Avant-propos:
Troisième atelier "Fouille de Données Complexes dans un processus d'extraction des connaissances"

1 Présentation

L'atelier sur la fouille de données complexes dans un processus d'extraction de connaissances est organisé à l'instigation du groupe de travail "Fouilles de Données Complexes" GT FDC et s'inscrit dans le cadre de la conférence EGC. Cet atelier se veut être un lieu de rencontre annuel où chercheurs/industriels peuvent partager leurs expériences et expertises dans le domaine de la fouille de données. L'atelier se veut ouvert en terme de propositions. On pourra y présenter aussi bien un travail abouti, des réflexions sur la fouille de données complexes ou un travail préliminaire (qui présentera davantage un problème qu'une solution). Enfin, les discussions sur les liens entre différentes disciplines sont également bienvenues.

Les deux premières éditions de cet atelier au sein des deux précédentes conférences EGC (2004 à Clermont-Ferrand et 2005 à Paris) furent une réelle réussite, accueillant des chercheurs/doctorants représentant plus de 30 laboratoires francophones différents. Ils auront permis d'avancer sur la compréhension de la complexité d'un processus d'extraction de connaissances à partir de bases de données et d'initier de nouveaux échanges scientifiques entre chercheurs.

La troisième édition de cet atelier a lieu dans le cadre de EGC06 (Lille) et est constitué de seize exposés scientifiques (13 articles longs, 2 courts et 1 poster). Les articles retenus ont fait l'objet de rapports de lecture dans le double but d'améliorer leur qualité et de donner des conseils aux auteurs. Une place privilégiée a été accordée aux jeunes chercheurs et à la présentation des travaux en cours dans les différentes équipes. Cela peut être l'occasion pour un doctorant de présenter son projet de recherche. Cette partie est particulièrement importante pour les travaux qui commencent et pour la mise en place de groupes de recherche sur des thèmes partagés.

Dans cette troisième édition, on note de nombreux travaux sur trois thèmes principalement, deux relatifs à la fouille de données (classification, approches hybrides) et un plus orienté sur le prétraitement de données complexes

- Classification non supervisée d'objets complexes : citons la pondération des attributs des objets, la prise en compte d'objets incomplètement définis, la classification de flots de séquences, la classification d'images, la fusion de classifieurs, la classification de données Web en commerce électronique.
- Approches hybrides en fouille de données complexes : classification en vue d'une extraction de motifs séquentiels dans les data streams, classification en pré-traitement en vue

d'extraire des motifs séquentiels de faible support, approche multi agent pour améliorer la classification et enfin approche RàPC pour la classification.
– Pré-traitement d'objets complexes : il est à noter cette année de nombreux articles sur la prise en compte de connaissances hétérogènes, de connaissances voire d'ontologies dans la constitution d'entrepots de données, dans le pré-traitement en vue de faciliter la fouille de données ainsi que l'interpétation des résultats.

Une discussion sur cette troisième édition de l'atelier est prévue en fin de journée.

2 Quelques mots sur la fouille de données complexes

Dans tous les domaines, les données à traiter pour y extraire de la connaissance utilisable sont de plus en plus complexes et volumineuses. Ainsi est-on amené à devoir manipuler des données : Souvent peu ou non structurées ; Issues de plusieurs sources comme des capteurs ou sources physiques d'informations variées ; Représentant la même information à des dates différentes ; Regroupant différents types d'informations (images, textes, video, son,...) ; ou regroupant encore des données de différentes natures (logs, contenu de documents, connaissances/ontologies, etc.). Aussi la fouille de données complexes ne doit plus être considérée comme un processus isolé mais davantage comme une des étapes du processus plus général d'extraction de connaissances à partir de données (ECD). En effet, les difficultés liées à la complexité des données se répercutent sur toutes les étapes du processus d'ECD : avant d'appliquer des techniques de fouille dans les données complexes, ces dernières nécessitent un travail préparatoire (principalement de structuration et d'organisation de ces données complexes). Parallèlement, de nouvelles méthodes de fouille (classification, catégorisation, recherche de motifs fréquents, etc.) doivent également être définies dans ce contexte de la complexité des données. Enfin la notion d'utilité des paradigmes extraits (anticiper la pertinence des résultats de la fouille) est également un problème à étudier.

Les thèmes liés à la fouille de données complexes peuvent donc comprendre :
– Pré-traitement, structuration et organisation de données complexes ;
  – Données inférées, Modélisation guidée par les résultats,
  – Enrichissement des données, Sélection, nettoyage des données, codage, transformation des données, ETL (Extracting, Transforming and Loading) ;
  – Cubes de données pour la fouille des données ;
  – OLAP et fouille de données ;
  – Intégration des données complexes ;
  – Modélisation des données complexes et XML ;
  – Métadonnées ;
  – Espaces de représentation des données complexes ;
  – Fusion de données ;
– Processus et méthodes de fouille de données complexes ;
  – Evaluation des méthodes actuelles ;
  – Proposition d'approches nouvelles (par exemple hybrides ou multi-stratégies) ;
  – Sélection de sources des données et d'attributs ;
  – Utilisation de relations spatiales ou temporelles entre les données ;
  – Utilisation de connaissances du domaine pour optimiser l'extraction ;

  - Post-traitement ;
  - Visualisation et aide à l'interprétation des résultats ;
  - Validation des motifs extraits ;
  - Mise à jour des connaissances ;
- Rôle des Connaissances, Ontologies, Méta données en ECD complexe ;
  - Utilisation de connaissances du domaine analysé ;
  - Utilisation de connaissances du domaine de l'analyste ;
  - Rôle des Métadonnées dans un processus global ECD ;
  - Aide à la réutilisation d'un processus ECD, Web sémantique en ECD ;
- Retours d'expériences (Web, sciences du vivant) etc.

3 Responsables

Boussaid Omar
(Laboratoire ERIC, Lyon)
email : omar.boussaid@univ-Lyon2.fr
Tel : 04 78 77 23 77

Trousse Brigitte
(Equipe-Projet Axis, Inria Sophia-Antipolis)
email : trousse@sophia.inria.fr
Tel : 04 92 38 77 45

4 Comité de lecture

Le comité de lecture est composé de membres du GT "Fouilles de Données Complexes" et d'experts du domaine.

Aufaure Marie-Aude (SUPELEC)
Bentayeb F. (ERIC)
Bouet Marinette (LIMOS)
Boussaid Omar (ERIC)
Briand Henri (IRIN)
Collard Martine (I3S)
Darmont Jérome (ERIC)
Djeraba Chabane (LIFL)
Elfaouzi Nour-eddin (Inrets)
Fertil Bernard (IMED)
Gancarski Pierre (LSIIT)
Gallinari Patrick (LIP6)
Hacid Mohand-Said (LIRIS)
Jomier Genevie‘ve (Lasmade)
Lechevallier Yves (INRIA)
Martin Arnaud (ENSIETA)
Masséglia Florent (INRIA)
Morin Annie (IRISA)
Napoli Amedeo (LORIA)
Nugier Sylvaine (EDF)
Petit Jean-Marc (LIMOS)
Poncelet Pascal (LGI2P)
Saidi-Glandus Alexandre (LIRIS
Sheeren David (LSIIT)
Teisseire Maguelonne (LIRMM)
Trousse Brigitte (INRIA)
Vrain Christel (LIFO)
Wemmert Cedric (LSIIT)
Zeitouni Karine (PRISM)
Zighed Djamel (ERIC)

5 Remerciements

Les responsables de l'atelier tiennent à remercier chaleureusement :
– les auteurs pour la qualité de leurs contributions,
– les membres du comité de lecture pour leur travail indispensable à la qualité de cet atelier
– Sophie Honnorat pour son aide précieuse dans la constitution des actes de l'atelier ainsi que Sergiu Chelcea pour son soutien au niveau informatique,
– Philippe Preux et Fatima Belkouche, responsables des ateliers pour EGC 2006 pour leur gentillesse,
– Chabane Djeraba, président du comité d'organisation d'EGC 2006 de nous accueillir à Lille.

PROGRAMME

Le 17 janvier 2006

8h30-9h Accueil

9h-9h10 Présentation de l'atelier

9h10-10h30 Fouille de données complexes : classification

Trois stratégies d'évolution pour la pondération automatique d'attributs en classification non supervisée d'objets complexes
Pierre Gançarski, Alexandre Blansché (LSIIT-AFD, Univ. Strasbourg – Illkirch)

Evaluation d'une approche probabiliste pour le classement d'objets incomplètement connus dans un arbre de décision
Lamis Hawarah, Ana Simonet, Michel Simonet (TIMC-IN3S – La Tronche)

LSA : les limites d'une approche statistique
Mathieu Roche, Jacques Chauché (LIRMM – Montpellier)

Fouille d'images IRMf : alogirthme CURE
Jerzy Korczak, Aurélie Bertaux (LSIIT, Univ. Strasbourg – Illkirch)

10h30-11h Pause

11h-11h20 Fouille de données complexes : classification (suite)

Estimation et fusion des temps de parcours routiers par la théorie de l'évidence
Eric Lefevre (LGI2A, Univ. Artois -Béthune),
Nour-Eddin El Faouzi (LICIT INRETS-ENTPE – Bron)

11h20-12h30 Fouille de données complexes : approches hybrides

Classification de flots de séquences basée sur une approche centroïde
Alice Marascu, Florent Masseglia (INRIA – Sophia Antipolis)

Techniques de généralisation des URLs pour l'analyse des usages du Web
Yves Lechevallier, Florent Masseglia, Doru Tanasa,
Brigitte Trousse (INRIA Rocquencourt and Sophia Antipolis)

Vers un algorithme RASMA : RAS basé multi-agent
Radhia Ben Hamed (SOIE ISG – Tunis), Hajer Baazaoui (Riadi GDL, Tunisie),
Sami Faiz (INSAT – Tunis)

Optimisation de la technique de RBC pour la classification dans un processus de data mining (10')
Mounir Ben Ayed, Issam Feki, Adel Alimi (NSE – Tunis)

12h30-14h Déjeuner

14h-15h30 Prétraitement : ontologies, entrepôt de données

Vers une approche de construction de composants ontologiques pour le Web sémantique – synthèse et discussion
Nesrine Ben Mustapha (RIADI ENSI - Tunisie), Marie-Aude Aufaure (Supélec, Gif sur Yvette), Hajer Baazaoui Zghal (RIADI ENSI - Tunisie

Ontologie et base de connaissances pour le pré-traitement et post-traitement en fouilles de données
Laurent Brisson, Martine Collard , Nicolas Pasquier (I3S UNSA – Sophia Antipolis

Un système de médiation base sur les ontologies
Nora Maiz, Omar Boussaid, Fadila Bentayed (ERIC, Univ. Lyon 2 – Bron)

Modèle d'entrepôt de données à base de règles
Cécile Favre, Fadila Bentayeb, Omar Boussaid (ERIC – Bron)

15h30-16h Pause

16h-16h30 Prétraitement : ontologies, entrepôt de données

Prétraitement et classification des données complexes dans le domaine du e-commerce
Sergiu Chelcea, Alzennyr da Silva, Yves Lechevallier, Doru Tanasa,
Brigitte Trousse (INRIA Rocquencourt and Sophia Antipolis)

Saada : un générateur de bases de données astronomiques (10')
Ngoc Hoan Nguyen, Laurent Michel, Christian Motch (Observatoire Astronomique, Strasbourg)

16h30-16h35 Poster

Fouille de données en appui des analyses métabonomiques
F. Jourdan, F. Vinson, C. Canlet, N. Priymenko, G. Gottardi, J. Molina
et A. Paris (INRA-ENVT, Toulouse)

16h35-18h00 Discussion

- Synthèse de l'atelier
- Groupe FDC
    - Présentation des thèmes du groupe
    - Discussion sur els actions futures du groupe FDC

TABLE DES MATIERES

Vers une approche de construction de composants ontologiques pour le Web sémantique – synthèse et discussion

Nesrine Ben Mustapha[1], Marie-Aude Aufaure[2], Hajer Baazaoui Zghal[1]
[1]Laboratoire RIADI – ENSI Campus Universitaire de la Manouba
2010
[nesrine.benmustapha, hajer.baazaouizghal]@riadi.rnu.tn
[2]Supélec – Plateau du Moulon – 3, rue Joliot Curie
91 192 Gif sur Yvette Cedex
Marie-Aude.Aufaure@supelec.fr

Résumé. Cet article propose une approche de construction de composants ontologiques pour le Web sémantique, basée sur l'étude des travaux proposés dans la littérature. De nombreuses méthodes et méthodologies ont été proposées dans la littérature cette dernière décennie. Celles-ci sont, en règle générale, dédiées à un type de données, comme par exemple du texte, des données semi-structurées, des données relationnelles, etc. Notre travail concerne les pages Web, nous nous sommes donc particulièrement intéressées aux méthodes s'appliquant aux textes, et plus particulièrement, aux méthodes d'apprentissage d'ontologies à partir de textes. Après une présentation détaillée de l'état de l'art en la matière, cet article présente une ébauche de construction de composants ontologiques pour le Web sémantique.

1 Introduction

De par le développement rapide d'Internet, des téléphones cellulaires, des assistants personnels, la quantité d'informations disponible sur le Web double d'une façon exponentielle. Par conséquence, le problème d'intégration des différentes sources hétérogènes des données se pose et la recherche d'information devient de plus en plus complexe. Aujourd'hui le Web est utilisé pour rechercher des informations via les moteurs de recherche grâce à l'indexation des pages, des procédures d'extraction d'information, des procédures d'analyse d'informations. Cependant, on note un manque d'exploitation des connaissances disponibles sur le Web pour une gestion plus intelligente des informations.

D'où l'apparition de la notion du Web sémantique dont le succès dépendra de la rapidité de déploiement des ontologies. Ces dernières constituent une brique de base du Web sémantique ; en effet, elles permettent d'améliorer la recherche dans le Web, de partager des connaissances relatives à un domaine et agréées par une communauté de personnes et d'assurer la gestion et l'intégration des connaissances Web. Toutefois, la construction manuelle d'une ontologie est une tâche longue et fastidieuse. Des approches ont été proposées ces dernières années dans ce domaine pour pallier à la difficulté de cette tâche. La description de ces approches fait l'objet de la première partie de notre article où nous approfondissons essentiellement les méthodologies d'apprentissage d'ontologies. Ensuite, nous présentons une ébauche de notre proposition de construction de composants ontologiques pour le Web sémantique. Finalement, nous concluons et donnons les perspectives de ce travail.

2 Classification des méthodologies de construction d'ontologies

De nombreuses méthodologies de construction d'ontologies ont été définies depuis une décennie. Elles peuvent être classifiées en fonction de l'utilisation ou non de connaissances à priori ainsi que de techniques d'apprentissage. Les premières méthodologies visaient à construire des ontologies sans connaissance à priori dans un but de développement d'ontologies d'entreprises, et étaient essentiellement manuelles. Des travaux concernant la construction coopérative d'ontologies ont été développés à savoir l'approche CO4 (Eueznat, 1995) et (KA)2 (Decker et al., 1999), sachant que peu d'outils à l'heure actuelle intègrent ce travail collaboratif pourtant nécessaire dans le contexte des ontologies. La rétro-conception des ontologies (Gomez-Perez et Rojas, 1999) est un autre axe de recherche basé sur le mapping d'un modèle conceptuel d'une ontologie implémentée avec un autre modèle plus valide pour le ré-impléménter. Les méthodologies d'apprentissage se distinguent selon les types de données en entrée : textes, dictionnaires, bases de connaissances, schémas relationnels et semi-structurés, sources de données hétérogènes. Nous insisterons sur la construction à partir de textes. Les ressources linguistiques brutes sont filtrées soit manuellement (acquisition de connaissances à partir de textes), soit automatiquement (fouille de textes). Les méthodologies de fusion d'ontologies sont également plus ou moins automatisées. L'évolution et l'évaluation d'ontologies sortent du cadre de cet état de l'art.

2.1 Les méthodologies de construction d'ontologies «from scratch »

Les méthodologies de construction d'ontologies « from scratch » furent parmi les premières réflexions dans le domaine de l'ingénierie d'ontologie et visent à concevoir un processus de construction d'ontologies en l'absence de connaissances à priori (d'où la signification du terme en anglais «from scratch »). Les premières méthodologies (Gruninger et Fox, 1995), (Ushold et King, 1995) ont été définies dans un but de développement d'ontologies d'entreprise. Elles ont été inspirées par le développement de systèmes à base de connaissances basés sur la logique du 1er ordre. La méthodologie de **Gruninger et Fox** (1995) consiste à construire un modèle logique des connaissances communes d'entreprise spécifié via l'ontologie en se basant sur les étapes suivantes : La première étape est la capture des « scénarios motivants » La deuxième étape est la détermination des exigences de l'ontologie à construire sous la forme des questions informelles de compétences. La troisième étape représente la formalisation de la terminologie extraite antérieurement à l'aide d'un formalisme bien déterminé comme Knowledge Interchange Format (KIF) ou la logique du premier ordre. Les deux étapes suivantes sont respectivement la spécification formelle des questions de compétence en utilisant la terminologie de l'ontologie et la spécification des axiomes et des définitions relatives aux termes de l'ontologie dans un langage formel comme la logique du premier ordre. La dernière étape concerne la spécification des théorèmes de complétude de l'ontologie qui vont représenter les conditions sous lesquelles les solutions des questions données seront complètes. Bien que cette méthodologie soit à la base de la construction et de la validation d'ontologie, elle est incomplète étant donné l'absence des activités d'intégration, d'acquisition ainsi que des fonctions de gestion comme la planification et le contrôle de qualité. **Methontology** (Fernandez et al., 1997, 1999) est une méthodologie qui a été étendue en 2001 (Arpirez et al., 2001) avec son application à

l'environnement de développement d'ontologie WebODE. Elle autorise la rétro-conception des ontologies. Elle préconise un processus de développement des ontologies dont les principales phases sont la spécification, la conceptualisation, la formalisation, l'implémentation et la maintenance. Les activités relatives à chacune des phases sont l'acquisition des connaissances, l'intégration, l'évaluation, la documentation, la gestion de la configuration et l'assurance de qualité. Dans ce processus, la phase la plus importante est celle de la conceptualisation pour laquelle l'activité d'acquisition des connaissances est une étape fondamentale. Cette phase de conceptualisation convertit une vue informelle du domaine en une spécification semi informelle en utilisant un ensemble de représentations intermédiaires basées sur des notations tabulaires et graphiques.

2.2 Les méthodologies d'apprentissage d'ontologies à partir de textes

La notion d'apprentissage renforce l'idée de la construction des ontologies sur la base de connaissances a priori ; celles-ci permettent d'automatiser l'enrichissement d'ontologie par des méthodes d'apprentissage. Selon **Maedche et Sttab** (2001), il existe autant d'approches d'apprentissage des ontologies que de types d'entrées. Nous distinguons les approches d'apprentissage à partir de textes, de dictionnaires (Hearst, 1992), (Jannink, 1999), de bases de connaissances (Suryanto et Compton, 2001), de schémas semi structurés (Deitel et al., 2001), (Doan et al., 2000), (Papatheodrou et al., 2002) et de schémas relationnels (Johannesson, 1994), (Kashyap, 1999), (Runin et al., 2002). Dans la suite de cette section, nous décrivons des méthodes d'apprentissage à partir de textes, qui sont relatives à nos travaux de construction de composants ontologiques pour le Web sémantique. Plusieurs techniques sont mises en jeu dans l'apprentissage d'ontologies comme les patrons lexico-syntaxiques, l'extraction basée sur les règles d'association, l'extraction basée sur le clustering, l'extraction basée sur le calcul des fréquences et l'extraction basée sur des techniques hybrides.

2.2.1 Apprentissage fondé sur des techniques de traitement automatique de langage naturel

Hearst (1998) a proposé une approche qui permet l'apprentissage automatique des relations d'hyponymie en extrayant l'ensemble des paires de concepts liés par une relation dans une ontologie existante pour construire des patrons lexico-syntaxiques. Ces patrons vont servir pour la découverte d'autres relations, basées sur les patrons appris, entre les concepts de l'ontologie existante. Cette méthode a été utilisée pour étendre des ontologies lexicales, les enrichir avec de nouveaux concepts et ajouter de nouvelles relations. Hearst, Alfonseca et Manandhar (2002a) ont travaillé sur l'ontologie WordNet et Kietz et al. (2000) ont travaillé sur GermaNet. Cette approche est aussi utilisée dans les outils Prométhée et Caméléon pour identifier des nouveaux modèles lexico-syntaxiques dans un corpus. Dans ces deux outils, l'approche de Hearst est appliquée à un domaine spécifique pour apprendre des modèles spécifiques en analysant les contextes dans lesquels des couples de termes liés apparaissent. Alfonseca et Manandhar (2002b) proposent une voie où cette méthode peut être combinée avec la méthode des signatures contextuelles afin d'améliorer la classification de nouveaux concepts à l'intérieur d'une ontologie existante. La méthodologie de **Aussenac-Gilles** et al. (2000a, 2000b) est basée sur la découverte des connaissances à partir de documents techniques. Elle permet la création d'un modèle du domaine par l'analyse d'un

corpus en utilisant des outils de traitement automatique du langage naturel (TALN) et des techniques linguistiques. Cette méthode peut avoir recours à des ontologies existantes ou à des ressources terminologiques pour la construction des ontologies. Elle comprend quatre activités principales à savoir la constitution du corpus, l'étude linguistique, la normalisation et la formalisation. **Nobécourt** (2000) a traité la construction d'une ontologie du domaine à partir de textes en se servant des techniques de traitement du langage naturel et d'un corpus de textes pré analysés. Cette approche adopte une méthodologie proposant deux phases : la modélisation et la représentation. La première phase comprend une première activité linguistique dont l'objectif est d'extraire les principaux termes relatifs au domaine ("les primitives conceptuelles") à partir du corpus. La deuxième activité est conceptuelle et consiste à sélectionner les termes les plus pertinents du domaine en vue d'induire des sous-domaines de l'ontologie. Ces termes sont modélisés en concepts ou propriétés constituant ainsi le premier squelette de l'ontologie. Ces concepts sont associés à leurs descriptions écrites en langage naturel. Ces descriptions constituent aussi de nouveaux documents qui peuvent être utilisés comme entrée de cette méthode pour découvrir une liste de nouvelles primitives exprimant les relations entre les concepts. Le processus proposé raffine itérativement le squelette. Quand à la deuxième phase, elle consiste en la formalisation des schémas de modélisation avec un langage d'implémentation. Cette méthode est assistée par l'outil TERMINAE (Biebow et Szulman, 1999). **Moldovan et Girju** (2000) proposent le système KAT (*« Knowledge Acquisition from Text* »). Il consiste à découvrir des concepts et des relations spécifiques à un domaine via l'enrichissement d'une ontologie existante avec de nouvelles connaissances acquises à partir des textes analysés. Elle comprend quatre phases principales à savoir : une phase d'apprentissage de nouveaux concepts, une phase de classification des concepts, une autre d'apprentissage des relations et une phase d'intégration et d'évaluation de l'ontologie construite. L'apprentissage de nouveaux concepts consiste à sélectionner les concepts granulaires qui sont jugés pertinents au domaine d'étude, extraire des phrases contenant des concepts granulaires à partir des documents Web avec un nombre maximum fixé arbitrairement de 500 phrases par concept et extraire des nouveaux concepts reliant les concepts granulaires par un verbe dans les phrases extraites des textes. La classification des concepts consiste à analyser les adjectifs apparus dans les expressions des concepts candidats et classifier les concepts de la forme [mot, concept granulaire] où le mot peut être un nom ou adjectif. Cette classification consiste à dire que le concept [mot, concept granulaire] subsume le concept granulaire, autrement dit à ajouter la relation d'hyperonymie. **Bachimont** (2000) a adopté la notion « d'ontologie différentielle » (Maliazé et al., 2004) qui « peut être interprétée comme une structure hiérarchique des termes, des signifiés normés assortis de leurs définitions encyclopédiques et systémiques ». Pour chaque terme, on précise un ensemble d'éléments sémantiques (similarité sémantique et différence avec le « père ontologique », similarité/différence avec son ou ses co-hyponymes). La méthodologie préconisée propose trois principales phases à savoir l'engagement sémantique, fixant le sens linguistique des concepts, l'engagement ontologique fixant leur sens formel et l'opérationnalisation. La première phase consiste à construire le corpus de documents et le repérage des énoncés définitoires au moyen des patrons lexico-syntaxiques, dans la lignée des travaux de (Hearst, 1992), (Aussenac-Gilles et al., 2000) et (Morin, 1999). Une structuration globale des concepts en un arbre ontologique, qui modélise les différents concepts en fonction de leurs «sèmes », est réalisée. La structuration locale, selon le paradigme différentiel adopté (communauté/différence avec le père et les frères), détermine la signification d'un nœud en fonction de ses plus proches voisins. L'engagement

ontologique ou la formalisation des connaissances vise à formaliser l'arbre des concepts en une structure algébrique des ensembles, voire en un treillis. Cette étape permet d'obtenir une ontologie formelle. Finalement, l'opérationnalisation consiste à coder l'ontologie avec un langage spécifique de représentation des connaissances. Cette méthodologie est partialement supportée par l'outil DOE. Le processus proposé par **Lonsdale** et al. (2002) comprend cinq étapes à savoir le prétraitement des sources de connaissances ; la sélection des concepts via la mise en correspondance entre le contenu textuel de ces sources de connaissances et les données ontologiques ; la découverte des relations en générant des schémas de représentation de ces dernières par le biais des techniques de la théorie des graphes et l'algorithme OntoSearch utilisé dans le projet MikroKosmos ; la découverte et la spécification des contraintes sur les relations en suivant les conventions adoptées dans (Embley et David, 1998) ; le raffinement et l'évaluation des résultats sont élaborés par l'utilisateur en vue d'éviter les erreurs d'inconsistance. La qualité de cette construction dépend de la qualité des sources de données utilisées dans le processus.

2.2.2 Apprentissage fondé sur des techniques de clustering

Aguire et al. (2000) propose une méthodologie qui consiste à classifier les documents en collections par sens de mots en utilisant un corpus étiqueté et Wordnet, extraire pour chacune des collections les mots et leurs fréquences respectives et les comparer aux autres collections, et enfin, construire des Topic signatures[1] en se basant sur une fonction $\chi 2$ de signature et le filtrage des signatures. La méthodologie proposée dans (Khan et Luo, 2002) construit une ontologie du domaine à partir des documents textes en utilisant des techniques de clustering et WordNet. La construction de cette l'ontologie est réalisée dans un mode ascendant en suivant trois étapes à savoir la sélection du corpus de documents, la construction de la hiérarchie des concepts et l'affectation d'un concept à chaque noeud de la hiérarchie construite. La hiérarchie des concepts est élaborée suite au partitionnement de l'ensemble des documents du corpus en un nombre de clusters. Afin d'étendre la hiérarchie des clusters en plusieurs niveaux, les auteurs ont appliqué quelques techniques existantes à savoir le clustering hiérarchique agglomératif, SOM (Self-organizing Map) et l'algorithme SOTA modifié (Sef-Organizing Tree Algorithm).

2.2.3 Apprentissage multi stratégies à partir des sources de données hétérogènes

Maedche et Staab (2001) ont proposé quelques techniques exemplaires dans le cycle d'ingénierie d'ontologie qu'ils ont mis en application dans l'environnement Text-To-Onto. Dans ce contexte, ils ont cité principalement l'algorithme généralisé, basé sur les règles d'association permettant la découverte des relations non taxonomiques et la détermination du niveau d'abstraction approprié pour la définition de ces relations. Les principaux composants de cette architecture sont un composant générique de gestion se chargeant de la délégation des tâches et la constitution de la base de l'infrastructure ; un composant de traitement des ressources ; une librairie d'algorithmes travaillant sur les résultats du composant précédant ainsi que les structures d'ontologie esquissées ; et enfin, une interface graphique d'ingénierie d'ontologie.

[1] Une Topic signature TS est l'ensemble de termes reliés à un Topic, où le Topic est le concept cible.

2.2.4 Apprentissage basé le calcul des fréquences

Kietz et al. (2000) ont adopté une méthode générique de découverte d'une ontologie du domaine à partir des ressources hétérogènes en utilisant des techniques d'analyse du langage naturel. Pour cette méthode, ils ont adopté une approche de modélisation coopérative équilibrée (Morik, 1991) où la construction de l'ontologie est distribuée entre plusieurs algorithmes d'apprentissage et l'utilisateur. Le processus de construction de l'ontologie se réduit principalement à quatre étapes à savoir la sélection des sources de données non structurées et semi structurées, l'apprentissage des concepts, l'apprentissage des relations et l'évaluation. L'apprentissage des concepts est basé sur l'analyse des fréquences des termes : les termes dont les fréquences dans un corpus spécifique à un domaine sont considérablement supérieures à celles d'un corpus générique vont être présentés à l'utilisateur qui va décider de leur ajout ou non à l'ontologie. De même, l'analyse des fréquences peut aider à apprendre les relations ad hoc en découvrant les corrélations fréquentes entre les concepts en se basant sur l'algorithme des règles d'associations proposé dans (Srikant et Agrawal, 1995). Cette méthodologie est supportée par l'outil Text-To-Onto (Maedche et Volz, 2001).

2.2.5 Apprentissage basé sur des techniques statistiques et des techniques linguistiques

Faatz et Steinmetz (2002) présentent une approche destinée à enrichir une ontologie par l'extraction de la sémantique à partir du Web. Le processus d'enrichissement est basé sur la comparaison entre l'information statistique d'utilisation de mot dans un corpus et la structure de l'ontologie elle-même. Chaque concept dans le corpus doit être associé à un(e) ou plusieurs phrase(s) ou mot(s) du langage naturel. Cette approche propose une méthode de calcul de la similarité entre les mots dans le but d'enrichir la définition des concepts et de créer des clusters de mots reliés au nouveau concept. Ces nouveaux concepts seront proposés à l'expert qui décidera de leur l'ajout dans l'ontologie. La méthodologie de **Alfonseca et Manandhar** (2002a, 2002b) consiste à acquérir automatiquement les propriétés contextuelles des mots qui sont en cooccurrence avec un ensemble de concepts. Elle peut être utilisée pour grouper des concepts à l'intérieur d'une ontologie ainsi que pour raffiner l'ontologie en ajoutant de nouveaux concepts. Le principe est basé sur l'hypothèse de la sémantique distributive qui admet que "la signification d'un mot est fortement corrélée aux contextes dans lesquels il apparaît". Ils proposent méthode permettant de combiner les différents types de signatures dans un système.

En résumé les différentes approches se distinguent suivant les critères suivants : les sources d'apprentissage, le type d'ontologie construite, les techniques d'extraction des concepts, des relations et des axiomes ainsi que les outils disponibles. A partir de cet état de l'art, nous pouvons conclure que la construction d'ontologie est une tache fastidieuse et complexe impliquant plusieurs domaines. De même, les méthodologies les plus récentes sont orientées vers l'utilisation des connaissances à priori qui peuvent être de différents types. Toutefois, chacune des ces méthodologies propose l'application des différentes techniques au niveau de l'extraction des concepts et/ou l'extraction des relations mais rarement pour l'extraction des axiomes. Ces derniers ne représentent pas seulement des contraintes de restrictions de sélection ou des spécification des cardinalités des relations mais aussi des connaissances inférentielles du domaine. Quant à l'extraction des instances, nous citons des

techniques basées sur l'apprentissage avec la logique du premier ordre (Junker et al., 1999) ou sur l'apprentissage bayésien (Craven et al., 2000).

3 Présentation d'une ébauche de l'approche proposée

Après avoir étudié les travaux relatifs à la construction d'ontologie, nous avons défini une approche hybride dans le but d'obtenir une ontologie spécifique à un domaine et prête à être intégrée dans le web. Au cours de cette spécification, nous avons identifié deux idées directrices. Tout d'abord, la construction d'ontologie à partir du Web évoque la même problématique que celle dégagée dans le travail fait au niveau des ontologies destinées aux systèmes d'information traditionnels. Toutefois, des caractéristiques spécifiques à cette première reflètent un manque et une primitivité des méthodes et des outils concernant l'extraction des d'informations et de connaissances à partir du Web (problème de pertinence, de redondance, de contradiction des informations sur le Web et d'hétérogénéité des structures des documents). De même, les mesures de similarité d'ordre sémantique ne sont pas fiables et la complexité des algorithmes actuels de clustering conceptuel s'accroît au fur et à mesure que l'ensemble de données s'agrandit. En deuxième lieu, le fait de proposer une approche de construction totalement automatique reste de nos jours un objectif relativement utopique qui dépend essentiellement de l'évolution de l'intelligence artificielle ainsi que des méthodes et des algorithmes d'apprentissage. Notre approche émane de la relation cyclique entre le Web Mining, le Web sémantique et l'ontologie résumée dans la figure 1 construite à partir de (Berendt et al., 2002). Notre proposition doit satisfaire le fait que la finalité de l'approche de construction d'ontologie est d'accroître la capacité de cette dernière à spécifier et extraire les connaissances à partir du Web, de construire une symbiose entre la sémantique du contenu et de la structure des documents Web et de combiner les techniques de traitement automatique du langage naturel avec les techniques d'apprentissage en prenant en compte l'enrichissement et le passage à l'échelle de l'ontologie construite. Etant donné que l'ontologie est construite dans le but de faciliter la mise en place du Web sémantique, celle-ci ne doit pas renfermer seulement la sémantique des documents Web du domaine mais aussi la sémantique de la structure de ces documents. Dans notre proposition, l'ontologie à construire sera le résultat du processus d'extraction des connaissances à partir du Web dont la phase le plus importante est la fouille du web. Ce dernier comprend la fouille du contenu du Web, la fouille de la structure du Web et la fouille d'usage du web. Nous allons nous appuyer dans un premier temps sur les deux premiers types de fouille. Toutefois, nous sommes confrontés à deux types de problèmes ; le premier vient de l'hétérogénéité des documents Web qui sont à des degrés de structuration différents; le deuxième est d'ordre technique et concerne le choix des techniques d'extraction des concepts, des relations et des axiomes, le choix des sources d'apprentissage, le suivi des contraintes de passage à l'échelle. Vu qu'une ontologie ne peut pas spécifier à la fois la sémantique du domaine et de la structure des documents Web, nous avons construit une architecture de composants ontologiques pour le Web sémantique, intégrant également la notion de service. La seule construction d'une ontologie du domaine n'est pas suffisante pour faciliter la recherche d'information. La structure des sites Web doit également être prise en compte, ainsi que la définition d'un ensemble de services relatifs au domaine considéré. Nous proposons l'ensemble d'ontologies interdépendantes représentées dans la figure 1. Nous distinguons trois ontologies à savoir une

ontologie générique des structures des sites Web, une ontologie de domaine et une ontologie des services.

