stock ainsi que les information qui les accompagnent, la deuxième est la mise à jour des bases de données dans le cas d'arriver de nouveaux matériel au magasin , et la dernière c'est le retrait de produits.

5.4 L'agent planification

Cet agent est destiné à établir les délais de production, et indiquer le manque en matière première. Le fonctionnement de l'agent de planification s'effectue au niveau des les deux modules suivants :

- **Un module interface :** qui saisie les information nécessaire à la planification
- ***Un module planification:*** servant à calculer les délais de production d'un produit en utilisant les informations contenu dans la table des produits et celle des matières premières.

5.5 L'agent GRH

L'agent de gestion des ressources humaine est responsable de ce qui concerne le personnel de l'entreprise, ils se charge de la gestion du payement ainsi que les augmentation de grade.

Les modules fonctionnels propre à l'agent GRH sont les suivants :

- **Un module interface :** à pour rôle l'acquisition des informations saisie par l'utilisateur, et qui sont les informations concernant les employés de l'entreprise il permet aussi de consulter ces informations à la demande de l'utilisateur par l'affichage de la table des historiques, ou celle des employés.
- ***Un module gestion du personnel:*** ce module à pour charge l'ajout d'enregistrement lors du recrutement d'un nouveau employé , cet enregistrement contient les informations nécessaires à l'identification de l'employé et son état dans l'entreprise, nous parlons de son grade et le département au quel il appartient. Aussi il fait appel à une fonction qui gère les absences et les sauvegarde dans la table des historiques. En plus il sert dans le cas d'augmentation du grade d'un employé
- **Un module payement :** responsable du calcule de la paye en fonction de grade de l'employé et s'il a reçu une prime, et s'il a absenté durant le mois. De plus il garde un historique des primes virer et la paye dans une table de la base de donnée.

6 Fonctionnement général du système

Pour mieux expliquer la structure ainsi que le fonctionnement de notre système nous avons fait appel au langage AUML, pour la modélisation de notre système d'agent :

En premier lieu, nous allons décrire les classes d'agents existants et les relations qui les lient par le diagramme de classe AUML suivant :

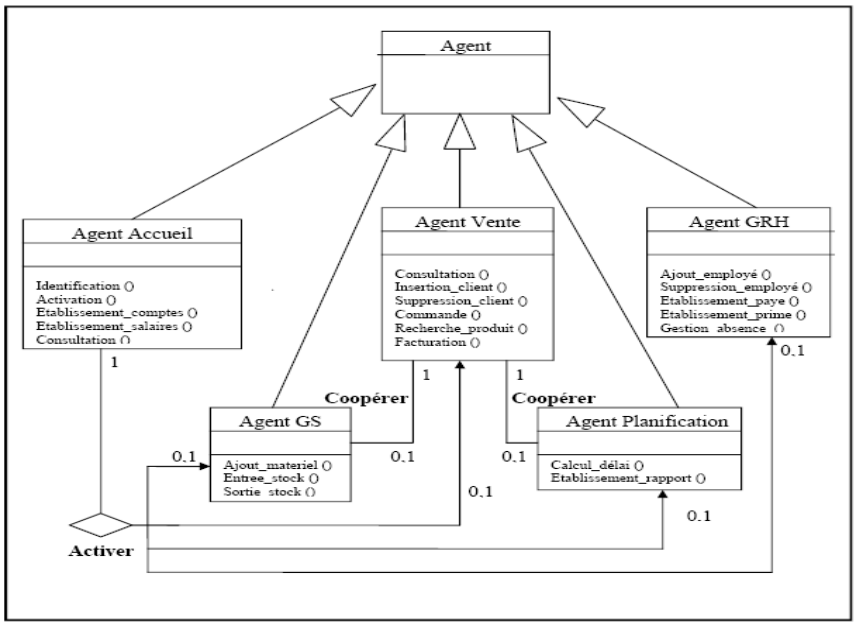


FIG. 3 – *diagramme de classe AUML du modèle proposé*

Par la suite nous allons décrire le processus de fonctionnement général du système en faisant appel au diagramme de séquence AUML.

Le premier contact de l'utilisateur avec le système sera via l'agent accueil qui va contrôler l'accès aux autres agents pour qu'ils accomplissent les taches qu'ils ont en charge, ce processus est décrit dans la figure suivante :

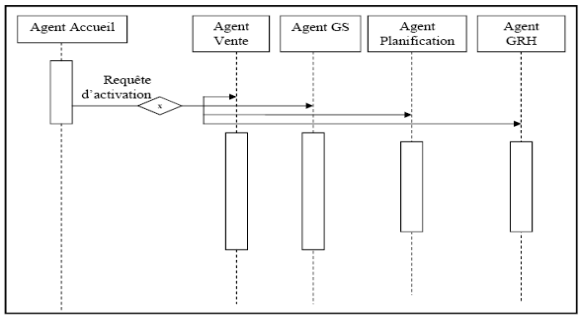


FIG. 4 – *diagramme de séquence AUML centré sur l'agent accueil*

A l'arrivée d'une commande l'utilisateur active l'agent vente à travers l'agent accueil bien sure, et suit le processus illustré dans le diagramme suivant :

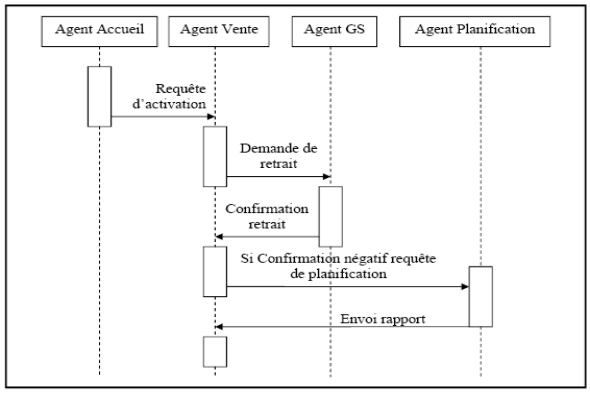


FIG.5 – *diagramme de séquence AUML dans le cas d'une commande*

Dans cette partie nous avons renforcer la présentation du processus principal du fonctionnement de notre modèle par l'utilisation des diagrammes AUML, qui ont montrés que les agents de ce modèle sont dotés d'un bon degré d'indépendance, le nombre d'interactions est minimum due à la centralisation de la base de données et son unicité, et donc la complexité du système sera minimiser.

7 Validation du modèle

Pour la validation de l'approche que nous avons proposée, nous avons établi une étude de cas portant sur l'entreprise Technique Sud Construction. Nous utilisons pour le développement de notre système un environnement JAVA, accompagné du système de gestion de base de données MySQL.

Notre système multi agents est réalisé et géré par la plateforme JADE (Java Agent Development Environment) qui est une plateforme destinée à la création d'agents logiciels et leurs gestion tout en respectant la norme FIPA, cette plateforme est implémentée en langage java ce qui garantie sa portabilité ; elle fournit un ensemble d'outils facilitant la mise en œuvre des systèmes multi agents.

Le fonctionnement de l'entreprise que nous avons étudiée repose sur trois points essentiels celui de production, de vente, et la gestion du personnel. Dans cette partie nous allons prendre comme exemple l'agent vente que nous avons définit dans le modèle que nous avons proposé de telle façon qu'ils répondent aux besoins fonctionnels de l'entreprise que nous étudions.

Cette entreprise a une gamme variée de produit qu'elle met en disposition de ses clients et qui est divisé en trois groupes, nous avons mis ces produits dans des tables accompagnés par les prix ; donc les tables de produits sont comme décrites ci dessous.

Une table de produit comporte six champs comme dans l'exemple suivant:

N° d'ordre	Désignation	*Prix H.T*	*T.V.A*	*Prix T.T.C*	*NBR*	Délai
01	*Bordures de trottoirs 0,20/100*	*160,00*	*27,20*	*187,20*		
02	*Bordures de trottoirs 0,25/100*	*190,00*	*32,30*	*222,30*		
03	*Bordures de trottoirs 0,28/100*	*210,00*	*35,70*	*245,70*		

TAB. 1 – *table des bordures en ciment CPA*

L'agent vente saisie Prix H.T entrée par l'utilisateur, et calcule les deux autre champs T.V.A et Prix T.T.C. Le dernier champ est mis à jour après chaque vente; c'est la fonction de retrait qui s'en charge et qui est invoqué par l'agent GS suite à une requête envoyée par notre agent vente.

De plus, l'activité principale dans tout système multi agent est la communication, et d'après ce que nous avons expliqué auparavant l'agent VENTE communique avec les deux agents GS et GRH, nous allons illustrer l'échange de message de notre système en prenant comme exemple la partie suivante du code de l'agent VENTE.

```
public class VENAgent extends GuiAgent {
...
// déclaration de  l'objet à envoyer dans le message
private Object[] obj=null;
string designation;
float quantité ;
...
protected void setup() {
try {
DFAgentDescription dfd = new DFAgentDescription();
dfd.setName(getAID());
DFService.register(this, dfd);
} catch (FIPAException e) {
e.printStackTrace();}
// Agent Vente envoi une requête de retrait à l'agent GS
 // donc le message doit contenir le nom du produit vendu ainsi que sa quantité
Object[] obj={"désignation","quantité"} ;
// Préparation du message
ACLMessage msg1 = new ACLMessage(ACLMessage.INFORM);
try {
msg1.setContentObject(obj);
msg1.addReceiver(new AID("GS", AID.ISLOCALNAME));
send(msg1);}
catch (IOException e) {
e.printStackTrace();}
} catch (UnreadableException e) {
e.printStackTrace();}}}}
// Reception de la réponse de l'agent GS
addBehaviour(new CyclicBehaviour(this) {
public void action() {
// Attente de la réponse de l'agent GS
ACLMessagemsg =receive(MessageTemplate.MatchPerformative(ACLMessage.INFORM));
if (msg = = null) {
// Blocage   en attente du message
block();}
protected void takeDown() {
try {
DFService.deregister(this);
} catch (FIPAException e) {
e.printStackTrace();}}...}
```

L'avantage majeur d'un système ERP réside dans sa possession d'une base de données centralisée celle du modèle proposé est schématisée par le diagramme de classe UML dans la figure suivante, ce diagramme explicite l'organisation des tables de cette base ainsi que les relations qui les lient.

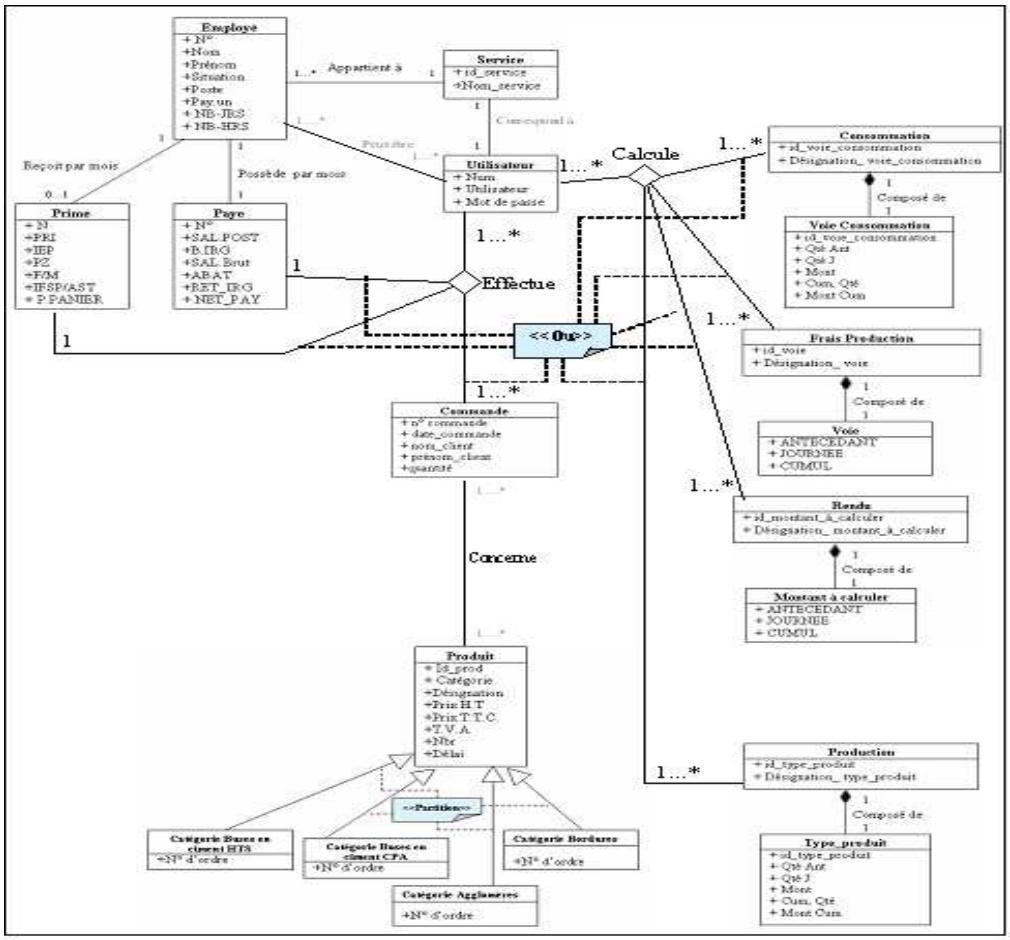


FIG.6 – *diagramme de classe UML de la base de données centralisé de l'ERP*

8 Conclusion

De par son objectif, l'ERP est amené à être modifié périodiquement pour prendre en compte toutes les évolutions et les nouvelles orientations de l'entreprise. D'où l'importance d'un système flexible qui ne court pas de risque lors du changement, et nous avons éclairci que les systèmes multi agent sont une bonne solution pour relevé se défi .

Dans ce but nous avons présenté un modèle qui incarne une approche d'intégration d'agents dans les ERP, tout en argumentant notre opinion.

L'étude du cas de l'entreprise Technique sud construction nous a permis l'exploitation de notre approche ; et sa validation à travers un système ERP composé de cinq agents, qui remplissent des fonctions parmi les plus nécessaires pour une entreprise. Ce système est réalisé dans un environnement assez performant pour la modélisation de systèmes multi agent qui est la plateforme JADE.

Nos perspectives dans ce domaine de recherches sont les suivantes :

- Rendre notre système plus distribué en introduisant des agents mobiles qui suivent les produits le long de leurs processus de fabrication.
- Nous avons pensé à apporter plus d'intelligence à notre système en introduisant un agent pour la prise de décision.
- Et pourquoi pas augmenter le niveau de raisonnement de nos agents en utilisant par exemples un raisonnement à base de cas.

Références

P. W. BLEVINS. "Enterprise Resource Planning (ERP): An Executive Perspective -- An Update". GLOVIA International, United states.

F. DARRAS (octobre 2004). "Proposition d'un cadre de référence pour la conception et l'exploitation d'un progiciel de gestion intégré". Thèse Présentée en vue de l'obtention du grade de docteur, Institut National Polytechnique De Toulouse.

Bih-Ru Lea, Wen-Bin Yu (2002), "A MULTI-AGENT BASED ERP ARCHITECHTURE", Decision Sciences Institute, Annual Meeting Proceedings.