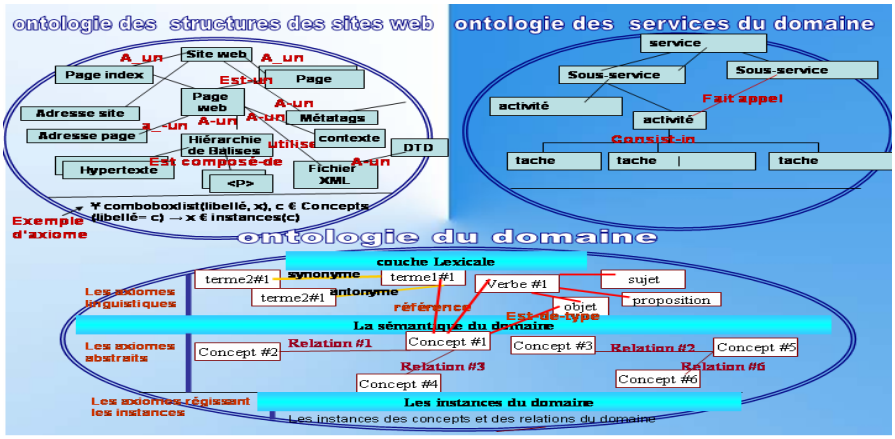


Figure 1 Les différents composants ontologiques

L'ontologie générique des structures des sites Web contient les concepts et les relations qui permettent une description commune des structures des pages Web HTML, XML et des DTD. La particularité de cette ontologie est de permettre l'apprentissage des axiomes qui spécifient la sémantique des formes des documents Web (exemple les menus hypertextuels, les mots au sein d'une balise «
 » et la liste des mots dans un « combobox » permettent d'enrichir l'ontologie du domaine en associant les concepts correspondants avec leurs instances). Le but est d'établir des axiomes spécifiant les résultats des techniques de fouille dans la structure des pages Web dont le résultat peut alimenter les instances de l'ontologie de domaine. **L'ontologie du domaine** est semblable aux ontologies d'application ou d'entreprise. Nous distinguons dans cette ontologie trois couches ordonnées suivant leur degré d'abstraction. La première couche est lexicale ; elle spécifie les connaissances linguistiques (exemple les patrons de verbes avec les propositions adéquates, les dérivation des noms, adjectifs et adverbes à partir des verbes) de haut niveau d'abstraction qui peuvent aider dans l'apprentissage de patrons lexico-syntaxiques ; ces connaissances lexicales couvrent les concepts, les relations et les axiomes d'ordre général de la couche centrale de l'ontologie du domaine. Quant à la dernière couche, elle est la plus opérationnelle étant donné qu'elle comprend, en plus des instances des concepts et des relations, les connaissances inférentielles ou les axiomes. Ces derniers servent à spécifier les contraintes sur chacune des couches de l'ontologie du domaine, ainsi qu'elles renferment des règles d'extraction des instances associées à un coefficient de confiance. Ces règles sont les résultats combinés de l'application des patrons lexico-syntaxtiques et du clustering basé sur des mesures de similarités relativement à un espace multidimensionnel de mot (Yamaguchi, 2001) spécifiant les connaissances du domaine. Ces axiomes sont enrichis d'une manière incrémentale par la fouille des contenus des pages Web visitées et la fouille des structures des sites Web. **L'ontologie de services** est définie en s'inspirant d'un type d'ontologie appelé « ontologie de tâches » (Gomez-Perez et Rojas, 1999). Dans le contexte du Web, nous ne parlons pas de tâches mais plutôt de « services Web ». Cette ontologie spécifie les services du domaine et sera utile pour la transformation des connaissances Web en un

ensemble de services Web interdépendants. Cette ontologie sert à avoir une vue macroscopique du domaine. Elle est structurée hiérarchiquement : la racine est le service « Root » et les feuilles sont des taches élémentaires dont chacune est associée à un triplet « concept-Relation-concept » issu de l'ontologie du domaine. Cette architecture implique de concevoir trois méthodologies différentes pour chacun des composants ontologiques, ainsi qu'une quatrième pour la mise en correspondance des trois ontologies. Cette mise en correspondance s'effectue par l'application de méthodes de fouille au niveau des instances, des techniques avancées du Web Mining (fouille d'hyperliens et hypertextes) et enrichissent les axiomes de l'ontologie générique des structures des sites Web et de l'ontologie de services. Cette architecture est doublement fonctionnelle. En effet, ces composants ontologiques formeront une base de connaissances Web d'un domaine spécifique et chacune des ontologies s'enrichissent mutuellement en s'appuyant sur les axiomes d'une autre. Dans ce cadre de notre expérimentation, nous allons nous intéresser dans un premier temps à la construction de l'ontologie du domaine du tourisme. Pour ce faire, nous proposons une approche incrémentale ayant comme sources de données une ontologie générale «Wordnet », le thésaurus du tourisme et des loisirs de l'OMT (l'Organisation Mondiale du Tourisme), un corpus textuel et un corpus des pages Web. Notre approche se base donc sur l'utilisation de connaissance à priori et est incrémentale. Dans cette phase, nous allons construire l'ontologie du domaine qui comporte essentiellement 3 niveaux d'abstraction à savoir un niveau linguistiques, un deuxième comprenant les concepts effectifs du domaine et les relations qui les lient et un dernier qui va être utile à spécifier les instances des concepts du domaine extraite des exemples du monde réel. Ce dernier niveau sera utile pour l'extraction des connaissances inférencielles ou de raisonnement servant à dégager les relations de dépendance entre les différents contextes sémantiques, sachant un contexte sémantique est définit comme étant un sous-ensemble des concepts et les relations qui les lient. En suivant la majorité des méthodologies de construction d'ontologie, nous allons commencer par la construction d'une ontologie minimale qui va être étendue suivant un processus d'enrichissement incrémental. L'idée est de construire une ontologie minimale qui comporte essentiellement trois niveaux d'abstraction à savoir un niveau linguistique, un deuxième comprenant les concepts effectifs du domaine et les relations qui les lient et un dernier qui va être utile pour spécifier les instances des concepts du domaine extraites des exemples du monde réel. Ce dernier niveau permettra l'extraction des connaissances inférencielles ou de raisonnement servant à dégager les relations de dépendance entre les différents contextes sémantiques, sachant un contexte sémantique est définit comme étant un sous-ensemble des concepts et les relations qui les lient. Quand à l'extraction des concepts, des relations et des axiomes, l'idée est de combiner les techniques linguistiques et les techniques d'apprentissage. Dans un premier temps, nous envisageons d'utiliser l'outil Text-To-Onto qui adopte une approche multistratégique qui met à notre disposition une librairie d'algorithmes d'extraction des concepts et des relations, pour enrichir notre ontologie minimale. Puis, dans un second temps, nous développerons nos propres algorithmes d'extraction de concepts (techniques de clustering), de relations taxonomiques (utilisation de la connaissance à priori) et de relations non taxonomiques (patrons lexico-syntaxiques), pour construire une ontologie de domaine complète.

4 Conclusion et perspectives

Les différentes méthodologies de construction d'ontologies décrites dans cet article sont relativement complémentaires. En effet, les méthodologies «from scratch » sont orientées vers l'étude des processus d'ingénierie ontologique et des cycles de vie de ces dernières en s'appuyant sur les méthodologies de développement d'un système d'information. L'apprentissage d'ontologie essaie de résoudre le problème de la construction manuelle des ontologies. Plusieurs auteurs appliquent des techniques d'apprentissage dont certaines sont symboliques et d'autres numériques. Ces techniques ont été exploitées pour rendre semi-automatique quelques tâches de base à savoir la construction d'une hiérarchie de concepts, l'extraction des relations taxonomiques, l'apprentissage des relations non taxonomiques, etc. Les méthodologies de rétro-conception viennent dans le cadre de la maintenance d'une ontologie existante ou de sa remise en cause afin d'en construire une autre adaptée au domaine cible. Tous ces travaux constituent une boite à outils méthodologiques pour aborder le problème de la semi-automatisation des ontologies. Toutefois, la construction d'ontologie pour le Web sémantique nous incite non seulement à résoudre les problèmes évoqués par les expériences antérieures mais aussi à concevoir une vision de ces ontologies en fixant comme perspectives, la fouille du Web sémantique et une restructuration des différentes pages Web, pour la mise en œuvre du Web adaptatif, en se basant sur la sémantique de la structure, du contenu et des services. Dans ce cadre, nous avons présenté une architecture de composants ontologiques. Dans un premier temps, nous allons nous attacher à automatiser la construction de l'ontologie du domaine.

Références

Arpirez, J., O. Corcho, M. Fernandez-Lopez and A. Gomez-Perez (2001). *WebODE: a Workbench for Ontological Engineering*. In First international Conference on Knowledge Capture, Victoria, Canada.

Agirre, E., O. Ansa, E. Hovy and D. Matinez (2000). *Enriching very large ontologies using the WWW*. In Proceedings of the Workshop on Ontology Construction of the European Conference of AI.

Alfonseca, E. and S. Manandhar (2002a). *An unsupervised method for general named entity recognition and automated concept discovery*. In Proceedings of the 1st International Conference on General WordNet, Mysore, India.

Alfonseca, E. and S. Manandhar (2002b). *Improving an Ontology Refinement Method with Hyponymy Patterns*. Language Resources and Evaluation (LREC-), Las Palmas, Spain.

Aussenac-Gilles, N., Biébow, B. and S. Szulman (2000a). *Corpus Analysis For Conceptual Modelling*.Workshop on Ontologies and Text, Knowledge Engineering and Knowledge Management, 12th International Conference EKAW, Juan-les-pins, France.

Aussenac-Gilles, N., B. Biébow and S. Szulman (2000b). *Revisiting Ontology Design: A MethodologyBased on Corpus Analysis*. In 12th International Conference in Knowledge Engineering and Knowledge Management (EKAW). Juan-Les-Pins, France.

Bachimont B. (2000). *Engagement sémantique et engagement ontologique*. In Eds., *Ingénierie des connaissances : évolutions récentes et nouveaux défi* , chapitre 19. Paris: Eyrolles.

Berendt, B., A. Hotho and G. Stumme (2002). *Towards semantic web mining*. In. International Semantic Web Conference, volume 2342 of Lecture Notes in Computer Science, Springer, 264–278.

Biébow, B. and S. Szulman. (1999) *TERMINAE: a linguistic-based tool for the building of a domain ontology*. In 11th European Workshop on Knowledge Acquisition, Modelling and management. Dagstuhl, Germany, LCNS, Berlin, Springer-Verlag, 49-66.

Craven, M., D. Dipasquo, D. Freitag, A. Mccallum, T. Mitchell., K. Nigam and S. Slattery (2000). *Learning to construct knowledge bases from the World Wide Web*. Artificial Intelligence.

Decker, S., M. Erdmann, D. Fensel. and R. Studer (1999). *Ontobroker: Ontology Based Access to Distributed and Semi-Structured Information*. In Semantic Issues in Multimedia Systems. Proceedings of DS-8. Kluwer Academic Publisher, Boston, 351-369.

Deitel, A., C. Faron,, and R. Dieng (2001). *Learning ontologies from RDF annotations*. In Proceedings of the IJCAI Workshop in Ontology Learning, Seattle.

Doan, A., P. Domingos, and A. Levy (2000). *Learning Source Descriptions for Data Integration*. Proceedings of the Third International Workshop on the Web and Databases.

Embley, D.W and W. David (1998). *Object Database Development: Concept and Principles*. Addison-Wesley, Reading, MA.

Eusenat, J. (1995). Building *consensual knowledge bases: context and architecture*, in Proceedings of 2nd international conference on *building and sharing very large-scale knowledge bases*, Enschede, IOS press, Amsterdam (NL).

Faatz, A. and Steinmetz, R. (2002). *Ontology enrichment with texts from the WWW*. Semantic Web Mining 2nd Workshop at ECML/PKDD-2002, Helsinki, Finland.

Fernandez, M., A. Gòmez-Pérez and N Juristo (1997). *METHONTOLOGY: From Ontological Art Towards Ontological Engineering*. Symposium on Ontological Engineering of AAAI. Stanford ,California.

Fernandez, M., A. Gòmez-Pérez, A. Pazos-Sierra and J. Pazos-Sierra (1999). *Building a Chemical Ontology Using METHONTOLOGY and the Ontology Design Environment*. IEEE Intelligent Systems & their applications.

Gòmez-Pérez, A. et M.D. Rojas (1999). *Ontological Reengineering and Reuse*. European Knowledge Acquisition Workshop (EKAW).

Grüninger, M. and M.S Fox (1995). *Methodology for the design and evaluation of ontologies*. Workshop on Basic Ontological Issues in Knowledge Sharing, Montreal, Canada.

Hearst, M.A. (1992). *Automatic acquisition of Hyponyms from large text corpora*. In Proceedings of the Fourteenth International Conference on Computational Linguistic, Nantes, France.

Hearst, M.A. (1998). *Automated Discovery of WordNet Relations*. In WordNet: In C Fellbaum ed. "*Wordnet An Electronic Lexical Database*". MIT Press, Cambridge, MA, 132-152.

Jannink, J. (1999). *Thesaurus Entry Extraction from an On-line Dictionary*. In Proceedings of Fusion '99, Sunnyvale CA.

Johannesson, P. (1994). *A Method for Transforming Relational Schemas into Conceptual Schemas*. In 10th International Conference on Data Engineering, Ed. M. Rusinkiewicz, Houston, 115 - 122.

Junker, M., M. Sintek, and M. RINCK (1999). *Learning for Text Categorization and Information Extraction with ILP*. In: J. Cussens (eds.), Proceedings of the 1st Workshop on Learning Language in Logic, Bled, Slovenia, 84-93.

Khan, L., and F. Luo (2002). *Ontology Construction for Information Selection*. In Proc. of 14th IEEE International Conference on Tools with Artificial Intelligence, Washington DC,122-127.

Kashyap, V. (1999). *Design and Creation of Ontologies for Environmental Information Retrieval*. Twelfth Workshop on Knowledge Acquisition, Modelling and Management Voyager Inn, Banff, Alberta, Canada.

Kiez J., A. Maedche and R. Volz (2000). *A Method for Semi-automatic Ontology Acquisition from a Corporate Intranet*, Workshop ``Ontologies and text'', co-located with EKAW'2000.

Lonsdale, D., Y. Ding, D.W. Embley and A. Melby (2002). *Peppering Knowledge Sources with SALT. Boosting Conceptual Content for Ontology Generation*. Proceedings of the AAAI Workshop on Semantic Web Meets Language Resources, Edmonton, Alberta, Canada.

Maedche, A. and S. Staab (2001). *Ontology Learning for the Semantic Web*. IEEE Intelligent Systems,Special Issue on the Semantic Web, 16(2):72–79.

Maedche, A. and R. Volz (2001). The Text-To-Onto Ontology Extraction and Maintenance Environment.Proceedings of the ICDM Workshop on integrating data mining and knowledge management, San Jose, California, USA.

Malaisé, V., P. Zweigenbaum and B. Bachimont (2004). *Extraction d'informations sémantiques pour l'aide à la construction d'ontologies différentielles*, In Actes Journées d'étude Terminologie, Ontologie et Représentation des Connaissances, Lyon.

Moldovan, D. I. and R. Girju (2000). *Domain-Specic Knowledge Acquisition and Classication using WordNet*, In proceeding of FLAIRS2000 Conference, ORLANDO.

Morik, K. (1991). *Balanced Cooperative Modeling Using MOBAL - An Introduction*. Technical Report (GMD-F3-Nachrichten AC Special Nr. 3), GMD, St. Augustin

Morin, E. (1999). *Des patrons lexico-syntaxiques pour aider au dépouillement terminologique*, Traitement Automatique des Langues, vol 40(1), 143-166.

Nobécourt, J. (2000). *A method to build formal ontologies from text*. In: EKAW-2000 Workshop on ontologies and text, Juan-Les-Pins, France.

Papatheodrou, C., A. Vassiliou and B. Simon (2002). *Discovery of Ontologies for Learning Resources Using Word-based Clustering*, EDMEDIA 2002, Copyright by AACE, Reprinted, Denver, USA.

Rubin D.L., M. Hewett, D.E Oliver., T.E Klein, and R.B Altman (2002). *Automatic data acquisition into ontologies from pharmacogenetics relational data sources using declarative object definitions and XML*. In: Proceedings of the Pacific Symposium on Biology, Lihue, HI.

Suryanto, H. and P. Compton (2001). *Discovery of Ontologies from Knowledge Bases*. Proceedings of the First International Conference on Knowledge Capture, The Association for Computing Machinery,New York, USA, pp171-178.

Srikant, R., and R. Agrawal (1995). *Mining generalized association rules*. In Dayal, Conférence Knowledge Discovery and Data Mining, Newport Beach, CA.

Uschold, M. and M. KING (1995). *Towards a Methodology for Building Ontologies*. Workshop on Basic Ontological Issues in Knowledge Sharing.

Yamaguchi T. (2001). *Acquiring Conceptual Relationships from Domain-Specific Texts*, Proceedings of the Second Workshop on Ontology Learning OL'2001 Seattle,

Summary

This paper presents an architecture of ontological components for the Semantic Web. Many methods and methodologies can be found in the literature. Generally, they are dedicated to a particular data type like text, semi-structured data, relational data, etc. Our work deals with web pages. We first study the state of the art of methodologies defined to learn ontologies from texts. Then, our architecture of ontological components for the Semantic web is defined.

Ontologie et base de connaissances pour le pré-traitement et post-traitement en fouille de données

Laurent BRISSON,
Martine COLLARD,
Nicolas PASQUIER

Laboratoire I3S - Université de Nice
2000 route des Lucioles
06903 Sophia-Antipolis, France
{brisson,mcollard,pasquier}@i3s.unice.fr
http://www.i3s.unice.fr/execo/

Résumé. Dans cet article, nous présentons la méthodologie EXCIS (Extraction using a Conceptual Information System) orientée ontologie qui permet d'intégrer la connaissance des experts dans un processus de fouille de données. EXCIS décrit les étapes d'un processus à la manière de la méthodologie CRISP-DM, mais son originalité réside dans la construction d'un système d'information conceptuel (CIS) lié au domaine d'application qui permet d'améliorer le pré-traitement des données et l'interprétation des résultats. ExCIS est en cours de développement et cet article présente uniquement la construction du système d'information qui consiste en la création de trois éléments : une ontologie extraite à partir des données brutes, une base de données orientée "fouille" dont les attributs sont les concepts de l'ontologie et une base de connaissance.

1 Introduction

Un des défis de la fouille de données est d'extraire de l'information qui soit intéressante et utile pour les utilisateurs experts. De nombreux algorithmes ont été construits pour extraire les modèles les meilleurs au sens de critères comme la précision, la surface de ROC, le lift ou d'autres mesures. Un certain nombre de travaux portent sur des indices qui mesurent l'intéressabilité des modèles extraits (Hilderman et al., 2001; Liu et al., 1999). Ils distinguent en général l'intérêt objectif de l'intérêt subjectif. La méthode développée par Liu et al. (1999) repose sur la prise en compte des attentes de l'utilisateur. Silberschatz et al. (1995) ont proposé une méthode qui définit l'inattendu à l'aide d'un système de croyances ; dans cette approche, sont définies les croyances faibles que l'utilisateur peut changer si de nouveaux motifs sont découverts et des croyances fortes qui ne peuvent pas être remises en cause.

Dans la plupart des projets de fouille de données, la connaissance a priori est soit implicite, soit organisée dans un système conceptuel structuré. EXCIS est dédié aux situations de fouille de données dans lesquelles la connaissance de l'expert est cruciale pour l'interprétation des motifs extraits, aucune représentation conceptuelle de cette connaissance n'existe et le processus de fouille ne dispose que de bases de données opérationnelles. Dans cette approche, une

ontologie du domaine est construite en analysant les données existantes à l'aide des experts dont le rôle est très important. Le processus de conception de l'ontologie est dirigé en vue de faciliter non seulement la préparation des jeux de données à fouiller, mais également l'interprétation des résultats. L'objectif central dans EXCIS, est de fournir une solution de manière à ce que le processus d'extraction puisse faire usage d'un système d'information conceptuel (CIS : Conceptual Information System) pour optimiser la qualité de la connaissance extraite. Nous considérons le paradigme de CIS comme défini par Stumme (2000). Le CIS fournit la structure d'information utile pour les tâches de fouille qui se succèdent. Il contient un schéma conceptuel définissant une *Ontologie* étendue par une *Base de Connaissance* (ensemble d'informations factuelles sur le domaine d'intérêt) et une *Base de données orientée fouille*(MOBD : Mining-Oriented relational DataBase). L'extraction de l'ontologie et la construction de la base sont également dirigées par l'objectif de fouille.

Les ontologies (Gruber, 2002) fournissent un support formel pour exprimer les croyances (Silberschatz et al., 1995) et la connaissance a priori sur un domaine. Des ontologies ne sont pas disponibles sur tous les domaines ; elles doivent être construites spécifiquement en interrogeant les experts et en analysant les données existantes. L'extraction de structures ontologiques à partir des données est très similaire au processus de recherche d'un schéma conceptuel dans une base de données opérationelle. Différentes méthodes ont été proposées par Kashyap (1999), Johannesson (1994), Stojanovic et al. (2002) et Rubin et al. (2002). Elles sont basées sur le postulat selon lequel la connaissance stockée dans les données relationnelles est suffisante pour produire une construction intelligente de l'ontologie. Elles appliquent en général une correspondance entre concepts ontologiques et tables relationnelles de sorte que l'ontologie est très similaire au schéma conceptuel de la base de données. Dans EXCIS, l'ontologie fournit une représentation conceptuelle du domaine d'application principalement extraite par l'analyse des données opérationnelles. Les caractéristiques principales de la méthodologie sont :

- Conceptualisation de la connaissance implicite : le CIS est conçu pour que les tâches de fouille utilisent l'ontologie, la base de connaissance et la MODB
- Adaptation à la méthodologie CRISP-DM avec :
  - le pré-traitement des jeux de données à l'aide du CIS.
  - le post-traitement des modèles extraits pour filtrer les informations surprenantes et/ou utilisables.
  - l'évolution incrémentale de la connaissance stockée dans le CIS.

Ce projet est actuellement mis en oeuvre dans le cadre d'une étude sur des données de la Caisse Nationale d'Allocations Familiales (CNAF). L'objectif de l'étude est l'amélioration des relations entre l'organisme et les allocataires. Nous disposons de deux sources d'information : une base de données des allocataires et de leurs contacts avec les caisses d'une part et la connaissance des agents des caisses, experts en matière de processus de gestion, comportements et habitudes au sein de l'organisme.

Cet article est organisé comme suit. La Section 2 donne une vue générale de l'approche EXCIS. La Section 3 décrit les structures conceptuelles de l'ontologie. Dans la Section 4, nous donnons une description détaillée de la construction du CIS. La Section 5 est dédiée à la construction de la base de connaissance et aux mécanismes d'inférence à l'aide du serveur Cyc. La Section 6 donne une conclusion.

2 Vue générale de l'approche EXCIS

ExCIS intègre la connaissance experte tout au long du processus de fouille : dans une première phase, la connaissance est structurée et organisée dans le CIS, puis dans les phases qui suivent, elle est exploitée et étendue.

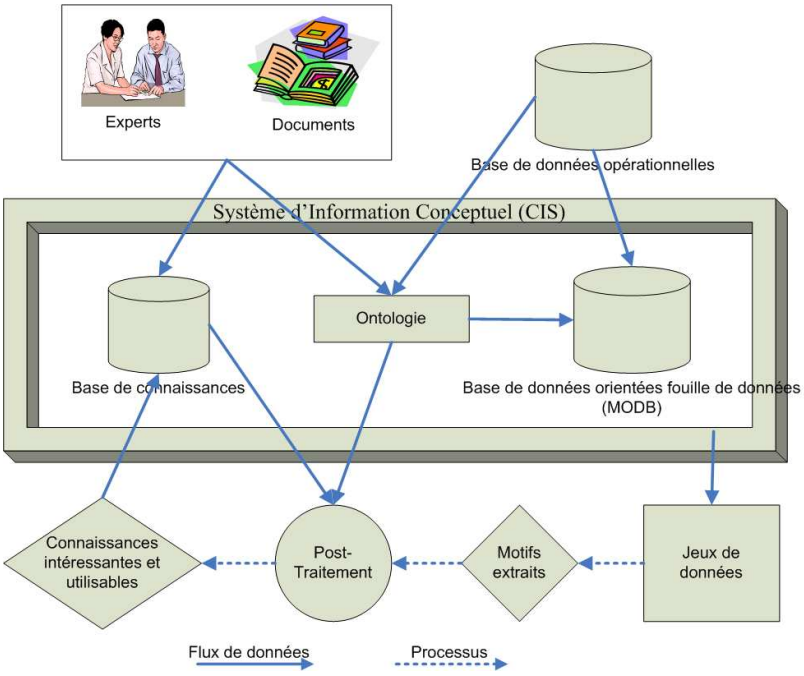


FIG. 1 – *Clustering hiérarchique des conditions.*

Le processus global présenté par la figure 1 montre :

- La construction du CIS où :
  - L'ontologie est extraite en analysant la base de données et en dialoguant avec les experts
  - La base de connaissance est déduite dans un premier temps des dialogues avec les experts
  - La nouvelle base de données MODB est construite.
- L'étape de pré-traitement dans laquelle des jeux de données spécifiques sont construits pour des tâches de fouille particulières
- L'étape de fouille standard où les motifs sont extraits
- L'étape de post-traitement où les motifs découverts peuvent être interprétés et/ou filtrés par comparaison à la fois avec la connaissance a priori stockée dans le CIS et les attentes individuelles des experts.

La MODB est dite générique car elle tient le rôle d'une sorte de réservoir de données à partir duquel des jeux spécifiques peuvent être générés. L'idée sous-jacente dans le CIS est de construire des structures qui procurent plus de souplesse à la fois dans les travaux de pré-

traitement et dans le post-traitement des modèles découverts. Les structures hiérarchiques et les liens de généralisation/spécialisation entre les concepts ontologiques jouent un rôle central :

- Ils permettent de réduire la taille des modèles extraits comme les ensembles de règles souvent volumineux
- Ils fournissent également un outil pour l'interprétation des résultats (classes) d'algorithmes de classification non supervisée.

Pour des données numériques ou catégorielles, ils fournissent des niveaux de granularité différents très utiles dans le pré-traitement et dans le post-traitement.

Exemple Supposons que 10 et 11 soient définis dans l'ontologie comme des concepts (NUMJOUR) qui héritent d'un concept plus général «NUMSEMAINE=2», alors les règles Règle1 et Règle2 peuvent être généralisées et remplacées par la Règle3 dont la partie gauche est plus générale.

```
Règle 1: If NUMJOUR = 10 and MOTIF = "Appel Entrant"
            then OBJECTIF = "Demande credit"
Règle 2: If NUMJOUR = 11 and MOTIF = "Appel Entrant"
            then OBJECTIF = "Demande credit"
Règle 3: If NUMSEMAINE = 2 and MOTIF = 'Appel Entrant"
            then OBJECTIF = "Demande credit"
```

3 Structures conceptuelles de l'ontologie

3.1 Ontologie

Dans l'approche EXCIS, l'ontologie du domaine est un outil essentiel à la fois pour améliorer le processus de fouille et pour interpréter ses résultats. L'ontologie est définie par un ensemble de concepts et de relations entre concepts qui sont découverts en analysant les données. Elle apporte un soutien dans l'étape de pré-traitement pour construire la MODB et dans le post-traitement pour raffiner les motifs extraits. Comme montré par la figure 2, les relations de généralisation/spécialisation entre concepts ontologiques fournissent une information importante ; ils peuvent être largement exploités pour réduire la taille des motifs découverts. Par exemple, un ensemble de règles de dépendances (règles attribut-valeur) peut être réduit par généralisation sur les attributs ou par généralisation sur les valeurs. Aussi, les règles de construction de l'ontologie sont les suivantes :

- Distinguer concept-attribut et concept-valeur.
- Etablir une correspondance entre attributs sources et concepts attributs d'une part et valeurs sources et concepts valeurs d'autre part
- Définir des hiérarchies de concepts.

Cette ontologie a deux caractéristiques importantes inhérentes à l'objectif de foullle de données :

- Elle ne contient aucune instance puisque les valeurs sont organisées en hiérarchies et considérées comme des concepts ; les instances sont uniquement présentes dans la base de données finale MODB.
- Chaque concept a uniquement deux propriétés génériques.

La MODB est une base de données relationnelle dont le rôle est de stocker les données de granularité plus fine issues de la base de données opérationnelle. Les attributs de la MODB sont ceux qui sont identifiés comme pertinents a priori pour la fouille et ses instances sont des enregistrements de valeurs de granularité le plus fine.

3.2 Notion de concept

Dans une ontologie EXCIS, un concept doit faire référence à un paradigme du domaine utile pour le processus de fouille. Un concept d'EXCIS est caractérisé par les deux propriétés suivantes : son rôle dans un motif extrait (attribut ou valeur) et une propriété booléenne qui indique sa présence dans la MODB. Un *concept-attribut* correspond à une propriété d'une donnée initiale et un *concept-valeur* correspond aux valeurs d'une telle propriété. Un concept qui est présent dans la MODB est appelé un *concept concret*. Un concept qui n'est pas concret, mais est utile pendant l'étape de post-traitement est appelé un *concept abstrait*. Par exemple sur la figure 2 «Nombre Enfants» est un concept-attribut abstrait et «3 Enfants» est un concept-valeur concret.

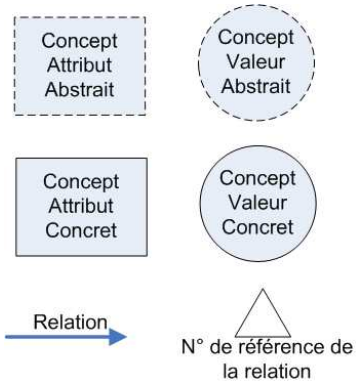


FIG. 2 – *Légende*

3.3 Relation entre concepts

Une relation est un lien orienté entre deux concepts. Etant donné que 4 types différents de concepts existent dans EXCIS et que nous distinguons les relations entre concepts d'une même hiérarchie et entre concepts de hiérarchies différentes, nous avons 32 sortes de relations entre concepts. Parmi ces relations, se distinguent particulièrement 3 cas :

- Relations de généralisation/spécialisation entre deux concepts-valeur qui sont des liens «est-une-sorte-de» entre 2 concepts-valeur (voir la relation 2 sur la figure 3)
- Relations de généralisation/spécialisation entre deux concepts-attribut qui sont des liens «est-une-sorte-de» entre 2 concepts-attribut (voir la relation 8 sur la figure 3)
- Relations de généralisation/spécialisation entre un concept-attribut et un concept-valeur qui sont des liens «est-une-valeur-de» (voir la relation 6 sur la figure 3)

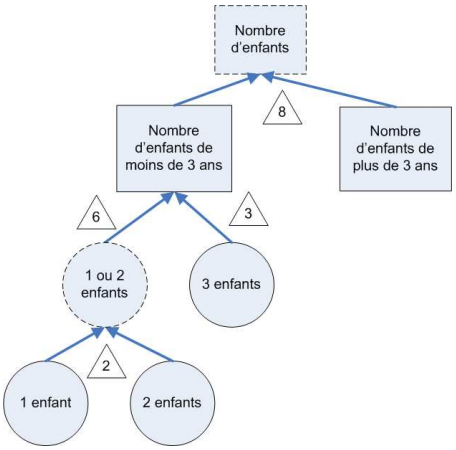


FIG. 3 – *Concepts relatifs aux enfants*

Les relations entre concepts à l'intérieur de la même hiérarchie sont énumérées dans la table 1 et les relations entre concepts de hiérarchies différentes sont énumérées dans la table 1. Les relations autorisées sont numérotées, les relations interdites sont représentées par une lettre.

TAB. 1 – *Relations entre concepts de la même hiérarchie*

Concept	Concret Valeur	Abstrait Valeur	Concret Attribut	Abstrait Attribut
Valeur Concret	1	2	3	D
Valeur Abstrait	B	2	6	5
Attribut Concret	C	C	9	7,8
Attribut Abstrait	C	C	A	10

TAB. 2 – *Relations entre concepts de hiérarchies différentes*

Concept	Concret Valeur	Abstrait Valeur	Concret Attribut	Abstrait Attribut
Valeur Concret	4	4	D	5
Valeur Abstrait	B	4	D	5
Attribut Concret	C	C	D	D
Attribut Abstrait	C	C	A	D

3.4 Description et utilisation des relations ontologiques

Avant tout, deux relations sont interdites dans EXCIS : toute relation de généralisation/spécialisation d'un concept-abstrait vers un concept concret qui a le même rôle (attribut ou valeur) car les

concepts abstraits ont été définis pour être plus généraux que les concepts concrets (voir les relations A ou B sur la figure 4), et toute relation de généralisation/spécialisation d'un concept-attribut vers un concept-valeur (voir la relation C sur la figure 4) car la relation «est-une-valeur-de» n'a pas de signification dans ce cas.

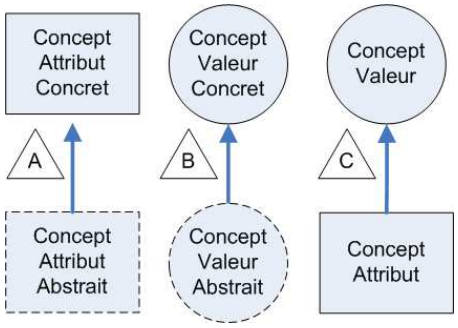


FIG. 4 – *Relations interdites*

3.4.1 Relations entre concepts-valeur

La généralisation ou la spécialisation entre concepts-valeur (voir relation 2 sur la figure 3) sont utiles de manière à généraliser des motifs pendant la phase de post-traitement. De plus, les relations entre deux concepts-valeur concrets de la même hiérarchie sont essentiels car ils permettent de choisir différents grains dans les données issues de la MODB. Si, par exemple, dans une session du processus de fouille, nous nous intéressons plus particulièrement aux types d'allocations, le grain des données sera choisi au niveau «Allocation Logement» (voir relation 1 sur la figure 5).

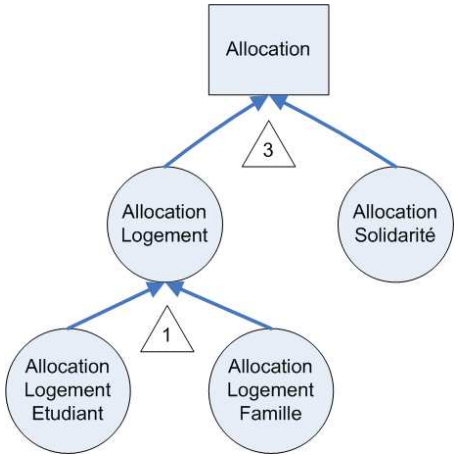


FIG. 5 – *Concepts relatifs aux allocations*

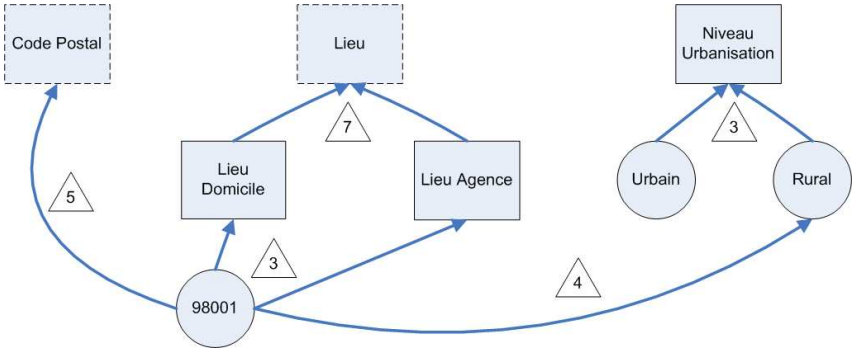


FIG. 6 – *Concepts relatifs à la localisation*

3.4.2 Relations entre concepts-attribut

La généralisation ou la spécialisation entre concepts-attribut sont également utiles de manière à généraliser des motifs pendant la phase de post-traitement. Cependant, ceci demande de procéder avec précaution car dans certains cas un attribut doit être remplacé par un autre attribut plus général (voir relation 7 sur la figure 6) et dans d'autres cas, une nouvelle valeur doit être calculée (voir relation 8 sur la figure 3). Par exemple, sur la figure 3 «Nombre Enfants» est la somme des valeurs de tous ses sous-concepts.

Les relations entre deux concepts-attribut concrets d'une même hiérarchie sont spécifiques ; elles doivent en effet être vérifiées pendant la génération des jeux de données.

La méthode EXCIS interdit des relations entre concepts-attribut de différentes hiérarchies Les relations entre concepts-attribut sont uniquement des relations de généralisation ; les concepts qui sont sémantiquement proches doivent être placés dans la même hiérarchie. Par exemple, «Lieu Domicile» et «Lieu Agence» sont dans la même hiérarchie (voir relation 7 sur la figure 6).

3.4.3 Relations entre concepts-valeur et concepts-attribut

Ces relations sont essentielles pour construire les données et fournir des vues sémantiques différentes pendant la phase de post-traitement (voir relation 5 sur la figure 6). Par exemple «98001» est à la fois un «Lieu Domicile» et un «Code Postal» (voir relation 3 sur la figure 6).

Chaque concept-valeur est lié à des concepts-attribut dans la même hiérarchie. EXCIS interdit des relations entre un concept-valeur et des concepts-attribut concrets de différentes hiérarchies ; en effet, si une telle relation existait, cela signifierait qu'un concept-valeur «est-une-valeur-de» deux différents concepts-attribut. Si ces concepts-attribut sont sémantiquement proches, ils doivent être placés dans une même hiérarchie et s'ils sont totalement différents, ils ne peuvent pas être en relation avec les même concepts-valeur.

4 Le Système d'Information Conceptuel (CIS)

Soit A l'ensemble des attributs de la base de données d'origine, C l'ensemble des concepts de l'ontologie et C_z l'ensemble des concepts associés à l'attribut $z \in A$. C est défini par $\bigcup_{z \in A} C_z$.

ExCIS diffère de CRISP-DM lors de la phase de préparation des données. Dans cette phase CRISP-DM décrit cinq tâches : sélection, nettoyage, construction, intégration et formatage des données. Les tâches de sélection et de formattage sont identiques dans les deux méthodes. Cependant dans ExCIS les tâches de nettoyage, construction et intégration des données sont fusionnées afin de pouvoir créer les concepts de l'ontologie et construire la MODB.

4.1 Définition de la portée et sélection des attribut d'origine

Les premières étapes de la méthode ExCIS correspondent aux étapes de compréhension du domaine et de compréhension des données de la méthode CRISP-DM. Ces étapes nécessitent une interaction importante avec les experts du domaine.

1. Déterminer les objectifs : par exemple dans notre cas d'application l'objectif est d'améliorer les «relations avec l'allocataire».

2. Définir des thèmes : en analysant les données nous pouvons les regrouper en ensembles sémantiques que l'on appelle thèmes. Par exemple nous avons créé trois thèmes : les profils allocataire, les contacts (par téléphone, courier, couriel, à l'agence, ...) et les évèvements (vacances, rentrée scolaire, naissance, mariage, ...).

3. Associer à chaque thème un ensemble d'attributs avec l'aide des experts.

4.2 Analyse des données et création des concepts attributs

4. Pour chaque attribut z :

5. Considérer son nom et sa description pour :
- Associer n concepts à l'attribut.
- Nettoyer C des homonymes (concepts différents avec le même nom), des synonymes (même concepts mais noms différents comme l'âge et la date de naissance par exemple), et des attributs inutiles selon nos objectifs.

6. Etudier les valeurs (distribution, valeurs manquantes, doublons, ...) afin de :
- Raffiner C_z (ajout ou suppression de concepts) selon les informations obtenues lors de l'étude.
- Nettoyer de nouveau les homonymes, synonymes et attributs inutiles. Par exemple, en analysant les valeurs nous nous sommes rendus compte que «Allocations» était en fait 2 concepts homonymes. Ainsi nous avons créé le concept «Montant de l'allocation» et le concept «Bénéficiaire de l'allocation».