J. Kim (2003), "Multi-Agent Based ERP". MAI Lab.

P. MASSOTTE, D. DIEP, R. BATAILLE, V. CHAPURLAT, A.MEIMOUNI, J. REAIDY (2003). "PABADIS : Plant Automation Based on Distributed Systems", Projet européen IST.

M. Pechoucek, A. Riha, J. Vokrinek, V. Marik, V. Prazma, (2005). "ExPlanTech: Applying Multi-agent Systems in Production Planning", Czech Republic.

Artem Katasonov (2008). "Introduction to JADE", Université de Jyväskylä.

Summary

Work in the field of ERP «Enterprise Resource Planning », offered the companies a strengthened control of its production activities by optimizing the use of resources. But they encounter a problem with their implementation, for this reason we used a branch of artificial intelligence which is multi agent systems to handle the shortcomings of ERP, the result of our research is modeling an ERP system consisting of agents, our approach is a cognitive approach based on knowledge to ensure the proper functioning of the system and its reliability.

Une meta-heuristique appliquée au problème d'ordonnancement avec contraintes d'indisponibilité

Azedine Later *, Ali Melit **
Mohamed benmohamed ***

* LAMEL, Université de Jijel Algérie
az_later@yahoo.fr
** LAMEL, Université de Jijel Algérie
ali_melit@yahoo.fr
*** LIRE, Université de Constantine Algérie
ibnm@yahoo.fr

Résumé. Notre étude portera sur un problème de base abondamment étudié par la théorie de l'ordonnancement, appelé *problème à m machines* : les ressources sont des machines identiques mises en parallèle, les tâches sont soumises à des contraintes de précédence et l'exécution de ces tâches nécessite des ressources qui dans notre cas seront les disponibilités des machines. Un ordonnancement alloue à chaque tâche une machine pendant un certain laps de temps en respectant les contraintes entre tâches et les périodes de disponibilité des machines. Ce modèle où les machines ne sont pas supposées constamment disponibles permet de tenir compte des tâches prioritaires comme la révision ou la maintenance d'une machine. Nous étudierons tout particulièrement les problèmes d'ordonnancement non préemptifs pour minimiser la durée totale. Pour cela nous développons une métaheuristique, basée sur des techniques dérivées de la génétique et des mécanismes d'évolution de la nature : sélections, croisements, mutations, etc.
Mots Clés : ordonnancement, indisponibilité, optimisation, métaheuristique, algorithme génétique.

1 Introduction

La plupart des travaux trouvés dans la littérature supposent que les machines sont disponibles simultanément à tout moment. Cependant, cette hypothèse peut ne pas être vérifiée, et particulièrement dans l'industrie où les ressources utilisées peuvent avoir des pannes (cas stochastique) ou être programmées pour la maintenance préventive (cas déterministe). Notre étude portera sur le problème d'ordonnancement d'un ensemble de travaux non préemptifs sur *m* machines parallèles identiques. Chaque machine possède plusieurs périodes d'indisponibilité connues a priori avec, comme critère, la minimisation de la plus grande date d'achèvement dite Makespan, notée Cmax (Maximum Completion Time). Les travaux trouvés dans la littérature se limitent essentiellement en une ou deux machines :

Pour une machine, Lee (*1996)* a montré que le problème non résumable minimisant le C_{max} est NP–difficile; cela est valable pour une ou plusieurs périodes d'indisponibilité. L'algorithme LPT (Longest Processing Times) a un rapport entre l'erreur commise et la

valeur optimale inférieur à *1/3*. Le même problème, dont l'objectif est de minimiser $\sum C_i$, a été étudié par Adiri et al (*1989)*, Lee et Liman (*1992)*, et ils ont montré que le problème est NP–difficile. Lee et Liman (*1992)* ont montré que l'algorithme SPT (Shortest Processing Times) a un rapport d'erreur inférieur à *2/7*. Imed et al *(2008)* proposent trois méthodes exactes minimisant la somme des durées pondérées pour résoudre le problème avec une seule période d'indisponibilité : une méthode branch-and-bound, un modèle de programmation en nombres entiers mixtes et une méthode de programmation dynamique.

Pour deux machines, Lin et Liao (2007) proposent une solution optimale pour un problème minimisant le C_{Max} où chaque machine peut être indisponible pendant une période fixe et connue. Pour les deux cas non résumable et résumable, *Liao et al (2005)* divisent le même problème en quatre sous problèmes dont chacun est résolu de façon optimale par un algorithme. Dans le cas ou une machine est périodiquement indisponible, *Dehua et al (2009)* ont montré que l'algorithme classique *LPT* et l'algorithme de liste (*LS*) ont respectivement des rapports d'erreur dans les pires des cas de *3/2* et *2*. *Lee et Liman (1993)* ont étudié un problème permettant de minimiser la somme des dates d'accomplissement des travaux, en autorisant une période de disponibilité sur l'une des deux machines. Ils ont proposé un algorithme pseudo-polynomial basé sur la programmation dynamique.

Lee (*1996)* a étudié le problème d'ordonnancement non résumable utilisant *m* machines parallèles, dont chacune possède une seule période d'indisponibilité, pour minimiser la durée totale. Il a montré qu'il est NP–difficile, même si $m = 2$. L'algorithme LPT et l'algorithme de liste (LS) permettant d'affecter le job à la machine qui peut l'exécuter le plut tôt possible, ont respectivement des rapports d'erreur inférieur à $(m + 1)/2$ et m.

2 Notations et présentation du problème

Une tâche représente une opération de durée entière, qui peut s'exécuter en mode non préemptif sur une et une seule machine ayant un nombre connu a priori de périodes d'indisponibilité.

Soient : $J = \{Ji,\ i = 1,2,\ ..,\ n\}$ l'ensemble des tâches à ordonnancer, Et $M = \{M_j,\ j = 1,2,\ldots..,\ m\}$ l'ensemble des machines.

- i : indice de la tache J_i
- j : indice de la machine M_j
- P_i: durée d'exécution de la tâche J_i
- K_j : le nombre de périodes d'indisponibilité sur la machine j
- $R_{j,t}$: date de début de la t^{ieme} période d'indisponibilité de la machine M_j ; avec : $j= 1\ldots m$; $t= 1\ldots K_j$.
- $\xi_{j,t}$: la t^{ieme} fenêtre de disponibilité sur la machine j : $\xi_{j,t} = [0, R_{j,1}],\ldots, [D_{j,kj}, +\infty]$
- d_i : date de début d'exécution de la tache J_i.
- C_i : date de fin d'exécution de la tâche J_i (completion time).
- C_{max} : représente la date de fin de l'ordonnancement $C_{max}=\max(C_i)$

3 Modélisation du problème

Le problème d'ordonnancement est modélisé par un graphe orienté acyclique $G= (T, E)$ avec :

- T : ensemble des tâches, $T\{T_i ; i = 0,\ldots n+1\}$, où n est le nombre des tâches.
- E : ensemble des arcs de dépendances de données entre les tâches c'est-à-dire $E=\{ e_{ij}=(T_i, T_j) ; i, j= 0,\ldots n+1 \}$. L'arc $e_{i,j}$ relie la tâche T_i à la tâche T_j , si la donnée résultat de la tâche T_i est utilisée par la tâche T_j.

Un ordonnancement est définit par une date t de début d'exécution pour chaque tâche i, et une machine m sur laquelle elle s'exécute. Les tâches T_0 et T_{n+1} sont des opérations fictives. T_0 est reliée à chaque opération sans prédécesseur ; et chaque opération sans successeur est reliée à T_{n+1}.

4 Algorithme génétique

Nous allons présenter nos choix concernant le codage des chromosomes, la fitness et les opérateurs de reproduction, adaptés à notre problème d'ordonnancement. Le schéma général de l'algorithme génétique, est donné dans FIG. 1.

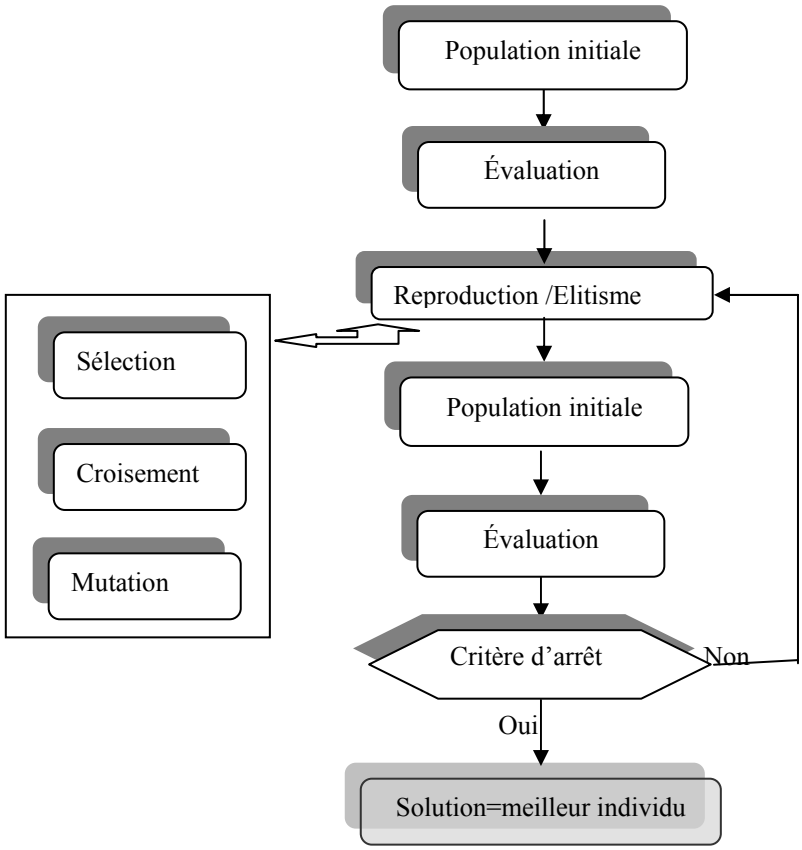


FIG. 1 - *schéma général de l'algorithme génétique*

4.1 Codage proposé pour le problème

Un individu caractérise une solution. Pour notre cas, l'individu est une chaîne de *n* string, où chaque string identifie le nom de la tâche. La chaîne représente quant à elle la séquence d'exécution des tâches, dont son ordre respecte les contraintes de précédence entres les différentes tâches. Il s'agit d'une suite de taches décomposée par niveau.

4.2 Algorithme de génération d'un individu

Pour créer un individu, on génère aléatoirement des entier entre 1 et *n* (*n* : nombre de tâches), où chaque entier correspond à l'indice d'une tâche. On place la tache générée dans l'individu tout en respectant son niveau. On répète cette étape jusqu'à l'ajout de toutes les tâches.

Début
n: nombre de tâches.
Pour chaque tâche on associe un indice *i*, tel que *i=1…n*
Tantque n≠0 ***Faire***
Générer un nombre aléatoire *r* entre 1, *n*
Si la tâche$_r$ n'est pas encore placée ***Alors***
Ajouter la tâche$_r$ à l'individu en respectant son niveau
n=n-1
Fin si
Fin Tantque
Fin

4.3 Fonction d'adaptation

Pour comparer en terme de qualité deux individus d'une population *P*, on utilise une fonction de coût (*fitness*) qui représente dans notre cas la date de fin de l'ordonnancement. Le calcul de cette fonction pour chaque individu de la population repose sur l'exploration des taches, une par une dans l'ordre généré. Chaque tâche est affectée à une période de disponibilité, donc on doit préciser la date début d'exécution de la tache et la machine qui l'exécute. On incrémente le temps à partir de la première date de disponibilité

Début
Pour chaque tâche$_i$ de l'individu ***Faire***
Si tâche$_i$ ne possède pas des prédécesseurs ***Alors***
$T \leftarrow 0$
Sinon
$T \leftarrow \max C_{i'}$ où *i'* est un prédécesseur de *i*
FinSi
Pour j=1, m ***Faire*** // *m* est le nombre de ressources //$\xi_{j,k} = [D_{j,k}, R_{j,k+1}]$
Chercher la première fenêtre de disponibilité $\xi_{j,k}$ qui vérifie $T \geq D_{j,k}$
Si *fenêtre inexistante* ***Alors*** $d_{i,,j} \leftarrow \infty$
Sinon
Si $P_i + T \leq R_{j,k+1}$ ***Alors*** $d_{i,,j} \leftarrow T$

Sinon
Poursuivre la recherche de la première fenêtre de disponibilité dont
$R_{j,k+1} - D_{j,k} \geq Pi$
Si *fenêtre trouvée* ***alors*** $d_{i,,j} \leftarrow D_{j,k}$
Sinon
$d_{i,,j} \leftarrow \infty$
Fins Si
Fin Si
Fin Si
Fin pour
$d_i = \min(d_{i,,j})$, j=1..m ; $C_i = d_i + P_i$
Affecter la *tâche*$_i$ à la *machine j* entre les instants d_i et C_i
Mettre à jour les périodes d'indisponibilité de M_j
Fin Pour
Fin

Exemple. Soient : $J = \{0, 1, 2, 3, 4, 5\}$ l'ensemble des tâches et $M = \{R1, R2\}$ l'ensemble des machines. *FIG. 2* représente l'affectation des tâches de l'individu [3, 2, 4, 5, 1, 0] aux ressources R_1, R_2.

A la date 0, R_1 est la seule ressource disponible. La tâche$_3$ est la première tâche de l'individu qui sera affectée à la ressource R_2 à la date 1, car la fenêtre de disponibilité de *R1* < à la durée de la tâche$_3$. Ensuite nous affecterons la tâche$_2$ à la ressource R_1 à la date T=0, la tâche$_4$ à R_1 à la date 4, la tâche$_5$ à R_1 à la date 1, la tâche$_1$ à R_2 à la date 5 et enfin la tâche$_0$ sera affectée à la ressource R_1 à la date 6.

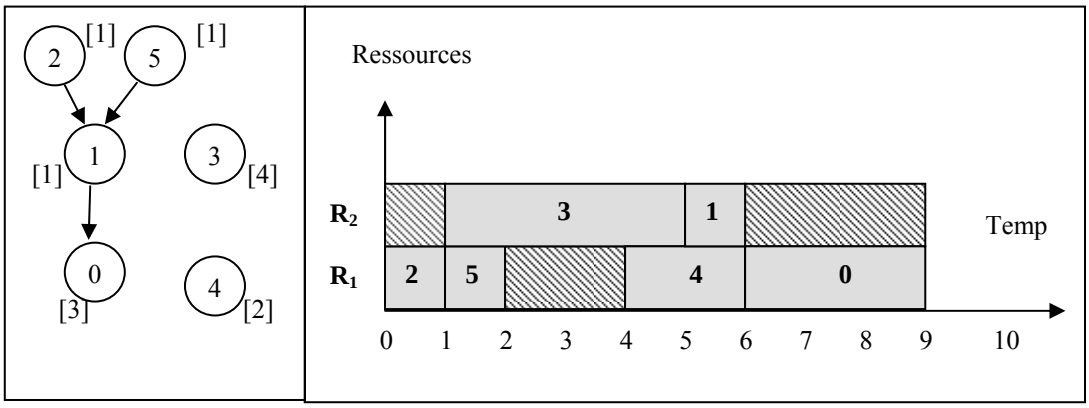


FIG. 2 - *Affectation des taches aux ressources*

4.4 Les opérateurs de reproduction

La méthode de sélection (La roulette). La sélection par la roulette consiste à diviser la roue en autant de secteurs que d'individus, où chaque individu occupe un secteur de taille proportionnelle à son adaptation. En faisant tourner la roue, l'individu pointé à l'arrêt de la

boule est sélectionné. Les individus les mieux adaptés ont donc plus de chance d'être tirés au sort lors du déroulement du jeu. Pour chaque individu, on associé une probabilité *Pi* de sélection correspondant à son adaptation dans la population. $P_i = V_i / T$, *où* V_i et *T* sont calculés comme suit: $b = \sum_{i=0}^{n} C_{\max}(Individu_i)$, $V_i = b - C_{\max}(Individu_i)$, $T = \sum_{j=0}^{i} V_j$. Pour sélectionner un individu, on doit calculer la probabilité cumulée comme suit : $Q_i = \sum_{j=0}^{i} P_j$. Par exemple si on effectue la sélection sur une population de 3 individus : A, B, C, où chaque individu correspond à une séquence d'exécution des tâches, pour chaque individu on associe un entier correspondant à son évaluation ($C_{\max}$), 8, 5, 7 respectivement.