7. Pour chaque concept associé à z, créer la procédure qui générera les concepts valeurs.

Lors de l'étape 7, si le concept attribut n'existe pas nous devons créer une table avec quatre champs. Ces champs sont l'attribut associé au concept, le nom de la table contenant l'attribut dans la base de données d'origine, le domaine de valeur de l'attribut et la référence à la procédure qui générera les concepts valeurs. Il y a une unique procédure par enregistrement

dans la table. Un domaine de valeur peut être une valeur distincte ou une expression régulière ; c'est également l'entrée de la procédure. En sortie la procédure retourne des références sur les concepts valeurs. Cette procédure peut être une requête SQL ou un programme externe (script shell, C, ...). Cependant, si le concept attribut existe déjà nous avons juste à ajouter un enregistrement à la table et créer une nouvelle procédure.

4.3 Création des concepts valeurs

Lors de cette étape, toutes les procédures pour générer les concepts valeurs sont créées.

8. Donner un nom à chaque concept valeur.

9. Nettoyer homonymes et synonymes parmi les concepts valeurs.

4.4 Construction de l'ontologie

10. Identifier les relations de généralisation parmi les concepts valeurs (voir figure 5).

11. Créer les concepts abstraits et réorganiser l'ontologie avec ces nouveaux concepts. Par exemple le concept «Location» sur la figure 6.

12. Créer les relations entre concepts valeurs de hiérarchies différentes (voir relation 4 figure 6).

4.5 Construction de la base de données orientée pour la fouille (MODB)

13. Générer la base de données en utilisant les procédures définies lors de l'étape 7.

Lors de cette dernière étape un programme lit les tables créées pour chaque concept attribut et exécute chacune des procédures afin de générer la MODB.

5 La Base de Connaissances

5.1 Le serveur de connaissances Cyc

Le serveur de connaissances Cyc est constitué d'une base de connaissances et d'un moteur d'inférence développé par Cycorp[1]. Le but de Cyc est de pouvoir rassembler un grand nombre de connaissances de «sens commun» afin d'aider toutes les applications d'intelligence artificielle et de traitement du langage naturel. Le serveur Cyc est construit autour des composants suivants : une base de connaissances, un moteur d'inférences et le langage de représentation CycL.

Le serveur Cyc nous permet de construire et de gérer l'ontologie ainsi que la base de connaissance du CIS. Nous avons choisi la technologie Cyc après une étude de langages (OWL, DAML+OIL, Frame Logic, ...) existants et de leurs outils. Cyc repose sur deux notions fondamentales : les collections et les individus. Une collection est une type de chose, une classe de choses. Les choses qui appartiennent à une collection sont appellées ses instances. A l'opposé,

[1] http ://www.cyc.com/

un individu est une chose atomique. Ces deux notions correspondent à nos besoins car dans notre ontologie l'instance d'un concept peut avoir ses propres instances. Une autre notion clef pour gérer les connaissances avec CycL est la notion de Microthéorie. Une microthéorie est un ensemble d'assertions extraites de la base de connaissances. Un des principales fonctions des microthéories est de séparer les assertions en ensembles consistants dans lesquels il n'existe aucune contradiction. Toutefois il possible qu'il y ait des contradictions entre assertions de microthéories différentes. Par conséquence, nous pouvons utiliser les microthéories afin de gérer les contradictions entre experts et de prendre en compte les opinions divergeantes des utilisateurs. Pour finir, la base de connaissance de «sens commun» de Cyc est une fonctionnalité intéressante puisque nous travaillons actuellement sur des données «sociales» fournit par les caisse d'allocation familiales.

5.2 Définition des relations entre concepts en CycL

Dans la section 3.4 nous avons montré qu'il existe trois sortes de relations utiles à l'analyse des résultats de la fouille de données. Ces relations peuvent être définies dans le serveur de connaissances Cyc en utilisant les relations binaires prédéfinies (isa, $genls$) ou des relations nouvelles à définir comme $valeurDe$ et $relationAvec$ présentées ci-dessous. Pour chacune d'entre elles nous pouvons choisir définir propriétés parmi : la reflexivité, l'irréflexivité, la symétrie, l'anti-symétrie, l'assymétrie et la transitivité.

5.2.1 isa : Une relation pour décrire les propriétés des concepts

En CycL, le prédicat isa est utilisé pour exprimer qu'une chose est instance d'une collection. Une expression de la forme *(isa X Y)* signifie que X est une instance de la collection Y. Nous utilisons la relation isa afin de décrire les concepts des propriétés : leur rôle (attribut/valeur) et leur présence dans le MODB (abstrait/concret). Ainsi, nous avons créé tous les concepts de la figure 2 dans une microthéorie appellée DataMiningMt qui diffère de la microtheorie du domaine étudié.

5.2.2 genls : Relations entre concepts attributs et concepts valeurs de la même hiérarchie

En CycL, $genls$ est utilisé pour dire qu'une collection est inclue dans une autre. Une expression de la forme *(genls X Y)* signifie que chaque instance de la collection X est aussi une instance de la collection Y. Nous utilisons $genls$ afin de définir les relations entre concepts attributs (voir les relations 7,8,9,10 dans la table 1) et les concepts valeurs de la même hiérarchie (voir les relations 1,2 dans la table 1).

5.2.3 valeurDe : Relations entre concepts valeurs et concepts attributs

Nous avons défini la relation $valeurDe$, irreflexive et assymétrique, pour représenter les relations 3,5,6 (voir table 1 et 2). Afin de permettre l'héritage à travers les relations $genls$ nous

avons dû définir de nouvelles assertions pour le moteur d'inférences Cyc à l'aide de la relation $implies$

5.2.4 relationAvec : Relations entre concepts valeurs de différentes hiérarchies

Nous avons défini la relation $relationAvec$, qui est uniquement transitive, pour représenter la relation 4 de la table 2. Afin de permettre au moteur d'inférence de traiter tous les concepts liés successivement par cette relation il est également nécessaire de créer de nouvelles assertions.

5.3 Un moteur d'inférence pour améliorer les résultats de la fouille de données

Notre objectif principal en développant le CIS est de fournir un ensemble d'outils afin d'analyser les résultats d'une fouille de données à un niveau sémantique. Une première étape est de réécrire les règles générées afin de les simplifier selon les relations définies dans l'ontologie et d'offrir à l'utilisateur un outil intuitif pour explorer ces règles. Dans une seconde étape, l'utilisateur devra exprimer ses connaissances au sein d'une microthéorie et notre algorithme sélectionnara les règles les plus intéressantes *en fonction des connaissances de l'utilisateur*. Actuellement la première étape est en cours de developpement. Pour l'illustrer voici un exemple :

Considérons les règles définies en section 2 où «NUMSEMAINE=2» est un sur-ensemble de «NUMJOUR=10» et «NUMJOUR=11» :

Afin de simplifier l'écriture de nos règles nous utilisons la notation suivante : soit A l'item NUMSEMAINE=2, A1 l'item NUMJOUR=10, A2 l'item NUMJOUR=11, B l'item MOTIF="APPEL ENTRANT" et C l'item OBJECTIF="DEMANDE CREDIT".

Les règles 1 et 2 sont défines par les axiomes suivants :

```
(rule (TheList A1 B C))
(rule (TheList A2 B C))
```

Définissons l'assertion suivante :

```
(implies
       (and
           (rule (TheList ?X B C))
           (rule (TheList ?Y B C))
           (genls ?X ?Z)
           (genls ?Y ?Z))
       (rule (TheList ?Z B C)))
```

Avec cette requête *(rule ?X)* le système Cyc retourne le résultat suivant, qui est la règle 3 déduite des règles 1 et 2 :

```
?X : (TheList A B C)
```

6 Conclusion

Nous avons présenté une nouvelle méthodologie ExCIS qui permet l'intégration de la connaissance des experts d'un domaine dans le processus de fouille de données. L'objectif principal est d'améliorer la qualité de la connaissance extraite et de faciliter son interprétation. ExCIS est basé sur un système d'information conceptuel (CIS) qui stocke la connaissance des experts. Le CIS joue un rôle central dans la méthodologie car il est utilisé pour générer des jeux de données avant la fouille, pour filtrer et interpréter les modèles obtenus et pour mettre à jour la connaissance des experts. Ce papier est essentiellement consacré à la description de la structure du CIS et de sa construction et évoque la manière dont il peut être utilisé pour améliorer les résultats de la fouille de données. Nous avons montré les structures ontologiques du CIS, et décrit les choix effectués pour identifier les concepts et les relations de l'ontologie en analysant des données opérationnelles. Nos travaux futurs seront consacrés au développement des techniques exploitant les informations de l'ontologie afin d'interpréter les résultats de la fouille.

7 Remerciements

Nous désirons remercier la CNAF et plus spécialement Pierre Bourgeot, Cyril Broilliard, Jacques Faveeuw, Hugues Saniel et le BGPEO pour avoir soutenu ce travail.

Références

T. Gruber (2002). *What is an Ontology ?*. http ://www-ksl.stanford.edu/kst/what-is-an-ontology.htm.

R.J. Hilderman et H.J. Hamilton (2001). *Evaluation of Interestingness Measures for Ranking Discovered Knowledge*. Proceedings 5th PAKDD conference, Lecture Notes in Computer Science 2035 :247-259.

P. Johannesson (1994). *A Method for Transforming Relational Schemas into Conceptual Schemas*. Proceedings 10th ICDE conference, M. Rusinkiewicz editor, pp. 115-122, IEEE Press.

V. Kashyap (1999). *Design and Creation of Ontologies for Environmental Information Retrieval*. Proceedings 12th workshop on Knowledge Acquisition, Modelling and Management.

B. Liu, W. Hsu, L.-F. Mun et H.-Y. Lee (1999). *Finding Interesting Patterns using User Expectations*. Knowledge and Data Engineering, 11(6) :817-832.

D.L. Rubin, M. Hewett, D.E. Oliver, T.E. Klein et R.B. Altman (2002). *Automatic Data Acquisition into Ontologies from Pharmacogenetics Relational Data Sources using Declarative Object Definitions and XML*. Proceedings 7th Pacific Symposium on Biocomputing, pp. 88-99.

A. Silberschatz et A. Tuzhilin (1995). *On Subjective Measures of Interestingness in Knowledge Discovery*. Proceedings of the First International Conference on Knowledge Discovery and Data Mining, 275-281.

L. Stojanovic, N. Stojanovic et R. Volz (2002). *Migrating Data-intensive Web Sites into the Semantic Web*. Proceedings 17th ACM Symposium on Applied Computing, pp. 1100-1107, ACM Press.

G. Stumme (2000). *Conceptual On-Line Analytical Processing*. K. Tanaka, S. Ghandeharizadeh et Y. Kambayashi editors. Information Organization and Databases, chpt. 14, Kluwer Academic Publishers, pp 191-203.

Summary

In this paper, we present the new ontology-based methodology ExCIS (Extraction using a Conceptual Information System) for integrating expert prior knowledge in a data mining process. This methodology describes guidelines for a data mining process like CRISP-DM. Its originality is to build a specific Conceptual Information System related to the application domain in order to improve datasets preparation and results interpretation.

Un système de médiation basé sur les ontologies

Nora Maiz*, Omar Boussaid**, Fadila Bentayeb***
Laboratoire ERIC, Université Lumière Lyon2
5, avenue Pierre Mendés France
69676, Bron Cedex, France
*nmaiz@eric.univ-lyon2.fr
** boussaid@univ-lyon2.fr
***bentayeb@eric.univ-lyon2.fr

Résumé. L'intégration de données par médiation est une solution pour les entreprises ayant des sources de données réparties et hétérogènes. L'utilisation des ontologies dans la médiation permet de gérer le problème d'hétérogénéité structurelle et sémantique. Dans cet article, nous proposons une méthode ascendante de construction des ontologies qui consiste à construire l'ontologie globale pour le médiateur et cela à partir des ontologies locales correspondantes aux sources de données afin de remédier le problème de l'hétérogénéité sémantique. Pour surmonter le problème de l'hétérogénéité structurelle, nous proposons un langage de requêtes et un algorithme de réécriture de requêtes exprimées dans ce langage. Notre langage permet d'exprimer une requête en termes de l'ontologie globale et les ontologies locales. L'algorithme de réécriture, que nous proposons, permet d'obtenir une requête cohérente.

1 Introduction

Les grandes entreprises modernes sont dotées d'organisations qui utilisent différents systèmes pour stocker et rechercher les données. La concurrence, la distribution géographique et la croissance de l'inévitable décentralisation contribuent à cette diversité. Ces systèmes sont indépendants, conçus avec des modèles et des langages différents. La plupart d'entre eux n'ont pas été créés pour être interopérables. Mais les besoins de l'entreprise incitent ces systèmes hétérogènes à l'être. Afin de réaliser cette interopérabilité, il faut procéder à une intégration des sources de données. Cette intégration se fait de différentes manières : intégration sémantique et intégration structurelle. L'intégration structurelle reste au cœur du fonctionnement du système, et l'intégration sémantique se situe au niveau des descriptions des sources.

Un système d'intégration de données peut être caractérisé par son architecture et son modèle d'intégration. On distingue deux architectures fondamentales pour l'intégration de données : (1) l'approche entrepôt [25, 26] qui applique le principe des vues matérialisées et intègre les données en accord avec les schémas globaux, le résultat est une base de donnée centralisée. (2) l'approche médiateur [20, 27] qui effectue une intégration virtuelle des sources de données à l'aide d'une couche supplémentaire dite médiateur qui assure la communication entre les utilisateurs et les différentes sources. L'utilisation des ontologies dans le processus de médiation est une approche prometteuse. Elle permet de mettre en œuvre une intégration sémantique et son utilisation comme un modèle de requêtes permet une intégration structurelle[16, 24]. Le rôle du modèle d'intégration est de décrire le schéma global et son intégration avec les schémas locaux des sources. En fait, concernant l'architecture avec plusieurs ontologies, plusieurs modèles structurels peuvent y être appliqués : le modèle *GAV* (*Global As View*) [21], le modèle *LAV* (*Local As View*) et plus

récemment, le modèle *GLAV* est apparu [5, 11, 21].

Récemment, quelques méthodologies ont été proposées pour soutenir le développement des ontologies et les mettre en place dans un système d'intégration [3, 22, 24]. Cependant, ces méthodes sont très générales et ne prennent pas en compte le contexte d'intégration. Elles proposent une démarche de construction descendante, c'est à dire construire l'ontologie globale puis les ontologies locales. Cette démarche ne simplifie pas la résolution des problèmes d'hétérogénéité sémantique. En parallèle, le traitement de requêtes dans le cadre de l'architecture avec plusieurs ontologies modélisées selon *GLAV* est possible, uniquement si la requête est exprimée dans un langage de requêtes qui prend en charge les niveaux global et local, un point important n'est pas détaillé dans la majorité des systèmes existants.

Dans cet article, nous proposons un nouveau système de médiation selon GLAV en utilisant les ontologies. Dans ce cadre, nous proposons une méthode de construction des ontologies, nous définissons un langage de requêtes spécifique pour formaliser les requêtes et nous définissons une stratégie de décomposition de requêtes à base d'ontologies. Ce travail s'inscrit dans le cadre d'un projet d'Entreposage Virtuel de Données Bancaires en collaboration avec LCL -Crédit Lyonnais- (Direction d'Exploitation Rhône-Alpes Auvergne).

La suite de cet article est organisée comme suit. un état de l'art présente les systèmes de médiation utilisant les ontologies dans la section 2. La section 3 présente notre approche de création des différentes ontologies appliquée au cas des sources du Crédit Lyonnais. La section 4 présente notre langage de requêtes et notre algorithme de réécriture de requêtes. L'architecture et l'implémentation de notre médiateur seront exposées dans la section 5. Nous finirons cet article par la section 6 qui conclut sur les travaux réalisés et présente les perspectives sur de nouvelles problématiques engendrées.

2 Etat de l'art

Le problème central pour les médiateurs est l'hétérogénéité des sources de données. Celle-ci peut être classifiée en deux catégories [8, 15, 17] : structurelle et sémantique. Les conflits sémantiques se produisent lorsque deux contextes n'emploient pas la même interprétation d'informations. GOH [15] identifie trois causes principales pour l'hétérogénéité sémantique : les conflits de confusion, les conflits de graduation et les conflits de nom.

Le terme « ontologie » est employé dans plusieurs domaines et de manières différentes. Dans [13], les ontologies sont présentées par Gruber comme «spécifications explicites d'une conceptualisation », ça signifie que les concepts et les relations du modèle abstrait ont été exprimés par des noms et des définitions explicites [22]. Par conséquent, une ontologie pourrait être employée pour des tâches d'intégration en raison de son potentiel de description de la sémantique des sources d'informations [3, 8, 15]. Mais la méthode de conception des ontologies et la manière dont elles sont utilisées peuvent être différentes. [29] fournit une vue sur les méthodes existantes, par exemple METHONTODOLOGY [12] et le projet KRAFT [19] . Dans [24], trois directions différentes sont identifiées : (1) l'approche avec simple ontologie, (2) l'approche avec multiples ontologies et (3) l'approche hybride. La première approche utilise une ontologie globale qui fournit un vocabulaire partagé pour la spécification de la sémantique de toutes les sources de données, par exemple le système *SIMS* [1]. Dans la deuxième approche avec de multiples ontologies comme dans *OBSERVER* [18], chaque source est décrite par sa propre ontologie. L'avantage de cette approche est que l'ontologie n'a aucun besoin d'engagement commun et minimal envers une ontologie globale [9] mais la définition des correspondances entre les ontologies apparaît comme un processus délicat dans la pratique [2]. L'approche hybride consiste à décrire la sémantique de chaque source par sa propre ontologie qui est construite à partir d'un vocabulaire partagé global [4, 5, 6,10, 14, 23]. D'autres approches de développement d'ontologies

sous l'architecture hybride sont proposées dans [22, 3]. Ces approches sont descendantes, l'ontologie globale est créée en amont de la construction des ontologies locales. Dans [3], la même démarche est suivie. L'ontologie globale et les ontologies locales sont décrites avec un même langage. Cette méthode a trois étapes principales : la construction d'un vocabulaire partagé (une ontologie globale), des ontologies locales et la définition des correspondances entre les ontologies locales et l'ontologie globale.

Les langages de représentation des ontologies ont suscité un grand intérêt. Beaucoup de langages ont été proposés lors de la dernière décennie. Au début des 90, un ensemble de langages d'implémentation des ontologies est apparu. Principalement, ces langages sont basés sur la logique du premier ordre (*KIF*), sur les frames combinées avec la logique du premier ordre (Ontolingua, *OCM*L et *FLOGIC*) ou bien sur la logique de description (*Loom*) [7].

L'avènement d'Internet a nécessité la création des langages de représentation des ontologies qui exploitent les caractéristiques du web. *SHOE* [2] est construit comme une extension de *HTML*. Après l'apparition de *XML* qui est utilisé comme un langage standard pour l'échange de l'information sur le web, la syntaxe de *SHOE* a été adaptée à XML afin de l'utiliser comme un langage récent et riche de représentation des ontologies . *XOL* est développé comme une XMLisation d'un sous ensemble de primitives du protocole *OKBC* (*Open Knowledge Base Connectivity*). *RDF* [28] a été développé par *W3C* comme un langage basé sur la sémantique pour la description des sources web. *RDF* Schéma[1] (*RDF/S*) est une extension de *RDF* avec des primitives basées sur les frames. Ces langages ont établi la base du web sémantique. Dans ce contexte, trois autre langages ont été développés comme une extension de *RDF/S* : *OIL, DAML+OIL*, et finalement *OWL*[2] qui est mis en œuvre en 2001 par un groupe de travail nommé WebOnto formé par W3C.

3 Système de médiation basé sur les ontologies

Dans cette section, nous présentons notre approche qui va d'une part, traiter l'intégration sémantique en proposant une méthode de développement des ontologies ; et d'autre part, l'intégration structurelle en proposant un langage et un algorithme de réécriture de requêtes.

L'architecture hybride est la mieux adaptée pour un système de médiation basé sur plusieurs ontologies. Les approches *GAV, LAV* et plus récemment *GLAV* indiquent la façon de répondre à une requête. Dans GAV, on procède à un dépliement de la requête, dans *LAV*, on effectue une réécriture de requête [20] et dans *GLAV*, le traitement de la requête passe par une réécriture puis un dépliement.

3.1 Approche de développement des ontologies

L'approche que nous proposons consiste à créer les ontologies locales, et à extraire l'ontologie globale à partir des différents concepts utilisés dans celles-ci. Une phase de validation par un expert est requise. Notre approche est volontairement voulue être semi automatique. La tâche de l'expert consiste à classifier les concepts et à résoudre les conflits sémantiques (synonymes, homonymes, …). Notre approche contrairement à [3, 22] prend en charge le contexte de la connaissance (la sémantique des concepts). Elle a pour conséquence la précision et l'augmentation de la qualité de la connaissance et surtout la diminution de la complication de sa réutilisation. En bref, la construction de l'ontologie globale sera plus facile puisque la sémantique des termes est contenue dans leur contexte (ontologies locales). Notre méthode a la caractéristique d'être ascendante la chose qui facilitera la réconciliation sémantique entre les différents concepts, en plus

[1] http://www.w3.org/TR/rdf-schema/
[2] http:/www.w3.org/2001/swWebOnt/.

de ça l'intervention de l'expert pour la vérification de la classification des concepts et la résolution des conflits sémantiques permet de réaliser une intégration sémantique.

Trois étapes principales (figure1) constituent notre méthode : (1) construction des ontologies locales ; (2) construction de l'ontologie globale ; (3) définition des différentes correspondances entre les sources, les ontologies locales et l'ontologie globale.

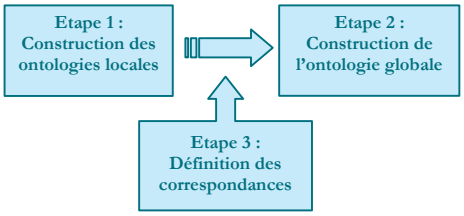


Figure 1: *Méthode de construction des ontologies*

3.2 Application aux données bancaires de LCL

Etape 1 : Construction des ontologies locales

Cette première étape contient deux phases principales : analyse des source de données et définition des ontologies locales. La première phase est une analyse complète de chaque source indépendamment des autres sources. Cette analyse consiste à rechercher les termes (ou primitives) utilisés dans la source. Par exemple quelle est l'information stockée, comment elle est stockée, quelle est la signification de cette donnée (sa sémantique).

Le LCL dispose de différentes sources de données hétérogènes. Nous ne nous intéressons qu'aux sources relationnelles dans le cadre de cet article. Le LCL dispose de deux sources relationnelles appartenant à deux départements indépendants. La première source contient les informations relatives aux collaborateurs ainsi qu'aux demandes marketing. Une demande marketing est la formulation d'une demande de ciblage de clients lors d'un événement particulier ou une opération spécifique de marketing. La seconde source de données simplifiée contient les informations sur les personnes et leur profil dans l'entreprise. Une partie du schéma des deux sources de données relationnelles est décrite ci-dessous. Chacune est composée de deux relations :
*Collaborateur (**IdCollaborateur**, nomCollaborateur)*

*DemandeMarketing (**IdDemande**, LibDemande, IdCollaborateur)*
*Personne (**IdPersonne**, NomPersonne)*
*Profil (**IdProfil**, LibProfil, TypeProfil, IdPersonne)*

Les relations sont représentées par des classes *OWL* (Ontology Web Language)(figure 2).

Tables	OWL équivalent
Collaborateur	<owl:Class rdf:ID="Collaborateur">
DemandeMarketing	<owl:Class rdf:ID="DemandeMarketing">
Personne	<owl:Class rdf:ID="Personne">
Profil	<owl:Class rdf:ID="Profil">

Figure 2 : *Représentation des tables relationnelles en OWL*

Les relations entre classes sont représentées dans *OWL* par *owl:ObjectProperty* et *owl:DatatypeProperty*. Les propriétés *OWL* peuvent représenter les différents attributs et contraintes dans le schéma. Les propriétés *OWL* représentent les attributs par des Datatype. Si l'attribut est contraint d'être une clé primaire alors la caractéristique fonctionnelle sera ajoutée.

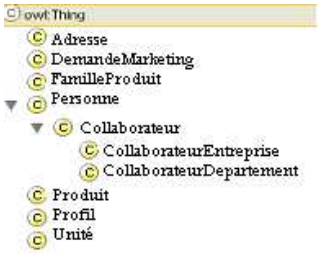
D'autre part, on utilise *owl:ObjectProperty* pour représenter les attributs qui sont sous une contrainte de clé étrangère. Nous obtenons ainsi deux ontologies représentant les deux bases de données relationnelles comme sources de données.

Etape 2 : Construction de l'ontologie globale

La construction du vocabulaire partagé contient deux étapes principales : (1) analyse des ontologies locales ; (2) sélection de tous les concepts et résolution des problèmes d'hétérogénéité sémantique. La première étape implique une analyse complète des ontologies déjà construites indépendamment. Après sélection des concepts, il faut localiser les problèmes d'hétérogénéité sémantique. Nous avons trois causes principales pour l'hétérogénéité sémantique : les conflits de confusion, les conflits de graduation, les conflits de nom [15]. Ces derniers sont les plus probables dans cette étape., deux solutions sont proposées :

- Cas d'homonymes : on trouve deux concepts identiques mais sémantiquement différents. Nous suggérons que les deux concepts doivent être renommés au niveau de l'ontologie globale. Par exemple : on peut trouver le concept « *Collaborateur* » qui représente les personnes travaillant dans l'entreprise et un autre concept « *Collaborateur* » qui représente les personnes affectées au département des ressources humaines. Dans ce cas, il faut différencier les deux concepts en les renommant en « *CollaborateurEntreprise* » et « *CollaborateurDepartement* ». Ces deux concepts peuvent être généralisés par le concept «*Collaborateur* » qui aura comme sous classes ces deux classes (figure 3).

- Cas de synonymes : on trouve deux concepts différents mais sémantiquement identiques. Il faudra dans l'ontologie globale exprimer une équivalence entre ces deux concepts. Par exemple « *Collaborateur* » dans une source a le même sens que « *Responsable* » dans une autre source.



.

Figure 3 : *Ontologie globale*

Etape 3 : Correspondances entre l'ontologie globale et les ontologies locales

L'ontologie globale est construite à partir des ontologies locales. Pour identifier la source ontologique originelle des concepts dans l'ontologie globale, nous utilisons des annotations. Nous avons vu que le langage *OWL* permet d'annoter des concepts et des propriétés selon un schéma de méta-données prédéfini. Nous annotons les concepts de l'ontologie globale avec un schéma d'annotation qui contient trois propriétés, Ontologie locale originelle du concept à annoter (*Uri_ Local_Ontology*), L'agent responsable de cette ontologie locale (*Agent_Local_Ontology*), Le nom du concept dans l'ontologie locale (*Concept_Local_Ontology*).
La figure 4 montre le schéma *RDF/S* pour l'annotation des concepts de l'ontologie globale, Une fois que les ontologies locales et globale sont construites, nous utilisons leurs concepts pour formuler des requêtes et les exécuter sur les sources de données. Pour réaliser ce travail, nous proposons un langage de requêtes et un algorithme de réécriture des requêtes.

```
<rdf:RDF>
 <rdfs:Class rdf:about="#Local_Ontology" rdfs:label="Local_Ontology">
  <rdfs:subClassOf rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
 </rdfs:Class>
<rdf:Property rdf:about="#URI_Local_Ontology" rdfs:label="URI_Local_Ontology">
  <rdfs:domain rdf:resource="#Local_Ontology"/>
  <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
 </rdf:Property>
<rdf:Property rdf:about=" #Agent_Local_Ontology" rdfs:label="Agent_Local_Ontology">
  <rdfs:domain rdf:resource="#Local_Ontology"/>
  <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
</rdf:Property>
<rdf:Property rdf:about=" #Concept_Local_Ontology" rdfs:label="Concept_Local_Ontology">
  <rdfs:domain rdf:resource="#Local_Ontology"/>
  <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
 </rdf:Property>
</rdf:RDF>
```

Figure 4 : Schéma RDF/S des annotations

4 Réécriture de requêtes

le système proposé se base sur les ontologies qui servent à décrire les sources de données, c'est à dire effectuer une intégration sémantique des différentes sources. Pour que le système fonctionne, autrement dit accéder aux données de sources et les exploiter, nous proposons un langage de requêtes qui utilise les termes (concepts) de l'ontologie et un algorithme de réécriture de requêtes qui assure la cohérence des requêtes après leur décomposition en sous requêtes et la combinaison des résultats provenant des différentes sources. Ce langage et cet algorithme de réécriture réalisent l'intégration structurelle.

Il existe plusieurs manières d'établir la correspondance entre le schéma global et les schémas des sources de données à intégrer. L'approche *GAV* fait correspondre à chaque concept $Concept_G$ de l'ontologie globale une vue *VueL* sur l'ontologie locale. Une requête exprimée en termes de l'ontologie globale peut être reformulée en vue. Une vue est une conjonction de concepts et de propriétés. On procède à un dépliement de requêtes. L'approche *LAV* fait correspondre chaque concept $Concept_L$ de l'ontologie locale une vue Vue_G sur l'ontologie globale. Une requête exprimée en terme de l'ontologie globale peut être réécrite en vue. On procède à une réécriture de requêtes. L'approche *GLAV* fait correspondre à chaque concept $Concept_G$ ou Vue_G de l'ontologie globale un concept $Concept_L$ ou une vue Vue_L sur l'ontologie locale. Une requête exprimée en terme de l'ontologie globale ne peut être toujours reformulée en vue sur l'ontologie locale [5, 11, 21]. Par exemple :

$$Q \rightarrow (Collaborateur\ (x) \wedge aUneAdresse\ (x,y) \wedge Adresse(y)) \wedge (Adresse\ (z) \wedge aPourVille\ (z, «Lyon»)) \wedge (Profil\ (p))$$

est une requête qui concerne tous les collaborateurs ayant une adresse à *Lyon* et un certain profil. Le médiateur la décompose en trois sous requêtes. Les deux premières sont envoyées à l'adaptateur pour leur exécution. Le concept *Adresse* leur sert de liant, elles peuvent donc être recomposées par une jointure. Cependant, la troisième sous requête n'a pas de lien avec les deux sous requêtes précédentes. Le médiateur doit trouver un lien entre le concept *Profil* et les concepts des sous requêtes précédentes, sinon il exclut ce concept. Dans cet exemple, le médiateur doit trouver un lien entre « *Collaborateur* » et « *Adresse* », qui est une propriété regroupant directement ces concepts au concept *Profil* ou bien une propriété qui lie le concept *Profil* à un autre concept équivalent, subsumé ou subsumant l'un des deux concepts précédents *Collaborateur ou Adresse*. Dans notre cas, le concept *Personne* est le concept subsumant *Collaborateur,* et il a un lien avec *Profil*. Le médiateur doit donc réécrire la troisième sous requête « *Profil (p)* » en « *Personne (r)* $\wedge$ *aUnProfil (r,p)* $\wedge$ *Profil(p)* ». Il doit ajouter dans sa table de correspondances que « *Collaborateur* » de la sous requête1 correspond à « *Personne* » de la sous requête 3.

Notre système de médiation permet de poser des requêtes en utilisant l'ontologie globale et éventuellement les ontologies locales. L'ontologie globale est un ensemble de concepts classifiés en une hiérarchie. Cette hiérarchie ne contient pas de propriété reliant les différents concepts. Ces propriétés existent chacune dans leur ontologie locale. L'utilisateur peut donc créer une requête qui contient des concepts de l'ontologie globale et éventuellement des propriétés des ontologies locales. La requête exprimée en termes de concepts et propriétés doit être réécrite de manière à obtenir des résultats qu'on peut agréger. Tout concept qui ne garantit pas la combinaison des données obtenues est exclu. Du point de vue sémantique, cette exclusion tend à rendre une requête cohérente. Une requête cohérente est une requête décomposable en sous requêtes exécutables et dont les résultats sont composables. Cette réécriture peut être vue comme une correspondance entre l'ontologie globale et les ontologies locales. Une requête utilisateur de base peut être formulée selon le modèle :

$\wedge_{i=1\ldots n,\ j=1\ldots m}$ ($Concept_i \wedge Propriété_j$) ou tout simplement : **Concept.**

C'est le langage de requêtes que nous préconisons. L'utilisateur peut formuler plusieurs requêtes de base, ou combinaison de requêtes.

4.1 Exemple

Dans ce qui suit, nous décrivons des requêtes exprimées dans le langage basé sur le modèle présenté ci-dessus. Les exemples sont donnés dans le contexte des sources relationnelles du Crédit Lyonnais que nous avons décrit précédemment. La première source est décrite par l'ontologie « Ontologie marketing ». La seconde source est décrite par l'ontologie *« Ontology habilitation »*. Considérons maintenant un médiateur portant sur le domaine bancaire dont l'ontologie globale est décrite par la figure 3. Supposons que nous voulons poser la requête suivante : « on veut tous les collaborateurs ayant émis une demande marketing ». A partir de l'ontologie globale, l'utilisateur choisit le concept « collaborateur » et le caractérise par la propriété *aEmitUneDemande*, extraite de l'ontologie locale qui est connue à travers l'annotation faite sur ce concept. On aura la requête Q suivante :

$Q \rightarrow Collaborateur(x) \wedge aEmitUneDemande(x, y) \wedge DemandeMarketing(y).$

Le médiateur connaît l'ontologie qui prend en charge cette requête. Cette ontologie a déjà été sollicitée pour spécifier le concept. L'annotation sur ce dernier permet au médiateur d'envoyer cette requête à l'adaptateur responsable de cette ontologie. Ce dernier identifie les concepts de la requête et la propriété, puis calcule le script adéquat à exécuter pour obtenir les résultats. La requête à exécuter est obtenue à l'aide des annotations faites sur les concepts dans l'ontologie locale. Par exemple, pour la requête *Q* l'adaptateur obtient la requête SQL suivante :

```
Select *
    From Collaborateur, DemandeMarketing
    Where Collaborateur.id_c = DemandeMarketing .id_c
```

L'adaptateur renvoie ensuite le résultat de cette requête au médiateur.

Un autre exemple de requête, dans le même langage, porte sur tous les collaborateurs ayant une adresse à « *Lyon* ». A partir de l'ontologie globale, l'utilisateur choisit le concept « *Collaborateur* » et le caractérise par la propriété *aUneAdresse*. Puis l'utilisateur caractérise le concept « *Adresse* » par la propriété *aPourVille* en attribuant la valeur « *Lyon* ». On aura donc la requête suivante : $Q \rightarrow (Collaborateur\ (x) \wedge aUneAdresse\ (x,y) \wedge Adresse(y)) \wedge (Adresse\ (y) \wedge aPourVille\ (y, «Lyon»)).$

Le médiateur dispose alors de deux sous requêtes qui sont respectivement :

$Collaborateur\ (x) \wedge aUneAdresse\ (x,y) \wedge Adresse(y)$ et $Adresse\ (y) \wedge aPourVille\ (y, «Lyon»).$

4.2 Cadre formel

Formellement, un système *O* de médiation basé sur les ontologies est un triplet *(G, S, M(G,s))* où G est l'ontologie globale, *S* est un ensemble d'ontologies locales et *M(G,s)* est la correspondance entre l'ontologie globale *G* et les ontologies locales *S* dans le système *O*.

- Ontologie globale. Soit *Cg* l'ensemble de termes de l'ontologie globale. Supposons que l'ontologie globale *G* est exprimée dans un langage de logique descriptive *Lg*. *Cg* est une hiérarchie de concepts Cg = {cg | cg est un terme de l'ontologie}. An_g l'ensemble des annotations. Une annotation *An* est une connaissance sur un terme. ***An : Cg*** ⟶ ***An_g, An = Annotation(cg)***

- Ontologie locale. Soit S un ensemble de n ontologies locales *S*1, *S*2,…, *S*n. Nous notons A_{Si} l'ensemble des termes d'une ontologie locale. $A_s = \bigcup_{i=1...n} A_{Si}$ des n ontologies. Nous supposons que ces ensembles de termes A_{Si} sont mutuellement disjoints et que chaque ontologie locale est exprimée dans un langage de logique descriptive L_s et chaque A_{Si} est une hiérarchie de concepts liés éventuellement entre eux par des rôles. Nous notons R_{Si} l'ensemble des rôles définis dans A_{Si} x A_{Si}, $R_s = \bigcup_{i=1...n} R_{Si}$. Soit An_s l'ensemble des annotations.

- Les correspondances. La correspondance est le cœur du système. La correspondance M(G,s) spécifie comment les concepts dans l'ontologie globale G et les concepts de l'ontologie locale Si peuvent être liés. M(G,s) est une fonction de Cg ⟶ Si.

- Le langage de requêtes *Qg*. Composé des concepts Cg de l'ontologie globale G et des rôles (propriétés) des ontologies locales, par exemple, ***Collaborateur(x)*** (requête Q_{LAV}) ou bien ***Collaborateur (x)*** ∧ ***aUneAdresse (x, y)*** ∧ ***Adresse (y)*** (requête Q_{GAV})**.**

Une requête de l'utilisateur est une conjonction de plusieurs sous requêtes ***Q*** **=** $(\wedge_{i=1..n}$ ***Qi*** **).** Une sous requête est une conjonction de requête Q_{GAV} et/ou Q_{LAV} : ***Qi*** **:** $(\wedge_{i=1..n}$ ***Q_{GAV}-i)*** ∨ $(\wedge_{j=1..m}$ ***Q_{LAV}-j)***

- Réécriture de requêtes. L'idée générale est que le médiateur doit avoir une conjonction de sous requêtes Q_{GAV} et une table de correspondances entre les différentes sous requêtes. Si la requête contient une ou plusieurs sous requêtes Q_{LAV} alors il doit reformuler tous les Q_{LAV} en sous requêtes Q_{GAV} de façon à permettre de construire la table de correspondances. L'algorithme est le suivant :

Fonction CorrespondreSousRequête(Q_i , Q_j)	**Fonction CorrespondreConcept(Ci , Cj)**
Entrée **:** Qi - Sous requête Qj - - Sous requête G, S, R *Sortie* : (Ci , Cj) – la correspondance entre Qi , Qj Pour chaque $C_k \in Q_i$ faire Pour chaque $C_h \in Q_i$ faire Si $C_k = C_h$ alors retourner (C_k , C_h) Sinon Si CorrespondreConcept(C_k , C_h) alors retourner (C_k , C_h) Sinon retourner Ø sFinSi FinSi FinPour fin pour	Entrée : C_i - Sous requête Q_{GAV} de l'utilisateur C_j - - Sous requête Q_{GAV} de l'utilisateur G, S, R Sortie : r – le rôle qui lie les deux concepts C_i , C_j Si r(C_i , C_j) ∈ R ou r(C_j , C_i) ∈ R alors retourner r Sinon retourner Ø FinSi

Algorithme 1 : Réécriture (Q)

Entrée : Q - La requête de l'utilisateur : $Q=(\Lambda\ Q_{i\ i=1...n})$
$G = \{C_{g1}, C_{g2}, C_{g3},..., C_{gn}\}$, $S = \{ C_{s1}, C_{s2}, C_{s3},..., C_{sn}\}$, $R = \{r_{s1}, r_{s2}, r_{s3},..., r_{sm}\}$

Résultat : Q_d [k] - l'ensemble de k requêtes déduites de Q
T [k] - l'ensemble des tables de correspondances pour les k requêtes déduites

1. k = 1;
2. Q_d [k]= Q ;
3. pasDeCorrespondant : = Vrai;
4. Pour chaque u = 1..NombreCase(Q_d) faire
5. Pour chaque $Q_{i\ (i=1...n)} \in Q_d$ [u] faire
6. Si Q_i est Q_{LAV} alors
7. Début
8. Pour chaque $Q_{j\ (j=i-1..1)}$ faire
9. -Obtenir l'ensemble Ψ par les concepts Subsumés, subsumants ou équivalents à Q_i en utilisant le
10. raisonnement
11. -Obtenir l'ensemble Ω par les concepts C_s de Ψ tel que Annotation(C_s)=Annotation(C_j) et le C_j est
12. le concept de Q_j
13. Pour chaque $C_j \in Q_j$ faire
14. Pour chaque $C_h \in \Omega$ faire
15. Si CorrespondreConcept(C_j ,C_h) ≠ Ø alors
16. Début
17. aUnCorrespondant := faux;
18. k++;
19. $Q_i \leftarrow C_i \cup r_h \cup C_h$ // r_h est le rôle qui lie C_i a C_h
20. Q_d [k] ← Q_d [k] ∪ Q_i
21. T[k] ← T[k-1]
22. Ajouter la correspondance (C_j ,C_h) dans T[k];
23. Fin
24. FinSi
25. FinPour
26. FinPour
27. FinPour
28. Si pasDeCorrespondant alors Q ← Q - Q_i FinSi
29. Fin
30. Sinon // Q_i est Q_{GAV}
31. Pour chaque $Q_{j\ (j=1..i-1)}$ faire
32. Si CorrespondreSousRequête(Q_i ,Q_j) ≠ Ø alors Ajouter la correspondance (C_i ,C_j)dans
33. T[k];
34. Sinon Q ← Q - Q_i
35. FinSi
36. FinPour
37. FinSi
38. FinPour
3 FinPour
9. Retourner (Q_d , T)

Le but est d'éliminer les requêtes Q_{LAV} qui ne correspondent pas avec les autres sous requêtes. A la requête en entrée, l'algorithme de réécriture trouve plusieurs requêtes sémantiquement équivalentes, en se basant sur les mécanismes de raisonnement de *OWL*.

Nous explicitons dans ce qui suit l'algorithme que nous proposons pour la réécriture des requêtes. Nous utilisons la notation Q - Q_i pour exprimer le fait de supprimer le sous ensemble de sous requêtes Q_i de la requête Q. La fonction *CorrespondreConcept* trouve un rôle (s'il existe) reliant deux concepts donnés en entrée. La fonction *CorrespondreSousRequête* retourne le couple de concepts qui se correspondent, c'est-à-dire les concepts qui sont égaux ou qui sont liés par un rôle.

L'algorithme présenté nécessite en entrée les concepts présents dans l'ontologie globale ainsi que les concepts et rôles présents dans les ontologies locales. En sortie, il produit un ensemble de requêtes sémantiquement équivalentes. Pour chaque sous requête $\boldsymbol{Q_i}$ de la requête Q (ligne L5) :

- Si *Qi* est un Q_{LAV}, c'est-à-dire constitué d'un seul concept. Pour le $\boldsymbol{C}_{\mathbf{i}}$ appartenant à la sous requête *Qi*, l'algorithme sélectionne à partir des concepts des sous requêtes précédentes $Q_{\mathbf{j}}$, les concepts ayant les mêmes annotations, c'est-à-dire appartenant à la même ontologie, (L 8-13). On a un ensemble $\boldsymbol{\Omega}$ de concepts candidats.
- L'algorithme vérifie s'il y a une correspondance entre les concepts candidats et le concept $\boldsymbol{C}_{\mathbf{i}}$ de la sous requête à réécrire. A chaque concept correspondant, l'algorithme crée une nouvelle requête réécrite avec ce concept (L 14-27).
- Dans le cas où il n'y a pas de concept correspondant, le concept traité Ci sera exclu (L 29)
- Si *Qi* est un Q_{GAV}, c'est-à-dire constitué de deux concepts et un rôle. L'algorithme cherche une correspondance entre cette sous requête et les sous requêtes précédentes. S'il n'y a pas de correspondance, la sous requête est exclue (L 31-37).
- L'algorithme fait ce traitement pour chaque sous requête de la requête initiale (L5), puis traite les nouvelles requêtes réécrites comme il a traité la requête initiale (L 6).

5 Implémentation du médiateur

Pour valider notre approche, nous proposons un environnement *INMEA* (Intégration par Médiation et Entreposage basé sur les Agents) qui implémente notre architecture de médiation. Cet environnement se distingue des autres environnements d'intégration par médiation existants par le fait qu'il permet d'exprimer les descriptions des sources par la récente recommandation W3C pour la description des ontologies *OWL*. *OWL* est déjà utilisé pour l'intégration des ressources Web dans le cadre du Web sémantique. Il offre des possibilités très intéressantes de descriptions et de raisonnements. Notre objectif est aussi de combiner la puissance d'expression et de description du langage *OWL* avec l'aspect communiquant et coopératif des Systèmes Multi Agents (*SMA*). Le médiateur est un agent qui communique avec les autres agents et dispose de l'ontologie globale. Les autres agents sont les agents sources ou adaptateurs qui sont responsables des ontologies locales. Le processus de création ou de réécriture de requête dans *INMEA* se fait par un dialogue entre l'agent médiateur et les autres agents.

Pour le développement de l'environnement *INMEA*, nous avons utilisé un certain nombre d'outils : l'éditeur d'ontologie Protégé2000[3], la plateforme *JADE*[4] pour la plateforme d'agents, la plateforme *Jena*[5] pour la manipulation des ontologies *OWL*. *Jena* est un projet de code source libre développé chez *HP* pour le Web sémantique. Cette plateforme nous offre beaucoup d'avantages, il nous permet d'avoir un accès uniforme à diverses ontologies car toutes les informations sont stockées dans un modèle Jena. Pour le raisonnement sur les ontologies *OWL*, nous utilisons le raisonneur libre *Pellet*[6] qui offre la possibilité de raisonner sur la partie terminologique. L'interface de requêtes se présente comme une application Web Java basée sur la plateforme *Struts*[7], qui implémente l'architecture « Modèle Vue Contrôleur ».

6 Conclusion et perspectives

Dans cet article, nous avons proposé une approche de médiation basée sur les ontologies dans le contexte de l'intégration des sources de données hétérogènes. Cette approche repose sur l'architecture hybride, utilisant une ontologie globale pour le médiateur et des ontologies locales au niveau des sources. Il s'agit de créer l'ontologie globale en aval, et ce, à partir des ontologies

[3] http://protégé.standford.org
[4] http://jade.tilab.com
[5] http://jena.sourceforge.net
[6] http://www.mindswap.org/2003/pellet
[7] http://jakarta.apache.org/struts

locales, ce qui facilite et améliore la résolution de l'hétérogénéité sémantique entre les ontologies de source. Nous avons défini une démarche de construction des ontologies, un langage qui garantit le traitement correct de la requête, en permettant l'expression de requêtes en termes des ontologies globale et locales. Nous avons également proposé un algorithme de réécriture de requêtes qui permet d'assurer la cohérence d'une requête, en garantissant une réponse globale. Nous avons appliqué notre approche de création des ontologies sur les sources relationnelles du Crédit Lyonnais. Ces ontologies sont utilisées dans notre système de médiation *INMEA*, et ont servi dans la phase de création et de réécriture de requêtes. Le langage de requêtes et l'algorithme de réécriture proposés ont été validés par une implémentation au cœur du système *INMEA*.

Différentes perspectives sont envisagées. En premier lieu, les adaptateurs qui ont été considérés comme une boîte noire dans cet article doivent être plus optimisés. Dans notre projet, ce module est fonctionnel et traite toutes les requêtes possibles. Cependant le langage de requête généré n'est pas optimisé (dans notre cas c'est le SQL). Cette partie doit faire l'objet, en plus d'une optimisation, d'une étude plus profonde pour prendre en charge des sources de données non relationnelles. De plus, le langage de requêtes doit être plus expressif, en permettant l'expression d'opérateurs plus complexes que l'égalité ou l'équivalence (comme l'inégalité, l'approximation, les comparaisons). Il faudra revoir également l'algorithme de réécriture en tenant compte du raisonnement avec des informations incomplètes. D'autre part, la génération d'un contexte d'analyse doit être considérée. Cette deuxième partie du projet d'Entreposage Virtuel de Données Bancaires pose de nombreux problèmes : le choix du modèle qui contient les données rapatriées, le méta modèle à instancier pour avoir le cube de données, la matérialisation du cube de données, la spécification des règles de calcul à effectuer au sein des cubes de données, et enfin, la capitalisation de la connaissance obtenue et la rendre opérationnelle dans le processus d'entreposage virtuel.

Références

[1] Y. Arens, Chun-Nan Hsu, C. A. Knoblock. (1996) : *Query processing in the SIMS information mediator*. In Advanced Planning Technology. AAAI Press, California, USA.

[2] S. Luke, J. Heflin (2000) *, SHOE 1.01. Proposed Specification, SHOE project technical report*, University of Maryland, available from : http://www.cs.umd.edu/projects/plus/SHOE/spec1.01.htm>

[3] A. Buccella, Cechich A. and Brisaboa N.R. (2003) : *An Ontology Approach to Data Integration*. Journal of Computer Science and Technology. Vol.3 (2). Available at http://journal.info.unlp.edu.ar/default.html, , (pp. 62-68).

[4] D. Calvanese, G. De Giacomo, M. Lenzerini. (2001) *Description logics for information integration. In Computational Logic: From Logic Programming into the Future (In honour of Bob Kowalski*), Lecture Notes in Computer Science. Springer-Verlag,.

[5] Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini (2001) : *A Framework for Ontology Integration*. SWWS 2001: 303-316.

[6] Vinay K. Chaudhri, Adam Farquhar, Richard Fikes, Peter D. Karp, and James P. Rice.(1998) : *Open knowledge base connectivity (okbc) specification document 2.0.3*. Technical report, SRI International and Stanford University (KSL), April.

[7] O.Corcho, M.Fernandez et A.Gomez-Perez (2003) : *Methodologies, Tools, and languages for building ontologies. Where is their meting point?*. D& k engineering, volume46 issue1. Elsevier Science PublishersB. V.

[8] Z.Cui, P. O'Brien.(2000) : *Domain Ontology Management Environment*. In Proceedings of the 33rd Hawaii International Conference on System Sciences –,Processing Society of Japan, Tokyo, Japan, 1994.

[9] Cécile Favre, Fadila Bentayeb, Omar Boussaid et Nicolas Nicoloyannis.(2005) : *Entreposage virtuel de demande marketing: de l'acquisition des objets complexes à la capitalisation des connaissances*, FDC (EGC'05) Paris.

[10] Mark S. Fox , Michael Gruninger. (1998) : *Enterprise modelling*, fall 1998, pp. 109-121.AI Magazine, 19(3):109–121.

[11] Marc Friedman, Alon Levy, and Todd Millstein.(1999) : *Navigational plans for data integration*. In Proc. of the 16th Nat. Conf. on Artificial Intellig (AAAI'99), pages 67–73. AAAI Press/The MIT Press.

[12] Ascuncion Gomez-Perez, M. Fernandez, A. de Vicente. (1996) : *Towards a method to conceptualize domain ontologies*. In Workshop on OntologicalEngineering, ECAI '96, pages 41–52, Budapest, Hungary.

[13] T. Gruber. (1995) *Toward principles for the design of ontologies used for knowledge sharing.*

[14] Chung Hee Hwang. (1999) :*Incompletely and imprecisely speaking: Using dynamic ontologies for representing and retrieving information.* Technical, Microelectronics and Computer Technology Corporation (MCC),

[15] Cheng Hian Goh. (1997) : *Representing and Reasoning about Semantic Conflicts in Heterogeneous Information Sources*. Phd, MIT.

[16] V. Kashyap, A. Sheth. (1996) : *Schematic and semantic similarities between database objects: A context-based approach*. The International Journal on Very Large Data Bases, 5(4):276–304.

[17] W.Kim, Jungyun Seo.(1991) : *Classifying schematic and data heterogeneity in multidatabase systems*. IEEE Computer, 24(12):12–18.

[18] E. Mena, V. Kashyap, A. Sheth, and A. Illarramendi. (1996) : *Observer: An approach for query processing in global information systems based on interoperability between pre-existing ontologies*. In Proceedings 1st IFCIS International Conference on Cooperative Information Systems (CoopIS '96). Brussels.

[19] J-C.R. Pazzaglia , S.M. Embury. (1998) : *Bottom-up integration of ontologies in a database context*. In KRDB'98 Workshop on Innovative Application Programming and Query Interfaces, Seattle, WA, USA.

[20]M.-C. Rousset, A. Bidault, C. Froidevaux, H. Gagliardi, F. Goasdoué, C. Reynaud, B. Safar. (2002) *Construction de médiateurs pour intégrer des sources d'information multiples et hétérogènes : le projet PICSEL*, in Revue I3 (Information – Interaction – Intelligence), Vol.2, N°1, p. 9-59.

[21] Marco Ruzzi, (2004) : *Data Integration : state of the art, new issues and research plan.*

[22] H. Stuckenschmidt. (2003) : *Ontology-Based Information Sharing in Weakly-Structure Environments*. Ph.D. Thesis, Faculty of Sciences, Vrije Universiteit Amsterdam, January.

[23] M. Uschold, M. Gruniger. (1996) : *Ontologies: Principles, methods and applications*. KnowledgeEngineering Review, 11(2):93–155.

[24] H.Wache, T. Vogele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, S. Hubner. (2001) : *Ontology-Based Integration – A Survey of existing Approaches*, in proceeding of IJCAI-01 Workshop: Ontologies and information Sharing, Seattle, WA, pp 108-117.

[25] R. Kimball, L. Reeves. W. Thornthwaite, M. Ross, W. Thornwaite (1998): *The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing and Deploying Data Warehouses*. New York, NY, USA.

[26] W.H. Inmon (2002) : *Building the Data Warehouse*,3rd Edition, 3rd edition. New York, NY, USA

[27] G.Wiederhold. (1995) : *Mediation in information systems*. ACM Computing Surveys, 27(2):265–267,June.

[28] O. Lassila, R. Swick (1999) : *Ressource description framework (RDF) model and syntax specification, W3C Recommandation*, <http://www.w3.org/TR/REC-rdf-syntax/>.

[29] D. M. Jones, T.J.M. Bench-Capon, and P.R.S Visser. (1998) : *Methodologies for ontology development*. In Proc. IT&KNOWS Conference of the 15th IFIP WorldComputer Congress, Budapest, Chapman-Hall.

Abstract. Data integration by mediation is a solution for the companies having distributed and heterogeneous data sources. The use of ontologies in the mediation allows the management of structural and semantic heterogeneity problem. In this paper, we propose an ascending method for ontologies construction which consists in building global ontology starting from local ontologies. In addition, to cure the structural heterogeneity problem we propose a query language and a query rewriting algorithm expressed in this language to express and treat correctly the queries.

Modèle d'entrepôt de données à base de règles

Cécile Favre, Fadila Bentayeb, Omar Boussaïd

Laboratoire ERIC, Université Lumière Lyon2
5, avenue Pierre Mendès-France
69676, Bron Cedex, France
{cfavre, bentayeb}@eric.univ-lyon2.fr
boussaid@univ-lyon2.fr
http://eric.univ-lyon2.fr

Résumé. Les entrepôts de données constituent une réponse aux besoins d'analyse des entreprises. Mais, cette solution est lourde à mettre en œuvre et à maintenir, en particulier au niveau de l'évolution des besoins d'analyse. Pour répondre à ce problème, nous proposons un nouveau modèle d'entrepôt de données à base de règles *R-DW*. Les règles permettent d'intégrer les connaissances de l'utilisateur dans l'entrepôt. Ce modèle est composé de deux parties : une partie fixe, définie en extension, comprenant une table des faits et les dimensions de premier niveau ; une partie évolutive, définie en intension par des règles. Grâce à ces règles, notre modèle *R-DW* permet de créer des hiérarchies de dimension de façon dynamique, rendant ainsi possible l'évolution des contextes d'analyse et renforçant l'interaction entre l'utilisateur et le système d'aide à la décision.

1 Introduction

Pour gérer une masse de données de plus en plus conséquente, provenant de sources hétérogènes, la mise en place d'un processus décisionnel est devenue nécessaire. Le stockage et la centralisation de ces données dans un entrepôt constituent un support efficace pour l'analyse. Un entrepôt de données est défini par une collection de données orientées sujet, intégrées, non volatiles et historisées, organisées pour le support d'aide à la décision (Inmon, 1996). À partir d'un entrepôt de données, des contextes d'analyse multidimensionnels ciblés, appelés communément cubes de données, sont construits. Ces cubes répondent à des besoins d'analyse prédéfinis en amont, l'objectif étant d'observer des faits, à travers une ou plusieurs mesures, en fonction de différentes dimensions. Il s'agit par exemple d'observer les niveaux de ventes en fonction des produits, des magasins et de la période d'achat. L'analyse en ligne OLAP (*On Line Analytical Processing*) est un outil basé sur la visualisation permettant aux décideurs d'une entreprise la navigation et l'exploration de ces cubes de données.

Mais si l'entreposage de données a connu un essor important grâce aux possibilités qu'il offre, sa mise en place et sa maintenance ne sont pas des tâches faciles. D'une part, la mise en œuvre d'un entrepôt de données nécessite un important travail préalable à la fois d'étude de l'existant et des besoins d'analyse, mais aussi de modélisation. D'autre part, la maintenance de l'entrepôt nécessite, non seulement un rafraîchissement périodique des données, mais également une évolution du schéma pour répondre à de nouveaux besoins d'analyse.

Les modèles multidimensionnels (Cabibbo et Torlone, 1998; Kimball, 1996) considèrent les faits comme la partie dynamique des entrepôts de données et les dimensions comme des entités statiques. L'historisation des données est assurée par la dimension *Temps*. Les autres dimensions sont supposées temporellement invariantes. Cependant, en pratique, des changements peuvent se produire dans le schéma des dimensions. Pour les prendre en compte, deux alternatives existent : la première propose la mise à jour du schéma (Blaschka et al., 1999; Hurtado et al., 1999) et la seconde consiste à gérer différents schémas en les historisant (Bliujute et al., 1998; Body et al., 2002; Chamoni et Stock, 1999; Eder et Koncilia, 2001). Ces deux approches constituent une réponse au problème de l'évolution des dimensions, lorsque cette dernière est orientée par l'évolution des données elles-mêmes. En revanche, elles n'apportent pas de solution à l'émergence de nouveaux besoins d'analyse qui sont orientés, non pas par l'évolution des données, mais par l'expression de nouvelles connaissances. En effet, une fois l'entrepôt constitué, l'utilisateur ne peut qu'effectuer des analyses prévues par le modèle.

Le travail de recherche exposé ici est réalisé en collaboration avec LCL - Le Crédit Lyonnais[1]. Pour compléter la politique marketing nationale, les responsables commerciaux sont amenés à faire, au niveau local, des demandes marketing. Une demande marketing est la formulation d'une demande de ciblage pour une action marketing ponctuelle (opération spécifique à un produit ou à un événement). Des données hétérogènes sont amenées à enrichir nos connaissances sur les demandes marketing. La demande marketing constitue alors un objet d'étude complexe. Pour analyser ces données complexes, il est nécessaire de les intégrer. Dans (Favre et al., 2005), nous avons proposé une architecture d'entreposage virtuel combinant la médiation et l'entreposage. Du fait que les applications considérées génèrent des informations de façon fréquente, la médiation nous dispense des tâches de rafraîchissement. Cependant, ces données sont sous un format non approprié à l'analyse. L'approche d'entreposage contribue à remédier à ce problème. L'analyse permet de générer des connaissances qu'il faut réintroduire comme source d'informations dans notre dispositif d'entreposage virtuel. Par ailleurs, établir de façon exhaustive les besoins de l'ensemble des utilisateurs est une tâche complexe. Parfois, les utilisateurs disposent de connaissances qui ne sont pas représentées dans l'entrepôt et qui sont susceptibles d'orienter l'analyse des données. Il est donc intéressant que l'utilisateur puisse exprimer ses connaissances pour définir de nouveaux axes d'analyse. Le modèle de l'entrepôt doit donc permettre l'évolution des contextes d'analyse grâce à des connaissances du domaine que l'utilisateur pourra intégrer dans l'entrepôt de données. Mais ce type d'évolution n'est pas facile à réaliser dans les modèles classiques des entrepôts qui sont peu flexibles.

Les travaux ayant porté sur l'augmentation de la flexibilité dans les entrepôts de données font généralement recours à des langages à base de règles. Certains auteurs se sont attachés à rendre la définition du schéma évolutive (Kim et al., 2003; Peralta et al., 2003). D'autres ont proposé des modèles supportant différents types de contraintes (Carpani et Ruggia, 2001; Hurtado et Mendelzon, 2002; Ghozzi et al., 2003) pour résoudre le problème de la cohérence des données. D'autres encore ont apporté une réponse au traitement des exceptions dans le processus d'agrégation, rendant ce dernier plus souple (Espil et Vaisman, 2001). Les langages à base de règles ont ainsi permis de rendre plus flexible l'entrepôt de données, mais les contextes d'analyse fournis par le modèle n'en demeurent pas moins figés.

[1]Collaboration avec la Direction d'Exploitation Rhône-Alpes Auvergne de LCL-Le Crédit Lyonnais dans le cadre d'une Convention Industrielle de Formation par la Recherche (CIFRE)

Dans cet article, nous proposons un nouveau modèle d'entrepôt de données à base de règles baptisé *R-DW* (*Rule-based Data Warehouse*). Il s'agit d'une solution conceptuelle prometteuse qui pose différents problèmes tels que la performance et l'optimisation, l'évolution du modèle. Dans cet article, nous nous attachons à présenter notre modèle et son apport. Notre modèle *R-DW* est composé de deux parties : une partie "fixe", définie en extension, et une partie "évolutive", définie en intension grâce à des règles. La partie fixe comprend une table de faits et les dimensions de premier niveau. La partie évolutive comprend un ensemble de règles qui définissent de nouvelles hiérarchies de dimension basées sur les dimensions existantes et des nouvelles connaissances. Les règles permettent alors d'établir les liens sémantiques entre les données, en définissant le passage entre deux niveaux de granularité.

Notre modèle *R-DW* présente plusieurs avantages par rapport aux modèles d'entrepôt existants. Il permet de :

- créer des hiérarchies de dimension de façon dynamique ;
- faire évoluer les contextes d'analyse ;
- renforcer l'interaction entre l'utilisateur et le système d'aide à la décision en permettant à celui-ci d'intégrer ses propres connaissances.

Cet article est organisé de la façon suivante. Nous introduisons tout d'abord un exemple motivant notre approche dans la section 2. Puis, nous présentons dans la section 3 un état de l'art sur l'évolution de schéma et la flexibilité apportée par l'utilisation des langages à base de règles dans les entrepôts de données. Nous présentons ensuite notre modèle *R-DW* dans la section 4 et définissons un cadre formel pour celui-ci dans la section 5. Nous exposons ensuite la mise en œuvre de notre modèle *R-DW* et l'application de celui-ci aux données bancaires dans la section 6, avant de conclure et d'indiquer les perspectives dans la section 7.

2 Exemple introductif

Pour illustrer notre approche de modélisation d'entrepôts de données à base de règles, nous utilisons, tout au long de cet article, le cas réel de LCL-Le Crédit Lyonnais. Le PNB annuel (Produit Net Bancaire) correspond à ce que rapporte un client à l'établissement. C'est donc une mesure intéressante qu'il convient d'étudier selon différents axes que peuvent être les caractéristiques de la clientèle (situation familiale, âge...), la structure commerciale de l'établissement, l'année... Le modèle multidimensionnel présenté dans la Figure 1 répond à ce besoin d'analyse.

Prenons le cas de la personne en charge de la clientèle étudiante au Crédit Lyonnais. Cette personne sait que certaines agences ouvertes récemment ne regroupent que des étudiants. Mais cette connaissance n'est pas visible dans le modèle et ne peut donc pas être utilisée a priori pour réaliser une analyse prenant en compte le type d'agence (agence dédiée ou non aux étudiants).

Notre modèle *R-DW* permet d'apporter une réponse à ce besoin d'analyse. La partie fixe du modèle est composée de la table des faits *TF_PNB* et des tables de dimension *CLIENT*, *ANNEE* et *AGENCE* (Figure 1). Nous ajoutons à cette partie fixe une partie évolutive contenant un ensemble de règles qui traduisent la connaissance de l'utilisateur. La connaissance sur les agences étudiantes peut être présentée par les règles suivantes :
$(R1) si\ idAgence \in$ {'01903','01905','02256'} $alors\ dim\_type\_agence =$ 'étudiant'
$(R2) si\ idAgence \notin$ {'01903','01905','02256'} $alors\ dim\_type\_agence =$ 'non étudiant'

Le modèle conceptuel induit (Figure 2) permet alors d'effectuer de nouvelles analyses générées par les connaissances de l'utilisateur. Grâce à notre modèle *R-DW*, il est donc possible

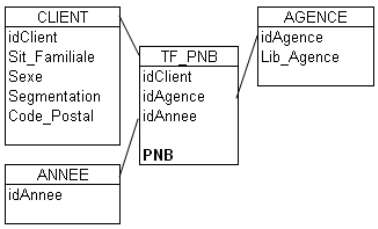


FIG. 1 – *Modèle conceptuel d'entrepôt de données pour l'analyse du PNB.*

de construire des agrégats, en considérant que les faits à agréger relèvent d'une agence étudiante $(R1)$, ou au contraire d'une agence non dédiée aux étudiants $(R2)$.

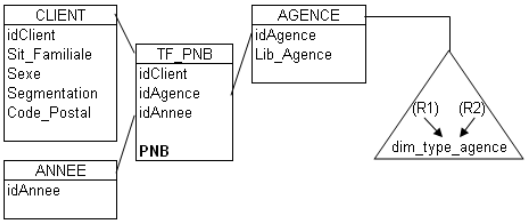


FIG. 2 – *Modèle conceptuel d'entrepôt de données à base de règles pour l'analyse du PNB.*

3 Etat de l'art

La prise en compte de nouveaux besoins d'analyse nécessite l'évolution de l'entrepôt de données. Deux alternatives peuvent être suivies. La première alternative propose la mise à jour du schéma, la seconde consiste à gérer différents schémas, en les historisant. Les travaux qui s'inscrivent dans la première alternative (Blaschka et al., 1999; Hurtado et al., 1999) consistent à migrer les données vers le schéma le plus récent. Dans ce cas, un seul schéma est supporté. L'avantage de cette approche est qu'elle fournit une comparaison des données dans le temps. Cependant, les données peuvent perdre de leur sens, voire disparaître (dans le cas d'une suppression d'un niveau de granularité dans une dimension). De plus, travailler avec la version la plus récente du schéma masque l'existence d'évolutions. Une analyse peut alors fournir de fausses conclusions. La deuxième alternative consiste à permettre une historisation des dimensions (Bliujute et al., 1998; Body et al., 2002; Chamoni et Stock, 1999; Eder et Koncilia, 2001) qui correspond à un versionnement de schémas. En effet, il s'agit de conserver chacune des versions du schéma. Le problème posé est alors de ne pas pouvoir avoir des comparaisons des données dans le temps, transversales aux différents schémas. Ces deux alternatives apportent une réponse à l'évolution des données. Mais elles n'apportent pas de solution à l'émergence de nouveaux besoins d'analyse qui sont orientés, non pas par l'évolution des données, mais par l'expression de nouvelles connaissances. Ces deux types de modélisation, malgré leurs inconvénients, permettent néanmoins d'apporter une certaine flexibilité temporelle au modèle.

Cette notion de flexibilité concerne l'ensemble de l'entrepôt. Différents travaux ont porté sur l'utilisation de langages à base de règles dans les entrepôts de données, pour apporter une flexibilité, que ce soit pour la définition de leur schéma, leur administration ou leur utilisation.

Différents travaux se sont intéressés à l'utilisation de langages à base de règles pour rendre la définition du schéma de l'entrepôt plus flexible. Deux alternatives sont possibles : utiliser les règles pour exprimer soit les besoins d'analyse, soit les connaissances sur la construction des entrepôts. Les travaux présentés dans (Kim et al., 2003) s'inscrivent dans la première alternative. Dans le contexte de la CRM (*Customer Relationship Management*), des règles de type "*si-alors*" permettent de représenter les campagnes marketing à analyser. La clause "*si*" présente les caractéristiques déterminant la population cible de la campagne marketing, et la clause "*alors*" comprend les caractéristiques de celle-ci. Le schéma de l'entrepôt est généré grâce à un algorithme qui extrait les mesures et les dimensions dans les clauses de la règle. Les travaux présentés dans (Peralta et al., 2003) s'inscrivent dans la deuxième alternative. Chacune des règles, qui représente une connaissance sur la construction de l'entrepôt de données, est spécifiée par une description, des structures cibles, un état de départ, des conditions d'application et un état final. Un algorithme gère l'ordre d'exécution des règles qui vont permettre une succession de transformations sur le schéma source pour obtenir le modèle logique final de l'entrepôt. Si cet ensemble de travaux propose de générer de façon automatique un schéma d'entrepôt en utilisant des règles, le problème de l'évolution n'est pas pris en compte.

Une des tâches importantes liées à l'administration de l'entrepôt est d'assurer la cohérence des données. En effet, si l'analyse est l'objectif primordial du processus décisionnel, celle-ci doit être faite avec cohérence. Ainsi, il est capital que l'administrateur puisse définir des contraintes d'intégrité pour assurer la cohérence à la fois de l'alimentation de l'entrepôt et de l'analyse des données. Dans (Carpani et Ruggia, 2001), les auteurs présentent un modèle conceptuel de données qui supporte un langage de contraintes, permettant d'assurer la cohérence des données. Les contraintes considérées peuvent porter sur des instances d'une dimension, ou sur différents niveaux d'une dimension, ou encore sur différentes dimensions simultanément. Dans (Hurtado et Mendelzon, 2002), les auteurs proposent des contraintes de dimension qui portent sur les chemins d'agrégation. Il s'agit d'exprimer, par exemple, que si le pays de vente est le Canada, alors les ventes seront agrégées par ville, puis par province. Le schéma enrichi de ces contraintes constitue alors un bon modèle pour inférer sur l'additivité, facilitant ainsi le processus d'analyse. Pour compléter ces approches, dans (Ghozzi et al., 2003), les auteurs proposent un modèle à contraintes pour les bases multidimensionnelles présentant un schéma en constellation. Ils définissent une typologie des contraintes sémantiques pouvant apparaître, non seulement au sein d'une même dimension, mais également entre les dimensions. Ils étendent les opérateurs de manipulation des données, en tenant compte de ces contraintes. Dans ces travaux, l'expression des contraintes d'intégrité permet d'utiliser la sémantique des données pour la gestion des incohérences dans les analyses. La sémantique des données n'est cependant pas exploitée pour l'analyse elle-même. Et si les approches proposées permettent des analyses cohérentes, l'évolution de ces dernières n'est pas évoquée.

Afin de pouvoir rendre l'analyse plus flexible, un langage à base de règles a été développé dans (Espil et Vaisman, 2001) pour la gestion des exceptions dans le processus d'agrégation. Le langage IRAH (*Intensional Redefinition of Aggregation Hierarchies*) permet de redéfinir des chemins d'agrégation pour exprimer des exceptions dans les hiérarchies de dimension du modèle. Ce langage constitue une alternative à la rigidité du schéma multidimensionnel lors

du processus d'agrégation, mais il ne fait qu'en modifier les chemins.

L'ensemble des travaux utilisant des langages à base de règles apporte une flexibilité, que ce soit dans la définition du schéma de l'entrepôt, dans son administration ou son utilisation. C'est précisément cette flexibilité que nous recherchons au niveau de l'analyse. L'évolution de l'analyse est conditionnée par celle des dimensions. La mise à jour du schéma de l'entrepôt ou le versionnement de schémas constituent une réponse au problème de l'évolution des dimensions, lorsque cette dernière est orientée par l'évolution des données elles-mêmes. En revanche, elles n'apportent pas de solution à l'émergence de nouveaux besoins d'analyse qui sont orientés par l'expression de nouvelles connaissances. Ainsi, pour répondre à notre objectif de faire évoluer les possibilités d'analyse de l'entrepôt en intégrant les connaissances de l'utilisateur, nous proposons un modèle d'entrepôt de données à base de règles.

4 Présentation du modèle *R-DW*

Un entrepôt de données présente une modélisation dite "dimensionnelle", qui répond à l'objectif d'observer les faits, à travers des mesures, en fonction des dimensions. Le schéma en étoile, qui constitue le schéma conceptuel de base pour un entrepôt de données, se compose classiquement d'une table des faits centrale et d'un ensemble de tables de dimension. D'un point de vue conceptuel, pour répondre à des besoins de performance et aux spécificités des données, le modèle en étoile a été étendu. D'une part, il a donné lieu à un schéma en constellation pour prendre en compte la nécessité de faire coexister plusieurs tables de fait qui partagent des dimensions. D'autre part, il a évolué vers un schéma en flocon de neige, dans lequel les dimensions ont été hiérarchisées. Cependant, ces modélisations ne constituent pas une réponse suffisamment flexible face à l'émergence de nouveaux besoins d'analyse. Dans ce travail, nous proposons alors un nouveau modèle d'entrepôt de données basé sur les règles (*R-DW*).

Le modèle *R-DW* est un modèle d'entrepôt de données composé de deux parties : une partie fixe, définie en extension, et une partie évolutive, définie en intension par des règles (Figure 3). La partie fixe peut être vue comme un schéma en étoile puisqu'elle comprend une table de faits et les dimensions de premier niveau (dimensions ayant un lien direct avec la table des faits). La partie évolutive comprend un ensemble de règles qui génèrent de nouvelles hiérarchies de dimension qui se basent sur la connaissance de l'utilisateur et sur les dimensions existantes.

Le métamodèle présenté dans la Figure 4 permet de généraliser le modèle *R-DW*. Il reprend en effet les classes de *Table de faits*, *Dimension* pour définir la partie fixe du modèle. Il comprend également la représentation de la partie évolutive, avec, en particulier, les classes *Règle définie en extension* et *Règle définie en intension* qui héritent de la classe *Règle*. En effet, il existe deux méthodes pour exprimer les règles qui génèrent les hiérarchies de dimension. Les règles sont exprimées en extension lorsqu'elles sont basées sur des valeurs connues qu'il est possible d'énumérer ; dans le cas contraire, elles sont définies en intension grâce à des fonctions.

Les règles exprimées en extension sont des règles de type "*si-alors*". Dans la clause "*alors*" figure la définition du niveau de granularité supérieur, en fonction de conditions exprimées dans la clause "*si*" qui portent sur les niveaux de granularité inférieurs et les connaissances de l'utilisateur. La règle suivante définit le niveau hiérarchique type_agence à travers l'attribut *dim_type_agence* en extension :
$si\ idAgence \in \{$'01903','01905','02256'$\}\ alors\ dim\_type\_agence =$ 'étudiant'
Les règles exprimées en intension sont des règles de calcul qui permettent d'inférer sur le ni-

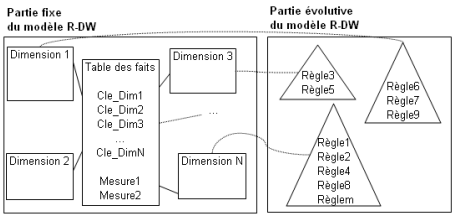


FIG. 3 – *Modèle R-DW.*

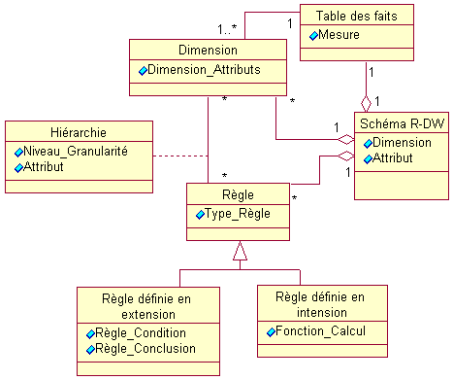


FIG. 4 – *Métamodèle de l'entrepôt de données R-DW.*

veau de granularité en fonction des niveaux inférieurs. Cette règle de calcul peut correspondre à une fonction de scoring, une extraction de caractères... Considérons par exemple le code postal du client dont on observe le PNB. En extrayant les deux premiers caractères du code postal, on obtient l'indicatif du département. À partir du code postal, on peut donc inférer sur le niveau hiérarchique département. Un autre exemple concerne les dates : il est possible d'extraire facilement les informations concernant les jour, mois, et année à partir de leur décomposition. Ce type d'extraction peut également être pertinent pour les données qui répondent à des normes : l'identification de produits, de magasins... C'est le cas par exemple au Crédit Lyonnais où la segmentation des clients comprend deux caractères. Le premier est un chiffre qui correspond au segment du client, le second est une lettre qui représente son potentiel. Il est alors facile, à partir de la segmentation du client, de définir son segment.