Donc b= 8+5+7=20.
V_A=20 - 8=12, V_B=20 -5=15, V_C =20 -7=13, T=12+15+13=40.
Enfin la probabilité de chaque individu P_A=12/40=0.3, P_B=15/40=0.375, P_C=13/40=0.325.
La probabilité cumulée pour chaque individu : $Q_{A=}$=0.3, $Q_{B=}$=0.3+0.375=0.675, $Q_{C=}$=0.3+0.375+0.325=1.

Une fois les probabilités cumulées sont calculées, on génère aléatoirement un réel r sur l'intervalle [0,1] N fois (N étant la taille de la population précédente). L'individu V_i est sélectionne lorsque : Qi-1 <r< Qi. Voici l'algorithme de la roulette :

Roulette (*population*$_i$ **,** *liste_sélect*, *N*)
Début
Population$_i$ **:** population à l'itération *i*.
Liste_sélect : la liste des individus sélectionnés
Qi : la probabilité cumulée de l'individu$_i$
N : nombre d'individus à sélectionner
Pour i=1, N ***Faire*** // sélectionner *N* individus
Générer un nombre aléatoire *U* ;
Sélectionner un *individu*$_i$ telle que :
Qi-1<U <Qi
Ajouter l'*individu*$_i$ à *liste_sélect* ;
Fin Pour
Fin

Le croisement. Cet opérateur choisit deux individus parents *père1* et *père2*, pour générer deux autres individus enfants *fils1* et *fils2* à partir d'un seul ou plusieurs points de croisement choisi, aléatoirement. Dans notre cas on utilise le croisement à 1 point avec le codage par valeur.

Croisement (*père1*, *père2*, *fils1*, *fils2*)
Début
père1, père2 : deux individus père
fils1, fils2 : deux individus fils à créer
Générer un nombre aléatoire *i* entre *1, n-1* // *n* : nombre total des tâches.
Pour j=1 à i ***Faire***
Copier la $tâche_j$ du *père1* dans *fils1*
Copier la $tâche_j$ du *père2* dans *fils2*
Fin pour
Pour j=1 à N ***Faire***
Si $tâche_j$ du *père1* n'existe pas dans *fils2* ***Alors***
Copier la $tâche_j$ du *père1* dans *fils2*
Fin Si
Si $tâche_j$ du *père2* n'existe pas dans *fils1* ***Alors***
Copier la $tâche_j$ du *père2* dans *fils1*
Fin Si
Fin pour
Fin

La mutation. Cet opérateur choisit aléatoirement deux points d'un même individu (*père*), si les deux tâches se trouvent dans le même niveau, alors on génère un autre individu (*fils*) en faisant des permutations.

Mutation (*père*, *fils*)
père: individus père
fils: fils à créer

Début
Père : individu père
Étape 1 **:** Choisir aléatoirement deux positions *i* et *j* tq $0 \le i, j \le$ taille du *père* et $i \neq j$
Étape 2 **:** Permuter entre la $tâche_i$, et la $tâche_j$ du *père*
Étape 3 **:** ***Si*** le nouvel individu respecte les contraintes de précédence ***Alors***
fils ← *père muté*
Sinon
Annuler la permulation
Aller à Étape1
Fin Si
Fin

L'élitisme. L'élitisme permet de garder les individus les mieux adaptés d'une génération à la suivante. A chaque génération, le meilleur individu, ayant le C_{max} minimum, passera à la génération suivante.

5 L'algorithme général

Pour un problème d'ordonnancement, un individu représente un point de l'espace de recherche, une solution potentielle ainsi que la valeur du critère à optimiser (*son adaptation)*, lui est associée (C_{max}). La reproduction des individus d'une population repose sur les processus de sélection, de croisement et de mutation.

Au début, une population initiale d'individus est générée aléatoirement. Le passage d'une génération *i* à la génération *i+1 se fait comme suit :* Dans un premier temps, la population est reproduite par *sélection* où les « bons » individus se reproduisent mieux que les mauvais, au sens du critère considéré. Ensuite, un *croisement* aux paires d'individus (les parents) d'une certaine proportion de la population (une probabilité *Pc* généralement autour de 0.6) est appliqué pour en produire des nouveaux (les enfants). Un opérateur de *mutation* est également appliqué à une certaine proportion de la population (probabilité *Pm*, généralement très inférieure à *Pc*). Un opérateur d'élitisme est appliqué pour faire passer le meilleur individu. Enfin, les nouveaux individus sont évalués et intégrés à la population de la génération suivante. Le critère d'arrêt de l'algorithme est basé sur le nombre de générations donné comme paramètre.

Population $_i$: la population à l'itération *i*
T_i : taille de la *population*$_i$.
Nbr : nombre de génération.
Liste_sélecte : liste contient les individus sélectionnés par la roulette.
P_c : probabilité de croisement.
P_m : probabilité de mutation.
C **:** *C*max pour le meilleur individu de la population$_i$.

Début
Etape1:
Générer aléatoirement une population initiale (*population*$_0$) de taille T_0
Calculer la fitness de chaque individu de population$_{i-1}$;
Calculer les probabilités cumulées
C=Cmax(le meilleur individu)
i :=1
Etape2 :
Tanque $i \leq Nbr$ ***Faire***

La sélection
Roulette (*population*$_{i-1}$**,** *liste_sélect*, T_{i-1}**)** //Sélectionner T_{i-1} individu à partir de la *population* $_{i-1}$ où T_{i-1} est la taille de la *population*$_{i-1}$
Le croisement
Pour j=1 à $P_c * n/2$ ***Faire*** // n : taille de *liste_select*
Choisir deux individus aléatoires *père1*, *père2* à partir de *liste_sélect*
Croisement (*père1*, *père2*, *fils1*, *fils2*)
Ajouter *fils1*, f*ils2* à la *population*$_i$
Supprimer *père1*, *père2* de *liste_sélect*
Fin pour

Transférer les individus sélectionnés ne participant pas au croisement vers la *population$_i$*

La mutation

Pour *j*=1 à $p_m * T_i$ ***Faire*** // T_i *:* la taille de la *population$_i$*
Choisir un individu aléatoire *indiv*, à partir du *population$_i$*
mutation (*indiv ,indiv_muté*) ;
Supprimer indiv de *population$_i$*
Ajouter *indiv_muté* au *population$_i$*
Fin Pour

L'élitisme

Faire passer le meilleur individu de *population$_{i-1}$* vers *population$_i$*

L'évaluation des l'individus de population$_i$

Calculer la fitness de chaque individu de population$_{i-1}$;
Calculer les probabilités cumulées
C=Cmax(le meilleur individu)
i := i+1

Fin Tanque

Solution_Optimale ← le meilleur individu de la *population$_i$*

Fin

6 Discussion des résultats expérimentaux

Notre algorithme possède quatre paramètres : la taille de la population, la probabilité de croisement, la probabilité de mutation et enfin le nombre d'itérations. Nous avons utilisé comme langage de programmation, le langage Java. Nous avons utilisés un jeu de test varié selon le nombre de taches avec leurs contraintes de précédente et le nombre de ressources avec leurs contraintes d'indisponibilité. Pour assurer l'existence d'une solution au problème, nous supposons qu'il existe au moins une machine dont toutes ses périodes d'indisponibilité sont limitées. En faisant varier les paramètres, nous avons remarqué que plus le nombre d'itération s'élève, plus on se rapproche mieux vers la solution optimale et plus qu'on favorise les opérations de croisement par rapport au mutation plus qu'on se rapproche mieux vers la solution optimale. Nous pouvons expliquer ces résultats par le passage de la meilleure solution de la population courante vers la population suivante d'une part et d'autre part par la prise en compte des deux stratégies d'intensification et de diversification.

7 Conclusion

En ce document nous avons traité un problème de m machines parallèles avec plusieurs périodes d'indisponible sur chaque machine. Nous avons proposé une méthode évolutive permettant de minimiser le C_{Max}. Nous savons que la solution donnée par l'algorithme, si elle n'est optimale elle représente une borne supérieure au problème. Une direction évidente

de la future recherche est de développer un algorithme permettant de trouver une borne inférieure qui sera utilisée dans le test d'arrêt de l'algorithme génétique. Ça va nous servir, d'une part pour d'améliorer la performance et d'autre part pour vérifier la convergence en utilisant des tests sur des instances qui seront générées aléatoirement.

Références

Adiri, I. Bruno J., E. Frostig, and A.H. G Rinnooy Kan. (1989). Single machine flow-time scheduling with a single breakdown. Acta Informatica, 26 :679-696

Dehua Xu, Zhenmin Cheng, Yunqiang Yin and Hongxing Li(2009). Makespan minimization for two parallel machines scheduling with a periodic availability constraint. Computers and Operations Research 36: 1809-1812

Imed Kacem ,Chengbin Chu et Ahmed Souissi (2008). Single-machine scheduling with an availability constraint to minimize the weighted sum of the completion times. Computers and Operations Research 35: 827-844

Lee C. Y. and S. D. Liman.(1992). Single machine flow-time scheduling with scheduled maintenance. Acta Informatica 29: 375-382.

Lee C. Y. Surya Danusaputro Liman (1993). Capacitated two-parallel machines scheduling to minimize sum of job completion time., Discrete applied mathematics 41: 211-222

Lee C. Y.(1996). Machine scheduling with an availability constraint. Journal of Global

Optimization 9:395-416.

Liao Ching-Jong Der-Lin Shyur and Chien-Hung Lin (2005). Makespan minimization for two parallel machines with an availability constraint. European Journal of Operational Research 160: 445-456

Lin Chien-Hung Liao Ching-Jong (2007). Makespan minimization for two parallel machines with an unavailable period on each machine. Int J Adv Manuf Technol 33: 1024–1030

Moschiov. G.(1994). Minimizing the sum of job completion times on capacitated parallel machines. Mathl. Comput. Modeling, 20: 91-99.

Summary

Our study will relate to a basic problem abundantly studied by the theory of scheduling, called problem with m machines: the resources are identical machines put in parallel, the tasks are subjected to constraints of precedence and the execution of these tasks requires resources which in our case will be the availabilities of the machines. A scheduling allocates with each task a machine during a certain amount of time by respecting the constraints between tasks and the periods of availability of the machines. This model where the machines are not supposed constantly available makes it possible to take account of the priority tasks like the revision or the maintenance of a machine. We will study the problems of scheduling particularly nonpreemptive to minimize the total duration. For that we develop

metaheuristic,based on techniques derived from the genetics and mechanisms from evolution of nature: selections, crossings, changes, etc

Key words: scheduling, unavailability, optimization, metaheuristic, genetic algorithm.

A Novel Decisional Clonal Selection Artificial Immune Support: Applied for Ultrasonic Motor Speed Control

Mehdi DJAGHLOUL

Department of Electrotechnic, UFAS University, Setif, Algeria.
Djaghloul_mehdi@yahoo.fr
http://membres.lycos.fr/djaghloulmehdi/

Abstract—In this paper we propose a novel decisional support based on a Clonal Selection Artificial Immune System (CSAIS). This system represents our own vision of decision task, applied for non linear system behaviour control. Based on the basic version of CSAIS, we propose a modified version (MCSAIS) to accomplish an improved controller applied for strongly nonlinear system control. Our modifications on CSAIS are done after a detailed study on its operators' parameters' influences. This last is acquired from our simulation of CSAIS using Netlogo tool for bi-dimensional optimization. The main objective of this study is to present the feasibility of nonlinear system control using the studied decisional system MCSAIS approach by reducing some control constraints as convergence speed, and control error order. Our control approach is applied for Senshai USR-60 travelling wave USM speed control, which is considered as strong nonlinear system behaviour. The whole control strategy and model simulation is realized on Matlab software. Simulation results, especially fluctuation test ones, are very satisfactory, and give more ways to the AIS as decisional system to be used as efficient intelligent control approach for the new nonlinear systems.

1 Introduction

In actuator domain, classic actuators are surpassed by a newer generation based on piezoelectric material properties. This generation completes the motorization application field, and represents a new interesting alternative element of development and evolution for this domain (Bekiroglu, 2008), (Wojciech, 2006). Among different existing actuator classes; ultrasonic piezoelectric motors (USMs) class is considered as one of the best evolved by several use properties view point (Bekiroglu, 2008). Different from their classic electromagnetic congeners, they have: special functional principles performing physical phenomena exploitation (Sashida, 1982, 1993),(Ueha et al, 1993), mechanisms work ensuring moves orientation and motorization (Flynn, 1997),(Kandare et al, 2002),(Fernandez et al, 2004),(Tsai et al, 2003),(Xu et al, 2003), an own modelling describing intern behaviours and functional correspondences (Kandare et al, 2002), (Fernandez et al, 2004), (Tsai et al, 2003), (Xu et al, 2003), (Hagood et al, 1995),(Matteo, 2005),(Bigdeli et al, 2005), (Senjyu et al, 1998a ,2006), and practical conditions use (Boumous et al, 2007), (Huafeng et al, 2004),(Djaghloul et al, 2007,2008), (Storck et al, 2002).

Ultrasonic motor is a newly developed motor. It is an exceptional type motor which it is a device that transforms vibration and wave motions of solids into progressive or rotational motions by contact frictional forces (Huafeng et al, 2004),(Djaghloul et al, 2008).The detailed principle is ensured by the ultrasonic vibration force of piezoelectric elements constructing this motor (Sashida, 1982),(Flynn, 1997). Due to its principles and mechanisms work, USM motor has an excellent performance and many other useful features. The most important of its characteristics are: high torque at low speeds, compactness in size, no electromagnetic interference, short start–stop times, and many others (Ueha et al, 1993),(Storck et al, 2002), (Djaghloul et al, 2007),(Kebbab et al, 2008).