L'intérêt de ce modèle est donc d'offrir à l'utilisateur la possibilité de définir ses propres règles pour déterminer de nouvelles hiérarchies de dimension. L'utilisateur peut ensuite effectuer une analyse selon les niveaux de granularité définis par les règles.Le modèle devient plus flexible. La modélisation sous forme de règles offre de réelles possibilités en terme de contenu des niveaux de granularité. En effet, pour générer un niveau de granularité, il est possible :

- d'intégrer les connaissances de l'utilisateur (Exemple : cas des agences étudiantes) ;
- d'introduire les informations disponibles dans d'autres sources de données (Exemple : si nous avons une source où figure le nombre d'employés par agence, il est possible de faire l'analyse du PNB selon le nombre d'employés de l'agence) ;

– de combiner les attributs d'une même dimension. (Exemple : analyse du PNB en fonction de l'appartenance du client aux groupes 'femmes mariées' ou 'hommes mariés', constitués à partir de la combinaison des attributs *Sit_Familiale* et *Sexe*).

5 Cadre formel de l'approche

5.1 Modèle *R-DW*

Nous représentons le modèle d'entrepôt à base de règles *R-DW* par le triplet suivant :

$$\textit{R-DW} = (\mathcal{F}, \mathcal{E}, \mathcal{U})$$

où $\mathcal{F}$ est la partie fixe de *R-DW*, $\mathcal{E}$ la partie évolutive et $\mathcal{U}$ l'univers de l'entrepôt *R-DW*.

Définition 1. *Univers de l'entrepôt*

L'*univers de l'entrepôt* $\mathcal{U}$ est un ensemble d'attributs, tel que :

$\mathcal{U} = \{B_1, B_2, \ldots, B_z, C_1, C_2, \ldots\}$, où $B_\alpha, 1 \leq \alpha \leq z$, est un attribut prédéfini et $C_\beta, \beta \geq 1$, est un attribut généré.

Définition 2. *Partie fixe de R-DW*

La *partie fixe de R-DW* est représentée par $\mathcal{F} = (F, \mathcal{D})$, où F est une table de faits, et $\mathcal{D} = \{D_s, 1 \leq s \leq t\}$ est l'ensemble des dimensions de premier niveau qui ont un lien direct avec la table des faits F. Nous supposons que ces tables de dimension sont indépendantes.

Une table de dimension D_s contient une clé primaire $D_s.PK$ et un ensemble de n_s attributs $\{D_s.Z_n, 1 \leq n \leq n_s\}$.

La table de faits F contient une clé composée d'un ensemble de t ($t \geq 1$) clés étrangères $\{F.K_s, 1 \leq s \leq t\}$ où $F.K_s = D_s.PK$, et un ensemble de m mesures notées $\{F.M_q, 1 \leq q \leq m\}$.

Exemple 2. Dans la Figure 1, (TF_PNB,$\{AGENCE, ANNEE, CLIENT\}$) constitue la partie fixe de l'entrepôt *R-DW* pour l'analyse du PNB.

L'expression de nouveaux besoins d'analyse se traduit par la définition de niveaux de granularité dans les hiérarchies de dimension.

Définition 3 *Hiérarchie de dimension et niveau de granularité*

Soit *R-DW* $= (< F, \mathcal{D} >, \mathcal{E}, \mathcal{U})$ un entrepôt de données.

Soit $D_s.H_k, k \geq 1$ une *hiérarchie de la dimension* $D_s \in \mathcal{D}$.

$\{L_1, L_2, \ldots, L_i, \ldots, L_w, w \geq 1\}$ forme la hiérarchie de dimension $D_s.H_k$, avec $L_1 \prec L_2 \prec \ldots \prec L_i \prec \ldots \prec L_w$.

$L_i, 1 \leq i \leq w$ est appelé *niveau de granularité* de la hiérarchie H_k de la dimension D_s et est noté $D_s.H_k.L_i$ ou L_i^{sk}. Les niveaux de granularité sont définis par des attributs générés.

Définition 4 *Attribut généré*

Nous appelons *attribut généré* un attribut qui caractérise un niveau de granularité dans une hiérarchie de dimension. Les modalités de cet attribut sont définies grâce à des règles qui seront présentées par la suite.

$\{D_s.H_k.L_i.A_b, 1 \leq b \leq d\}$ est l'ensemble des d attributs générés, qui caractérisent le niveau de granularité L_i de la hiérarchie H_k de la dimension D_s.

Pour des raisons de simplification, nous supposons que chaque niveau d'une hiérarchie de dimension ne contient qu'un seul attribut généré, même s'il est possible de générer plusieurs attributs par niveau de granularité. Ainsi, l'attribut généré caractérisant le niveau de granularité L_i de la hiérarchie H_k de la dimension D_s est noté $L_i^{sk}.A$. Ces attributs générés sont définis grâce à la partie évolutive de *R-DW*.

Définition 5. *Partie évolutive de R-DW*

La *partie évolutive de R-DW* est représentée par $\mathcal{E} =< \mathcal{R}, \mathcal{L} >$ avec $< \mathcal{R}, \mathcal{L} >= \{< \mathcal{R}_i, L_i^{sk}.A >\}$ où $\mathcal{R}_i = \{r_{ij}, 1 \leq j \leq v, 1 \leq i \leq w\}$ est un ensemble de v règles définissant les modalités de l'attribut généré $L_i^{sk}.A$ qui représente le niveau de granularité Li de la hiérarchie H_k de la dimension D_s.

En fonction des connaissances dont nous disposons, nous exprimons les règles soit en extension, soit en intension. Nous allons définir les règles exprimées en extension (resp. intension) qui vont permettre de créer le lien entre les différents niveaux de granularité dans les dimensions grâce à l'expression de conditions (resp. de fonctions de calcul).

Définition 6. *Règle définie en extension*

Une *règle définie en extension* est une règle de type "*si-alors*", qui permet de marquer le lien sémantique qui existe entre deux niveaux de granularité dans une hiérarchie de dimension.

Elle est basée sur des termes de règle, notés RT_p tel que $RT_p = U_r\ op\ \{ens|val\}, 1 \leq p \leq n$ où $U_r \in \mathcal{U}$ l'univers de l'entrepôt ; *op* est un opérateur soit relationnel (=, <, >, ≤, ≥, ≠, ...), soit ensembliste (∈, ∉, ...) ; ens est un ensemble de valeurs et val est une valeur finie.

Exemple 6a. Les expressions $idAgence \in$ {'01903','01905','02256'}, $idAnnee < 2001$ et $Sexe =$ 'F' constituent des termes de règle.

Une règle définie en extension est basée sur une composition de *(conjonctions|disjonctions)* de ces termes de règles : $r_{ij} : si\ RT_1\ (and|or)\ RT_2 \ldots (and|or)\ RT_n\ alors\ L_i^{sk}.A = val$

Exemple 6b. Les règles suivantes définissent les modalités de l'attribut dim_type_agence en extension :

r_{11} : $si\ idAgence \in$ {'01903','01905','02256'} $alors\ dim\_type\_agence =$ 'étudiant'

r_{12} : $si\ idAgence \notin$ {'01903','01905','02256'} $alors\ dim\_type\_agence =$ 'non étudiant'

Les règles définies en extension permettent donc de créer ou d'enrichir une hiérarchie de dimension en définissant les modalités de l'attribut en fonction d'une condition ou de composition de conditions basées sur les attributs des niveaux inférieurs de la dimension.

Exemple 6c. La règle suivante permet de définir la valeur *'femmes_mariées'* de l'attribut *dim_groupe_personnes* à partir des attributs *Sit_Familiale* et *Sexe* :

$si\ Sit\_Familiale =$ 'Marié' $and\ Sexe =$ 'F' $alors\ dim\_groupe\_personne =$ 'femmes mariées'

Définition 7. *Règle définie en intension*

Une *règle définie en intension* est une règle qui permet de calculer l'attribut qui caractérise le niveau de granularité ajouté dans la hiérarchie de dimension, à partir d'attributs de niveaux inférieurs : $r_{ij} : L_i^{sk}.A = f(\mathcal{U})$

où $f(\mathcal{U})$ désigne une fonction quelconque *(extraction de caractères|fonction de scoring| ...)* pouvant s'appliquer sur un ou plusieurs attributs de l'univers de l'entrepôt $\mathcal{U}$.

Exemple 7. La règle suivante définit l'attribut $dim\_departement$ en intension :

$dim\_departement = gauche(Code\_Postal, 0, 2)$

où $gauche(chr, x, y)$ est une fonction qui permet d'extraire y caractères de la chaîne de caractères chr à partir de la position x.

5.2 Processus d'agrégation

$\mathcal{RE}$ désigne l'ensemble des règles définies en extension.Dans la règles r_{ij} définie en extension, la condition dans la clause "*si*" est notée $body(r_{ij})$, et la conclusion dans la clause "*alors*" est notée $head(r_{ij})$.

Pour réaliser une analyse à partir de l'entrepôt *R-DW*, il est nécessaire de prendre en compte les règles. Nous avons donc défini un algorithme qui permet le calcul d'agrégats (Figure 5).

Pour des raisons de clarté, nous nous restreignons à une agrégation selon un attribut déterminé par un ensemble de règles définies en extension . Cet algorithme permet de construire une table d'agrégats $TAgreg$ à partir de l'entrepôt et des caractéristiques de l'analyse (attribut A selon lequel est faite l'agrégation, mesure M_q, opérateur d'agrégation op).

Cet algorithme peut être utilisé pour répondre à la requête décisionnelle "Quel est le PNB moyen par type d'agence ?". Dans ce cas, $TAgreg$ est une table contenant deux tuples où figurent le PNB moyen pour les agences de type 'étudiant' d'une part, et de type 'non étudiant' d'autre part. Les valeurs correspondantes ont été obtenues par l'exécution des requêtes suivantes :

(1) SELECT MOY(PNB) FROM TF_PNB WHERE idAgence IN ('01903','01905','02256') ;
(2) SELECT MOY(PNB) FROM TF_PNB WHERE idAgence NOT IN ('01903','01905','02256') ;

```
Algorithme Calcul_Agreg
Input :  table des faits F,
             ensemble des règles définies en extension RE,
             attribut A,
             mesure F.M_q,
             opérateur d'agrégat op,
Output : table des agrégats TAgreg
Début
    Pour chaque r_ij ∈ RE
       Si A ∈ head(r_ij) Alors
          TAgreg =' SELECT op(F.M_q) FROM F WHERE body(r_ij)'
       Fin Si
    Fin pour
Fin
```

FIG. 5 – *Algorithme de calcul d'agrégats*

6 Mise en œuvre et application aux données bancaires

Pour valider notre approche, nous avons réalisé une implémentation du modèle *R-DW*. Nous avons développé une plateforme Web (HTML/PHP), qui interface le SGBD Oracle. La table de faits et les tables de dimension sont définies dans Oracle. Deux tables permettent de regrouper respectivement les règles définies en extension et les règles définies en intension. La plateforme Web permet à l'utilisateur de visualiser et de définir les règles qui génèrent des axes d'analyse. Elle permet également la visualisation des résultats d'analyse. Concernant l'analyse, nous nous sommes restreints, dans un premier temps, à une requête décisionnelle mettant en jeu une agrégation selon un niveau hiérarchique d'une dimension donnée. Cette agrégation est implémentée sous la forme d'une procédure stockée en PL/SQL du SGBD.

Nous avons appliqué notre modélisation aux données bancaires qui concernent l'analyse du PNB. La partie fixe du schéma est représentée dans la Figure 1. À partir de ces dimensions, et des connaissances de l'utilisateur, différentes hiérarchies de dimension ont été représentées par l'ensemble de la Figure 6, qui constitue la partie évolutive du schéma *R-DW*. Ainsi, à partir de ce nouveau modèle, l'utilisateur pourra effectuer des analyses sur le PNB, non seulement en utilisant les dimensions de premier niveau, mais également en faisant intervenir des niveaux de granularité comme le type d'agence, la période, le département, les classes d'âge...

FIG. 6 – *Partie évolutive de l'entrepôt R-DW pour l'analyse du PNB.*

7 Conclusion

Dans cet article, nous avons proposé un nouveau modèle d'entrepôt de données à base de règles nommé *R-DW*. Les règles permettent d'intégrer de nouvelles connaissances de l'utilisateur dans l'entrepôt. Notre modèle *R-DW* est composé de deux parties : une partie fixe, définie en extension, comprenant une table des faits et les dimensions de premier niveau ; une partie évolutive, définie en intension par des règles, qui détermine les niveaux de granularité dans les hiérarchies de dimension. Notre modèle *R-DW* présente plusieurs avantages. Il permet de créer des hiérarchies de dimension de façon dynamique en renforçant l'interaction entre l'utilisateur et le système d'aide à la décision. En effet, les hiérarchies de dimension sont générées grâce aux propres connaissances de l'utilisateur. Cette génération de hiérarchies de dimension "à la demande" permet de faire évoluer les contextes d'analyse. Par ailleurs, nous avons proposé un métamodèle qui décrit le modèle *R-DW*. L'implémentation que nous avons réalisée a permis d'appliquer notre modèle aux données bancaires réelles de LCL-Le Crédit Lyonnais.

Ce travail ouvre différentes perspectives. Tout d'abord, nous voulons mesurer la performance de notre approche en termes d'espace de stockage et de temps de réponse. En terme de performance, il s'agit également d'étudier le problème de la matérialisation des vues et des structures d'index. Ensuite, concernant la définition des règles, l'atout de notre approche est de pouvoir faire intervenir l'utilisateur en lui laissant la possibilité d'introduire ses connaissances dans le système. Mais il nous semble intéressant, en parallèle, de pouvoir l'aider à découvrir de nouveaux axes d'analyse. Pour ce faire, nous pensons que des méthodes d'apprentissage non supervisé peuvent être utilisées. Il s'agit également de définir un langage qui permette de valider les règles utilisées, que ce soit pour la gestion des conflits entre les règles, ou pour la vérification de contraintes sur celles-ci.

Références

Blaschka, M., C. Sapia, et G. Höfling (1999). On Schema Evolution in Multidimensional Databases. In *DaWaK'99 : 1st International Conference on Data Warehousing and Knowledge Discovery*, pp. 153–164.

Bliujute, R., S. Saltenis, G. Slivinskas, et C. Jensen (1998). Systematic Change Management in Dimensional Data Warehousing. In *3rd International Baltic Workshop on DB and IS*.

Body, M., M. Miquel, Y. Bédard, et A. Tchounikine (2002). A Multidimensional and Multiversion Structure for OLAP Applications. In *DOLAP'02 : 5th ACM International Workshop on Data Warehousing and OLAP*.

Cabibbo, L. et R. Torlone (1998). A Logical Approach to Multidimensional Databases. In *EDBT'98 : 6th International Conference on Extending Database Technology*, pp. 183–197.

Carpani, F. et R. Ruggia (2001). An Integrity Constraints Language for a Conceptual Multidimensional Data Model. In *SEKE'01 : XIII International Conference on Software Engineering Knowledge Engineering*.

Chamoni, P. et S. Stock (1999). Temporal Structures in Data Warehousing. In *DaWaK'99 : 1st International Conference on Data Warehousing and Knowledge Discovery*, pp. 353–358.

Eder, J. et C. Koncilia (2001). Changes of dimension data in temporal data warehouses. In *DaWaK'01 : 3rd International Conference on Data Warehousing and Knowledge Discovery*.

Espil, M. M. et A. A. Vaisman (2001). Efficient Intensional Redefinition of Aggregation Hierarchies in Multidimensional Databases. In *DOLAP'01 : 4th ACM International Workshop on Data Warehousing and OLAP*.

Favre, C., F. Bentayeb, O. Boussaid, et N. Nicoloyannis (2005). Entreposage virtuel de demandes marketing : de l'acquisition des objets complexes à la capitalisation des connaissances. In *2ème atelier FDC de EGC05, Paris*, pp. 65–68.

Ghozzi, F., F. Ravat, O. Teste, et G. Zurfluh (2003). Constraints and Multidimensional Databases. In *ICEIS'03 : 5th International Conference on Enterprise Information Systems*.

Hurtado, C. A. et A. O. Mendelzon (2002). OLAP Dimension Constraints. In *PODS'02 : 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*.

Hurtado, C. A., A. O. Mendelzon, et A. A. Vaisman (1999). Updating OLAP Dimensions. In *DOLAP'99 : 2nd ACM International Workshop on Data Warehousing and OLAP*.

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

Kim, H. J., T. H. Lee, S. G. Lee, et J. Chun (2003). Automated Data Warehousing for Rule-Based CRM Systems. In *14th Australasian Database Conference on Database Technologies*, pp. 67–73.

Kimball, R. (1996). *The Data Warehouse Toolkit*. John Wiley & Sons.

Peralta, V., A. Illarze, et R. Ruggia (2003). On the Applicability of Rules to Automate Data Warehouse Logical Design. In *CAiSE Workshops*.

Summary

Data warehouses are an answer to enterprises'analysis needs. However, such a system is hard to build and maintain, particularly when these analysis needs evolve. To solve this problem, we propose a new data warehouse model based on rules called Rule-based Data Warehouse (*R-DW*). The rules are used to integrate user's knowledge in the data warehouse. This model is composed of two parts : one fixed part, defined extensionally, composed of a fact table and dimensions of the first level ; a second evolving part, defined intentionally with rules. Having these rules we are able to dynamically create dimension hierarchies. It makes thus possible the contexts of analysis evolution, and it increases the interaction between the user and the decision support system.

Prétraitement et classification de données complexes dans le domaine du commerce électronique

Sergiu Chelcea*, Alzennyr Da Silva* & **, Yves Lechevallier**,
Doru Tanasa*, Brigitte Trousse*

*AxIS, INRIA Sophia-Antipolis
2004, Route des Lucioles, B.P. 93
06902 Sophia Antipolis Cedex, France
{Sergiu.Chelcea, Doru.Tanasa, Brigitte.Trousse}@sophia.inria.fr
http://www-sop.inria.fr/axis/
**AxIS, INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, B.P. 105
78153 Le Chesnay Cedex, France
{Alzennyr.Da_Silva, Yves.Lechevallier}@inria.fr

Résumé. L'objectif de cet article est de proposer une méthode originale de prétraitement associée à une classification non supervisée dans le cadre de données complexes relatives aux activités d'internautes sur plusieurs sites de commerce électronique. La première contribution concerne le prétraitement visant la construction d'un entrepôt de données intégrant le concept de « visite muti-sites » et offrant une structure riche des données. La deuxième contribution consiste à présenter quelques analyses statistiques. Ces analyses ont été appliquées sur des données structurées basées sur la notion de « période de temps ». Une méthode de classification croisée a été ensuite appliquée sur ces données extraites de notre entrepôt, application originale dans le contexte du Web Usage Mining. Nos résultats préliminaires sont prometteurs et illustrent les avantages de notre approche dans le domaine du WUM.

1 Introduction

Cet article traite le problème des données complexes du Web dans le domaine du e-commerce. La complexité de ces données est liée aux particularités suivantes: volume important, sources multiples (des fichiers logs multi-sites), hétérogénéité (tables contenant les descriptions de produit, fichiers logs) et temporalité (des clicks enregistrés en périodes temporelles) des données.

Le nombre d'accès par jour à un site Web peut aujourd'hui facilement atteindre plusieurs millions. Les contenus des pages différents et les objectifs hétérogènes constituent la réalité du Web. Avec la croissance explosive des données disponibles sur Internet, la découverte et l'analyse d'information utile du Web devient une nécessité. La fouille de donnée d'usage du Web (« Web Usage Mining » en anglais) correspond à l'application de techniques de fouille de donnée sur des fichiers log de grande taille qui représentent les accès des utilisateurs aux

serveurs Web. Le Web Usage Mining est l'instanciation d'un processus ECD sur les données provenant des fichiers log Web. Les objectifs sont l'amélioration de l'architecture des sites Web, l'analyse de la performance des systèmes, la compréhension des réactions et des motivations des utilisateurs, la personnalisation des sites Web, et la construction de sites Web adaptatifs (Jaczynski (1998), Jaczynski et Trousse (1998), Mobasher (2000), Srivastava et al. (2000), Trousse et al. (1999)).

Les gérants de sites d'e-commerce sont particulièrement intéressés par les résultats obtenus par le Web Usage Mining (Perkowitz et Etzioni (1997), Kohavi (2001)). Les résultats aident non seulement à améliorer leur site Web, mais également leurs services et stratégies de vente (promotions, bannières, etc.). Cet article vise à analyser un ensemble de *clickstream* dans le domaine d'e-commerce avec le point de vue du concepteur du site.

Cet article est organisé comme suit. La section 2 présente l'ensemble de données analysées et notre prétraitement avancé sur les données multi-sites structurées en termes de visites de consommateurs sur plusieurs magasins virtuels. Ensuite, la section 3 donne les premiers résultats issus d'une analyse statistique préliminaire de cet ensemble de données. Après, avant la conclusion dans la section 5, la section 4 décrit notre méthode de classification, les données utilisées et les analyses respectives.

2 Prétraitement multi-site avancé

2.1 Description des données utilisées

Nous avons choisi l'ensemble de *clickstreams* proposé dans le *challenge* 2005 de la conférence PKDD et qui comprend 576 fichiers log de grande taille contenant 3 617 171 requêtes. Les requêtes étaient faites sur sept sites Web d'e-commerce différents situés en République Tchèque (voir le tableau 1). Chaque fichier log contient toutes les requêtes réalisées sur les sept sites Web durant une période continue de 24 jours débutant le 20 janvier à 09:00am et se terminant le 13 février 2004 à 08:59am.

ShopID	Nom du site (magasin)	#Requêtes
10	www.shop1.cz	509 688
11	www.shop2.cz	400 045
12	www.shop3.cz	645 724
14	www.shop4.cz	1 290 870
15	www.shop5.cz	308 367
16	www.shop6.cz	298 030
17	www.shop7.cz	164 447

TAB. 1 – *Nombre de requêtes par magasin.*

Chaque ligne des fichiers log contient le résultat d'une requête sur une page de l'un des 7 serveurs Web d'e-commerce et a les 6 champs suivants (tableau 2) :

- *ShopID* : l'identificateur du serveur Web d'e-commerce ayant reçu la demande ;
- *Date* : le temps Unix de la requête (nombre de secondes depuis le 1er janvier 1970) ;
- *IP address* : l'adresse IP de l'ordinateur de l'utilisateur;
- *SessionID* : l'identificateur de session PHP qui est produit de manière automatique pour chaque nouvelle visite sur chaque serveur ;

- *Page* : la ressource demandée (page) sur le serveur ;
- *Referrer* : l'adresse de la page demandée.

ShopID	Date	IP address	SessionID	Page	Referrer
11	1074585663	213.151.91.186	939dad92c4…84208dca	/	
11	1074585670	213.151.91.186	87ee02ddcff…7655bb9e	/ct/?c=148	http://www.shop2.cz

TAB. 2 – *Format des requêtes de page.*

Une table contenant les noms et les identificateurs des sept magasins a été fournie (tableau 1, les deux premières colonnes). Pour des raisons de confidentialité, les noms des sept sites Web d'e-commerce ont été rendus anonymes dans cette table aussi bien que dans le champ Referrer.

Les pages Web sur ces serveurs sont reliées par des liens hypertexte, donc les utilisateurs peuvent naviguer d'un magasin à l'autre en utilisant seulement ces liens (clics). Cependant, puisque le SessionID est produit à la première requête pour une page d'un site, un utilisateur obtiendra un nouveau SessionID lorsqu'il accédera à un nouveau site.

Le champ Page contient le chemin sur le serveur du champ ShopID de la page demandée. Les types de pages correspondent aux 21 catégories de page de premier niveau (voir le tableau 3 ci-dessous).

ID	Type de page	Description	#Requêtes	%
1	/ct	Catégorie de produit	228,991	6.33
2	/ls	Fiche produit	1,363,187	37.68
3	/dt	Détail du produit	1,233,570	34.1
4	/znacka	Liste des marques ou détails d'une marque	88,189	2.43
5	/akce	Offres actuelles	26,260	0.72
6	/df	Comparaison des paramètres produit	57,939	1.60
7	/findf	Recherche textuelle de produits et accessoires	55,139	1.52
8	/findp	Recherche basée sur les paramètres	93,455	2.58
9	/setp	Etablir les paramètres d'affichage	11,752	0.32
10	/poradna	Conseil en ligne	107,711	2.97
11	/kosik	Panier d'achat, details du contract, enregistrer une commande	35,487	0.98
12	/	Page principale	219,218	6.06
13	/obchody-elektro	Liste des magasins de produits électroniques	10,926	0.30
14	/kontakt	Information de contact	6,104	0.16
15	/faq	Foire aux questions	861	0.02
16	/onakupu	Informations sur l'achat	6,659	0.18
17	/splatky	Possibilités d'achat au crédit	2,846	0.07
18	/mailc	Disponibilité de produits	6,680	0.18
19	/mailp	Envoyez cette page	6,905	0.19
20	/mailf	Envoyez un feedback	1,855	0.05
21	/mailr	Formulaire de plainte	494	0.01
		Total	3,564,228	98,45

TAB. 3 – *Types de pages.*

En utilisant le type de ces pages, nous pouvons savoir si l'utilisateur a fait une requête pour un produit ou pour une catégorie ou pour un thème spécifique. Par exemple, dans la deuxième ligne du tableau 1, l'utilisateur a demandé les produits de la catégorie d'écouteurs (code 148). Pour décrire les variables présentes dans les pages demandées, quatre tables ont été sont disponibles :

1. ***kategorie*** : contient les 60 descriptions des catégories de produit, la table correspondante est appelée « category » ;
2. ***liste*** : contient les 157 descriptions des produits, la table correspondante est dénotée « product » ;
3. ***znacka*** : contient les 197 descriptions des marques de produit, la table correspondante est dénotée « brand » ;
4. ***tema*** : contient les 36 descriptions des thèmes de produit, la table correspondante est dénotée « theme ».

Le champ de Referrer représente l'URL de la page Web contenant le lien que l'utilisateur a suivi pour aboutir à la page courante, ce champ parfois peut être vide.

2.2 Méthode de prétraitement

Pour préparer l'ensemble de données, nous avons employé une méthodologie récemment proposée par (Tanasa et Trousse (2004), Tanasa (2005)), qui étend les travaux de Cooley (2000). Malheureusement, les données brutes n'ont pas été fournies dans le format CLF (« Common Log Format » W3C (1995)) et plusieurs champs n'étaient pas disponibles (le statut, l'agent d'utilisateur, le login). De ce fait nous n'avons employé que le champ *SessionID* pour identifier la visite d'un utilisateur.

Le prétraitement de données a été fait en quatre étapes : la fusion de données, le nettoyage de données, la structuration de données et l'agrégation de données.

Dans l'étape de fusion de données, les fichiers log de différents serveurs Web ont été réunis dans un seul fichier log. Nous avons également changé le format de date en temps grégorien afin de faciliter l'interprétation de nos analyses. Enfin, nous avons groupé la page et le champ *ShopID* dans un nouveau champ *URL* (voir le tableau 4) afin d'avoir le même format que pour le champ *Referrer*.

Datetime	IP	SessionID	URL	Referrer
2004-01-20 09:01:03	213.151.91.186	939dad92c4…84208dca	http://www.shop2.cz/	-
2004-01-20 09:01:10	213.151.91.186	87ee02ddcff…7655bb9e	http://www.shop2.cz/ct/?c=148	http://www.shop2.cz/

TAB. 4 – *Format du fichier log transformé (après la fusion).*

L'étape de nettoyage de données, comprend le filtrage des ressources Web qui ne sont pas utiles à l'analyse (par exemple les fichiers .jpg, .gif, .js, etc.). L'information sur le statut de la requête permet elle aussi de filtrer les données mais, ici, cette information était absente.

Dans l'étape de structuration de données, les requêtes sont groupées par l'utilisateur, par session utilisateur, et par visite. Comme nous l'avons mentionné précédemment, un utilisateur qui change de magasin change aussi de numéro *SessionID*, aussi nous avons décidé de grouper de tels *SessionID* qui appartiennent à un seul utilisateur (ayant le même IP) dans un groupe de sessions, correspondant une visite réelle d'un l'utilisateur. Ceci a été fait en comparant le *Referrer* à l'*URL* précédemment consultée. Si le *Referrer* (une page d'un autre magasin dans ce cas) était précédemment consulté, nous groupons les deux *SessionID* ensemble. Par exemple, les deux requêtes présentées dans le tableau 4, ayant des *SessionID* différents, ont été regroupées car le URL présent dans le champ *Referrer* de la deuxième requête (ligne 2) a été demandé peu de temps avant celui de la ligne 1.

En utilisant cet algorithme, nous avons groupé les 522 410 *SessionID* initiales dans 397 629 groupes, équivalent à une réduction de 23,88% dans le nombre des visites des utilisateurs. Ainsi, nous avons obtenu les visites d'utilisateur multi-sites qui seront employées les analyses globales(cf. sections 3 et 4).

Toutefois, de nos jours, pour des raisons de confidentialité, un grand nombre de logiciels de sécurité (parefeux, anti spyware, etc..) bloquent les « cookies » de l'utilisateur (utilisés pour *SessionID*) et/ou le *Referrer* des requêtes. Dans ce cas-ci, le blocage des « cookies » et du *Referrer* aura comme conséquence un *SessionID* différent pour chaque requête d'un utilisateur qu'il change ou pas de magasin. Ceci se produit également dans le cas d'un robot Web (ABCInteactive.com (2004)) employé pour indexer les pages sur ces sites Web ou dans le cas d'un logiciel de téléchargement de sites Web. Ainsi, nous avons constaté que 2.54% des adresses IP (2 020 sur 79 526) ont effectué 141 976 requêtes distinctes correspondant au même nombre de *SessionID* représentant 27.18% du nombre total des *SessionID* distincts. Avec cette information, nous pourrions réduire encore plus le nombre des visites.

Finalement, après l'identification de chacune des variables présentes dans l'URL consulté, nous avons défini un modèle de base de données relationnelle pour stocker ces informations structurées (voir la figure 1). Ensuite, les données prétraitées ont été stockées dans cette BD relationnelle, durant l'étape d'agrégation de données.

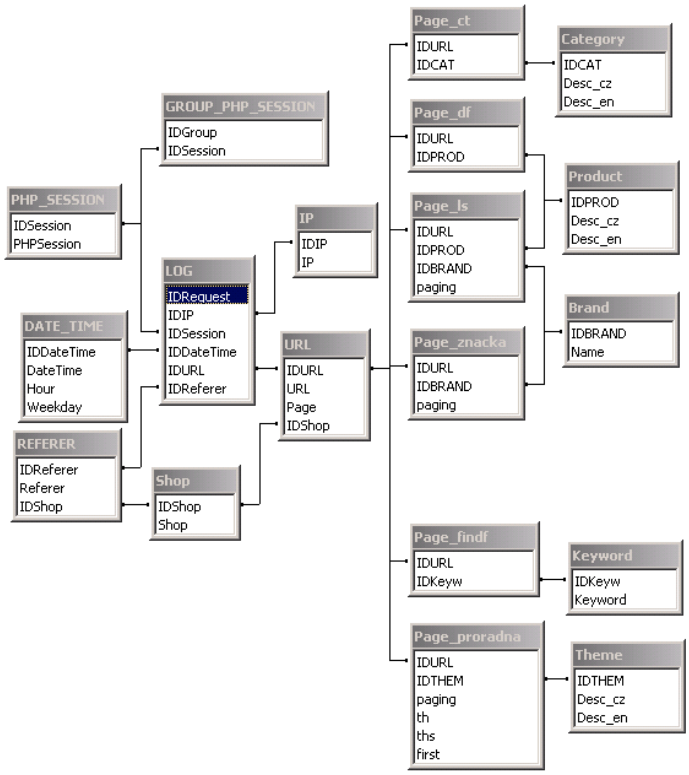


FIG. 1 – *Model relationnel de la base de donnée.*

Nous avons étendu le modèle présenté en Tanasa et Trousse (2004) et en Tanasa (2005) en ajoutant de nouvelles tables et des nouveaux attributs aux tables existantes. La table LOG est la table principale dans ce modèle et contient dans chaque enregistrement, les informations sur une requête de vue de page obtenues à partir des fichiers log fournis.

Pendant l'étape de prétraitement et l'analyse statistique (voir les tableaux 1 et 3), nous avons identifié plusieurs inconsistances dans l'ensemble de données. Par exemple, nous avons trouvé des entrées manquantes ou absentes dans les tables de description, des informations absentes sur les variables, des numéros *SessionID* mal formés, etc.

Dans les deux prochaines sections, nous montrons l'avantage d'avoir des données prétraitées et structurées ainsi que la flexibilité de notre modèle relationnel par l'intermédiaire d'une analyse statistique et de l'application de deux méthodes de classification.

Afin d'analyser la charge des sept serveurs Web d'e-commerce, nous avons décidé d'employer des unités classiques de temps : type de jour et heure.

3 Analyse statistique basée sur des périodes temporelles

L'objectif de cette section est de montrer les résultats obtenus lors d'une analyse multi-sites et en employant la notion de « groupe de *SessionsID* », qui représente réellement la visite d'un utilisateur.

Dans le tableau 5 on observe que le mercredi est le jour où les visites sont les plus nombreuses. De plus, on remarque que le mercredi et le dimanche les taux de réduction de *SessionsID* aux visites sont les plus importantes(plus de 35%). Nous croyons que ceci est lié aux reconnexions fréquentes des utilisateurs, ce qui est confirmé par le ratio élevé du nombre de *SessionID* par visite. Notons également que l'analyse multi-sites n'a pas été significative (le pourcentage des visites multi-sites varie entre 2.72% et 4.49% par jour).

Jour	#Requêtes	#SessionID	#Visites	Réduction (%)	#Visites multi-sites	#SessionID /Visites MS
Lundi	551 138	73 700	53 373	27,58	2 903	7
Mardi	675 984	80 649	66 565	17,46	3 753	3,75
Mercredi	677 243	112 580	73 126	35,04	3 576	11,03
Jeudi	612 158	76 211	64 338	15,57	3 410	3,48
Vendredi	461 607	64 737	57 065	11,85	2 430	3,15
Samedi	296 334	51 018	44 706	12,37	1 601	3,94
Dimanche	342 707	63 515	38 456	39,45	1 859	13,47
Total	3 617 171	522 410	397 629	23,88	19 532	6,39

TAB. 5 – *Analyse statistique par jour de la semaine.*

Quant aux heures analysées, le tableau 6 montre qu'il y a quatre périodes horaires (7-8, 12-13, 13-14 et 20-21) très importantes en termes de charge des serveurs Web (reconnexions). Dans la figure 2 nous présentons la distribution globale des visites par heure.

Heure	#Requêtes	#SessionID	#Visites	Réduction (%)	#Visites multi-sites	#SessionID/ Visites MS
0-1	59 205	12 804	9 407	26,53	274	12,39
1-2	32 110	9 352	8 309	11,15	165	6,32
2-3	19 183	6 628	6 376	3,80	90	2,8
3-4	13 302	5 937	5 815	2,05	58	2,1
4-5	14 082	6 999	6 743	3,65	51	5,01
5-6	15 691	7 772	7 265	6,52	65	7,8
6-7	43 459	11 178	10 161	9,09	258	3,94
7-8	103 445	22 827	14 589	36,08	521	15,81
8-9	156 642	24 805	17 913	27,78	899	7,66
9-10	200 170	24 746	21 006	15,11	1 152	3,24
10-11	228 906	26 685	23 112	13,38	1 286	2,77
11-12	246 296	29 661	22 967	22,56	1 332	5,02
12-13	264 805	43 493	24 598	43,44	1 424	13,26
13-14	275 854	36 981	24 767	33,02	1 454	8,4
14-15	264 876	31 040	24 788	20,14	1 433	4,36
15-16	242 962	29 220	23 092	20,97	1 304	4,69
16-17	206 331	23 841	20 531	13,88	1 179	2,8
17-18	179 601	21 241	18 259	14,03	957	3,11
18-19	185 730	23 714	20 086	15,29	1 068	3,39
19-20	187 077	22 933	19 158	16,46	1 051	3,59
20-21	219 793	37 663	20 599	45,30	1 174	14,53
21-22	200 343	27 394	19 612	28,40	1 048	7,42
22-23	154 582	20 645	15 462	25,10	762	6,8
23-0	102 726	14 851	13 014	12,36	527	3,48
Total	3 617 171	522 410	397 629	23,88	19 532	6,39

TAB. 6 – *Analyse statistique par heure.*

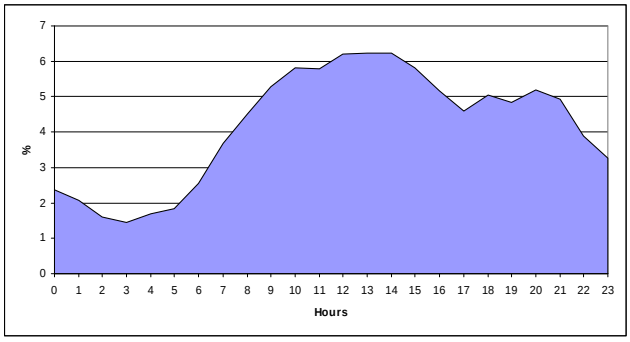


FIG. 2 – *Nombre de visites par heure.*

Les figures 3a (visites globales) et 3b (visites multi-sites) montrent clairement qu'il y a très peu de nouvelles visites le samedi, dimanche et pendant le déjeuner et beaucoup de visites le mardi et le mercredi.