Owing to the properties and the advantages mentioned above, the USM attracted considerable attention and has been used in many practical applications (Bekiroglu, 2008), (Ueha et al, 1993), (Maeno, 2006). It is used as MEMS, in robots (Sun et al, 2007), (Yamano, et al, 2005), in medical instruments, in cameras and aeronautics field. All this domains applications require a quick response and efficient speed or position control. They reflect the precise aspect use of this type of motors in the real applications.

Since this type of motors is a peculiar one, his driving principle is different from the other electromagnetic-type and its characteristics have not been elucidated in detail. It has strong nonlinear speed characteristics (Hong-Wei et al, 2007a), (Zhang et al, 2006). The properties of the speed vary with the driving frequency, voltage, load, solid state, and many other factors (Boumous et al, 2007), (Djaghloul et al, 2008). It is therefore, difficult to construct efficient control approach for the USM speed. According to the conventional control theory, an accurate mathematical form model should be set up and elaborated. For the USM motor, this idea is so difficult to be used for performing an effective control approach. In the same time, the simulation and control for USM are crucial in the actual use of such systems (Bigdeli et al, 2005), (Senjyu et al, 1998a, 2006), (Boumous et al, 2007), (Djaghloul et al, 2008), (Maeno, 2006).

In this application environment, and in order to guaranty a high-quality use of the USM motor, we must consider appropriate control approaches. These control approaches must compensate interne behaviour of this motor without prior knowledge (Hong-Wei et al, 2007a, 2007b) ,(Xu et al, 2005),(Zhang et al, 2006), and ensure a global functional stability. Indeed, the defined controls nature, for this actuator, must take in consideration its functional characteristics and its use limits (Senjyu et al, 2006), (Boumous et al, 2007). These controls are also confronted to the modelling complexity of this motor, due to its specificities, in usual cases (Boumous et al, 2007), (Djaghloul et al, 2007).

For our control view of USM, the use of an intelligent technique to accomplish the task of control is conceivable. It seems adequate following the depicted constraints, on the USM use, above (Zhang et al, 2006), (Xu et al, 2005), (Hong-Wei et al, 2007b). In this study, we opt to use the Artificial Immune Systems (AIS) to ensure the control task. Artificial immune systems (AIS) are adaptive systems inspired by theoretical immunology and observed immune functions principles and models (Nunes de Castro, 2000), (Aickelin et al, 2005). They are applied to problem solving (Nunes de Castro, 2000), (Nunes de Castro et al, 2000), (Joanne et al, 2003). Used in engineering, AIS systems with their different forms and

algorithms (Aickelin et al, 2005) represent an interesting technical solution provider (Nunes de Castro et al, 2000). Systems control is one important touched field by the AIS to be explained with (Joanne et al, 2003). As particular choice, the Clonal Selection kind is improved in our study, by its attitude to imitate the strong survival behaviour (Nunes de Castro et al, 2000) (Joanne et al, 2003), to be used as direct control approach.

We expose so in the present work, in first our technical view on Clonal Selection AIS modification, after detailed study and simulation of its basic version. As an improvement simulation test, Clonal Selection AIS is applied for a bi-dimensional recursive optimization task, and accomplished on the Netlogo tool (Wilensky,1999). Based on its simulated behaviour and a detailed study on their interne parameters, CSAIS is modified to MCSAIS, following our own functional orientation view point and the study technical needs. In this work MCSAIS is used as direct speed control approach for USM. In the section tree, we develop and detail this idea by proposing our structure conception and enhancement on this control use. For the section four, we present the simulation results of our control approach by two kinds of tests which are step type tracking tests and complex trajectory pursuit test. Finally, we conclude the study by the found results and many notices.

2 Clones Selection Immune System

2.1 The Basic Clonal Selection AIS

CSAIS is an adaptive system using the clonal selection to converge to the best solution (Joanne et al, 2003),(Nunes de Castro et al, 2002),(Yunyi et al, 2007), proposed for the first time as an optimization algorithm by Leandro Nunes de Castro and Fernando Von Zuben (Nunes de Castro, 2000),(Nunes de Castro et al, 2000), and named in their study (CLONal selection ALGorithm). It was applied for several tasks like the shape recognizing, global optimisation and other (Joanne et al, 2003).The basic clonal selection system as an algorithm evolution is described as follows:

1. Generate initial antibodies (each antibody represents a solution Ab).
2. Compute the fitness of each antibody.
3. Select antibodies from population which will be used to generate new antibodies Ab_n (the selection can be random or according to the fitness rank).
4. For each antibody, generate clones and mutate each clone according to fitness (maturation).
5. Eliminate antibodies with lower fitness form the population, and then add to the population the new antibodies Ab_d to complete the population size.
6. Repeat the steps from 2-5 until stop criterion is met. The number of iterations can be used as the stop criterion.

To ensure the good evolution of the showed functional diagram of the proposed system behaviour (CSAIS), several operators are detailed following our study needs.

2.1.1 Selection operator

This operator consists to select and order the best antibodies by the criteria of their affinity (fitness) to be used as support population treatment. The size of the selected population ns must be defined by the algorithm user relating to the problem solving knowledge.

2.1.2 Clonal operator

Following the antibodies fitness rank in the selected population, the clonal operator creates a number Nc_i of clones for every antibody from the selected antibodies population. The number Nc_i is the number of clones for the (i^{th}) antibody in selected antibodies population, and its formula is:

$$Nc_i = Round\left(\frac{\beta \cdot N}{i}\right) \tag{1}$$

- i : The rank of the ith antibody (1=the best, n_s =the worst),
- N : Total number of antibodies (selected and not selected),
- β : Clonal multiplicity factor (0< β <1),
- $Round$: For just round up to the nearest integer.

The total size of clones' population for all the antibodies of the selected population is C :

$$C = \sum_{i}^{n_s} Nc_i \tag{2}$$

2.1.3 Maturation operator

Relating to natural immune system, the change (mutation) on antibodies touches more the low-functional ones, according to their efficiency rank, and selects the best functional antibodies by the elimination of worst ones; this is the affinity (fitness) maturation of the antibodies population. Therefore, in the artificial form, the mutation is applied at a rate witch is inversely proportional to the affinity (better fitness, less mutation). The mutation rate α can be calculated by the formula (3):

$$\alpha = \exp\left(-\rho \cdot \mathrm{f}\right) \tag{3}$$

- f : Normalized affinity (fitness),
- ρ : Mutation (maturation) multiplicity factor (0< ρ <1).

Applied as additional term, the mutation rate α makes effect on total clones' population C, and the new ones C*can be calculated as:

$$C^* = C + \alpha \cdot ND(0,1) \quad (4)$$

- C^* : Clones' population after affinity (fitness) maturation ,
- C : Clones' population,
- ND : Normal Gaussian random variable (mean = 0 and standard deviation = 1).

2.2 Study on CSAIS operators' parameters

According to the presented operators above, Clonal Selection AIS (CSAIS) has four parameters, which influence its evolution, efficiency use and application. These parameters are: clonal factor β, mutation factor ρ, best selected population size n_s and the total population size N.

In order to test the influence of these parameters, on the CSAIS evolution and efficiency, we have simulated global behaviour of this last for a bi-dimensional recursive optimization on the Netlogo software (Wilensky,1999).The elaborated tests on CSAIS using Netlogo tool have revealed for each parameter specific influence properties. The properties of the defined parameters are detailed as follows:

- Total population size N : as moderated number, it allows population to cover a large area of the research space, to be near to the desired solution by the proposed candidate solutions (population).
- Best selected population size n_s : constrain the pre-selected candidate solutions to be modified by clonal and maturation (mutation) operators. As limited number, n_s should be inferior or equal to the total population size N , and can be evaluated following a wanted population quality to represent candidate solutions for the next iteration. More n_s value augments, and more the previous candidate solutions mark is lost, and this can cause a probably divergence. n_s and β values could be treated as combined values (formula (2)), for their influence, in the same time use , and for a best evaluation of each one of them.
- Clonal factor β : constrain the clones' number for each ordered candidate solutions, following the selection criterion, to ensure a new population generation.
- Mutation (maturation) factor ρ : limits mutation step, applied to each candidate solution (formula (4)). This step has a great impact on the rabidity convergence to attempt the desired final solution.

2.3 The Modified Clonal Selection AIS (MCSAIS)

The basic form of the CSAIS system includes tree main dynamic phase evolutions which are: clonal, selection, and maturation. Over than their mathematical objectives, the formulas describing these operators present two interne behaviours: collaboration and competition between antibodies.

The collaboration aspect is treated in the CSAIS using normalized affinity (fitness) to create a king of pursuit to the best candidate solution. The candidate solutions are maturated following this imposed pursuit attitude. By this orientation manner, the enhancement of all the candidate solutions is ensured recursively.

The competition aspect appears in the clonal operation by ordering the candidate solutions to be cloned following their performance ranks. The performance ranks represent the positions of the solutions affinities using a considered criterion test. Seeing the formula (1) the clonal operation uses these ranks to favourite a clonal force for each candidate solution.

In the follow of this section, we propose our own contribution on the operators form. We present a new view of competition and collaboration aspects in the CSAIS. The modified Clonal Selection AIS (MCSAIS) system keeps the same global structure as the basic one. Based on the presented study of the basic CSAIS parameters influence, done in the previous subsection, we opt to change operators' effect for the selection operator and the maturation one.

2.3.1 Competition by the selection effect

In the selection operator, competition aspect can be introduced by an adequate use of constrained number for the antibodies to be cloned. This action is done in order to control the antibodies clonal force distribution in the next population. For the basic version of CSAIS system, selection is inspired from the natural immune system behaviour according to the antibodies survival low. More the antibody is feeble and more it is a candidate to die. The same idea is present in the artificial version, but in this one the antibodies' number in the population is pre-defined.

To give more competition using this operator, we have limited the number of the new (cloned) antibodies in the next population by a percentage factor η (0 to 1). The other $(1-\eta)$ antibodies percentage is the feeblest ones treated in the previous iteration. This action is made to guaranty a kind of covering and antibodies' candidature diversity in the new search space (in the actual iteration). This idea is very useful to ensure antibodies competition inter generation in the global population evolution.

To calculus the adequate antibody rank limit (i^{th}), with out tacking care of a fixed n_s, we can use as base the formulas (1) and (2).The following formula describes the calculus relation of the used (i^{th}) to ensure the wanted competitive selection.

$$\sum_{i} \left(\frac{1}{i}\right) \underset{1^{st}}{\geq} \frac{(\eta \cdot N)}{Round\,(\beta \cdot N)} \tag{5}$$

We denote that in this formula, we seek the first i value ensuring the inequality relation. Practically, the calculated i value represents our dynamic n_s following β, N and η which are defined values.

2.3.2 Collaboration by the maturation effect

The maturation (mutation) operator represents, after the clonal action, one of original ideas of CSAIS system. It describes, with an explicit manner, the utility of the collaboration for the success of CSAIS system use. Maturation in the natural immune system is the behaviour of adaptation ensured by the antibodies to confront a new pathogen with no priori knowledge; it is the nature evolution.

The basic CSAIS system treats the idea of maturation by formula (3) and (4). Following the formula (4), maturation is ensured by a calculated rate and an affected random value to each antibody. Changes on the maturated antibodies are mainly related to the use of the affinity normalization, in maturation rate α (formula (3)). As their basic use, the random element and the maturation rate don't favourite collaboration between antibodies. Especially, in the particular case, when antibodies are static in the research space and far than the desired solution. In other words, when all antibodies aren't near of the desired solution, following the considered criterion and their affinities (fitness) are so near in value.

To favourite a collaborative behaviour in CSAIS, by this operator, we introduce a new king of following of all antibodies to the best one. We created an adaptive factor influencing the antibodies modifications. This factor represents a rapport between the best affinity (fitness) and the allowed calculated error in the CSAIS use, iteration by iteration.

As a description of this factor influence, it imposes collaboration on all antibodies by an adaptive change. More the best antibody is far than the desired solution, it contributes to great changes on the antibodies which are worse than it. And more this antibody is near from the desired solution, the presented factor is near in value from 1. In this changes case are just done by the classic maturation of CSAIS like in the formula (3). With a studied maturation factor, precision is ensured. The new formula of this operator, using our modification on, is:

$$C^* = C + \alpha \cdot ND(0,1) \cdot \left(\frac{Best\_fitness}{Tolerance} \right) \quad (6)$$

- C^* : Clones' population after affinity (fitness) maturation,
- C : Clones' population,
- ND : Normal Gaussian random variable (mean = 0 and standard deviation = 1),
- $Best\_fitness$: Best antibodies' affinity met in the iteration,
- $Tolerance$: The allowed error using the considered criterion.

3 USM Speed Control using MCSAIS

3.1 The control scheme specification

In this section, we elaborate and present a novel controller specially designed to control the strong nonlinear behaviour systems. Our control approach uses the MCSAIS as bases, we name it MCSAISC controller. The proposed online control strategy and model can be used for nonlinear systems control especially when a direct controller cannot be designed due to the complexity of related system model (Senjyu et al, 1998a), (Djaghloul et al, 2008), (Xu et al, 2005).

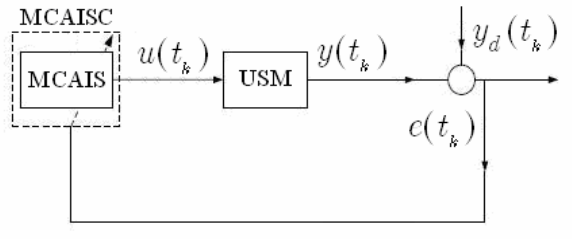


FIG.1 – *Adopted direct control scheme.*

The used system for this work, USR-60 ultrasonic motor (Shinsei, 2005), is considered as a system having strong nonlinear speed behaviour (Senjyu et al, 2006). In this work, USR-60 ultrasonic motor is used to be controlled and to test the performance of the proposed controller (MCSAISC controller). Global use of MCSAISC controller and the control strategy and model are illustrated in Fig.1.

For the presented control scheme, blocks elements are USM as ultrasonic motor, MCSAISC as modified Clone Selection AIS controller, $u(t_k)$ as system input, $y(t_k)$ as system output, $y_d(t_k)$ as desired output, $e(t_k)$ as control error.

3.2 USM functional parameters

Since we have to control USR-60 ultrasonic motor, we shall present its global fictional parameters and their adequate environment use, following the objectives of this study. For USR-60 dynamic evolution, we treat in this work as output the rotational speed, which represent the most interesting output to be controlled. As input, we treat the voltage driving frequency of this motor. Our knowledge, on the USR-60 motor speed dynamic behaviour, are acquired from a modified USM model based on Bigdeli and Senjyu groups works (Bigdeli et al, 2005), (Senjyu et al, 2006). Those works study the correspondence between the inputs constrain and the outputs in USM dynamic using the Hammerstein modelling .Hammerstein modelling is a nonlinear modelling in which we separate between the static correspondences and the dynamic ones (Bigdeli et al, 2005), (Zhang et al, 2008).

For the USM motor, this idea is applicable, where static correspondences can be doing between the inputs constrain (voltage driving frequency, voltage amplitude, phase difference and load torque effect), and the finale values of the outputs (speed and position). For the dynamic aspect, we can model the normalized outputs dynamic. Several like dynamic modelling can be used to ensure an adequate description of this aspect behaviour. We denote as most used ones: Ordinary Differential Equation system, Neural Network models or simple Fuzzy models (Zhang et al, 2008).The used USR-60 USM motor model description is presented as appendix part in this work.

USR-60 motor is alimented by two voltages sinusoidal signal forms, and they are characterized by three parameters: amplitude of the inputs voltage, driving frequency, and phase difference between the two sinusoidal signals. In our work, the amplitude of the input voltage and the used phase difference are fixed, as 100 Volt and 90 deg respectively. Following the used USM model, the adopted control input is the voltage driving frequency. The variation interval of this input is limited to [41..42] kHz.

By an analogy to the proposed control scheme (Fig.1), the bond-outputs, in this scheme, are defined as follows: $u(t_k)$ as driving frequency, $y(t_k)$ as USM speed, $y_d(t_k)$ as desired speed, $e(t_k)$ as speed control error.

3.3 CSAIS affinity function parameters adjustment

CSAIS is an intelligent system, able to be adaptive following desired behaviour objectives. These behaviour objectives describe the efficiency use of CSAIS for good identification or control approaches. For CSAIS, adaptive behaviour can be presented by its affinity function.

In this study, the affinity function used for MCSAIS is a time function exposing the Euclidian distance between the actual rotational USM speed and the desired. The rotational USM speed depends on the MCSAIS generated output, which is our estimated control input of USR-60 ultrasonic motor. The used model describing the USR-60 speed behaviour ensures perfectly the depiction of the correspondence between the driving frequency and the rotational speed. Therefore, the used affinity function can be presented as function using the mathematical model description of the input and its direct influence on rotational USR-60 speed behaviour.

As detail, the following formula describes the used affinity function, with : $y(t_k|\theta)$ as USR-60 rotational speed, $y_d(t_k)$ as desired rotational speed, t_k time sequence and θ as control input value following times sequence.

$$F_k(\theta, Z_k) = \sqrt{\left(y_d\left(t_k\right) - y\left(t_k|\theta\right)\right)^2} \tag{7}$$

The function $\hat{\theta}$ represents the search base function to find the optimum input control value $\hat{\theta}$. In our case study, finding $\hat{\theta}$ returns to minimize this function using θ and Z_k. Z_k is a training sequence test following t_k. The calculus of $\hat{\theta}$ is ensured by the formula in follow:

$$\begin{aligned} \hat{\theta} &= \arg\min\left(F_k(\theta, Z_k)\right), \\ \theta &= \left[u(t_k)\right], \\ Z_k &= \left[y(t_k), u(t_k)\right]. \end{aligned} \tag{8}$$

4 Numerical Simulation and Results Discusses

Numerical simulations, in this paper are performed using the proposed approach for the USR-60 USM motor speed control. Some fixed parameters of this USM model are taken as driving frequency 41 kHz, amplitude of driving voltage 100 Volt, phase difference 90 deg, maximum allowed load torque 0.2 Nm, rotation speed 3.5 rad/s.

In order to check the efficiency of the proposed control approach, firstly, the speed reference is chosen as simple and multilevel step type reference for the step tracking tests. In second, sinusoidal behaviour is used as a complex trajectory tracking test.

4.1 Step type tracking tests

As first treated tests, the step level control is realized in this study. Fig.2 and Fig.4 show the control results with two different speed level variations reference manner. In the Fig 2 one step level control is presented, where the dashed line represents the fixed speed reference at 2.5 rad/s and solid line represents the controlled rotational USM speed.

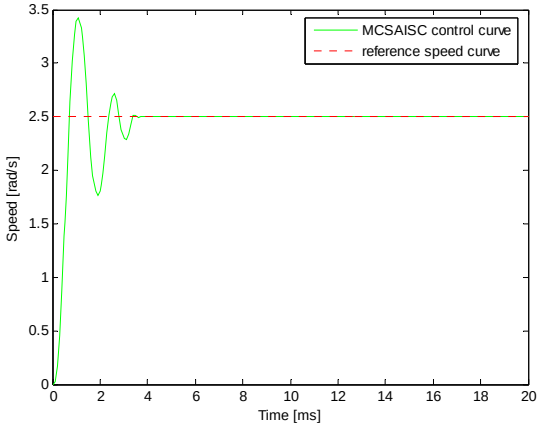


FIG.2 – *MCSAISC control result and reference speed curves for the simple step type reference.*

In this test the control error is limited to 0.001 following a defined tolerance on the affinity function of MCSAIS. This error limitation is depicted in Fig.3, which is an enlargement of the steady evolution in Fig.2.

From Fig.3, it can be seen that the amplitude of the speed fluctuation using the proposed approach is significantly very small at the steady state speed evolution .The fluctuation degree ζ is defined as:

$$\zeta = \left(V_{\max} - V_{\min}\right) / V_{ave} \times 100\% \tag{9}$$

Where $V_{\max}$, $V_{\min}$ and V_{ave} represent the maximum, minimum, and average values of the speeds. According to the control studies done by Senjyu (Senjyu et al, 1998b), Shi (Shi et al, 2004) and Hong-Wei(Hong-Wei et al, 2008), the calculated speed fluctuation ζ in these works are 5.7 %, 1.9 % and 0.1 % respectively (Hong-Wei et al, 2008). Using MCSAISC controller approach the value of the speed fluctuation ζ is reduced to less than 0.08. This comparison shows that our control is very appropriate and so efficient.

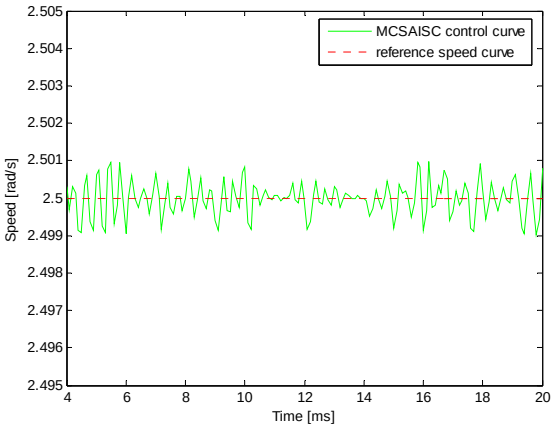


FIG.3 – *An enlargement of Fig.2 in the time window [4ms, 20ms].*

Fig.4 presents the multilevel step type test for the speed control, using MCSAISC controller. In this figure two level speed reference step type are used as reference. This test approves the ability and the adaptive behaviour of MCSAISC controller.

The enlargement of the control evolution in the time window [9ms,13ms] shows the rapidity of the proposed MCSAISC controller reaction. The present test makes in value the capacity of this controller to surmount the brisk jumps and variations of the reference. The controller imposed reaction time is reduced by more than 60% comparing to the time done for the initial motor start to steady state time (4[ms]) .

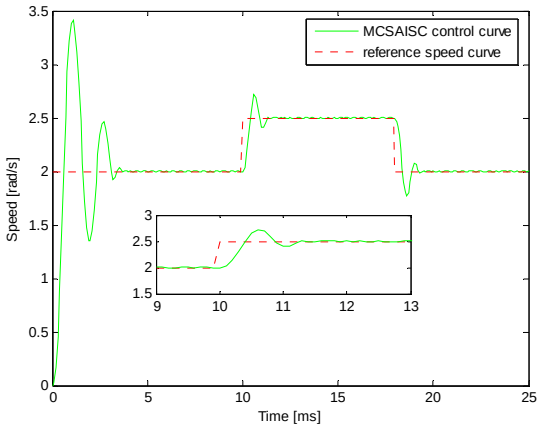


FIG.4 – *MCSAISC control result and reference speed curves for the multiple step type reference, with an enlargement of the evolution in the time window [9ms,13ms].*

From this figure we can see that our MCSAISC control strategy can ensure a very good pursuit to the reference just after the first 4 [ms], witch is very useful for the direct application of this motor.

4.2 Complex trajectory tracking test

As second test type of our proposed control approach, using MCSAISC controller, we propose a complex trajectory tracking. Sinusoidal behaviour is the used trajectory to be tracked in this test. Fig.5 shows the reference and speed control curves, where the reference speed evolution follows a sinusoidal behaviour: the dotted line represents the reference speed curve while the solid line represents the control result curve based on MCSAISC controller.

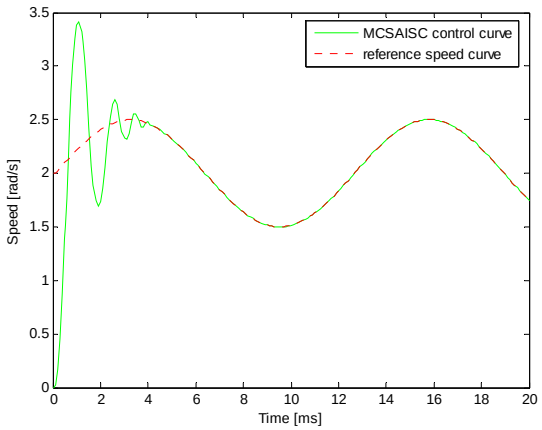


FIG.5 – *MCSAISC control result and reference speed curves for the sinusoidal reference.*

From the elaborated tests (see figures 2 to 5), it can be seen that the proposed controller performed successfully and efficiently, with rapid reaction and high control precision.

5 Conclusion

This paper has treated the strongly nonlinear system control using Modified decisional Clonal Selection Artificial Immune System (MCSAIS). CSAIS is chosen, in this work, relating to its interesting adaptive behaviour and its ability to surmount strong nonlinear variations. Seen as its basic form, CSAIS can be used for solving engineering problems, like identification and control systems, following its capacity to be modified and oriented. In this work, modifications are done on the parameters operators influence in this AIS system. In order to study the influence effect of these operators, we have simulated behaviour variation of CSAIS following to their parameters. This simulation was done using Netlogo software and applied for a recursive bi-dimensional optimization. Based on this simulation we have acquired more knowledge on these parameters influence and use. To improve the basic CSAIS we have proposed our own study and operators modifications. The enhancement view point, on CSAIS, favourites more the collaborative and competitive aspects in CSAIS. We named MCSAIS the modified CSAIS. MCSAIS possesses high speed convergence and good performance; it can be used as control approach of strongly nonlinear systems.

In this paper, main ideas are to describe, to analyze, and to discuss a controller based on MCSAIS. This controller is designed for USR-60 USM speed control. The application tests of MCSAISC controller for speed USM are: the step type reference and sinusoidal behaviour reference. The step type tracking test is done to depict the robustness of the proposed controller for the brisk variations, and the sinusoidal behaviour tracking is done to show the pursuit rapidity reaction of this controller and its precision.

Numerical simulations, of the detailed tests above, show that the proposed approach control has obtained satisfying results. These results approve the successes, performance of the controller as rapidity reaction and high precision esurient. Therefore, the designed controller is quite very well and can be applied to any nonlinear systems, point view behaviour, control.

Appendix A. The Hammerstein model of "USR-60" USM

The mathematic model of the USR- 60 USM (Senjyu et al, 2006) [14] used in this paper is represented by the following equations:

$$V_r \max(f) = 3.415 - 4.584 \cdot (f - 41.0) + 2.230 \cdot (f - 41.0)^2 \quad \text{(A.1)}$$

$$V_r dec(\tau) = 0.975 - 3.415 \cdot \tau \quad \text{(A.2)}$$

$$\phi dec(\tau) = 43 \cdot \tau + 100 \cdot \tau^2 \quad \text{(A.3)}$$

$$\mathrm{V_r}(f,\phi,\tau)=\mathrm{V_r}\max(\tau)\cdot \mathrm{V_r}dec(\tau)\sin\left(\frac{2\pi}{360}\cdot\frac{90\cdot\left(\phi-\phi dec(\tau)\right)}{90-\phi dec(\tau)}\right) \tag{A.4}$$

$$\begin{cases} V(s)=\dfrac{\omega_n^2}{s^2+2\zeta\omega_n s+\omega_n^2}, \\ \zeta=0.184595 \text{ , damping coefficient} \\ \omega_n=2.30856 \text{ , natural frequency} \end{cases} \tag{A.5}$$

Explanations of the above symbols are given below:

f	driving frequency	τ	load torque
$\mathrm{V_r}dec(\tau)$	speed drop following τ	$\mathrm{V_r}(f,\phi,\tau)$	static speed model
$\phi dec(\tau)$	dead zone width following τ	$V(s)$	dynamic speed model
$\mathrm{V_r}dec(\tau)$	speed drop following τ	$\mathrm{V_r}(f,\phi,\tau)$	static speed model

References

Aickelin, U. and Dasgupta, D, (2005). Artificial Immune Systems Tutorial. In: Introductory Tutorials in Optimisation, *Decision Support and Search Methodology*. Kluwer Academic Publishers, Dordrecht, Boston. 2005.

EJoanne H. Walker, Simon M. Garrett, (2003): Dynamic Function Optimisation: Comparing the Performance of Clonal Selection and Evolution Strategies. *ICARIS 2003:* 273-284.

Erdal Bekiroglu (2008) Ultrasonic motors: Their models, drives, controls and applications, *J Electroceram* 20:277–286 , *Springer* 2008.

F.Z.Kebbab, M.Djaghloul, Z.Boumous, S.Belkhiat (2008): Rotary Ultrasonic Motors: Daimler-Benz AWM 90–X TWUSM motor, Experimental and Simulation mechanical characteristics», *2nd International Conference on Electrical Engineering Design and Technologies ICEEDT*, 2008.to appear in JES journal 2009 issue.

Flynn, A.M (1997): "Piezoelectric Ultrasonic Micromotors" *MIT Artificial Intelligence Laboratory*, December, 1997.

G. Kandare and J. Wallaschek (2002) : "Derivation and validation of a mathematical model for traveling wave ultrasonic motors" *Smart Mater. Struct*.,vol.11, pp. 565-574, 2002.

H. Storck , W. Littmann, J. Wallaschek, M. Mracek (2002) : "The effect of friction reduction in presence of ultrasonic vibrations and its relevance to travelling wave ultrasonic motors", *Ultra-sonics* vol. 40,pp. 379–383.2002.

Hong-Wei Ge , Yan-Chun Liang , Maurizio Marchese , (2007):A modified particle swarm optimization-based dynamic recurrent neural network for identifying and controlling nonlinear systems" , *Computers and Structures* 85 , pp 1611–1622 ,2007.

Hong-Wei Ge, Feng Qian,Yan-Chun Liang,Wen-li Du, LuWang; (2008) Identification and control of nonlinear systems by a dissimilation particle swarm optimization-based Elman neural network nonlinear analysis - *real world applications* ;2008 ; 9(4) pp1345-1360 .

Hong-Wei Ge, Wenli Du, Feng Qian, Zhencheng Ye, (2007) Speed Identification of Ultrasonic Motors Based on Evolutionary Elman Network, icnc,pp.471-475, *Third International Conference on Natural Computation (ICNC 2007)*, 2007 .

Ikuo Yamano Takashi Maeno, (2005), "Five-fingered Robot Hand using Ultrasonic Motors and Elastic Elements", *International Conference on Robotics and Automation* ,Barcelona, Spain, April 2005

J M. Fernandez, M. Krummen, Y.Perriard (2004): "Analytical and Numerical Modeling of an Ultrasonic Stepping Motor Using Standing Waves", *Ultrasonics Symposium*,0-7803-8412-1/04 IEEE, 2004.