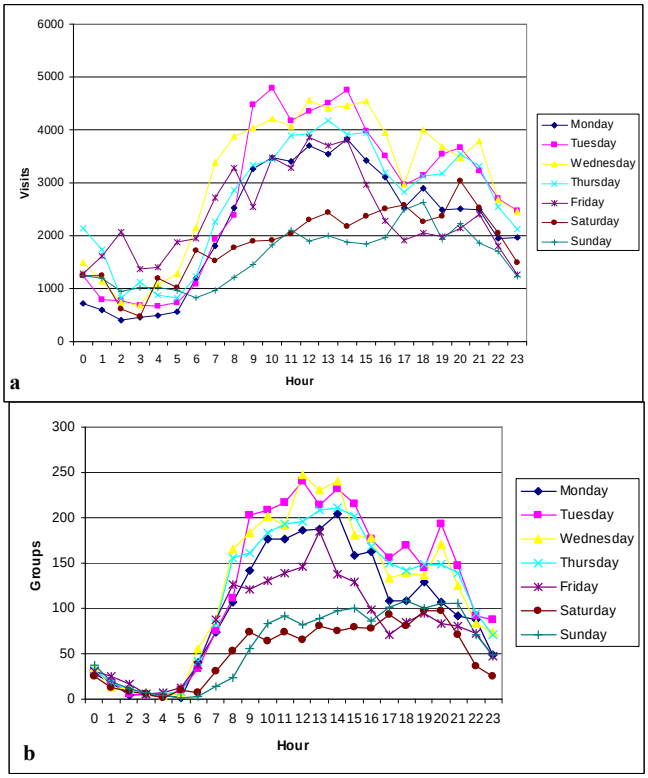


FIG. 3 – *Visites par jour et heur e: (a) globalement, (b) multi-site.*

4 Classification simultanée d'un ensemble de lignes et de colonnes d'un tableau de comptage

4.1 Le problème

L'utilisation d'un algorithme de classification est souvent une première étape pour l'extraction de connaissances à partir d'une base de données. Grouper, en fait, mène à la construction d'une structure classificatoire, c'est-à-dire, l'identification des sous-groupes homogènes et distincts d'objets (Gordon (1981), Bock (1993)), où la définition d'homogénéité est associée à l'algorithme de classification utilisé. En cas d'absence de connaissances a priori sur la forme multidimensionnelle de la représentation des objets, la recherche d'une partition peut être un point de départ raisonnable vers la découverte des structures plus riches et complexes. Malgré la grande richesse des algorithmes de classification existant, l'arrivée des nouvelles bases de données de grande dimension et de complexité croissante pose un certain nombre de nouveaux problèmes de classification.

Une caractéristique importante des nouvelles collections de données est la taille toujours croissante de ces bases de données. Il n'est plus maintenant rare de travailler avec des bases

de données contenant plusieurs millions d'individus et des centaines ou des milliers de variables. La plupart des algorithmes actuels de classification sont limités par le nombre d'individus qu'ils peuvent confortablement manipuler (de quelques centaines à quelques milliers).

Pour analyser la charge du trafic sur le site des sept magasins de e-commerce, nous avons groupé les requêtes en termes de périodes de temps. Nous avons limité notre analyse aux 490 883 requêtes enregistrées sur les pages de la rubrique « /ls » du site du magasin 4. Ce magasin 4 est le plus visité selon le tableau 1.

4.2 La méthode

Notre objectif est d'obtenir de manière simultanée une partition des lignes et une partition des colonnes d'une relation complexe. Les auteurs de Govaert (1977) et de Govaert et Nadif (2003) ont proposé la maximisation du critère chi-carré entre les lignes et les colonnes d'un tableau de contingence. L'avantage principal de cette approche est d'avoir comme résultat un nouveau tableau de contingence mais contenant moins de lignes et de colonnes que le tableau initial tout en conservant les interactions entre ces deux ensembles. Avec notre approche la caractérisation du groupe est basée sur les distributions des descripteurs à valeurs multiples et les interactions entre les descripteurs et les individus. Le critère optimisé est basé sur le meilleur ajustement entre les classes des objets et les classes des variables à valeurs multiples. Dans notre contexte d'analyse de ces relations entre les périodes de temps et des produits du magasin 4, nous proposons de représenter les classes par les prototypes qui résument l'information contenue dans périodes de temps appartenant à chacune des classes. Chaque prototype est modélisé comme un objet décrit par des variables à multi-catégories avec distributions associées.

Dans ce contexte, plusieurs distances et fonctions de dissimilarité peuvent être proposées. En particulier, si les objets et les prototypes sont décrits par des variables à multi-catégories, la mesure de dissimilarité peut être la distance classique entre les distributions (par exemple chi-carré). La convergence de l'algorithme vers une valeur stationnaire du critère est garantie par le meilleur ajustage de précision entre le type de représentation des classes et les propriétés de la fonction d'allocation.. L'algorithme dynamique décrit dans (Verde et Lechevallier (2003) et Lechevallier et Verde (2004)) peut être utilisé dans différents contextes d'analyse, par exemple: pour grouper des données archéologiques décrites par variables multi-catégoriques (Verde et al. (2000)); pour comparer des caractéristiques socio-économiques dans différents secteurs géographiques en ce qui concerne aux distributions des variables (par exemple: activités économiques; distribution de revenues; heures travaillées; etc.).

4.3 Analyse

4.3.1 Création d'un tableau de données en fonction des tranches horaires

Pour notre analyse, nous avons construit un tableau croisé où chaque ligne décrit un objet qui représente le couple *jour de la semaine* et l'*heure* de la requête réalisée sur une page du répertoire « /ls » du site du magasin 4 (le plus visité selon le tableau 1), et la colonne décrit une variable multi-valuée qui représente le nombre de produits demandés par des utilisateurs dans une tranche spécifique d'horaire (voir le tableau 7, où nous avons 7 x 24 objets). Nous avons limité notre analyse aux demandes enregistrées sur le site du magasin 4 même si cette approche peut être effectuée sur l'ensemble des magasins.

Jour de la semaine x Heure	Produit (nombre de requêtes)
Lundi_0	Built-in electric hobs (10), Built-in dish washers 60cm (64), Corner single sinks (50), ...
Lundi _1	Free standing combi refrigerators (44), Corner single sinks (50), Built-in hoods (60), ...
…	…
Samedi_22	Built-in microwave ovens (27), Built-in dish washers 45cm (38), Built-in dish washers 60cm (85), ...
Samedi _23	Built-in freezers (56), Kitchen taps with shower (45), Garbage disposers (32), ...

TAB. 7 – *Quantité de produits registrés sur shop 4 en fonction du jour de la semaine x heure.*

4.3.2 Résultats

Les 168 périodes du tableau 7 résument 490 883 requêtes sur tous les produits du magasin 4. Le tableau 8 présente les résultats après l'application de notre méthode de classification croisée indiquant 7 classes de périodes et 5 classes de produits.

	Produit_1	Produit _2	Produit _3	Produit _4	Produit _5	Total
Période_ 1	2847	5084	3284	2265	2471	15951
Période _ 2	11305	31492	12951	1895	9610	67253
Période _3	33107	55652	36699	5345	20370	151173
Période _4	22682	46322	30200	5165	27659	132028
Période _5	9576	20477	19721	2339	7551	59664
Période _6	1783	3515	2549	392	11240	19479
Période _7	15019	14297	8608	1397	6014	45335
Total	96319	176839	114012	18798	84915	490883

TAB. 8 – *Tableau de confusion.*

Le tableau 8 montre clairement que la classe de produits numéro 5 a été uniquement définie par un seul produit, à savoir « *Free standing combi refrigerators* » (voir le tableau 9) et a été consultée notamment le vendredi entre 17:00 et 20:00 heures (voir le tableau 10). Il est important de noter que bien que ce produit appartienne à la classe des produits responsable de seulement 17.3 % de toutes les interrogations sur le magasin 4, le pourcentage les interrogations de cette période sur ce produit est égal à 57.7 % (voir le tableau 10). En d'autres termes, il signifie que le produit « *Free standing combi refrigerators* » est davantage demandé les vendredis de 17:00 à 20:00. Une telle information pourra être employée sur les stratégies de vente et du marketing.

Produit_5 Cardinal: 1
/product/Free standing combi refrigerators

TAB. 9 – *Regroupement des produits.*

Période _6 Cardinal: 8
Vendredi_2, Vendredi _6, Vendredi _17, Vendredi _18, Vendredi _19, Vendredi _20, Samedi_5, Mardi_4

TAB. 10 – *Regroupement des tranches horaires*

5 Conclusions et travaux futurs

Dans cet article, nous avons proposé une méthode de prétraitement de données multi-site du Web dans le domaine du commerce électronique. Notre analyse sur les fichiers *log* proposés dans le cadre du *challenge* PKDD a montré que la grande flexibilité offerte par l'entrepôt

des données construit avantage une restructuration en termes d'interrogations des clients. Deuxièmement, nous avons présenté une analyse de regroupement basée sur les méthodes de classification croisée efficaces et appliquées aux *clickstreams* du Web basés sur des périodes de temps. Nos premiers résultats sur l'ensemble de données sont prometteurs. Ces analyses nous permettent d'identifier les meilleures heures pour certaines stratégies de vente, comme les promotions rapides, les conseils en ligne, les pubs par bannières, etc. D'autres analyses pourraient être réalisées (cf. le tableau 3) dans l'avenir, exploitant par exemple le lien entre les activités du consommateur et les périodes de temps par magasin ou se concentrant sur les visites d'utilisateurs en plusieurs magasins.

Références

ABCInteactive.com (2004): Spiders and Robots. http://www.abcinteractiveaudits.com/abci_iab_spidersandrobots/.

Ambroise, C., G. Seze, F. Badran, et S. Thiria, (2000). Hierarchical clustering of Self-Organizing Maps for cloud classification. *Neurocomputing* 30, pp. 47-52.

Arnoux, M., Y. Lechevallier, D. Tanasa, B. Trousse, et R. Verde (2003). Automatic Clustering for the Web Usage Mining. In *Proc. of the 5th Intl. Workshop on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC03)*, Ed. Mirton, Timisoara, pp. 54-66.

Bock, H. H. (1993). Classification and clustering: Problems for the future. In: E. Diday, Y. Lechevallier, M. Schader, P. Bertrand and B. Burtschy (eds.): *New Approaches in Classification and Data Analysis*, Springer, Heidelberg, pp. 3-24.

Ciampi, A. et Y. Lechevallier (2000). Clustering large: An approach based on Kohonen Self Organizing Maps, In *Proc. of PKKD 2000*, Springer-Verlag, Heidelberg, pp. 353-358.

Cooley, R. (2000): Web Usage Mining: Discovery and Application of Interesting Patterns From Web Data. PhD Thesis, Dept of Computer Science, Univ. of Minnesota.

Gordon, A. D. (1981). *Classification: Methods for the Exploratory Analysis of Multivariate Data*, Chapman& Hall, London.

Govaert, G. (1977): Algorithme de classification d'un tableau de contingence. In Proc. of first international symposium on Data Analysis and Informatics, INRIA, Versailles, pp. 487-500.

Govaert, G. et M. Nadif (2003). Clustering with block mixture models. *Pattern Recognition* 36, Elservier Science Publishers, pp. 463-473.

Hébrail, G. et A. Debregeas (1998). Interactive interpretation of Kohonen maps applied to curves. *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*. AAAI press, Menlo Park pp. 179-183.

Jaczynski, M. (1998): Scheme and Object-Oriented Framework for case Indexing By Behavioural Situations: Application in Assisted Web Browsing. PhD Thesis, University Nice of Sophia-Antipolis.

Jaczynski, M. and Trousse, B. (1998): WWW Assisted Browsing by Reusing Past Navigations of a Group of Users. In Advanced in Case-based Reasoning, 4th European Workshop on Case-Based Reasoning, LNAI 1488, pp. 160-171.

Kohavi, R. (2001). Mining E-Commerce Data. KDD 01, San Francisco CA, USA.

Kohonen, T. (1997). Self-Organizing Maps. Springer, New York.

Lechevallier, Y. et R. Verde (2004). Crossed Clustering method: An efficient Clustering Method for Web Usage Mining. In *Complex Data Analysis*, Pekin, Chine.

Mobasher, B. (2000): Mining Web Usage Data for Automatic Site Personalization. In: Proc. 24th Annual Conference of the Gesellschaft Fur Klassifikation E.V., University of Passau, pp. 15–17.

Perkowitz, M. et O. Etzioni (1997). Adaptive sites: Automatically learning from user access patterns. In *Proc. 6th Int'l World Wide Web Conf.*, Santa Clara, California.

Raghavan, S. NR. (2005). Data Mining in e-commerce: A survey. *Proceedings in Engineering Sciences*, v. 30, parts 2 & 3, Indian Academy of Sciences, pp. 275-289.

Sauberlich, F. et K.-P. Huber (2001). A Framework for Web Usage Mining on Anonymous Logfile Data. In *Exploratory Data Analysis in Empirical Research*, Springer-Verlag, Heidelberg, pp. 309-318.

Srivastava, J., R. Cooley, M. Deshpande et P.-N. Tan (2000): Web usage mining: Discovery and applications of usage patterns from Web data. *ACM SIGKDD Explorations*, v.1 (2).

Tanasa, D. et B. Trousse (2004). Advanced Data Preprocessing for Intersites Web Usage Mining. *IEEE Intelligent Systems*, v. 19(2):59-65.

Tanasa, D. (2005). *Web Usage Mining: Contributions to Intersites Logs Preprocessing and Sequential Pattern Extraction with Low Support*, Thèse de doctorat, Université de Nice Sophia Antipolis.

Trousse, B., M. Jaczynski et K. Kanawati (1999). Using User Behavior Similarity for Recommendation Computation: The Broadway Approach, In *Proceedings of 8th international conference on human computer interaction (HCI'99)*, Munich.

Verde, R., F.A.T. De Carvalho et Y. Lechevallier (2000). A Dynamical Clustering Algorithm for Multi-Nominal Data. In *Data Analysis, Classification, and Related Methods*, Springer-Verlag, Heidelberg, pp. 387-394.

Verde, R. et Y. Lechevallier (2003). Crossed Clustering method on Symbolic Data tables. In *New developments in Classification, and Data Analysis*, Springer-Verlag, Heidelberg, pp. 87-96.

W3C (1995). Logging Control In *W3C httpd*. http://www.w3.org/Damon/User/Cofig/Logging.html#common-logfile-format.

Summary

The objective of this article is to propose an original preprocessing method applied together with a clustering method on complex data represented by Web users' activities on several e-commerce Web sites. The first contribution concerns the preprocessing with the objective of building a rich data warehouse which integrates the intersites aspect and offers a better data structuration. The second contribution consists in presenting some statistical analyses. These analyses were applied on the structured data and were based on the notion of "time period". Then, an efficient crossed-clustering method was applied on these complex data extracted from our data warehouse, application which is original in the context of Web Usage Mining. Our preliminary results are promising and they illustrate the benefits of our approach in the WUM field.

Saada : Un Générateur Automatique de Bases de Données Astronomiques

Ngoc Hoan Nguyen, Laurent Michel, Christian Motch

11 Rue de l'université, 67000 Strasbourg
nguyen@astro.u-strasbg.fr
http://amwdb.u-strasbg/saada

Résumé. Cet article présente le générateur automatique de bases de données hétérogènes astronomiques nommé Saada (Système d'Archivage Automatique des Données Astronomiques). Chaque base générée qui est capable d'héberger simultanément des spectres, des catalogues, des séries temporelles et qui possède un moteur de requête permettant de chercher rapidement les objets astronomiques par des motifs de corrélation.

1 Introduction

L'amélioration des détecteurs astronomiques utilisés au sol et dans l'espace a conduit ces dernières années à une inflation importante du volume et de la complexité des données générées par les instruments d'observation. Les grands observatoires tels que l'Observatoire Européen Austral[1] maintiennent en général des bases de données d'archives contenant les observations faites par les plus grands instruments. Cependant, de par leurs structures, ces bases n'ont en général pas la capacité de représenter toute la complexité et la richesse des données. D'autre part, les données produites par de nombreuses instrumentations astronomiques, en particulier celles gérées par de petites équipes, restent souvent inaccessibles via un système moderne de base de données. Certains programmes génèrent aussi des ensembles de données d'observations à fortes valeurs scientifiques ajoutées grâce à une analyse spécifique ou par la création de relations entre elles comme des liens entre les entrées de catalogues et les images obtenues à différentes époques ou à différentes longueur d'ondes. En parallèle, l'augmentation de la puissance de calcul permet la génération de simulations (Teuben, et al., 2001) de plus en plus complexes qui doivent être comparées avec des données d'observations elles aussi de plus en plus riches. Enfin, pour assurer leur visibilité internationale, ces données devront être à terme facilement publiables (McDowell, 2003) dans le cadre émergeant de l'Observatoire Virtuel[2]. Le besoin d'un outil informatique aisément déployable, interfaçable avec les grandes bases de données astronomiques actuelles et capable de représenter et d'exploiter toute la richesse des données modernes se fait donc sentir de façon aiguë.

Ce besoin peut être délimité par les points suivants :

[1] ESO -http://www.eso.org
[2] International Virtual Observatory Alliance -http://www.ivoa.net

- Grâce aux performances atteintes par le réseau Internet, les chercheurs peuvent partager leurs données avec la communauté grâce à des bases de données locales.
- Les chercheurs n'ont pas seulement besoin de bases de données comme outils de recherches multicritères. Ces bases de données peuvent aussi être intégrées dans des chaînes d'archivage, de calculs ou de simulation. Le processus d'un chargement des données peut être rendu automatique en employant des scripts insérés dans une boucle de simulation ou dans un pipeline de traitement de données. Les produits ingérés peuvent alors être vérifiés directement par des requêtes complexes ou d'autres calculs utilisant les interfaces fournies par Saada.
- La gestion de données hétérogènes peut être facilitée par une implémentation performante des liens entre objets astronomiques tels que des relations entre des données observées et des données simulées ou entre des spectres et des images par exemple.
- Les astronomes ont besoin d'un moyen simple pour publier leurs données dans le monde de l'Observatoire Virtuel, grand projet mondial porteur de l'Astronomie de demain.

Le développement de bases de données contenant des fonctionnalités de haut niveau scientifique par des équipes d'astronomes dont ça n'est pas le métier n'est pas un travail simple. Ce type de développement n'est en particulier pas concevable pour des projets à court terme car il accaparerait trop de ressources. De plus, un tel travail se heurte aux difficultés suivantes:

- La modélisation efficace de données hétérogènes contenant des spectres, des images, des sources et des séries temporelles étant elles-mêmes de formats hétérogènes est une tâche délicate.
- La description des objets astronomiques peut être enrichie de relations entre données dont l'implémentation efficace reste très ardue.
- Il n'y a pas encore de moteur de requêtes de haut niveau capable de traiter rapidement et facilement des requêtes comprenant plusieurs contraintes astronomiques classiques et plusieurs contraintes sur des relations complexes. Ainsi, comment un utilisateur peut-il exprimer facilement des requêtes de ce type :

  " *Trouver des sources situées à moins de 60 secondes d'arc d'un objet donné qualifié par une vitesse radiale > 100 km/sec et corrélées à moins de 3 secondes d'arc avec au moins 1 source radio ou infrarouge".*
- La quatrième difficulté est le traitement performant des requêtes contenant des jointures entre deux tables de données relationnelles. En particulier quand les liens de corrélations entre deux tables s'accroissent, la performance des requêtes diminue proportionnellement.

Actuellement et à notre connaissance, aucun autre système généraliste n'a été conçu pour permettre aux astronomes de créer et de déployer facilement leurs bases de données hétérogènes locales contenant des fonctionnalités comme le chargement automatique des données, un moteur de requêtes de haut niveau ou encore des capacités d'interopérabilité.

2 Approche et Développement

Ce travail apporte dans ce contexte une solution originale. Saada[3] (Nguyen, 2006), (Nguyen et al., 2003) et (Michel et al., 2005) est un générateur automatique de bases de données astronomiques répondant aux problèmes évoqués ci-dessus. Saada vise à automatiser la procédure de création et d'exploitation de telles bases de données. La conception de ce système générique de gestion de données hétérogènes s'articule autour des composants suivants :

(i) Modèle de données commun flexible: il permet de mettre en valeur le contenu scientifique des données hébergées en proposant une organisation basée sur les notions de classes, de collections et en gérant des relations qualifiées entre enregistrements. Il permet également la gestion simultanée de spectres, d'images, de sources et de séries temporelles dans la même base de données. Ce modèle est extensible afin de pouvoir s'adapter aux besoins de l'utilisateur. De nouveaux attributs peuvent lui être ajoutés sans avoir à écrire de code.

(ii) Couche de gestion des données Objet-Relation: Cette couche contient un service de persistance d'objets qui établit la correspondance entre le modèle de données à objets, les tables relationnelles et un cache d'objets améliorant les performances d'accès aux données. Pour des raisons de performance et d'efficacité des fonctions de bas niveau nous avons développé notre propre service.

(iii) Auto-configuration: La première mission de ce module est de configurer et de générer automatique les composantes du modèle de données. Son principal rôle est de prendre en charge la classification et l'identification automatique des produits à archiver en analysant son contenu et en appliquant des règles données par l'opérateur. Il prend aussi en charge le contrôle du processus de chargement automatique des données.

(iv) Langage de requêtes SaadaQL[4]: Ce langage manipule aisément des requêtes incluant des contraintes astronomiques classiques et des motifs de corrélation sur le modèle de données généré sans qu'il soit nécessaire de connaître en détail la structure cachée du système relationnel. Il a également été conçu pour faciliter l'édition de requêtes d'intérêt astronomique à partir d'une interface Web.

(v) Moteur de requêtes SaadaQL: Ce moteur est au coeur de notre système. Il peut traiter rapidement des requêtes SaadaQL contenant plusieurs contraintes complexes et aussi portant sur les motifs de corrélation formés par les relations dans lesquelles les objets recherchés sont impliqués. Ce moteur est utilisé pour tous les accès aux données (Interface Web, API, et Interface avec l'Observatoire Virtuel). Une méthode originale d'indexation en mémoire lui évite d'avoir à traiter des jointures dans le traitement des liens de corrélations pour obtenir de bonnes performances de recherche.

(vi) Interface avec l'Observatoire Virtuel (OV): Un module externe permet d'accéder aux bases crées par Saada en utilisant les protocoles standard d'accès (Shirasaki, et al. 2005) développés dans le cadre de l'OV. Les requêtes de ces types sont converties en SaadaQL avant d'être traitées par le moteur de requêtes.

---

[3] Projet Saada : http://amwdb.u-strasbg.fr/saada. Ce travail est co-financé par le Centre National d'Etudes Spatiales-CNES et la Région Alsace.

[4] http://amwdb.u-strasbg.fr/saada/saadaql.html

Les bases de donnée générées par Saada sont des bases locales ayant une structure commune (modèle de données, langage de requêtes, Auto-Configuration,..) mais des fonctions scientifiques propres (simulation, statistique, calculs scientifique) définies par l'utilisateur. Ces bases possèdent une grande interopérabilité.

3 Génération d'une base de données hétérogènes

Afin de rendre le plus simple possible la génération d'une base de données locale nous allons essayer de masquer autant que possible à l'opérateur la structure logicielle au profit du contenu scientifique des produits.

Les bases générées par Saada sont nommées Saada-DBs. Les Saada-DBs ont toutes la même architecture d'intégration objet–relation. Les données peuvent être hébergées par tout système de gestion de base de données relationnel supportant JDBC[5]. La technologie à objets est utilisée pour le traitement des données hétérogènes. Les Saada-DBs peuvent gérer simultanément des images, des spectres, des séries temporelles et des listes de sources. Les données hébergées dans les Saada-DBs sont accessibles facilement par : Interface Web, API en Java et Protocole VO.

Etapes pour générer une Saada-DB :

1. Installation de Saada.
2. Génération de la Saada-DB vide.

Ici, la Saada-DB est indépendante de Saada qui n'aura plus à être utilisé.

3. Configuration du modèle de données.
4. Chargement et mise à jour des données.
5. Génération et déploiement de l'interface Web.

Ici, l'utilisateur peut exploiter la base générée en utilisant les requêtes SaadaQL

6. Définition et chargement des corrélations (optionnel)

Ici, les données sont interconnectées et l'utilisateur peut chercher des objets isolés ou corrélés

7. Publication des collections de données dans l'Observatoire Virtuel (optionnel)

Ici, Saada-DB peut être connectée à l'Observatoire Virtuel. L'utilisateur peut utiliser les protocoles standard d'accès à l'OV pour exploiter la base.

4 Conclusion

Saada est un outil de transformation rapide des données non-organisées en données recherchables et interconnectées. Saada est un système opérationnel qui a été évalué par l'European Southern Observatory, qui est utilisé pour l'exploitation des deux catalogues de données du satellite européenne XMM-Newton par l'équipe scientifique de Strasbourg, et qui propose les solutions réelles qui correspondent à des besoins réels.

[5] Java Database Connectivity (JDBC)

Références

McDowell, J.C. (2003), *Small Theory Data in the Virtual Observatory*, in ASP Conf. Ser., Vol. 295 ADASS XII (San Francisco: ASP), 61.

Michel .L, Nguyen N.H., and Motch C. (2005), *Saada: Astronomical Databases Made Easy*, ASP (San Francisco: ASP) (In press).

Nguyen N.H, Michel L., and Motch C. (2003*), Saada: An Automatic Archival System for Astronomical Data*, in ASP Conf. Ser., Vol. 314 ADASS XIII (San Francisco: ASP), 121

Nguyen N.H. (2006), *Conception et Réalisation d'un générateur de bases de données astronomiques: Saada*. Thèse de doctorat, Université de Louis Pasteur Strasbourg 1.

Teuben P., DeYoung D., Hut P., Levy S., Makino J., McMillan S., Zwart S. P., Slavin S. (2001), *Theory in a Virtual Observatory.* Astrophysic*s*, abstract astro-ph/0111478.

Shirasaki Y.,Ohishi M., Mizumoto Y., Tanaka M., Honda S., Oe M.(2005), *Structured Query Language for Virtual Observatory*. ADASS XIV P1-1-23 ASP Conf Ser., Vol. XXX.

Summary

We present new software called Saada-An Automatic Astronomical Database Generator. Project Saada aims to give to teams in astronomy a tool making it possible to build and deploy very easily databases containing heterogeneous data from observational and simulated data without creation code. Saada can simultaneously host spectra, images, source lists and plots by only providing a limited number of product mapping rules. Data sets can be correlated one with each other using qualified links. The query engine is based on a language well suited to the data model which can handle constraints on correlation patterns in addition to classical astronomical queries.

Trois stratégies d'évolution pour la pondération automatique d'attributs en classification non supervisée d'objets complexes

Pierre Gançarski*, Alexandre Blansché*

*LSIIT-AFD
UMR CNRS/ULP 7005 Strasbourg
Bd S. Brant, BP 10413, F-67412 Illkirch
{Pierre.Gancarski, Alexandre.Blansche}@lsiit.u-strasbg.fr,
http://lsiit.u-strasbg.fr/afd

Résumé. Les données devenant de plus en plus complexes, la sélection ou la pondération des attributs prennent une place de plus en plus importante dans le processus de fouille de données et en particulier dans la phase de classification automatique. Dans ce papier nous proposons une nouvelle méthode de clustering intégrant des mécanismes de pondération d'attributs basés sur des algorithmes génétiques utilisant des stratégies d'évolution différentes. Nous présentons des versions évolutionnaires construites sur des stratégies d'évolution inspirées des théories d'évolution darwinienne et lamarckienne ainsi que de l'effet baldwinien. Pour évaluer la qualité des solutions proposées lors de l'évolution, nous utilisons comme fonction de fitness la fonction de coût définie dans l'algorithme `Weighting-`Kmeans. Nous comparons ces méthodes à l'algorithme `Weighting-`Kmeans sur deux jeux de données. Les résultats montrent que sur ces données, nos approches sont toujours supérieures à l'algorithme `Weighting-`Kmeans.

1 Introduction

Les données devenant de plus en plus complexes, la sélection ou la pondération des attributs prennent une place de plus en plus importante dans le processus de fouille de données et en particulier dans la phase de classification automatique. En effet, lorsque les objets sont décrits par un grand nombre d'attributs plus ou moins complexes, beaucoup d'entre eux sont redondants, non pertinents et bien souvent bruités. De nombreuses méthodes de sélection de variables (angl. "feature selection") existent et consistent à extraire le sous-ensemble des attributs discriminant « au mieux » les classes. Ceci revient en fait à pondérer les attributs de façon binaire. Or dans Wettschereck et al. (1997) il a été montré que des méthodes utilisant des pondérations continues (angl. "feature weigthing") produisent de meilleurs résultats que les méthodes de sélection.

De nombreuses méthodes de pondération d'attributs ont été proposées avec différents aspects (Howe et Cardie, 1999; John et al., 1994; Wettschereck et Aha, 1995; Blum et Langley,

1997) mais la majorité des celles-ci sont supervisées ou utilisent une pondération globale sur l'ensemble des données. Or dans Howe et Cardie (1997) et Frigui et Nasraoui (2004) il a été montré que l'utilisation d'une pondération locale par classe à extraire est plus efficace : l'importance relative des attributs changeant en fonction de la classe à extraire. Par exemple, un géographe désirant mettre en évidence les classes d'eau dans une image de télédétection utilisera principalement des informations radiométriques. D'un autre côté, pour discriminer deux types de zones urbaines, il utilisera avant tout des informations de texture.

Enfin, il a été montré dans le cadre de la classification supervisée que l'approche intégrée (angl. "wrapper") permet d'obtenir de meilleurs résultats que l'approche par pré-traitement (angl. "filter"), grâce au retour d'informations de la classification (Wettschereck et Aha, 1995).

Il existe peu de méthodes intégrées de pondération locale d'attributs.

La méthode `Weighting-`Kmeans présentée dans Chan et al. (2004) est basée sur une mesure de (dis)simimilarité pondérée et utilise les principes de Kmeans pour classifier les objets et dans un même temps, déterminer les pondérations « optimales ».

Nous proposons une méthode de classification automatique originale qui combine l'approche Kmeans pour l'allocation des objets aux clusters et un algorithme génétique d'optimisation des pondérations, le critère d'optimisation globale étant la fonction de coût définie dans `Weighting-`Kmeans.

En fait, nous avons défini plusieurs algorithmes génétiques basés sur des stratégies d'évolution différentes : darwinienne, larmarckienne ou basée sur l'effet baldwinien.

Dans la suite du papier, après avoir succinctement décrit l'algorithme `Weighting-`Kmeans (section 2), nous présentons la version évolutionnaire (section 3) de plusieurs algorithmes de pondération inspirés de l'approche darwiennene, lamarckienne ou baldwinienne. Enfin, nous comparons toutes ces méthodes principalement sur le jeu de données Iris Plants de l'UCI Repository Datasets (section 4). Nous concluons ce papier dans la section 5.

Notation

- S est l'ensemble des données à classer ;
- K est le nombre de classes à extraire ;
- n est le nombre d'attributs.

2 `Weighting-`Kmeans

`Weighting-`Kmeans (Chan et al., 2004) est une méthode de classification basée sur Kmeans (MacQueen, 1965) intégrant une pondération automatique des attributs. L'idée est d'optimiser la fonction de coût $WCost$ définie comme suit :

$$WCost(c, W) = \sum_{k=1}^{K} \sum_{o \in S} \sum_{j=1}^{n} C_k(o) w_{(k,j)}^{\beta} d_j(o, c_k), \tag{1}$$

où

- $c = \{c_k\}_{k \in [1,K]}$ (ensemble des centres) et $W = \left\{w_{(k,j)}\right\}_{j \in [1,n], k \in [1,K]}$ sont les inconnus : $w_{(k,j)}$ étant le poids de l'attribut j pour le cluster k ;

- $\beta \geq 1$ (par défaut $\beta = 1.8$, voir (Chan et al., 2004; Huang et al., 2005) pour plus de détails) ;
- $d_j(o, c_k)$ est une mesure de dissimilarité entre l'objet o et le centre c_k sur le j-ème attribut ;
- $C = \{C_1, C_2, \ldots, C_K\}$ est l'ensemble des clusters donnés en extension. Il est calculé classiquement à partir de c en utilisant les pondérations sur les attributs lors de la mesure de distance entre un objet o et les centres des clusters comme suit :

$$C_k(o) = \begin{cases} 1 & \text{si } \sum_{j=1}^{n} w_{(k,j)} d_j(o, c_k) < \sum_{j=1}^{n} w_{(k',j)} d_j(o, c_{k'}) \quad \text{pour } 1 \leq k' \leq K, k' \neq k \\ 0 & \text{sinon} \end{cases} \tag{2}$$

L'optimisation se fait par itération de trois étapes.

Les deux premières sont identiques à Kmeans : affectation des objets aux clusters puis recalcul des centres. La seule différence réside dans le fait que l'algorithme utilise une distance pondérée par les poids des attributs donnés par W (Eq. 2). Par extension, dans la suite, chaque référence à Kmeans, référencera en fait cette version pondérée.

La dernière étape consiste à modifier les poids des attributs pour chaque cluster suivant la formule ci-dessous :

$$w_{(k,j)} = \begin{cases} 1/zero_k & \text{si } sum_k^j = 0 \\ 0 & \text{si } sum_k^j \neq 0, \text{ et } zero_k \neq 0 \\ 1/\sum_{t=1}^{n} \left[\frac{sum_k^j}{sum_k^t} \right]^{1/(\beta-1)} & \text{si } zero_k = 0. \end{cases} \tag{3}$$

où $sum_k^j = \sum_{o \in S} C_k(o) d_j(c_k, o)^2$ et $zero_k = |\{t \mid sum_k^t = 0\}|$.

De fait, une exécution de `Weighting`-Kmeans produit donc en plus d'un partitionnement, les poids « optimaux » (minimum local) utilisés pour le produire.

3 Approches génétiques pour la pondération d'attributs

Les premiers algorithmes génétiques ont été présentés par (Holland, 1975) comme étant une manière de résoudre des problèmes pour lesquels il n'existe pas de solutions classiques (réellement) implantables. Ce sont des algorithmes (Goldberg, 1989) basés sur des heuristiques de recherche et d'optimisation inspirées de l'évolution naturelle. A ce titre, plusieurs théories d'évolution peuvent être considérées :

- La théorie de Darwin dans laquelle les individus s'adaptent à leur environnement. Cette adaptation est sous contrôle des gènes hérités. Seuls les individus présentant une « bonne » adaptation survivent et se reproduisent

– L'approche de Lamarck est différente car elle considère que cette adaptation a des effets directs sur les gènes et donc qu'il existe un héritage de caractère acquis (cette approche est utilisée en algorithmie génétique bien que fausse dans la vie réelle)
– L'approche de Baldwin (Baldwin, 1896) est similaire à celle de Darwin, sauf qu'elle introduit en plus la notion de plasticité pour le phénotype (angl. "phenotypic plasticity") : cette plasticité est définie comme étant la flexibilité et la créativité de l'individu à s'adapter.

Dans notre approche, il n'y a qu'une population chargée de trouver la partition optimale. Cette population cherche :
– un ensemble de centres de classe
– un ensemble associé de pondérations locales
qui minimisent la fonction de coût $WCost$ (Eq. 1).

Les trois méthodes que nous proposons sont décrites dans les sections suivantes. Elles suivent toutes le schéma général suivant.

Après une initialisation :

1. Chaque individu i construit une solution globale : un chromosome $W^i(t)$ à la t-ième génération contient l'ensemble des pondérations nécessaires pour produire une telle solution via l'algorithme Kmeans. Un gène $w^i_{k,j}(t)$ est donc la pondération sur l'attribut j pour le cluster k utilisée par Kmeans (Fig. 1).
2. Les chromosomes issus de la phase de calcul précédente sont éventuellement affectés aux individus (approche lamarckienne)
3. Les individus sont évalués par la fonction $WCost$
4. Les chromosomes correspondant aux meilleurs individus sont utilisés dans une phase de reproduction classique

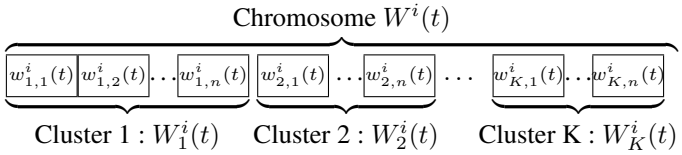


FIG. 1 – *Un chromosome*

Notons $c(t)$ l'ensemble de centres qui associé à l'ensemble des pondérations $W(t)$ fournit la partition « optimale » au début de la t-ième génération.

Dans la suite, le couple ($c(t)$,$W(t)$) sera appelée *meilleure partition courante*.

3.1 Approche évolutive darwinnienne

Notre processus évolutionnaire darwinien est implanté de façon classique :
– *Étape 0 - initialisation*
L'algorithme est initialisé par une exécution de Kmeans initialisé avec un vecteur de centres initiaux $c(0)$ aléatoirement choisi et un vecteur de poids $W(0)$ où tous les poids valent $\frac{1}{n}$.