L. Huafeng Z. Chunsheng G. Chenglin (2004) , Study On The Contact Model Of Ultrasonic Motor Considering Shearing Defomation ,*Journal of ELECTRICAL ENGINEERING*, VOL. 55, NO. 7-8, pp. 216-220. 2004.

Leandro N. de Castro & Jon Timmis (2002): An Artificial Immune Network for Multimodal Function Optimization, *Proceedings of IEEE Congress on Evolutionary Computation (CEC'02)*, vol. 1, pp. 699-674, Hawaii, May 2002,.

Leandro Nunes de Castro, (2000) : Artificial Immune Systems: Theory and Applications" ,*Brazilian Symposium on Neural Networks, SBRN* 2000.

Leandro Nunes de Castro, Fernando J. Von Zuben (2000):The Clonal Selection Algorithm with Engineering Applications, In Workshop Proceedings of GECCO, pp. 36-37, *Workshop on Artificial Immune Systems and Their Applications*, Las Vegas, USA, July 2000.

M.Djaghloul ,Z.Boumous, (2007) : Structure, Function, Existing Forces in the Ultrasonic Motor , *International conference on Sciences and Techniques of Automatic control STA*, pp 56-79 ,2007.

M.Djaghloul, Z.Boumous, S.Belkhiat (2008): Driving Forces Study In Ultrasonic Motor » , AIP Conf. Proc.Volume 1019,pp.420-426 in *1st Mediterranean Conference on Intelligent Systems and Automation(CISA08)* June 12,2008.

Matteo.B. Mémoire de doctorat (2005): Modélisation et commande de moteurs piézoélectriques à onde progressive , *Université Lausanne* 2005.

MS Tsai, CH Lee, and SH Hwang (2003): "Dynamic modeling and analysis of a bimodal ultrasonic motor", *IEEE Trans. Ultra-son., Freq. Contr*, vol. 50, no. 3, pp. 245-256, 2003

N. Hagood and A. J. McFarland (1995): Modelling of a piezoelectric rotary ultrasonic motor, *IEEE Tmns. Ultmson., Fermeleet . Freq., Gontr*., vol. 42, no. 2, pp. 210-224, 1995.

Nooshin Bigdeli, Mohammad Haeri (2005): Simplified modeling and generalized predictive position control of an ultrasonic motor, *ISA Transactions* 44 ~2005! 273–282.

Qiao Zhang, Xu Xu, Yanchun Liang, (2006) : Identification and Speed Control of Ultrasonic Motors Based on Modified Immune Algorithm and Elman Neural Networks. *RSCTC 2006*: 746-756

Senjyu T, Miyazato H, Yokoda S, Uezato K (1998): Speed control of ultrasonic motors using neural network. *IEEE Trans Power Electron* 1998;13(3):381–7.

Shi XH, Liang YC, Lee HP, Lin WZ, Xu X, Lim SP, (2004) Improved Elman networks and applications for controlling ultrasonic motors. *Appl Artif Intel* 2004;18(7):603–29.

Shinsei 2005;USM ultrasonic motor general catalogue ,2005.

T. Sashida (1982): Trial construction and operation of an ultrasonic vibration driven motor, OYO BUTURI, (in Japanese) vol. 51, no. 6, pp. 713-720,1982.

T. Sashida, T. Kenjo (1993), An Introduction to Ultrasonic Motors", *Clarendon, Oxford*, 1993.

T.Senjyu, S.Yokoda, H.Miazato, K.Uezato (1998) : Speed control on ultrasonic motors by adaptive control with a simplified mathematical model, *IEE Proc-Electr. Power Appl*, Vol 145, NO 3, may 1998.

Takashi Maeno (2006), Ultrasonic Motors and Their Applications, 1st International Symposium on Next-Generation Actuator Leading Breakthroughs, , pp. 123-126, 2006.

Tomonobu Senjyu , Toshihisa Funabashi (2006) ,Mathematical Model of Ultrasonic Motors for Speed Control, *IEEE* 2006.

Ueha, S. Tomikawa (1993) : "Ultrasonic Motors Theory and Applications", *Clarendon, Oxford*, 1993.

Wilensky, U. (1999). NetLogo. http://ccl.northwestern.edu/netlogo/. *Center for Connected Learning and Computer-Based Modeling, Northwestern University*, Evanston, IL.

Wojciech SZLABOWICZDEA, (2006) : Contribution au dimensionnement et à la réalisation d'actionneur piézoélectrique à rotation de mode fort couple pour applications aéronautiques, *Génie électrique de l'INPT*, 2006 .

X. Xu, Y.C. Liang, H.P. Lee, W.Z. Lin, S.P. Lime, X.Shi, (2005): A stable adaptive neural-network-based scheme for dynamical system control, *Journal of Sound and Vibration* 285 ,pp 653–667 ,2005.

Xin-liang ZHANG, Yong-hong TAN , (2008), Modelling of ultrasonic motor with dead-zone based on Hammerstein model structure' *Journal of Zhejiang University SCIENCE* A ISSN 1673-565X (Print); ISSN 1862-1775 (Online) ,58 9(1):58-64 , 2008.

Xu X, Liang YC, Lee HP, Lin WZ, Lim SP, Lee KH (2003). Mechanical modeling of a longitudinal oscillation ultrasonic motor and temperature effect analysis. *Smart Mater Struct 2003*;12(4):514–23.

Yunyi Zhu, Shangce Gao, Hong-Wei Dai, Fangjia Li, and Zheng Tang (2007) : Improved Clonal Algorithm and Its Application to Traveling Salesman Problem, *IJCSNS International Journal of Computer Science and Network Security*, VOL.7 No.8, August 2007.

Z.Boumous, M.Djaghloul, Z.E.Kheribeche, S.Boumous, S.Belkhiat (2007) : Simulation of Ultrasonic Piezoelectric Motor , *1st International Conference on Electrical Engineering Design and Technologies ICEEDT*, 2007.

Zhijun Sun , Rentao Xing , Chunsheng Zhao , Weiqing Huang (2007): Fuzzy auto-tuning PID control of multiple joint robot driven by ultrasonic motors, *Ultrasonics* 46 ,pp303–312, 2007

Summary

Résumé—Dans cet article on propose un nouveau support décisionnel basé sur le système immunitaire artificiel à sélection de clone (CSAIS). Ce system représente notre propre vision de la tache de décision appliquée pour la commande des systèmes à comportement non linéaire. Basée sue la version basic du (CSAIS), on propose une version modifiée (MCSAIS) afin d'accomplir un contrôleur amélioré appliqué pour la commande des systèmes fortement non linéaires. Nos modifications portées sur CSAIS sont faites après une étude détaillée sur les influences des paramètres de ses operateurs. Cette dernière est acquise de notre simulation de CSAIS en utilisant l'outil Netlogo pour une optimisation bidimensionnelle. L'objective principal de cette étude est de présenter la faisabilité de commande des systèmes non linéaires en utilisant l'approche du système décisionnel MSCAIS, tout en renduisant quelques contraintes de commande comme la rapidité de convergence et l'ordre de l'erreur de commande. Notre approche de commande est appliquée pour la commande de la vitesse du moteur ultrasonique à onde progressive USR-60 de chez Senshai, qui est considéré comme un système a comportement fortement non linéaire. La globale simulation de la stratégie de commande et du modèle est réalisée sous le logiciel Matlab. Les résultats de simulation, et particulièrement ceux du test de fluctuation, sont très satisfaisants et donnent plus d'opportunités à AIS, vu comme système décisionnel, a être utilisés comme approche de commande intelligente efficace pour les nouveaux systèmes non linéaires.

Index Terms— AIS system, Clonal Selection AIS, Travelling wave USM, Decisional system and support, Intelligent control, Nonlinear system control.

Paradigme structural pour l'alignement stratégique du système d'information

Azedine Boulmakoul[*], Noureddine Falih[*], Rabia Marghoubi[**]

[*] FST Mohammedia, Département informatique, B.P. 146 Mohammedia MAROC
INPT – 2, av ALLal EL Fasse - Madinat AL Irfane - Rabat – MAROC

Azedine.boulmakoul@yahoo.fr
nourfald@yahoo.fr
m.rabia@inpt.ac.ma

Résumé. Les systèmes d'information sont fortement sensibles aux évolutions stratégiques de l'entreprise : changement organisationnel, mutation des objectifs, variété modifiée, nouveaux objets et processus métier, etc. Dans l'objectif de maîtriser l'alignement stratégique du système d'information, nous proposons une approche centrée sur le méta-modèle de l'entreprise ISO/DSI 19440. A ce méta-modèle nous proposons une extension intégrant les structures nécessaires, empruntées au référentiel COBIT relatif aux processus IT. Afin de mieux conduire les évolutions du système d'information, cette extension permet l'intégration des outils systémiques, basés sur le paradigme structural. Les aspects *objectif* et *métriques décisionnelles* sont aussi considérés dans cette méta-modélisation.

1 Introduction

L'entreprise est un système sociotechnique dynamique complexe qui se définit comme une totalité organisée de composants en interactions, selon une finalité (Le Moigne 1983), (CIGREF 2002), (Vernadat 2001). L'entreprise moderne est fortement structurée par des processus informatiques répondant aux différents processus métiers. Le système d'information garantit la communication entre le système opérant et le système décisionnel ainsi que l'échange avec l'environnement. Le système d'information est fortement sensible aux évolutions stratégiques de l'entreprise, changement organisationnel, mutation des objectifs, variété modifiée, nouveaux objets et processus métier, etc. Selon Scott Morton (Scott et al. 1994), les facteurs influençant les orientations stratégiques d'une organisation sont sommairement présentés sur la figure 2. Dans l'objectif de maîtriser l'alignement stratégique du système d'information, nous proposons une approche centrée sur le méta modèle de l'entreprise ISO/DSI 19440 étendu (ISO/19440 2007). Cette extension intègre les structures nécessaires pour la création des outils systémiques, basés sur le paradigme structural, afin de mieux conduire les évolutions du système d'information. Un système d'information (SI) est un composant important d'une organisation. Les composantes d'un système d'information incluent le logiciel, le matériel, les procédures et les personnes. Le système d'information, coordonne grâce à l'information, les activités de l'organisation et lui permet ainsi d'atteindre ses objectifs. Il est le coordonnateur de la communication dans l'organisation. De plus, le SI

représente l'ensemble des ressources organisées pour : collecter, stocker, traiter et communiquer les informations (figure 1).

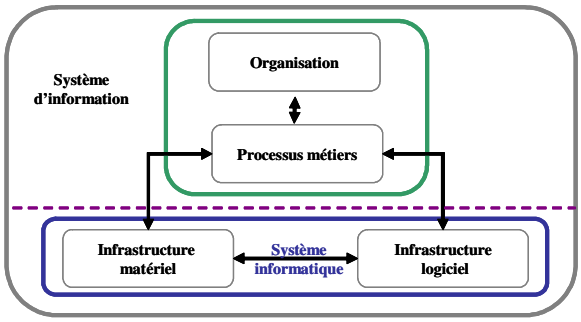


FIG 1 – *Système d'information/système informatique.*

L'étude de l'alignement stratégique des systèmes d'information a été largement étudiée par des chercheurs dans les deux dernières décennies (Lederer et Sethi 1992), (Earl 1993). La recherche dans ce domaine inclut des études prescrivant des méthodologies et des pratiques techniques, des études décrivant les modèles conceptuels, des études identifiant des facteurs de succès ou des problèmes/impacts, des études de cas pour la validation et le test d'hypothèses (Brown 2004). Cet article est structuré comme suit, dans la section 2, nous rappelons tout d'abord les fondements de l'alignement stratégique. Dans la section 3, nous donnons dans un premier temps, un aperçu sur la modélisation d'entreprise et les principaux courants ayant conduit au développement des langages de modélisation. Ensuite nous présentons les relations entre les différents composants de COBIT. La section 4 aborde le méta-modèle ISO 19440 pour appréhender les facettes de l'alignement stratégique et d'intégrer le point de vue structural pour l'édification des outils systémiques pour un meilleur pilotage de l'évolution du système d'information. Dans la section 5, une panoplie de structures algébriques est proposée. Pour chaque classe de structure nous précisons son rôle et sa contribution à la gouvernance du système d'information. La conclusion de ce travail situe le plan d'actions proposé dans cet article pour en souligner l'apport et les limites ainsi que les investigations futures à développer.

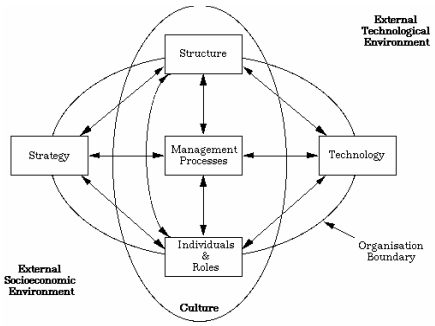


FIG 2 –. *Influence des objectifs des organisations.*

2 Alignement stratégique

L'utilisation stratégique de la technologie de l'information, mieux connue en tant qu'alignement stratégique, a sensiblement augmenté, en raison de la forte dépendance de l'organisation des activités à l'égard des systèmes d'information et leurs technologies supports. L'alignement stratégique est considéré comme un élément clé pour améliorer la performance des organisations dans l'objectif d'accroître l'efficacité et l'efficience et de permettre à des organisations d'être plus concurrentielles dans leurs industries respectives.

Traduit littéralement de «*strategic alignment*», l'expression « alignement stratégique » exprime l'idée d'établir et de suivre un cap. Il s'agit de mettre en cohérence la stratégie du système d'information avec la stratégie de l'entreprise sur le métier (CIGREF 2002), (Shimizu 2006). Lederer et Sethi (1992) définissent l'alignement stratégique des systèmes d'information comme *" 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 cette citation :

"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 et Venkatraman 1999); (Ward and Peppard 2002).

Il y a quatre buts importants pour s'engager dans la formulation de la planification stratégique des systèmes d'information (figure 3) (Lederer et Sethi 1988); (Ward et Peppard 2002) :

- Alignement : identifiant les applications informatiques susceptible d'aider l'entreprise à atteindre ses objectifs métiers.
- Impact : recherche des applications à impact important, qui aideraient l'organisation à gagner un avantage de compétitivité sur le marché.
- Développement d'une infrastructure technologique flexible et rentable,
- Développement des ressources et des compétences requises afin de déployer le système d'information avec succès à travers l'organisation.

Une des premières étapes vers l'alignement stratégique est de disposer d'outils permettant de le mesurer. Des approches courantes d'évaluation, bien que, principalement focalisées au niveau stratégique, fournissent peu de finesse aux niveaux tactiques et opérationnels, qui sont identifiés en tant que domaines importants pour réaliser l'alignement stratégique.

En outre, la plupart des approches existantes sont examinées dans de grands organismes et il y a peu de recherche pour évaluer l'efficacité de ces approches dans de petites et moyennes entreprises. Ce travail propose des instruments systémiques, à base de l'analyse structurale, plutôt que se focalisant seulement au niveau stratégique il vise aussi à mesurer l'alignement aux niveaux tactiques et opérationnels.