Puis, à chaque génération t, le processus consiste en trois phases :

– *Étape 1 - Production et évaluation des solutions proposées par les individus*

1. La solution $c^i(t)$ correspondant à un individu I^i est produite en exécutant Kmeans initialisé à partir de $c(t)$ de la meilleure partition courante et en utilisant le vecteur de poids $W^i(t)$ de l'individu.
2. Chaque solution (individu) est évaluée par la fonction $WCost$ (Eq. 1)
   Soit i_{max}, l'individu d'évaluation maximale et $(c^{i_{max}}(t), W^{i_{max}}(t)))$ sa partition.

– *Étape 2 - Construction de la meilleure solution courante.*

- Si cette partition $(c^{i_{max}}(t), W^{i_{max}}(t)))$ est meilleure que la partition $(c(t), W(t)))$ c'est-à-dire qu'elle minimise la fonction de coût $WCost$, alors elle est conservée comme meilleure partition pour la génération suivante :

$$(c(t+1), W(t+1)) = (c^{i_{max}}(t), W^{i_{max}}(t))$$

- sinon, la partition courante est conservée pour la génération suivante :

$$(c(t+1), W(t+1)) = (c(t), W(t))$$

– *Étape 3 - Reproduction* La reproduction se fait en utilisant les opérateurs et les taux classiques de croisement et de mutation.

3.2 Approche évolutionnaire lamarckienne

Nous avons étendu la méthode de pondération précédente à une approche lamarckienne. Pour donner aux individus une chance d'apprendre de nouvelles pondérations, nous avons choisi la méthode `Weighting-`Kmeans comme "lifetime learning method". Elle est utilisée lors de l'étape d'initialisation des individus et lors de la construction des solutions.

Ainsi, les individus modifient leur génotype via cet algorithme : pour chaque individu i les poids proposés par `Weighting-`Kmeans remplacent les gènes de l'indivu. La qualité d'un individu i est alors donnée par l'évaluation de la partition produite par `Weighting-`Kmeans $W(t)$:

$$WCost((c^i(t), W^i_{new}(t)))$$

où $W^i_{new}(t)$ est le nouveau chromosome de l'individu i. L'étape de reproduction se fait à partir de ce nouveau matériel génétique.

De fait notre processus d'apprentissage lamarckien est identique au processus darwinien excepté que :

– l'algorithme utilisé est `Weighting-`Kmeans et ce dans toutes les étapes ;
– à chaque génération, après calcul du résultat d'un individu, les nouveaux poids remplacent les pondérations initiales de l'individu : à chaque individu est associé un nouveau chromosome.

3.3 Approche évolutionnaire baldwinienne

Enfin nous avons combiné les deux méthodes précédentes pour définir une approche évolutionnaire baldwinienne. Pour évaluer la "phenotypic plasticity" de chaque individu, nous

utilisons la même fonction d'apprentissage `Weighting-`Kmeans que dans l'approche lamarckienne. La potentialité d'adaptation d'un individu i est alors donnée par l'évaluation de la partition produite par `Weighting-`Kmeans :

$$WCost((c^i(t), W_{new}(t)))$$

où $W_{new}(t)$ est le vecteur formé des poids proposés par `Weighting-`Kmeans.

Par contre, contrairement à l'approche lamarckienne, l'étape de reproduction se fait à partir du matériel génétique $\{W^i(t)\}$ initial.

De fait notre processus d'apprentissage baldwinien est identique au processus lamarckien excepté qu'à chaque génération, après calcul du résultat d'un individu, les nouveaux poids ne remplacent pas les pondérations initiales de l'individu : chaque individu conserve son chromosome initial. Ces nouveaux poids sont par contre utilisés pour évaluer l'individu.

4 Résultats

Nous avons testé nos solutions entre autres, sur deux jeux de données issus de l'UCI Repository Datasets (D.J. Newman et Merz, 1998) : le jeu de données sur les Iris et celui sur les "Balance Scale Weight and Distance". Comme toutes ces données sont à attributs numériques, nous avons utilisé une fonction de distance classique sur les attributs. Pour chaque jeu de données, nous avons fait les expériences sur les données brutes puis sur des données normalisées (les attributs sont alors sur [0,1]).

Dans cette section nous comparons les résultats obtenus avec `Weighting-`Kmeans avec ceux obtenus par les méthodes darwinienne (`DEA`), lamarckienne (`LEA`) et baldwinienne (`BEA`). A chaque test, les centres initiaux de Kmeans et `Weighting-`Kmeans ont été choisis aléatoirement parmi les objets du jeu de données. Les principaux paramètres utilisés ont été les suivants :

- pour les taux de mutation et de croisement nous avons pris les taux habituels ;
- le paramètre β de `Weighting-`Kmeans est égal à 1.8 ;
- le nombre de tours pour `Weighting-`Kmeans est égal à 1 ;
- le nombre K de clusters est donné avec le jeu de données.
- le nombre d'individus est égal à $K \times 20$

Dans la suite, toutes les valeurs sont des moyennes sur plus de 100 exécutions.

4.1 Les iris

Avec ce jeu de données (Iris Plants), les auteurs donnent quelques indications :

- sur la description des données : 3 classes (Iris Setosa, Iris Versicolour et Iris Virginica), 4 attributs numériques en cm (sepal length, sepal width, petal length, petal width) ; 150 instances (50 de chacune des trois classes)
- sur des statistiques concernant les données et les corrélations entre elles et les classes (voir Tab. 1) .

	Min	Max	Mean	σ	Indice de corrélation avec la classe
sepal length	4.3	7.9	5.84	0.83	0.7826
sepal width	2.0	4.4	3.05	0.43	-0.4194
petal length	1.0	6.9	3.76	1.76	**0.9490**
petal width	0.1	2.5	1.20	0.76	**0.9565**

TAB. 1 – *Iris Plants : Quelques statistiques*

4.1.1 Critères de qualité internes

De nombreux critères ont été proposés dans la littérature pour évaluer la qualité interne du résultat d'une classification non supervisée telles que la compacité ou l'inertie intra-classe (Bolshakova et Azuaje, 2003; Günter et Burke, 2001; Halkidi et al., 2001). D'autres méthodes (Bel Mufti et Bertrand, 1997; Levine et Domany, 2001; Tibshirani et al., 2000) telle celle basée sur la stabilité des clusters sont aussi souvent utilisées, en particulier pour trouver le bon nombre de clusters. Dans notre cas, d'une part nous disposons avec chaque jeu de données, du nombre de classes attendues. D'autre part, toutes les méthodes que ce soit `Weighting-`Kmeans ou les autres, utilisent un critère de qualité intrinsèque qu'est la fonction de coût WCost (Eq. 1) qui se rapproche beaucoup du critère habituel basé sur l'inertie intra-classe. Nous l'avons donc utiliser pour évaluer la qualité interne des résultats.

Sur la figure 2-(a), nous avons représenté les différentes valeurs de la fonction $WCost$ des résultats obtenus à partir de données brutes par `Weighting-`Kmeans et par nos différents méthodes (pour chaque algorithme génétique, nous avons sélectionné la partition produite par le meilleur individu). La courbe représente les 10 premières générations ou itérations sur les 100 calculées. Plus la valeur de la fonction est basse, meilleur est le résultat.

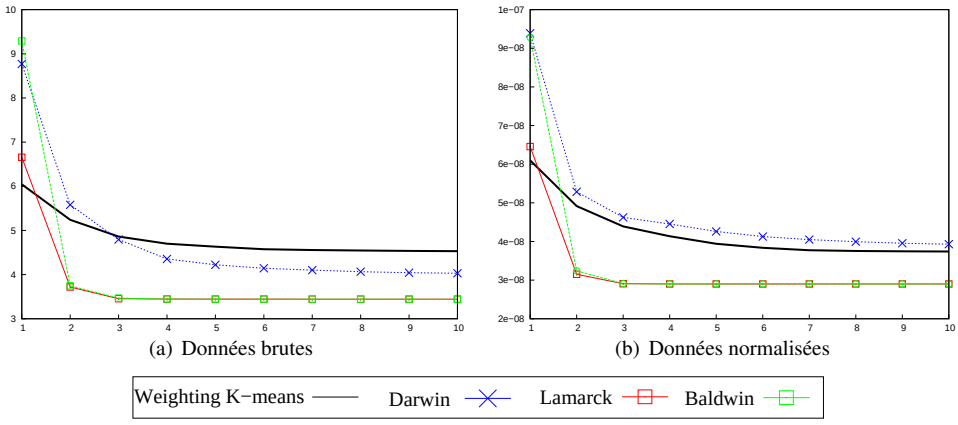


FIG. 2 – *Iris : Evolution de la fonction WCost pour les différents algorithmes*

La figure 2-(b) montre les valeurs obtenues à partir des données normalisées.

Nous avons aussi utilisé l'indice de qualité d'un partitionnement WG_q (Wemmert, 2000; Wemmert et al., 2000) défini à partir du critère de qualité des clusters r_k suivant :

$$r_k = \begin{cases} 0 \quad \text{si } \frac{1}{\|C_k\|} \sum_{o \in C_k} \sum_{j=1}^{n} \frac{w_{(k,j)} d_j(o, c_k)}{w_{(k',j)} d_j(o, c_{k'})} > 1 \\ 1 - \frac{1}{\|C_k\|} \sum_{o \in C_k} \sum_{j=1}^{n} \frac{w_{(k,j)} d_j(o, c_k)}{w_{(k',j)} d_j(o, c_{k'})} \quad \text{sinon} \end{cases}$$

où c_k est le centre de C_k et $c_{k'}$ est le centre centre le plus proche de o différent de C_k.

Cet indice tient donc compte pour chaque classe C_k de l'inertie intra-classe (numérateur) et de la distance entre les objets de la classe et les centres des autres classes (dénominateur) : cet indice r_k est d'autant meilleur que sa valeur est grande.

L'indice de qualité WG_q est alors défini par :

$$WG_q = \frac{1}{|S|} \sum_{i=k}^{K} |C_k| r_k$$

Le tableau Tab. 2 donnent les valeurs obtenues par chacun des algorithmes.

Méthodes	Données brute	Données normalisées
Kmeans	0.440 ± 0.060	0.411 ± 0.029
Weight.-Kmeans	0.519 ± 0.043	0.560 ± 0.025
DEA	0.552 ± 0.018	0.559 ± 0.018
LEA	0.554 ± 0.002	0.568 ± 0.001
BEA	0.554 ± 0.002	0.568 ± 0.001

TAB. 2 – *Iris Plants : Qualité interne* WG_q

Ces résultats montrent que nos méthodes semblent supérieures à `Weighting-`Kmeans. Il est à noter que `Weighting-`Kmeans est bloqué sur un minimum local dont il ne sortira plus quelque soit le nombre d'itérations supplémentaires. De plus, toutes nos méthodes produisent une meilleure solution dès la troisième génération exceptée la méthode darwinienne qui met 17 générations à surclasser `Weighting-`Kmeans sur les données normalisées. Les approches lamarckienne et baldwinienne sont toujours plus efficaces que l'approche darwinienne.

4.1.2 Etude des pondérations

Pour le jeu de données Iris, nous disposions des coefficients de corrélation des attributs avec les classes (cf. Tab.1). La figure 3 donne la moyenne et variance des poids trouvés pour les attributs sur 100 expériences pour chacune des méthodes. Les expériences montrent que les poids les plus importants sont presque toujours mis sur les deux derniers attributs dans les cas des données normalisées ce qui est conforme aux indications données par les auteurs. Pour les attributs non normalisés, cette étude n'a pas de sens dans le mesure où un poids important sur un attribut à petites valeurs (moyenne basse et variance faible) ne signale pas nécessairement que cet attribut est fortement discriminant. Et réciproquement pour des poids faibles sur des attributs à fortes valeurs et à faible variance.

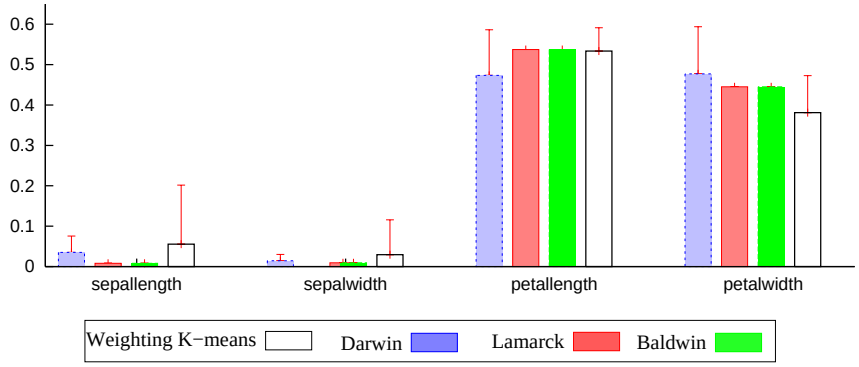


FIG. 3 – *Iris : Moyenne et variance des poids trouvés sur chacun des attributs (100 expériences)*

4.1.3 Critères externes

Enfin, nous avons comparé les clusters obtenus avec les labels donnés avec le jeu de données. Pour cela nous avons utilisé le Rand index et le coefficient de Jaccard décrits dans Halkidi et al. (2001) : ces index ont une valeur de 1 si les clusters trouvés peuvent être mis en relation bijective avec les classes réelles des objets. Les tableaux (voir Tab. 3) donnent les résultats obtenus à partir des données brutes (tableau de gauche) et normalisées (tableau de droite)

Méthodes	Rand Index	Jaccard Index
Kmeans	0.831 $\pm$ 0.068	0.629 $\pm$ 0.089
Weight.-Kmeans	0.825 $\pm$ 0.083	0.622 $\pm$ 0.144
DEA	0.948 $\pm$ 0.006	0.853 $\pm$ 0.017
LEA	0.949 $\pm$ 0.002	0.856 $\pm$ 0.005
BEA	0.949 $\pm$ 0.002	0.856 $\pm$ 0.005

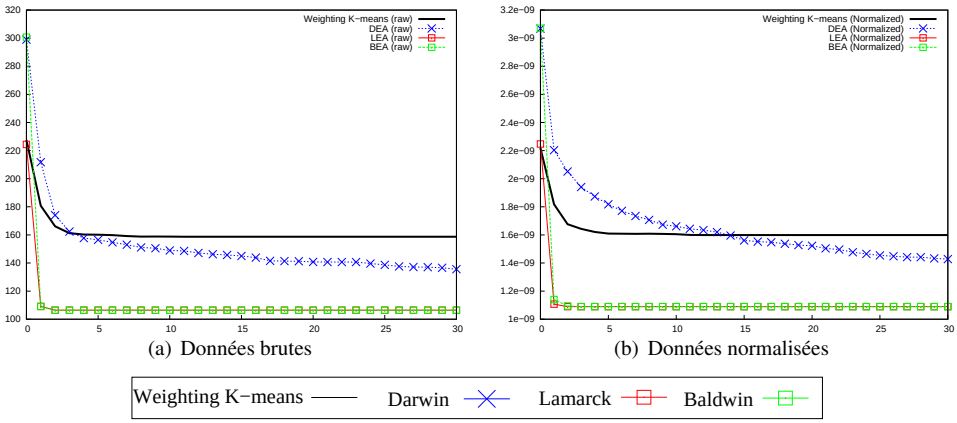
Méthodes	Rand Index	Jaccard Index
Kmeans	0.818$\pm$ 0.055	0.595 $\pm$ 0.064
Weight.-Kmeans	0.904 $\pm$ 0.088	0.782 $\pm$ 0.145
DEA	0.948 $\pm$ 0.008	0.854 $\pm$ 0.021
LEA	0.950 $\pm$ 0.001	0.858 $\pm$ 0.000
BEA	0.950 $\pm$ 0.001	0.858 $\pm$ 0.000

TAB. 3 – *Iris Plants : Qualités externes*

Nous remarquons que nos méthodes sont de qualités identiques et dans tous les cas supérieures à Kmeans et `Weighting`-Kmeans.

4.2 Balance Scale Weight and Distance

Les tests sur les données “Balance Scale Weight and Distance” ont produit des résultats en tous points similaires (Figure 4).



(a) Données brutes (b) Données normalisées

FIG. 4 – *Diabetes : Evolution de la fonction WCost pour les différents algorithmes*

5 Conclusion

Dans ce papier nous avons présenté trois approches génétiques pour la classification automatique intégrant un mécanisme de pondération automatique des attributs. À partir d'une « meilleure » partition courante, les algorithmes calculent un nouveau vecteur de poids via un mécanisme de reproduction guidé par la fonction de coût definie dans Chan et al. (2004) et détermine une nouvelle partition courante. A chaque génération :

– les opérations génétiques sont appliquées aux chromosomes initiaux (approches darwinienne et baldwinienne) ou à des nouveaux chromosomes issus d'une fonction d'apprentissage propre aux individus (approche lamarckienne)
– la fonction de coût prend en compte l'évolution de l'individu au courant de sa vie (approches lamarckienne et baldwinienne) ou non (approche darwinienne).

Les résultats expérimentaux ont montré que ces méthodes surpassent toujours `Weighting-`Kmeans aussi bien au niveau de la fonction de coût qu'en qualité externe. Nous avons aussi remarqué que les résultats sont meilleurs avec une approche lamarckienne ou baldwinienne qu'avec une approche darwinienne. Enfin, nous avons pu observer que ces méthodes, comme `Weighting-`Kmeans, étaient capables de trouver le « meilleur » vecteur de poids, et ce, en particulier pour les données normalisées.

Nous nous intéressons maintenant d'une part à améliorer nos méthodes en utilisant plusieurs populations spécialisées et d'autre part à définir une nouvelle fonction de coût ne nécessitant pas de mesure de distance. Enfin, nous utiliserons la capacité de nos méthodes à trouver les bons attributs pour réduire la dimentionalité des données et ainsi les intégrer dans un processus multi-étape plus général.

Références

Baldwin, J.-M. (1896). A new factor in evolution. *American Naturalist 30*, 441–451.

Bel Mufti, G. et P. Bertrand (1997). Validation d'une classe par rééchantillonnage. In *Cinquième rencontres de la Société Francophone de Classification, Lyon*, pp. 251–254.

Blum, A. et P. Langley (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence 97*(1–2), 245–271.

Bolshakova, N. et F. Azuaje (2003). Cluster validation techniques for genome expression data. *Signal Processing 83*(4), 825–833.

Chan, E., W. Ching, M. Ng, et J. Huang (2004). An optimization algorithm for clustering using weighted dissimilarity measures. *Pattern Recognition 37*, 943–952.

D.J. Newman, S. Hettich, C. B. et C. Merz (1998). UCI repository of machine learning databases.

Frigui, H. et O. Nasraoui (2004). Unsupervised learning of prototypes and attribute weights. *Pattern Recognition 34*, 567–581.

Goldberg, D. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc.

Günter, S. et H. Burke (2001). Validation indices for graph clustering. In *Proc. 3rd IAPR-TC15 Workshop on Graph-based Representations in Pattern Recognition*, pp. 229–238. J.-M. Jolion, W. Kropatsch, M. Vento.

Halkidi, M., Y. Batistakis, et M. Vazirgiannis (2001). On clustering validation techniques. *Journal of Intelligent Information Systems 17*(2-3), 107–145.

Holland, J. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press.

Howe, N. et C. Cardie (1997). Examining locally varying weights for nearest neighbor algorithms. In *ICCBR*, pp. 455–466.

Howe, N. et C. Cardie (1999). Weighting unusual feature types. Technical Report TR99-1735, Ithaca.

Huang, J. Z., M. K. Ng, H. Rong, et Z. Li (2005). Automated variable weighting in k-means type clustering. *IEEE Transactions on Pattern Analysis & Machine Intelligence 27*(5), 657–668.

John, G., R. Kohavi, et K. Pfleger (1994). Irrelevant features and the subset selection problem. In *Proceedings of the Eleventh International Conference on Machine Learning*, pp. 121–129.

Levine, E. et E. Domany (2001). Resampling method for unsupervised estimation of cluster validity. *Neural Computation 13*(11), 2573–2593.

MacQueen, J. (1965). Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA, pp. 281–297. University of California Press.

Tibshirani, R., G. Walther, et T. Hastie (2000). Estimating the number of clusters in a dataset via the Gap statistic. Technical Report 208, Department of Statistics, Stanford University.

Wemmert, C. (2000). *Classification hybride distribuée par collaboration de méthodes non supervisées*. Ph. D. thesis, Université Louis Pasteur, Strasbourg.

Wemmert, C., P. Gançarski, et J. Korczak (2000). A collaborative approach to combine multiple learning methods. *International Journal on Artificial Intelligence Tools 9*(1), 59–78.

Wettschereck, D. et D. Aha (1995). Weighting features. In M. Veloso et A. Aamodt (Eds.), *Case-Based Reasoning, Research and Development, First International Conference*, Berlin, pp. 347–358. Springer Verlag.

Wettschereck, D., D. W. Aha, et T. Mohri (1997). A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artificial Intelligence Review 11*(1-5), 273–314.

Summary

Feature weighting is a step of increasing importance in clustering, since data are becoming more and more complex. We propose a new feature weighting type method based on genetic algorithms. It uses the cost function defined in the embedded local feature weighting method `Weighting-`Kmeans as fitness function. We present new methods which are based on the Darwinian and Lamarckian evolution theories or on the Baldwin effect. For each evolution theory, we describe an evolutionary version. We compare them with hill-climbing optimization Kmeans type algorithms on different datasets. The results show that the proposed methods are always better than the `Weighting-`Kmeans algorithm.

Evaluation d'une approche probabiliste pour le classement d'objets incomplètement connus dans un arbre de décision

Lamis Hawarah, Ana Simonet, Michel Simonet
TIMC - IN3S (INstitut de l'INgenierie et de l'INformation de Sante)
Faculte de Medecine
38700 LA TRONCHE
{lamis.hawarah, ana.simonet,michel.simonet}@imag.fr
http://www-timc.imag.fr/

Résumé. Nous présentons une approche probabiliste pour déterminer les valeurs manquantes des objets incomplets pendant leur classement dans les arbres de décision. Cette approche est dérivée de la méthode d'apprentissage supervisé appelée Arbres d'Attributs Ordonnés proposée par Lobo et Numao en 2000. Notre travail prend en compte les dépendances entre les attributs en utilisant l'information mutuelle et fournit le résultat du classement d'un objet incomplet sous la forme d'une distribution de probabilités (au lieu de la classe la plus probable). Dans cet article, nous testons notre approche sur des bases réelles et nous comparons nos résultats avec ceux donnés par la méthode C4.5 de Quinlan. Nous étudions également les avantages et les inconvénients de chacune de ces méthodes.

1 Introduction

Les arbres de décision sont une des techniques de l'apprentissage automatique couramment utilisées pour la fouille de données. L'apprentissage par arbres de décision se situe dans le cadre de l'apprentissage supervisé, où l'espace des classes est bien défini. Un arbre de décision nous permet, d'une part d'extraire les attributs pertinents dans une base d'apprentissage, et d'autre part on peut utiliser cet arbre pour classer un nouveau cas dont la classe est inconnue. La présence d'une valeur manquante dans les données se traduit par l'absence de réponse pour un ou plusieurs attributs d'un objet. Nous avons étudié ce problème dans le cadre des arbres de décision, et plus précisément lors du classement d'objets avec des valeurs manquantes.

Plusieurs méthodes ont été proposées pour résoudre ce problème: White et al.(1997) et Quinlan (1989). La méthode la plus simple pendant la phase de construction d'un arbre de décision est d'ignorer les instances ayant une ou plusieurs valeurs manquantes. Certaines méthodes remplacent un attribut inconnu par sa valeur la plus probable comme la méthode de majorité (Kononenko et al., 1984). La méthode *surrogate splits,* proposée par Breiman et al. (1984), remplace l'attribut inconnu par un autre attribut, celui qui lui est le plus semblable. On trouve encore l'approche probabiliste proposée par Quinlan (1990, 1993) qui remplace la valeur manquante par une distribution de probabilité, la méthode *Lazy decision tree* (Friedman et al., 1996), la méthode de Shapiro, décrite par Quinlan (1986), et les *Arbres d'Attributs Ordonnés* (Lobo and Numao, 1999, 2000).

Notre travail se situe dans le cadre des arbres de décision probabilistes (Breiman et al.,1984) et (Quinlan, 1986, 1990, 1993). Dans le but d'aider l'utilisateur à prendre sa décision, nous avons choisi de garder sur chaque feuille dans l'arbre de décision toutes les valeurs de classe avec leurs probabilités, c'est à dire utiliser un arbre de décision probabiliste plutôt qu'un arbre de décision classique. Le fait de fournir un résultat probabiliste donne à l'utilisateur une information plus fine sur la proportion des valeurs de classe sur chaque feuille, au lieu de prendre la valeur de classe la plus probable et de considérer que les autres cas qui arrivent à cette feuille et qui n'appartiennent pas à sa classe sont mal classés. Par ailleurs, conserver toutes les valeurs de classe sur chaque feuille est particulièrement important lorsque la classe possède un nombre de valeurs supérieur à deux. En effet, dans le cas d'un arbre de décision classique, on connaît pour chaque feuille la valeur de classe la plus probable, le nombre d'objets arrivant à cette feuille ainsi que le nombre d'objets mal classés, mais on ne connaît pas la distribution des valeurs de classe considérées comme mal classées sur cette feuille (cf note de bas de page de la Fig. 2). Conserver sur chaque feuille toutes les valeurs de classe avec leurs probabilités rend l'arbre plus simple à interpréter.

Notre approche vise à une détermination probabiliste des valeurs manquantes, en prenant en compte les dépendances entre l'attribut manquant et les autres attributs de l'objet, ce qui permet d'utiliser le maximum de l'information contenue dans l'objet pour le calcul des valeurs manquantes. Parce que les arbres de décision sont capables de déterminer la classe d'une instance à partir des valeurs de ses attributs, on peut les utiliser pour déterminer les valeurs d'un attribut inconnu (qui joue alors le rôle de la classe) à partir des attributs dont il dépend. Nous nous intéressons aux méthodes qui utilisent les arbres de décision pour trouver l'attribut manquant, et en particulier à la méthode des Arbres d'Attributs Ordonnés de Lobo et Numao (1999, 2000), qui construit un arbre de décision pour chaque attribut dans la base selon un ordre de construction croissant en fonction de l'information mutuelle relativement à la classe. Nous expliquons cette méthode en détail dans la section suivante.

1.1 Arbres d'Attributs Ordonnés (AAO)

Les Arbres d'Attributs Ordonnés (AAO) sont une méthode d'apprentissage supervisé proposée par Lobo et Numao (1999, 2000) pour traiter le problème des valeurs manquantes, à la fois dans les phases de construction et de classement. L'idée générale de cette méthode est de construire un arbre de décision, appelé arbre d'attribut, pour chaque attribut dans la base en utilisant un sous-ensemble d'apprentissage contenant les instances ayant des valeurs connues pour cet attribut. Pour un attribut donné, son arbre d'attribut est un arbre de décision dont les feuilles représentent les valeurs de cet attribut. Ces arbres sont construits selon un ordre de construction croissant en fonction de l'Information Mutuelle (IM)[1] entre chaque attribut et la classe (Shannon et Weaver, 1949). L'arbre d'attribut est utilisé pour déterminer la valeur de l'attribut pour les instances où elle est inconnue. Il est utilisé dans deux cas distincts: 1) lors de la construction de l'arbre de décision, pour déterminer la valeur de l'attribut pour les

[1] L'Information Mutuelle mesure la force de la relation entre deux attributs ou entre un attribut et la classe. L'IM entre deux attributs catégoriels X et Y est définie comme suit:
$IM(X,Y) = - \sum_{x \in Dx} P(x) * \log_2 P(x) + \sum_{y \in Dy} P(y) * [\sum_{x \in Dx} P(x\backslash y) * \log_2 P(x\backslash y)]$
D_x et D_y sont les domaines des attributs catégoriels X et Y. P(x) et P(y) sont les probabilités de $x \in D_x$ et $y \in D_y$, respectivement. P(x\y) est la probabilité conditionnelle que X prenne la valeur x sachant que Y est connu et prend la valeur y.

instances de la base d'apprentissage où cet attribut est inconnu; 2) lors du classement d'instances incomplètes, pour déterminer la valeur de l'attribut lorsque celle-ci est manquante.

Après avoir calculé l'IM entre chaque attribut et la classe, les attributs sont ordonnés par ordre croissant d'IM. Le premier arbre d'attribut construit est celui qui correspond à l'attribut ayant l'IM minimale. Il est représenté par un seul nœud-feuille avec sa valeur la plus probable dans la base d'apprentissage. Pour les autres attributs, on fournit, à partir de l'ensemble d'apprentissage initial, le sous-ensemble d'apprentissage qui contient les instances ayant des valeurs connues pour cet attribut. Ces instances sont décrites seulement par les attributs qui ont déjà été traités (c'est-à-dire les attributs pour lesquels on a déjà construit les arbres d'attributs et déterminé leurs valeurs manquantes dans la base d'apprentissage). L'algorithme utilisé pour la construction est un algorithme standard de construction d'un arbre de décision. Lors d'un classement, les valeurs des attributs inconnus de l'objet sont calculées successivement, par ordre d'IM croissante. L'étude qui a été faite par Lobo et Numao (2001) a montré que les relations entre attributs d'une base d'apprentissage doivent vérifier certaines conditions pour que la méthode AAO soit applicable.

2 Approche probabiliste

Nous présentons rapidement dans cette section notre approche, qui se compose de deux parties: 1) les Arbres d'Attributs Ordonnés Probabilistes (AAOP) qui étendent la méthode précédente AAO en construisant un arbre de décision probabiliste pour chaque attribut au lieu d'un arbre de décision classique; 2) les Arbres d'Attributs Probabilistes (AAP) qui prennent en compte la dépendance entre les attributs lors de la construction des arbres de décision probabilistes et non selon l'ordre croissant de l'information mutuelle.

2.1 Arbres d'Attributs Ordonnés Probabilistes (AAOP)

Notre première proposition (Hawarah et al., 2004) est une extension de la méthode des Arbres d'Attributs Ordonnés (Lobo et Numao, 1999, 2000). Elle consiste à construire pour chaque attribut un arbre d'attribut selon la méthode AAO. Cependant, contrairement à Lobo, qui, en suivant la méthodologie classique, associe à chaque feuille la valeur la plus probable, nous conservons dans chaque feuille d'un arbre d'attribut la distribution des fréquences des valeurs de l'attribut en question. Les attributs utilisés pour construire un arbre d'attribut selon AAOP sont les attributs déjà traités et dépendants de l'attribut en question (au lieu d'utiliser tous les attributs déjà traités, comme dans la méthode AAO, cf paragraphe 1.1). La distribution de probabilité sur chaque feuille d'un arbre d'attribut probabiliste permet de déterminer le classement probabiliste des valeurs d'un attribut manquant. En conséquence, elle permet le classement probabiliste d'un objet avec des attributs manquants. On appelle cette proposition Arbres d'Attributs Ordonnés Probabilistes (AAOP). Le résultat du classement permettant de déterminer une valeur manquante est une distribution de probabilités des valeurs de l'attribut. Ainsi, le classement d'un objet incomplet en utilisant les AAOPs est une distribution probabiliste de classe au lieu d'une seule valeur de classe.

2.2 Arbres d'Attributs Probabilistes (AAP)

La méthodologie des arbres d'attributs probabilistes (AAP) que nous avons proposée par Hawarah et al. (2004, 2005) est une méthodologie qui, pour chaque attribut, construit un

arbre d'attribut probabiliste en utilisant les attributs dont il dépend. Afin de déterminer les dépendances entre les attributs, nous calculons l'IM entre chaque couple d'attributs de la base. En effet, l'IM entre deux attributs est la réduction moyenne de l'incertitude sur un attribut, sachant l'autre. Ainsi, pour un attribut A_i, les attributs dont il dépend sont calculés par l'expression :

$$Dep(A_i) = \{A_j \mid IM(A_i, A_j) \geq Seuil\}$$ (Seuil à fixer, voir section 4)

3 Classement d'une instance avec attributs manquants

Pour classer un objet ayant des valeurs manquantes dans l'arbre de décision probabiliste final[2], on commence par parcourir l'arbre de décision, en partant de sa racine jusqu'à ce qu'on arrive à une feuille en suivant les branches correspondant aux valeurs des attributs de l'instance à classer. Une fois que l'on rencontre une valeur manquante pour un attribut test (nœud), on explore toutes les branches correspondant aux valeurs de cet attribut. Dans ce cas, on arrive à plusieurs feuilles dans l'arbre, et pas seulement à une seule feuille comme dans le classement classique. Pour cela, il faut calculer les probabilités de classe sur chacune de ces feuilles.

Supposons que l'on ait deux valeurs de classe A et D et que, pour un chemin de la racine de l'arbre jusqu'à une feuille F, on passe par les branches $B_1, B_2,.., B_n$. Alors:

P(classe A sur une feuille F) = P(A\chemin de la racine à F) = P(A\ $B_1, B_2,.., B_n$)

P(classe D sur une feuille F) = P(D\chemin de la racine à F) = P(D\$B_1, B_2,.., B_n$)

P(A dans tout l'arbre) = $\sum_i P(A\backslash F_i) * P(F_i)$ où i = 1,..,m (m le nombre de feuilles dans l'arbre)

P(B dans tout l'arbre) = $\sum_i P(A\backslash F_i) * P(F_i)$ (la probabilité total de B)

Le calcul précédent est inspiré de la méthodologie utilisée par Quinlan (1986, 1990, 1994) lors du classement d'une instance avec valeurs manquantes dans l'arbre de décision.

La probabilité $P(A\backslash F_i)$ est la probabilité de classe A attachée à cette feuille; la probabilité $P(F_i)$ est la probabilité jointe d'attributs dans le chemin de la racine d'arbre jusqu'à F_i. Pour simplifier, considérons que le chemin de la racine jusqu'à F_i passe par les branches B_1,B_2:
$P(F_i) = P(B_1,B_2) = P(B_1) * P(B_2\backslash B_1)$ sachant que B_1 est moins dépendant de la classe que B_2[3].

Calcul de la probabilité jointe $P(B_1,B_2)$ dans notre approche. On distingue les cas suivants:

1. B_1 et B_2 sont indépendants[4]: $P(B_2\backslash B_1) = P(B_2)$, et $P(B_1,B_2) = P(B_1) * P(B_2)$
   en conséquence, l'AAP de B_1 est construit sans B_2 et l'AAP de B_2 est construit sans B_1. On appelle pour l'attribut B_1 son AAP et pour B_2 son AAP.
2. B_1 et B_2 sont dépendants mais l'AAOP de B_1 est construit sans B_2: $P(B_1\backslash B_2) \neq P(B_1)$. On appelle pour l'attribut B_1 son AAOP et pour B_2 son AAP.
3. B_1 et B_2 sont dépendants, B_1 est le moins dépendant de la classe, un autre attribut manquant G dépend de B_1 et B_2, G[5] est moins dépendant de la classe que B_1 et B_2.
   On peut donc écrire: $P(B_1)=\sum_i P(B_1\backslash G_i) * P(G_i)$

[2] L'arbre de décision final est l'arbre de classement qui correspond à toute la base d'apprentissage.

[3] On commence par calculer la probabilité de l'attribut le moins dépendant de la classe.

[4] B_1 et B_2 sont indépendants si $IM(B_1,B_2) < Seuil$.

[5] B_1 et B_2 sont deux valeurs possibles pour deux attributs différents; par exemple: B_1 correspond à (*Sensibilité* est *Normale)*, B_2 correspond à *(Epw_somnolent* est *non*). Par contre, G est un autre attribut comme *Alcool_cons* qui prend deux valeurs possibles (*non, oui*).

$$P(B_2\backslash B_1) = \sum_i P(B_2\backslash B_1, G_i) * P(G_i \backslash B_1)$$
$$P(B_1,B_2) = \sum i\ P(B_1,B_2,G_i) = \sum_i P(G_i)\ P(B_1\backslash G_i) * P(B_2\backslash B_1,G_i)$$[6]

4. B_1 et B_2 sont indépendants mais ils dépendent d'un autre attribut G. G est moins dépendant de la classe que B_1 et B_2: $P(B_1,B_2) = \sum_i P(B_1,B_2,G_i) = \sum_i P(G_i) * P(B_1\backslash G_i) * P(B_2\backslash Gi)$
   B_2 est indépendant de B_1 conditionnellement à G.

Le but de notre approche de classement est d'explorer toutes les instances que l'on peut générer à partir de l'instance à classer ayant des valeurs manquantes; les instances sont générées pendant le classement en remplissant seulement, dans l'instance à classer, les valeurs des attributs inconnus rencontrés: 1) soit dans l'arbre de décision final, 2) soit dans les arbres d'attributs probabilistes appelés au cours du classement. Par exemple: le classement d'un objet ayant 4 attributs manquants donné dans le Tab. 1 revient à étudier les 8 instances dans la Tab.3:

Age	Bmi3-obésité	Tabac	Alcool	Tabagisme	Spicy	Epw_somnolent	Sensibilité	Type_SDB
Jeune	?	Oui	Oui	Nul	?	?	?	?

TAB. 1 – *Instance à classer.*

Attributs	Age	Bmi3-obésité	Tabac	Alcool	Tabagisme	Spicy	Epw_somnolent	Sensibilité	Classe: Type_SDB
Valeurs	Jeune agé	oui non	oui non	oui non	nul continu	oui non	oui non	normale anormale	Control SDB

TAB. 2 – *Les attributs d'apnées du sommeil avec leurs valeurs.*

Le Tab. 2 contient les attributs de la base *apnées du sommeil* (Bounhoure et al., 2005) avec leurs valeurs, chaque attribut dans cette base prenant deux valeurs. La classe (l'attribut à prédire) est l'attribut *Type_SDB* qui prend deux valeurs: *Control et SDB*.