Paradigme structural pour l'alignement stratégique

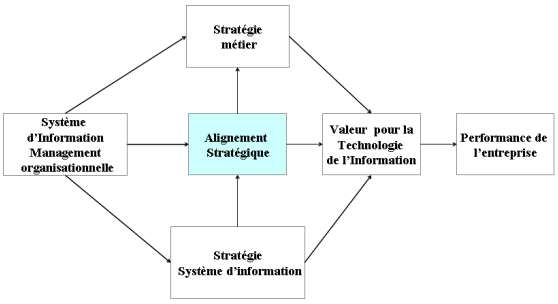


FIG 3 – *Contexte de l'alignement stratégique.*

3 Modélisation de l'entreprise

Le concept d'entreprise, telle qu'elle est comprise dans le cadre de la modélisation d'entreprise, réfère un ensemble organisé d'activités mises en oeuvre par des ressources socio-techniques, dans le cadre d'une finalité identifiée. Dans de tels systèmes, la dimension financière est généralement présente, que ce soit d'un point de vue gain, ou plutôt de consommation de ressources financières. Nous considérons l'entreprise comme un système, au sens systémique du terme. L'entreprise est un système qui évolue dans son environnement. Elle poursuit des buts (profit, puissance, pérennité ...), s'organise pour les atteindre (définition de plans d'action, de budgets ...), se dote de structures d'exécution, de direction et de contrôle.

L'entreprise est aussi un ensemble de sous-systèmes en interaction. La modélisation d'entreprise est un procédé incontournable d'étude des organisations dans l'objectif d'améliorer ses performances. La modélisation de l'entreprise permet de représenter le *système entreprise,* selon une abstraction multi points de vue. C'est une pratique qui assure à l'entreprise de s'auto informer et de conduire intelligemment l'alignement de ses objectifs en cohérence avec son environnement et en écoute de sa clientèle. Les efforts de recherche des années 1990 ont conduit à un cadre normalisé pour répondre aux besoins d'une approche systémique de l'entreprise.

De nombreux langages et méthodes, ont été développés, tels que CIMOSA (CIMOSA 1996), GERAM (GERAM 1988), IDEF suite (NIST 1993), GRAI (Schekkerman 2003), BPDM (OMG 2005), les standards : ISO 14258, ISO/15704, ISO/TR/10314, ENV/12204, ENV/40003. Actuellement, dans un souci d'unification, de nombreux travaux concourent à la définition d'un langage unifié de modélisation d'entreprise (Unified Entreprise Modeling Language) (Vernadat 2001) , (Gudas et al. 2005).

Ces approches de modélisation abordent les aspects à la fois organisationnel, informationnel et humain. A titre d'illustration, la norme ISO/14258, *Concepts and rules for enterprise models*, (ISO/14528 2003) propose une approche systémique de l'entreprise ; la norme ISO /15704 (ISO/15704 1998), *Requirements for enterprise-reference architectures and methodologies*, (ISO/15704 2005) aborde les exigences attendues d'une architecture de modélisation d'entreprise; la norme ENV/12204 propose une première spécification des éléments nécessaires à la modélisation des construits «Enterprise Activity », « Business Process », «Event», «Resource», «Enterprise Object», «Object View», «Object State».

3.1 Le méta-modèle ISO 19440

Le standard ISO 19440:2007 spécifie les caractéristiques du noyau des construits nécessaires à la modélisation informatique des entreprises conformément à l'ISO 19439. La norme ISO 19440 identifie sept phases dans le cycle de vie des modèles : la définition du domaine étudié, la définition des concepts nécessaires, la définition des besoins de l'entreprise, la conception du modèle, la mise en œuvre du modèle, l'usage du modèle dans les opérations, le retrait ou l'arrêt des opérations. Elle propose quatre vues sur ces modèles (figure 4) : la vue organisationnelle, la vue informationnelle, la vue fonctionnelle et la vue des ressources. La vue informationnelle porte sur la représentation des données du Système d'Information. La vue organisationnelle porte sur la stratégie d'entreprise. La vue fonctionnelle porte sur les processus. La vue des ressources porte sur les ressources utilisées par les processus métier de l'entreprise.

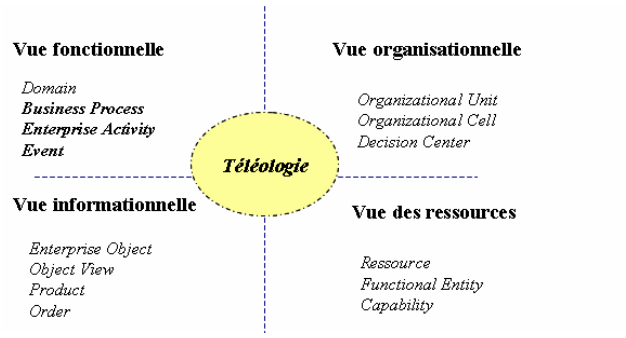


FIG 4 – *Vues ISO/SDI 19440 avec ancrage téléologique*

Le standard ISO/DSI 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 le cadre de l'ISO/DIS 19440. Ce modèle est donné dans la figure 6, il intègre les quatre points de vue présentés dans la figure 4. Un domaine représente la frontière et le contenu d'une entreprise ou d'une partie d'une entreprise. Un processus métier représente une certaine partie du comportement d'entreprise. Un processus métier est une agrégation du processus métier et/ou activité d'entreprise ainsi que l'information décrite par les règles de gestion. L'activité d'entreprise est la réalisation d'une transformation des entrées aux sorties par des ressources spécifiques. L'activité d'entreprise et le processus métier s'appellent collectivement Enterprise Function. Des règles de gestion sont employées pour définir le comportement d'un processus métier. Elles définissent les contraintes sur l'ordonnancement, et des dépendances entre les processus métiers et/ou activités d'entreprise. Un événement lance l'exécution d'un processus métier ou d'une 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 la sémantique de chacun des construits.

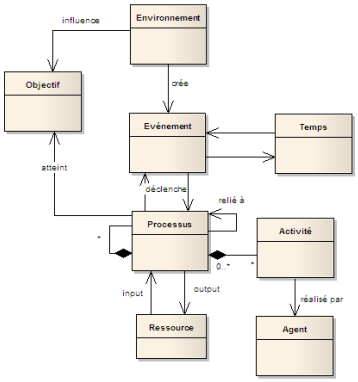


FIG 5 – *Modèle systémique du processus.*

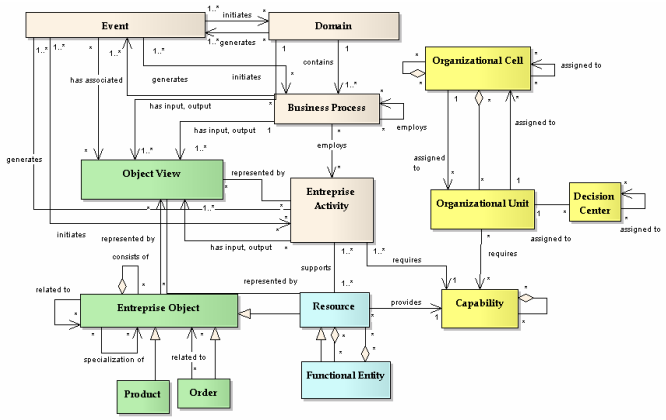


FIG 6 – *Méta-modèle ISO/DIS 19440.*

3.2 Le référentiel COBIT

Le référentiel COBIT (Control objectives for information and technology) (ISACA 2008), a été conçu en 1996 par l'ISACA (Information Systems Audit and Control Association). Ce référentiel constitue un cadre de référence ainsi qu'un ensemble d'outils pour assurer la maîtrise et le suivi de la gouvernance du Système d'Information. COBIT est fondé sur un ensemble de bonnes pratiques, qui se propose d'établir un cadre de pilotage orienté processus du SI afin de contribuer efficacement à l'alignement des technologies sur la stratégie d'entreprise. Le cadre de référence de COBIT répond aux besoins de l'entreprise par quatre caractéristiques principales : il est centré sur les métiers de l'entreprise, organisé par les processus, 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'assurances de différents acteurs, comme le montre la figure 7.

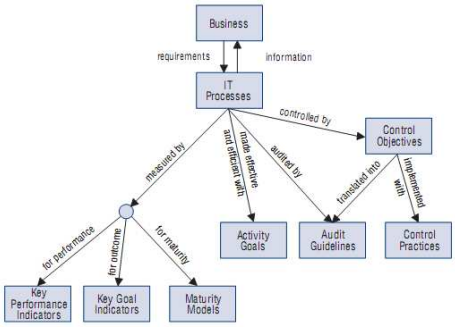


FIG 7 – *Interconnexion des composants COBIT.*

Pour que l'informatique réponde correctement aux attentes de l'entreprise, les dirigeants doivent mettre en place un système de contrôle ou un cadre de contrôle interne. Pour répondre à ce besoin, le cadre de référence de contrôle de CobiT :

- établit un lien avec les exigences métiers de l'entreprise,
- structure les activités informatiques selon un modèle de processus largement reconnu,
- identifie les principales ressources informatiques à mobiliser,
- définit les objectifs de contrôle à prendre en compte.

L'orientation métiers de CobiT consiste à lier les objectifs métiers aux objectifs informatiques, à fournir les métriques (ce qui doit être mesuré et comment) et les modèles de maturité pour faire apparaître leur degré de réussite et à identifier les responsabilités communes aux Propriétaires de processus métiers et aux propriétaires de processus informatiques. En résumé, pour fournir les informations dont l'entreprise a besoin pour réaliser ses objectifs, les ressources informatiques doivent être gérées par un ensemble de processus regroupés selon une certaine logique.

Dans la section 4, nous empruntons à COBIT, les éléments de contrôle et de mesures des processus IT. Ces éléments sont utilisés pour l'extension de certains points du méta-modèle ISO 19440, fort utiles à l'alignement stratégique du système d'information.

4 Méta-modélisation étendue

Dans cette section nous proposons de construire une extension du méta-modèle ISO 19440, de sorte que nous puissions traduire explicitement la problématique de l'alignement des divers aspects du système d'information. Nous développons d'abord l'analyse de la structure du méta-modèle de base.

Les frontières fondamentales de l'alignement se situent aux interactions et couplages entre les différents points de vue du méta-modèle. L'interaction entre les entités *entreprise activity* et *resource* manifeste l'alignement <processus, activité | ressource> ; le *couplage business process*, *entreprise activity* et *object view* relate l'alignement <processus, activité | information> ; 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 système d'information, dans les formes décrites ci-dessus. Cependant la

formulation de l'alignement stratégique au sens décisionnel n'est pas explicite au niveau de la modélisation des quatre points de vue. Nous proposons de reprendre les bonnes pratiques de COBIT pour le pilotage par les processus IT. Ainsi nous ajoutons le concept abstrait « objective» qui sera spécialisé selon le point de vue (figure 9). Le domaine d'activité de l'entreprise, les processus métiers, les activités, les centres de décisions sont contrôlés et pilotés par des objectifs (figure 10). Nous rappelons que le construit « objective » a été proposé dans l'iso 19440 *functional aspects*, cependant les liens avec les centres de décisions et les métriques de mesure ne sont pas explicites (voir figure 14). Nous allons ajouter aussi une spécialisation de *Functional Entity* pour modéliser les processus informationnel technologique (IT processes). Ces processus de technologie de l'information utilisent des ressources selon une connotation « technologie de l'information » (IT resource), cette entité *IT resource* est modélisée par une spécialisation de l'entité *resource* (figure 8).

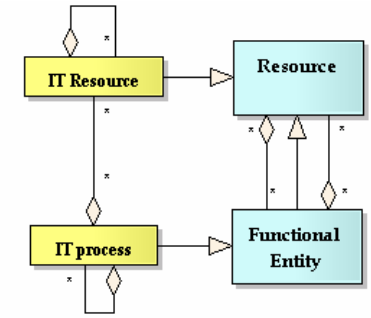


FIG 8 – *Intégration de IT Resource et de IT process.*

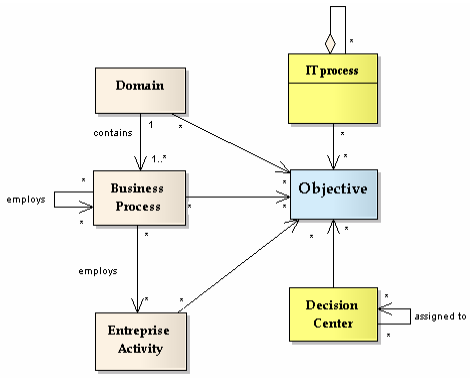


FIG 9 – *Intégration de l'entité objective.*

Nous ajoutons aussi les construits *indicators* et *metrics* pour la mesure de la performance (figure 10). Dans la figure 11, nous explicitons les construit *Analyse structural* et concepts dérivés pour l'évaluation de l'alignement avec des outils systémiques. La figure 13 situe le modèle ISO/19440 augmenté des construits proposés.

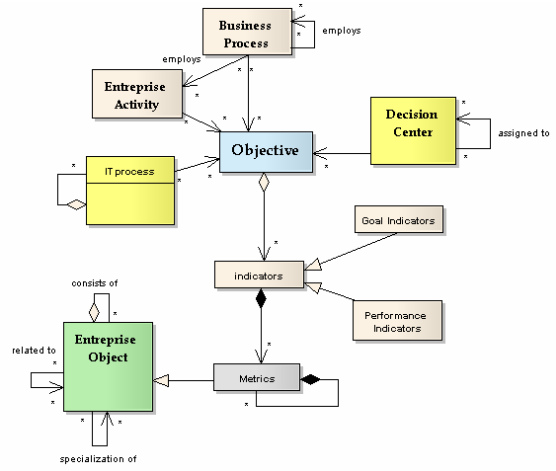


FIG 10 – *Objectif et indicateurs de mesures.*

5 Paradigme structural et outils systémiques

Les préceptes systémiques définissent un système comme une unité globale organisée d'éléments en interaction, fonctionnant et évoluant en fonction d'une finalité, plongée dans un environnement qui agit sur elle et sur lequel elle agit (Morin 1986), (Le Moigne 1983). La réécriture symbolique de la définition d'un système prend la forme suivante :

$\{S\} = \{E, R^i, O, R^e\}$, où E : l'ensemble des éléments constituants, R^i : ensemble des relations internes, O : ensemble des objectifs et R^e : ensemble des relations extérieures. Cette réécriture symbolique renvoie au concept de structure. La généalogie de la systémique recèle un apport important du paradigme structural (structuralisme), qui sous sa projection mathématique a donné naissance à plusieurs structures fédératrices : les structures algébriques (groupe, monoïde, dioïdes), les structures d'ordre (treillis), et les structures topologiques basées sur la notion de voisinage. Les outils systémiques à base du structuralisme puisent leurs forces de représentation dans ces trois types de structures, ou par combinaison de ces structures de références (c'est le cas de la topologie algébrique).

Dans les bonnes pratiques de la systémique, la structure fonctionnelle est décrite par les processus ; une question fondamentale émerge « comment s'agencent les processus ? ». Les matrices structurales ont été utilisées pour donner une réponse à cette question. L'analyse de ces matrices concernent les réseaux de processus et permet d'étudier l'arborescence des processus, les enchaînements linéaires, les rétroactions, etc.

Dans la même vision, pour les diverses problématiques d'alignement du système d'information : {Organisation, Activité, Processus} ×{Ressource} ; {Activité, Processus, Ressource}×{Information}; {Activité, Processus, Organisation}×{Information}. Nous proposons la construction des matrices structurales et par le biais des structures appropriées engager des analyses admises par ces structures.

Les structures que nous suggérons dans ce travail sont distinguées en deux catégories : les structures qui permettent une lecture unique de l'analyse des matrices structurales, à savoir les treillis de Galois (structure d'ordre avec le concept de fermeture) et la méthode *Q-analysis* (structure issue de la topologie algébrique) (Atkin 1974). L'autre catégorie dénom-

mée par « décomposition structurale » permet de hiérarchiser la matrice structurale (structure d'ordre ou de pré-ordre). Cette décomposition exploite des similarités ou des dissimilarités, des indices de couplage, ainsi que des algorithmes de hiérarchisation. Dans cette catégorie plusieurs algorithmes de décomposition sont référencés : analyse de similitude, arbre minimal, classification ascendante hiérarchique, etc. Divers types de couplage pourront être mesurés : couplage processus/processus, via des ressources, liaison activité/ressource, dépendance des processus par des mesures entropiques, liaison information/ressource, liaison processus/objectif, (table 1-3) etc.

		Processus			
		P_1	P_2	…	P_L
Processus	P_1				
	P_2				
	…			C_{xy}	
	P_L				

TAB 1 – *Matrice de couplage processus-processus.*

C_{xy} = couplage entre le processus x et y. Le couplage C_{xy} pourra être calculé selon une similarité, $C_{X,Y} = \frac{\left\|\varpi(X) \cap \varpi(Y)\right\|}{\left\|\varpi(X) \cup \varpi(Y)\right\|} \in [0,1]$ Où $\varpi(X)$ correspond à l'ensemble des ressources du processus X, $\|K\|$ dénote le cardinal de l'ensemble K.

		Objectifs stratégiques			
		O_1	O_2	…	O_N
Processus	P_1				
	P_2				
	…			C_{xy}	
	P_L				

TAB 2 – *Matrice de couplage processus-objectif.*

C_{xy} = mesure de la contribution du processus x à l'objectif y.

		Activité			
		1	A_2	…	A_N
Ressource	R_1				
	R_2				
	…			C_{xy}	
	R_L				

TAB 3 – *Matrice de couplage ressource activité : C_{xy} = l'activité X utilise la ressource Y ?*

D'autres structures d'ordre pourront être utilisées pour aborder d'autres problématiques structuralistes (exemple : le problème de *prioritisation* des processus).

Dans la suite de cette section, nous nous limitons à l'analyse structurale par les treillis de Galois. Les autres outils seront envisagés dans d'autres travaux futurs.

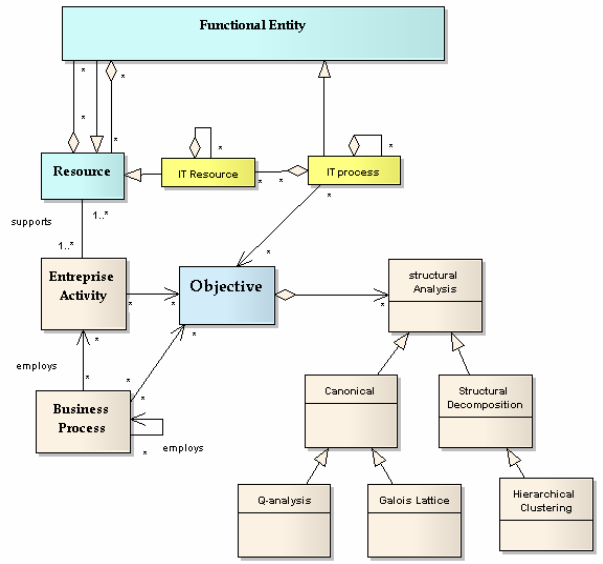


FIG 11 – *Intégration de l'analyse structurale.*

5.1 Analyse par les treillis de Galois

Dans ce paragraphe nous abordons les notions de base de treillis de Galois (Wile 1982) (Marghoubi et al. 2006), (Boulmakoul et al. 2007). Un contexte C est un triplet (O, A, R) où O, A sont des ensembles et R est une correspondance.

La table 4 montre un exemple de contexte représenté par C= (O; A; R) avec O= $\{r_1, r_2, r_3, r_4, r_5, r_6,\}$ et A = $\{p_1, p_2, p_3, p_4, p_5\}$. Ce contexte exprime le fait qu'une ressource X est utilisée ou non par le processus Y.

	p_1	p_2	p_3	p_4	p_5
r_1	1	1	0	1	1
r_2	0	1	1	0	1
r_3	1	1	0	1	1
r_4	1	1	1	0	1
r_5	1	1	1	1	1
r_6	0	1	1	1	0

TAB 4 – *La matrice binaire décrivant la correspondance R du contexte C = (O, A, R).*

5.1.1 Treillis de Galois

L'ensemble L de tous les concepts, muni de la relation d'ordre $\leq$, possède la structure mathématique de treillis et est appelé treillis de Galois L(C) du contexte C. Un treillis de Galois est un concept formel dérivé à partir d'une relation R. C'est une structure de graphe particulière. Un treillis étant un graphe orienté, sans cycle et comportant un noeud minimal et un noeud maximal. Le treillis de Galois correspond à un ordre partiel induit par une relation binaire R entre deux ensemble discrets, un ensemble d'objets O et un ensemble d'attributs A.

La correspondance de Galois

Deux fonctions Φ et Ψ permettent d'exprimer les *correspondances* entre les sous-ensembles d'objets *P(O)* et les sous-ensembles d'attributs *P(A)* induits par la relation R. La fonction Φ associe l'ensemble des attributs communs à un ensemble d'objets, tandis queΨ , la fonction duale de Φ, associe l'ensemble des objets communs à un ensemble d'attributs :

$$\Phi : P(O) \to P(A), \Phi(X) = X' = \left\{ a \in A / \forall o \in X, o \Re a \right\}$$

$$\Psi : P(A) \to P(O), \Psi(Y) = Y' = \left\{ o \in O / \forall a \in Y, a \Re o \right\}$$

Le couple (Φ, Ψ), ainsi présenté, définit la *correspondance de Galois* entre les sous-ensembles d'objets *P(O)* et d'attributs *P(A)* du contexte.

Fermeture de Galois

Une fermeture sur un ensemble ordonné (E, $\leq$) est une application $\Re : E \to E$ qui pour tout $x, y \in E$ vérifie les propriétés suivantes :

- $x \leq \Re(x)$ ($\Re$ est extensive)
- si $x \leq y$ alors $\Re(x) \leq \Re(y)$ ($\Re$ est monotone croissante)
-si $\Re(x) = \Re(\Re(x))$ ($\Re$ est idempotente)

Un élément x de E est fermé pour $\Re$ si est seulement si x= $\Re$ (x)

Les compositions h= $\Phi o \Psi$ et h'= $\Psi o \Phi$, constituent des opérateurs de fermeture de la connexion de Galois. L'opérateur $\Phi o \Psi$ génère des sous-ensembles fermés d'objets tandis que l'opérateur $\Psi o \Phi$ génère des sous-ensembles fermés d'attributs.

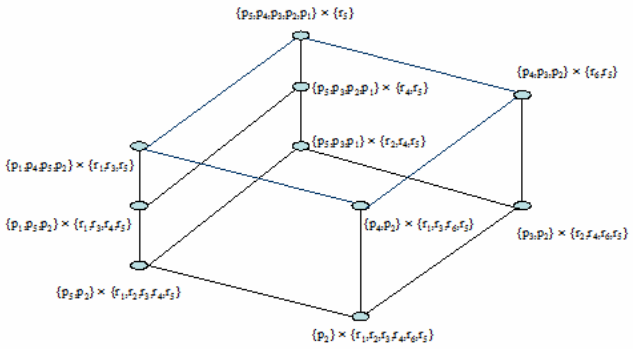


FIG 12 – *Treillis de Galois du contexte donné par la table 4.*

Dans une di-clique maximale (A, B) (fermé du treillis de Galois), B est l'ensemble de toutes les ressources utilisées au même temps par tous les processus de A, aucun autre processus ne les utilisant tous.

La structure de Galois peut être exploitée pour le découpage en domaines informationnels de l'entreprise : « on peut définir un domaine comme une activité ou un ensemble d'activités s'appuyant sur un ensemble d'informations communes et n'ayant que peu d'échanges avec les autres activités » (table 3).

Avec cet exemple simple (figure 12), la représentation laticielle des processus en relation avec les ressources est canonique. Une lecture unique est permise avec ce diagramme. Les échelles de *Gutman*, pourront être exploitées pour exhiber des classifications dites Galoisiennes. Avec cette même structure et avec des méthodes de générations des règles d'associations issues du data mining, il est possible de générer des associations entre processus.

Ces associations expriment des dépendances cachées existant entre processus : c'est des méta-processus. Nous obtenons à titre d'exemple deux associations fortes de confidence 100% $p_5 \rightarrow p_2$ et $(p_4,p_5) \rightarrow p_1$,

L'intention de toutes les instances des structures de références considère tous les points de vues de l'alignement du système d'information (table 5).

Structure de référence	*Type de représentation*	*Instance*
Structure d'ordre de type treillis	Représentation canonique	Treillis de Galois
Topologie algébrique	Lecture topologique unique	Q-Analysis
Structure d'ordre de type anneau et semi-anneau	Arbre et Graphe (pas d'unicité de lecture)	Algèbre de chemins
	Arbre ou graphe de décomposition	Analyse de similitude, Décomposition des systèmes

TAB 5 – *Structures de référence et instances.*

6 Conclusion

Dans cet article, nous avons apporté de nouveaux construits pour l'extension du métamodèle ISO 19440, d'une part pour appréhender les diverses facettes de l'alignement stratégique du système d'information, et d'évaluer cet alignement avec des outils systémiques, issus du paradigme structural, d'autre part. Nous avons montré comment les structures d'ordre de type treillis de Galois pourront être exploitées dans ce contexte. D'autres structures ont été discutées dans ce travail et qui font objet de nos recherches actuelles. Certes, le présent travail ne souligne que des aspects fondamentaux, nous avons mis en place une stratégie de déploiement des éléments énoncés dans ce travail, sur des sites réels (organisme public et un établissement privé). Nous sommes convaincus que l'implémentation de ce type de pratique serait d'un grand intérêt à l'ingénierie des systèmes d'information.

Références

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

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

Brown, I. T. J. (2004), "Testing and extending theory in Strategic information systems planning through literature analysis", *Information Resources Management Journal*, 17 (4): 20-48.

CIGREF (2002), "Alignement stratégique du système d'information", *Rapport en ligne www.cigref.fr.*

CIMOSA Association (1996), "CIMOSA - Open System Architecture for CIM", *Technical Baseline; Version 3.2, private publication*, March 1996

Earl, M. J. (1993), "Experiences in strategic information systems planning", *MIS Quarterly*, 17 (1): 1-24.

ENV 12204 (1995), "Advanced Manufacturing Technology - Systems Architecture - Constructs for Enterprise Modeling", *CEN TC 310/WG1*, 1995

ENV 40003 (1990), "Computer Integrated Manufacturing - Systems Architecture - Framework for Enterprise Modeling", *CEN/CENELEC*, 1990.

GERAM (1998), "Generalised Enterprise Reference Architecture and Methodologies", *Annex A to ISO 15704, IC184/SC5/WG1 N423*, 1998

Gudas S., Lopata A., Skersys T. (2005), "Approach to Enterprise Modelling for Information Systems Engineering", *Informatica*, 2005, Vol. 16, No. 2, 175–192

Henderson, J., and Venkatramen , N. (1992), "Strategic Alignment: A Model for Organisational Transformation: In Transforming Organizations". In: *T. Kochan and M. Unseem, Editors. Oxford, University Press*, USA, Janvier 11. 1992.

ISACA (2008), "Cobit 4.1", *http://www.isaca.org/cobit*

ISO 15704 (1998) "Requirements for Enterprise Reference Architectures and Methodologies" *ISO TC184/SC5/WG1N423*, 1998;

ISO 19440 (2007), "Enterprise integration -- Constructs for enterprise modelling", *Edition 1,* www.iso.org 2007.

ISO TR 10314-1 (1991), "Industrial Automation – Shop Floor Production Model", 1991.

Le Moigne J.L. (1983), « La théorie du système général », *édition PUF*.

Lederer, A. L. and Sethi, V. (1988),"The implementation of strategic information systems planning methodologies", *MIS Quarterly*, 12 (3): 445-461.

Lederer, A. L. and Sethi, V. (1992), "Root Causes of Strategic Information Systems Planning Implementation Problems", *Journal of Management Information Systems*, 9(1): 25-45.

Lorino P. (1997), “Méthodes et pratiques de la performance”, *Les édition d’organisation*, ISBN : 2-7081-1977-X, 1997.

Marghoubi R., Boulmakoul A., Zeitouni K. (2006) Utilisation des treillis de Galois pour l'extraction et la visualisation des règles d'association spatiales, Conférence INFORSID 2006, vol 2 pp. 703-718. ISBN: 2-906855-22-7, Hammamet, Tunisie.

Morin E. (1986), “La méthode: Tome 3 la connaissance de la connaissance”, édition Le Seuil.

NIST, F. (1993), “Publication 183: Integration Definition of Function Modeling (IDEF0)”. National Institute of Standards and Technology 128

OMG (2005), "BPDM: Business Process Definition Metamodel". http://www.omg.org (2005)

Schekkerman, J. (2003), “How to Survive in the Jungle of Enterprise Architecture Frameworks”. Trafford, Canada.

Scott Morton, Michael S., (with Thomas J. Allen) (1994), “Information Technology and the Corporation of the 1990s: Research Studies”; Oxford University Press.

Shimizu T., Monteiro de Carvalho M., Jose Barbin Laurindo F. (2006), "Strategic Alignment Process and Decision Support Systems - Theory and Case Studies", *IRM Press*, ISBN : 1591409764.

Vernadat, F. (2001). “UEML: towards a unif ed enterprise modelling language”. In Proceedings of International Conference on Industrial Systems Design, Analysis and Management (MOSIM’01, 2001), Troyes, France.

Ward, J. and Peppard, J. (2002), “Strategic planning for information systems”, *John Wiley & Sons Ltd*.

Wille R. (1982), “Restructuring lattice theory: an approach based on hierarchies of concepts”, *Ordered Sets* (I. Rival, ed.), pp. 445–470, *Reidel, Dordrecht-boston*, 1982.

Summary

Information systems are strongly sensitive to strategic evolutions of enterprise: organisational change, and change of objectives, modified variety, new objects and business process. In the objective to control strategic alignment of information systems, we propose an approach based on extended enterprise metamodel ISO/DSI 19440. This extension was borrowed from COBIT framework for IT processes. In order to better lead evolutions of information system, this extension integrates necessary structures for developing systemic tools, based on structural paradigm. Objective and decisional metrics aspects are also considered in this meta-modelling.

INDEX DES AUTEURS