Age	Bmi3-obésité	Tabac	Alcool	Tabagisme	Spicy	Epw_somnolent	Sensibilité	Type_SDB
Jeune	oui	Oui	Oui	Nul	?	oui	Normale	?
Jeune	non	Oui	Oui	Nul	?	oui	Normale	?
Jeune	oui	Oui	Oui	Nul	?	non	Normale	?
Jeune	non	Oui	Oui	Nul	?	non	Normale	?
Jeune	oui	Oui	Oui	Nul	?	oui	Anormale	?
Jeune	non	Oui	Oui	Nul	?	oui	Anormale	?
Jeune	oui	Oui	Oui	Nul	?	non	Anormale	?
Jeune	non	Oui	Oui	Nul	?	non	Anormale	?

TAB. 3 – *Les instances étudiées.*

On ne prend pas en compte l'attribut *Spicy* (Tab.3) qui est également manquant, parce qu'il ne dépend ni de la classe ni d'autres attributs manquants (*Sensibilité, BMI3_obésité, Epw_somnolent*) (voir Tab.4). En conséquence, le classement probabiliste de l'instance ayant

[6] $P(B_1,B_2) = P(B_1) * \sum_i P(B_2\backslash B_1,G_i) * P(G_i\backslash B_1)$
$= \sum_i P(B_2\backslash B_1,G_i) * P(G_i\backslash B_1) * P(B_1) = \sum_i P(B_2\backslash B_1,G_i) * P(B_1\backslash G_i) * P(G\ G_i)$

4 attributs manquants revient à calculer la probabilité que chacune de ces 8 instances possibles appartient à la classe *Control*, ainsi qu'à calculer la probabilité d'appartenance à la classe *SDB*.

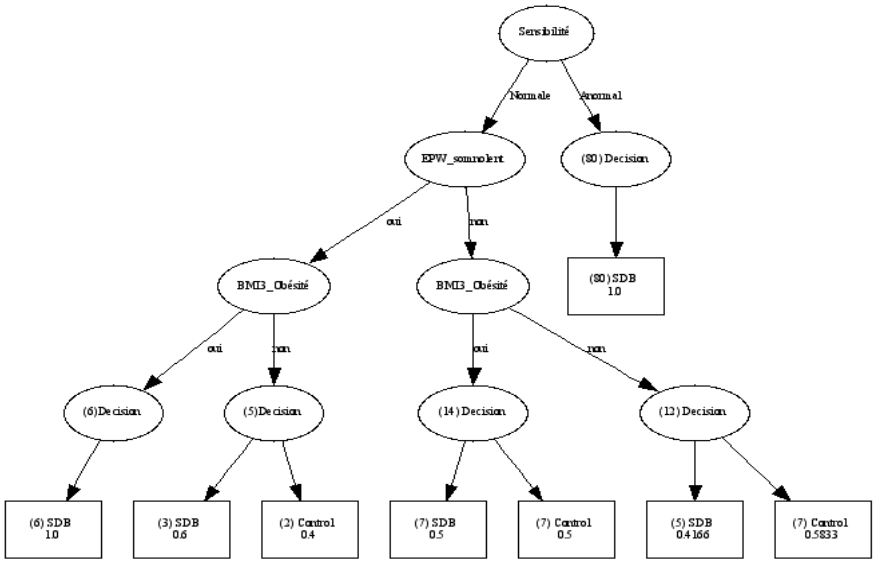


FIG. 1 – *L'Arbre de Décision Probabiliste AAP (base apnées du sommeil).*

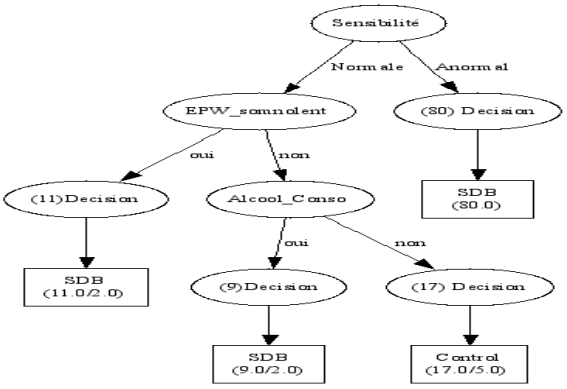


FIG. 2 – *L'Arbre de Décision Probabiliste AAP (base apnées du sommeil).*

L'arbre dans la Fig. 1 est un arbre de décision probabiliste et l'arbre dans la Fig. 2 est un arbre de décision classique[7].

[7] Dans la Fig. 2, C4.5 donne le résultat suivant: *Control(17.0/5.0)* qui signifie que la valeur de classe sur cette feuille est *Control* avec 17 objets arrivant à cette feuille dont 5 objets sont mal classés. Dans ce cas, nous déduisons que les 5 autres objets appartiennent à la classe *SDB*. Si la classe possède un nombre de valeurs supérieur à 2, nous ne savons pas à quelle classe appartiennent les 5 objets considérés comme mal classés.

4 Expérimentation

Choix d'un seuil. Le seuil[8] est fixé en calculant l'Information Mutuelle (IM) entre chaque attribut et la classe. Dans notre approche, on détermine un seuil qui est supérieur à l'information mutuelle minimale trouvée (non nulle) et proche de l'information mutuelle moyenne. Par exemple, si l'IM minimale est 0.01 et l'IM maximale est 0.2, et si la moyenne de toutes les valeurs d'IM est 0.05, on peut tester la méthode sur plusieurs seuils: 0.02, 0.03 et 0.04.

- Le fait de fixer un seuil conduit à éliminer avant la construction d'un arbre d'attribut probabiliste les attributs qui sont indépendants de l'attribut courant[9].
- Si on choisit un seuil très faible, on augmente le nombre d'attributs dépendants. De plus, on diminue le nombre d'instances par feuille. En revanche, le choix d'un seuil assez élevé nous conduit à diminuer le nombre d'attributs dépendants. Pour cela, nous testons notre approche sur plusieurs seuils pour la même base de test et nous comparons les résultats.
- Dans le cas où le nombre de valeurs par attribut est élevé, nous étudions l'information mutuelle normalisée et nous pouvons choisir un seuil égal ou supérieur à la moyenne.

Information Mutuelle Normalisée. En général, plus un attribut a de valeurs, plus son information mutuelle, relativement à la classe, a tendance à être élevée. Pour cela, le but de normaliser l'Information Mutuelle est de la rendre insensible au nombre de valeurs de l'attribut testé (Cremilleux, 1991). Pour normaliser le gain d'information, on peut diviser le gain par log m (m est le nombre de valeurs de l'attribut testé). Des résultats expérimentaux (Kononenko et al., 1984) ont montré que ce facteur semble pénaliser les attributs ayant un nombre élevé de valeurs. Quinlan (1986) utilise le gain de ratio qui consiste à diviser l'Information Mutuelle par l'entropie de l'attribut testé (à condition que celle-ci ne soit pas nulle). Nous avons préféré utiliser la normalisation de l'Information Mutuelle faite par Lobo et Numao (2001): l'Information Mutuelle entre deux attributs X et Y est symétrique, donc:

$$IM(X,Y) = IM(Y,X)$$
$$IM(X,Y) = H(X) - H(X\backslash Y) \leq H(X)$$
$$IM(Y,X) = H(Y) - H(Y\backslash X) \leq H(Y)$$

d'où:

$$IM(X,Y) \leqslant \frac{H(X)+H(Y)}{2}$$

En considérant que log $|D_x|$ et log $|D_y|$ sont les bornes supérieures de H(X) et H(Y) respectivement, la borne supérieure de l'Information Mutuelle devient:

$$IM(X,Y) \leqslant \frac{\log\|D_X\|+\log\|D_Y\|}{2} \quad \text{qui s'écrit donc} \quad \frac{2\ IM(X,Y)}{\log\|D_X\|+\log\|D_Y\|} \leqslant 1$$

D'où l'Information Mutuelle normalisée:

$$IM_N(X,Y) \equiv \frac{2\ IM(X,Y)}{\log\|D_X\|+\log\|D_Y\|}$$

[8] L'expert du domaine peut également fixer le seuil.

[9] L'attribut courant est l'attribut pour lequel on veut construire l'arbre d'attribut probabiliste et qui joue donc le rôle de classe.

4.1 Résultat de classement

Nous avons testé notre approche sur la base *apnées du sommeil* du CHU de Grenoble (Bounhoure et al., 2005) et (Dematteis et al., 2005). Nous avons construit nos arbres AAOPs et AAPs pour chaque attribut dans la base d'apprentissage. L'arbre dans la Fig. 1 est notre arbre de décision probabiliste correspondant à la base complète d'*apnées du sommeil* ayant 117 instances sans valeurs manquantes et 8 attributs discrets. La classe *Type-SDB* prend deux valeurs: *Control* et *SDB*. L'arbre dans la Fig. 2 correspond à l'arbre de décision élagué construit selon C4.5 pour la même base d'apprentissage.

Le tableau Tab. 4 représente les attributs utilisés pour construire l'arbre de décision final correspondant à la base d'apprentissage selon C4.5 et selon notre approche AAP dont le seuil est 0.04. Le choix de l'attribut *Alcool_conso* par C4.5 comme attribut-nœud dans l'arbre de décision est lié au fait que *Alcool_conso* a le gain d'information maximum dans la sous-base d'apprentissage correspondant à *Sensibilité* est *Normale* et *Epw_somnolent* est *non*. La racine dans les deux arbres de décision C4.5 et AAP est l'attribut *Sensibilité*, et le deuxième attribut utilisé dans les deux arbres est *Epw_somnolent.*

Base apnées du sommeil	Attributs utilisés	Caractéristiques d'attributs	IM_{base} (Attr,C)	Attributs dépendants
C4.5	Sensibilité	Racine de l'arbre		
	Epw_somnolent	Deuxième attribut dans l'arbre		
	Alcool_conso	GainInfoC4.5(Alcool_conso) est maximum ici		
AAPs	Sensibilité	Racine de l'arbre	0.26	BMI3_Obésité, EPW_somnolent
	Epw_somnolent	Deuxième attribut dans l'arbre	0.064	Sensibilité
	BMI3_Obésité	IM_{base}(BMI3_Obésité,C) > IM_{base}(Alcool_conso,C)	0.04	Sensibilité, Âgé

TAB. 4 – *Comparaison entre l'arbre de C4.5 et l'arbre de AAP pour un seuil 0.04.*

Le premier test de classement est effectué sur la base d'apprentissage *(apnées du sommeil)* entière ayant 117 instances sans valeurs manquantes. Dans Tab. 5, nous remarquons que notre résultat est proche de celui qui est donné par C4.5 quand le seuil est 0.02 ou 0.03.

Base d'apprentissage	Seuil	Bien classés	Mal classés	50,00%
AAPs	0.02	88,00%	7,00%	5,00%
	0.03	88,00%	7,00%	5,00%
	0.04	82,00%	6,00%	12,00%
C4.5		92,30%	7,69%	

TAB. 5 –*Test de AAP et C4.5 sur la base d'apprentissage d'apnées du sommeil.*

La colonne 50,00% dans Tab.5 contient le pourcentage des objets ayant la probabilité 0.5 pour chaque valeur de classe.

Nous avons ensuite généré une base de test qui possède 58 objets à partir de la base *apnées du sommeil* entière. Le taux de valeurs manquantes dans la base de test est 62% pour l'attribut *Sensibilité,*14% pour les attributs *Epw_somnolent et* Alc*ool_con*s, et 6,8% pour *Tabagisme* et *BMI3_Obésité*.

Age2	BMI3_Ob	Alcool_O	Tabac_O	Tabagism	Spicy_Co	EPW_so	Sensibili	Type_SD	SEUIL	Control	SDB	Control0	SDBC4.5
Agé	oui	non	oui	Nul	oui	?	?	SDB	0,04	0,08	0,92	0,19	0,81
Agé	non	non	oui	continu	non	oui	?	Control	0,04	0,11	0,89	0,06	0,94
Agé	non	non	oui	continu	non	?	?	Control	0,04	0,3	0,7	0,19	0,81
Agé	oui	non	oui	Nul	non	oui	?	SDB	0,04	0	1	0,06	0,94
Agé	non	non	oui	continu	non	?	Normale	Control	0,04	0,5	0,5	0,61	0,39
Agé	non	non	oui	Nul	non	non	?	Control	0,04	0,47	0,53	0,24	0,76
Jeune	oui	non	oui	continu	non	non	?	Control	0,04	0,15	0,85	0,24	0,76
Jeune	non	non	non	Nul	non	oui	?	SDB	0,04	0,11	0,89	0,06	0,94
Agé	non	oui	oui	Nul	non	non	?	SDB	0,04	0,47	0,53	0,07	0,93
Agé	oui	oui	oui	Nul	non	non	?	SDB	0,04	0,15	0,85	0,07	0,93
Jeune	non	non	oui	continu	oui	non	?	Control	0,04	0,47	0,53	0,24	0,76
Agé	non	oui	non	Nul	non	non	?	Control	0,04	0,47	0,53	0,07	0,93
Agé	non	non	non	Nul	non	non	?	Control	0,04	0,47	0,53	0,24	0,76
Jeune	non	non	non	Nul	non	oui	?	SDB	0,04	0,11	0,89	0,06	0,94
Jeune	oui	non	non	continu	non	oui	?	SDB	0,04	0	1	0,06	0,94
Jeune	non	non	oui	continu	non	oui	?	SDB	0,04	0,11	0,89	0,06	0,94
Agé	non	oui	oui	Nul	non	non	?	SDB	0,04	0,47	0,53	0,07	0,93
Jeune	oui	?	oui	continu	non	oui	?	SDB	0,04	0	1	0,06	0,94
Agé	oui	non	non	Nul	non	oui	?	SDB	0,04	0	1	0,06	0,94
Jeune	oui	oui	non	Nul	non	non	?	SDB	0,04	0,15	0,85	0,07	0,93
Agé	oui	oui	oui	continu	oui	oui	?	SDB	0,04	0	1	0,06	0,94
Agé	?	non	non	Nul	non	oui	?	SDB	0,04	0,03	0,97	0,06	0,94
?	non	non	non	?	non	oui	Anormal	SDB	0,04	0	1	0	1
?	oui	non	non	Nul	oui	non	?	SDB	0,04	0,15	0,85	0,24	0,76

TAB. 6 – *Partie du résultat de classement avec AAP et C4.5.*

Le tableau Tab.6 contient une partie de la base de test ainsi que le résultat du classement de chacun de ses objets; les deux avant-dernières colonnes contiennent le résultat du classement donné par AAP, et les deux dernières colonnes représentent celui donné par C4.5. À partir de Fig. 1, Fig. 2, Tab. 4, et Tab. 6 nous remarquons que l'attribut *Sensibilité* est la racine[10] de l'arbre et sa valeur *Anormale* est la plus fréquente dans la base. Chaque objet dont l'attribut *Sensibilité* est manquant et classé avec l'arbre de C4.5 comme malade (c'est à dire *Type-SDB* est *SDB*). En revanche, dans notre approche le résultat de classement est variable selon les valeurs de chaque attribut dont *Sensibilité* dépend; ces attributs sont *BMI3_Obésité, EPW_somnolent (*quand le seuil est 0.04).

Le résultat des tests de cette base, sur plusieurs seuils: 0.02, 0.03, 0.04, est donné dans Tab.7. Dans le même tableau, on trouve le résultat de classement donné par l'arbre de C4.5 pour la même base de test.

Base de test	Seuil	Bien classés	Mal classés	50,00%
AAPs	0.02	63,79%	15,51%	20,68%
	0.03	59,00%	24,00%	17,00%
	0.04	67,21%	29,31%	3,48%
C4.5		65,51%	34,48%	

TAB. 7 – *Comparaison entre AAP et C4.5.*

[10] La racine est l'attribut ayant l'influence la plus forte sur la décision.

Nous avons également testé notre approche ainsi que C4.5 sur la base *Breast-cancer* issue du (*UCI Repository of machine learning databases*) (Blake et al. 1998). La base possède 286 objets dont 9 objets ayant des valeurs manquantes. Elle contient 9 attributs discrets dont *age* qui possède 9 valeurs, *tumor-size* qui possède 12 valeurs, et *inv-nodes* qui possède 13 valeurs. Pour la construction des arbres AAOPs et AAPs, nous avons utilisé la base précédente sans les 9 objets ayant des valeurs manquantes. L'Information Mutuelle normalisée est utilisée pour étudier la dépendance entre les attributs ainsi que pendant la construction des arbres, car le nombre de valeurs pour certains attributs dans cette base est assez élevé.

Attribut	Information mutuelle	Information mutuelle normalisée
Age	0.020728822	0.009942060
menopause	0.011550725	0.008936861
tumor-size	0.061456810	0.026807988
inv-nodes	0.082421073	0.035069516
node-caps	0.055882090	0.055882090
deg-malig	0.088532845	0.068498359
breast	0.001232732	0.001232732
breast-quad	0.008641386	0.005202633
Irradiat	0.034703629	0.034703629
Class	0.871825271	0.871825271

TAB. 8 – *L'information mutuelle et l'information mutuelle normalisée les attributs et la classe.*

Le tableau Tab. 8 contient l'information mutuelle et l'information mutuelle normalisée sur la base *Breast-cancer*. On remarque que l'information mutuelle normalisée n'a pas trop pénalisé les attributs ayant un nombre élevé de valeurs. Le seuil choisi est 0.04 parce qu'il nous a donné le meilleur arbre de décision où les attributs utilisés sont deg-malig et node-caps. Pour la phase de classement, nous avons utilisé une base de test ayant 92 objets dont le taux de valeurs manquantes est 59% pour l'attribut *node-caps*, 37% pour l'attribut *deg-malig*, et 15% pour l'attribut *irradiat*. Le résultat est donné dans le Tab. 9.

Base de test	Seuil	Bien classés	Mal classés	50,00%
AAPs	0.04	72,00%	28,00%	0,00%
	0.03	71,00%	24,00%	5,00%
C4.5		70,65%	29,34%	

TAB. 9 – *Résultat des tests sur la base breast-cancer.*

Nous remarquons que nos résultats sont meilleurs par rapport aux résultats donnés par C4.5 même si nous considérons que les 5% d'objets qui possèdent une probabilité 0.5 pour chaque valeur de classe (avec le seuil 0.03) sont mal classés.

5 Conclusion et Perspectives

Dans cet article, nous avons utilisé une approche probabiliste pour résoudre le problème des valeurs manquantes dans les données. Nous avons proposé de remplacer une valeur manquante par une distribution de probabilités et un objet incomplet par une distribution de probabilités de classe. Un arbre de décision probabiliste est construit pour chaque attribut dans une base d'apprentissage en utilisant les attributs dont il dépend.

A partir du test que nous avons fait, nous avons remarqué que quelques attributs qui sont moins dépendants de la classe dans la base d'apprentissage pourraient devenir plus dépendants de la classe dans un sous-ensemble de cette base; il en est de même pour la dépendance entre les attributs eux-mêmes. Dans notre approche, nous avons éliminé avant la construction d'un arbre d'attribut probabiliste tous les attributs qui ne dépendent pas de l'attribut courant dans la base d'apprentissage entière. Nous ignorons les dépendances possibles entre l'attribut courant et les attributs éliminés dans un sous-ensemble d'apprentissage pendant la construction, parce que les attributs les moins dépendants de la classe ou de l'attribut courant ont moins de chances de participer à la construction de l'arbre d'attribut probabiliste (Lobo et Numao, 1999).

Le test que nous avons réalisé dans la section précédente a pour but de valider nos arbres probabilistes dans les deux phases (construction et classement). Le résultat du classement de chaque objet est encourageant et donné sous forme d'une distribution de probabilité (Tab. 6). Nous avons comparé notre résultat avec celui de la méthode C4.5, qui fournit également un résultat probabiliste mais qui donne comme seule réponse la valeur de classe la plus probable. Nous travaillons actuellement sur d'autres bases d'apprentissages réelles. Nous essayons également de trouver une autre façon de comparer nos résultats avec les résultats des autres méthodes de manière probabiliste, sans considérer les valeurs de classe les moins probables comme des mauvaises réponses. Dans le but d'améliorer la performance de notre approche, nous allons analyser le résultat du classement de chaque objet dans une base d'apprentissage.

Références

Blake C.L. and Merz C.J. (1998): UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

Breiman L., Friedman J.H., Olshen R.A. and Stone C.J. (1984): *Classification and Regression Trees*, Wadsworth and Brooks.

Bounhoure JP., Galinier M., Didier A., Leophonte P. (2005): *Sleep apnea syndromes and cardiovascular disease.* Bull Acad Natl Med. 2005 Mar;189(3):445-59.

Crémilleux B. (1991). *Induction automatique: aspects théoriques, le système ARBRE, Applications en médecine*. Thèse de doctorat, Université Joseph Fourier-Grenoble I.

Dematteis M., Levy P., and Pepin JL. (2005): *A simple procedure for measuring pharyngeal sensitivity: a contribution to the diagnosis of sleep apnoea*, may, Thorax.

Kononenko I., Bratko I. and Roskar E. (1984): Experiments in Automatic Learning of Medical Diagnostic Rules, Technical Report, Jozef Stefan Institute, Ljubljana, Yugoslavia.

Friedman J.H., Kohavi R. and Yun Y. (1996): Lazy Decision Trees, AAAI.

Hawarah L., Simonet A. and Simonet M. (2004): Une approche probabiliste pour le classement d'objets incomplets dans un arbre de décision, EGC 2004, poster.

Hawarah L., Simonet A. and Simonet M. (2004): A probabilistic approach to classify incomplete objects using decision trees, Spain, DEXA. LNCS 3180 pp. 549-558.

Hawarah L., Simonet A. and Simonet M. (2005): Classement d'objets incomplets dans un arbre de décision probabiliste, Deuxième atelier sur la "Fouille de données complexes dans un processus d'extraction des connaissances", EGC 2005, Paris.

Lobo O.O. and Numao M. (1999): Ordered estimation of missing values, Pacific-Asia Conference on Knowledge Discovery and Data Mining.

Lobo O.O. and Numao M. (2000): Ordered estimation of missing values for propositional learning, Japanese Society for Artificial Intelligence, JSAI, vol.15, no.1.

Lobo O.O. and Numao M. (2001): Suitable Domains for Using Ordered Attribute Trees to Impute Missing Values. IEICE TRANS. INF. & SYST., Vol.E84-D, NO.2.

Quinlan J.R. (1989): Unknown attribute values in induction. Proc. Sixth International Machine Learning Workshop, Morgan Kaufmann.

Quinlan J.R. (1986): Induction of decision trees. Machine Learning, 1, pp.81-106.

Quinlan J.R. (1990): Probabilistic decision trees, in Machine Learning: an Artificial Intelligence Approach, ed.Y.Kodratoff, vol.3, Morgan Kaufmann, San Mateo, pp.140-152.

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

Shannon C.E., Weaver W. (1949): Théorie Mathématique de la communication, les classiques des sciences humaines.

White A.P. Liu W.Z., Thompson S.G. and Bramer M.A. (1997): Techniques for Dealing with Missing Values in Classification. LNCS 1280, pp. 527-536.

Summary

We present a probabilistic approach to fill missing values in decision trees during classification. This approach is derived from a supervised learning method called Ordered Attributes Trees, proposed by Lobo and Numao in 2000. Our approach takes the dependences between the attributes into account when constructing the attributes trees by using mutual information, and provides a probability distribution as a result when classifying an incomplete object (instead of the most probable class). In this paper, we tested our approach on real datasets and we compared our results with those given by the C4.5 method. We also study the advantages and the disadvantages of each of these methods.

LSA : les limites d'une approche statistique

Mathieu Roche et Jacques Chauché

Équipe TAL, LIRMM - UMR 5506, Université Montpellier 2,
34392 Montpellier Cedex 5 - France
{mroche,chauche}@lirmm.fr

Résumé. Cet article propose une méthode de classification conceptuelle à partir de textes qui sont par nature des données complexes. Nous nous sommes intéressés à la méthode statistique appelée LSA (Latent Semantic Analysis) utilisée pour regrouper des termes et/ou des textes. Cet article met particulièrement en relief les limites de LSA sur des données réelles. Des propositions pour améliorer la qualité des résultats sont enfin proposées.

1 Introduction

Cet article s'intéresse au regroupement des termes extraits à partir de corpus spécialisés. Nous définissons un terme comme un groupe de mots ayant des propriétés syntaxiques (de type Nom-Nom, Nom-Adjectif, etc.) et qui représente une trace linguistique de concepts (Kodratoff (2004)). Les concepts sont définis par l'expert du domaine. Par exemple, les termes *génie logiciel* et *intelligence artificielle* pourraient être associés au concept de *Cours en informatique*. Dans cet article, nous ne détaillerons pas la méthode d'extraction de la terminologie qui a été utilisée (voir l'approche décrite dans les travaux de Roche (2004)).

Plusieurs méthodes de classification de termes à partir de textes existent dans la littérature. La plupart de ces systèmes sont fondés sur des méthodes mixtes : linguistiques et statistiques. Par exemple, le système LEXICLASS (Assadi (1997, 1998)) utilise des mesures de similarité pour regrouper les termes partageant souvent un même contexte (par exemple, les termes qui sont souvent en présence du même adjectif peuvent être regroupés). Le système ASIUM développé par Faure et Nédellec (1998) utilise une hypothèse similaire fondée sur le fait que le contexte permet de déterminer la sémantique. Le système ASIUM qui possède une approche coopérative avec l'expert utilise les connaissances syntaxiques (obtenues avec un analyseur syntaxique) et des mesures de similarité pour construire une classification conceptuelle. Le système ROWAN (Fontaine et Kodratoff (2002)) construit des classes de termes et de relations syntaxiques (Sujet-Verbe, Verbe-Objet, etc.). Outre l'approche coopérative avec l'expert de ROWAN (Fontaine et Kodratoff (2002)), un algorithme d'induction a également été proposé par Kodratoff (2004). Notons qu'un résumé de l'état de l'art des méthodes de classification de termes à partir de textes est présenté dans l'article de Aussenac-Gilles et Bourigault (2003).

Comme nous l'avons précisé, la plupart des systèmes de classification conceptuelle à partir de textes utilisent des approches mixtes. Dans cet article, nous allons nous appuyer sur la méthode appelée Latent Semantic Analysis (LSA) développée par Landauer et Dumais (1997);

Landauer et al. (1998)[1]. LSA est une méthode uniquement fondée sur une approche statistique appliquée à des corpus de grande dimension consistant à regrouper les termes (classification conceptuelle) ou les contextes (classification de textes).

Cet article présente la méthode LSA en mettant particulièrement en relief les limites de celle-ci. Ces limites sont dues à la complexité des données textuelles. Ainsi, nous souhaitons ici discuter des résultats obtenus avec LSA qui peuvent se révéler parfois décevants. Nous tenterons d'apporter des hypothèses afin d'expliquer et de discuter de ces limites. Enfin, nous proposerons des pistes de travail que nous souhaitons mettre en œuvre dans l'équipe TAL du LIRMM pour améliorer la qualité des résultats obtenus avec LSA. L'approche que nous souhaitons proposer consiste à apporter des informations syntaxiques à LSA.

2 Latent Semantic Analysis (LSA)

La méthode LSA qui s'appuie sur l'hypothèse « harrissienne », est fondée sur le fait que des mots qui apparaissent dans le même contexte sont sémantiquement proches. Le corpus est représenté sous forme matricielle. Les lignes sont relatives aux mots et les colonnes représentent les différents contextes choisis (un document, un paragraphe, une phrase, etc.). Chaque cellule de la matrice représente le nombre d'occurrences des mots dans chacun des contextes du corpus. Deux mots proches au niveau sémantique sont représentés par des vecteurs proches. La mesure de proximité est généralement définie par le cosinus de l'angle entre les deux vecteurs.

Caractéristiques théoriques de LSA

La théorie sur laquelle s'appuie LSA est la décomposition en valeurs singulières (SVD). Une matrice $A = [a_{ij}]$ où a_{ij} est la fréquence d'apparition du mot i dans le contexte j, se décompose en un produit de trois matrices USV^T. U et V sont des matrices orthogonales et S une matrice diagonale. La figure 1 représente le schéma bien connu d'une telle décomposition où r représente le rang de la matrice A.

Soit S_k où $k < r$ la matrice produite en enlevant de S les $r - k$ colonnes qui ont les plus petites valeurs singulières. Soit U_k et V_k les matrices obtenues en enlevant les colonnes correspondantes des matrices U et V. La matrice $U_k S_k V_k^T$ peut alors être considérée comme une version compressée de la matrice originale A.

Il est coutume de dire que LSA est une méthode statistique ou numérique car elle s'appuie sur une théorie mathématique bien connue. Cependant, on peut également dire que LSA est une méthode géométrique car seuls des résultats d'algèbre linéaire sont utilisés.

Nous précisons qu'avant d'effectuer la décomposition en valeurs singulières, nous effectuons une première étape de normalisation de la matrice d'origine A. Cette normalisation consiste à appliquer un logarithme et un calcul d'entropie sur la matrice A. Ainsi, plutôt que de se fonder directement sur le nombre d'occurrences de chacun des mots, une telle transformation permet de s'appuyer sur une estimation de l'importance de chacun des mots dans leur

[1] voir aussi, http ://www.msci.memphis.edu/∼wiemerhp/trg/lsa-followup.html

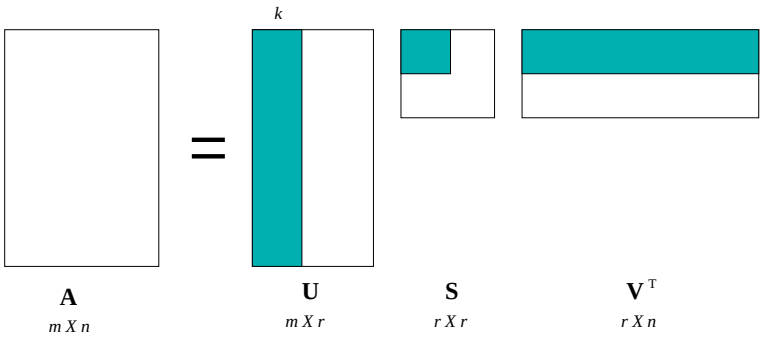


FIG. 1 – *Décomposition en valeurs singulières. La matrice A représente le corpus d'origine de m lignes (mots du corpus) et n colonnes (contextes).*

contexte. De manière similaire aux travaux de Turney (2001), cette étape de normalisation peut également s'appuyer sur la méthode du tf×idf, approche bien connue dans le domaine de la Recherche d'Information (Salton (1991)).

Précisons de plus que nous ne prenons pas en compte les ponctuations ainsi qu'un certain nombre de mots non significatifs du point de vue sémantique tels que les mots « et », « à », « le », etc. L'utilisation d'une telle liste de mots « vides » influence positivement le résultat final (Roche et Kodratoff (2003)).

Les expériences décrites dans la section suivante ont été menées avec un nombre de facteurs k égal à 200. Comme le montrent Landauer et Dumais (1997) le choix du nombre de facteurs de la matrice est un paramètre qui influence sensiblement le résultat final. Nous avons expérimenté différents paramètres, et comme dans les expériences de Wiemer-Hastings (2000), nous avons estimé qu'un nombre de facteurs égal à 200 donnait les meilleurs résultats sur notre corpus qui est décrit dans la section suivante.

3 Expérimentations

3.1 Caractéristiques du corpus étudié

Dans cet article, nous allons nous appuyer sur un corpus des Ressources Humaines de la société PerformanSe écrit en français[2]. Une caractéristique essentielle de ce corpus, d'une taille de 3784 Ko, est qu'il utilise un vocabulaire spécialisé comme la plupart des corpus que nous étudions. Cependant, une spécificité du corpus des Ressources Humaines est qu'il contient des tournures de phrases revenant souvent, ce qui peut influencer positivement le traitement avec LSA. À titre d'exemple, la meilleure similarité trouvée entre deux termes a été obtenue grâce à deux phrases ayant des tournures strictement identiques et pour lesquelles seul un terme différait. Dans nos expérimentations sur le corpus des Ressources Humaines, chaque phrase

[2]Fragment du corpus disponible à l'adresse : http ://www.lri.fr/∼roche/Recherche/corpusPsy.html

représente les contextes, c-à-d. les colonnes dans notre matrice relative au corpus. Nous précisons que pour d'autres corpus explicitement découpés en documents distincts (par exemple, des corpus de résumés), les contextes peuvent être représentés par les documents eux-mêmes.

3.2 Protocole expérimental

Afin d'effectuer des regroupements de termes, il est nécessaire d'obtenir des similarités entre les termes de bonne qualité deux à deux. Ainsi, dans les expérimentations que nous allons décrire, nous nous intéressons aux couples de termes trouvés automatiquement par LSA. Nous allons nous appuyer sur le corpus des Ressources Humaines pour lequel plus de 1800 termes présents ont été extraits dans le corpus à l'aide de la méthode d'extraction de la terminologie décrite dans les travaux de Roche (2004). Ces termes ont alors été associés manuellement à un concept par un expert du domaine. Par exemple, avec ce corpus, l'expert a défini le concept « Relationnel » dont les termes *confrontation-ouverte*, *contact-superficiel* et *entourage-compréhensif* sont des instances.

Afin d'évaluer la qualité des regroupements donnés par LSA, nous allons effectuer deux types de mesures. La première mesure correspond au pourcentage de couples corrects (c-à-d., les termes associés par LSA qui appartiennent à un même concept). Dans le domaine de l'apprentissage, une telle mesure correspond à la Précision. La deuxième mesure que nous allons utiliser est la couverture des termes classés par LSA. Cette mesure d'évaluation consiste à calculer le pourcentage de termes de la classification conceptuelle présents dans les couples formés par LSA.

4 Première limite majeure de LSA : l'influence de la taille des contextes

Nous avons alors effectué différentes expérimentations (voir TAB. 1) sur le corpus lemmatisé des Ressources Humaines. En effet, nous avons montré que le corpus lemmatisé donnait de meilleurs résultats que le corpus non lemmatisé (Roche et Kodratoff (2003)).

Les expérimentations effectuées sur ce corpus montrent que le nombre de couples de termes corrects, c-à-d. appartenant à un même concept, est globalement faible. Dans TAB. 1, nous pouvons également noter que plus la similarité entre les termes est proche de 1 (similarité maximale) et plus la qualité des couples de termes extraits est intéressante. En effet, en augmentant le seuil de similarité de 0.1, le pourcentage de couples de termes corrects augmente de plus de 10%.

Bestgen (2004) précise que la taille des contextes (documents) est primordiale pour obtenir une qualité des résultats satisfaisante. Cette affirmation confirme les travaux de Rehder et al. (1998) qui ont effectué des expérimentations pour estimer la taille minimale d'un contexte afin d'obtenir des résultats intéressants avec LSA. Ces expérimentations ont consisté à découper les documents d'un corpus correspondant à des essais d'étudiants en documents de 10 mots, 20 mots, et ceci jusqu'à 200 mots. Les expérimentations ont montré que si les contextes (do-

Similarité (cosinus)	0.3	0.4	0.5	0.6
% de termes correctement associés (c-à-d. % de couples corrects)	**19.2 %** (31/161)	**32.1 %** (9/28)	**42.9 %** (3/7)	**75.0%** (3/4)
% de termes de la classification manuelle présents dans les couples	**9.8 %** (180/1842)	**2.7 %** (49/1842)	**0.8 %** (14/1842)	**0.4 %** (8/1842)

TAB. 1 – *Expérimentations sur le corpus des Ressources Humaines de la société PerformanSe.*

cuments) possèdent moins de 60 mots alors la méthode LSA se révèle décevante.

Les contextes de notre corpus correspondent aux phrases qui sont composées de beaucoup moins de 60 mots. En effet, notre corpus possède, en moyenne, 27 mots par phrase. Cette situation pourrait donc expliquer les résultats pas toujours convaincants que nous avons obtenus avec LSA.

Notre motivation d'utiliser LSA repose sur une représentation mathématique solide des textes qui est indépendante des langues. La méthode LSA de base est automatique sans la nécessité de s'appuyer sur des connaissances linguistiques et du domaine (dictionnaires, ontologies, grammaires, étiquettes morpho-syntaxiques, etc.). Ceci est un avantage car la méthode LSA est largement généralisable selon les langues et les domaines des textes de spécialité étudiés. Mais cela peut aussi être une faiblesse de l'approche lorsque l'on a des objectifs pointus à atteindre comme nous allons le présenter dans la section suivante (section 5). La section 6 montrera enfin que l'ajout de connaissances syntaxiques à la méthode générale et automatique LSA pourrait améliorer les résultats.

5 Seconde limite majeure de LSA : difficulté dans le cas de la proximité du vocabulaire utilisé

Dans cette section, nous allons définir une autre tâche à effectuer prenant en compte des contextes de taille plus importante. Ainsi, nous allons mener des expérimentations consistant à classifier des textes. Pour cela, nous allons travailler à partir d'un autre corpus explicitement divisé en différentes thématiques. Nous proposons de travailler avec le corpus issu de la tâche Novelty de la compétition internationale TREC 2004 à laquelle l'axe *Fouille de Textes* de l'équipe IA[3] du LRI[4] mené par Yves Kodratoff a participé (Amrani et al. (2004)). Ce corpus est composé d'articles journalistiques divisés en cinquante thématiques distinctes.

Le but de la tâche Novelty était de retrouver les phrases pertinentes et nouvelles dans les articles journalistiques. Pour chacun des cinquante thèmes traités dans les articles du corpus, une description explicite de la pertinence et de la nouveauté était donnée aux participants de TREC Novelty. Lors de l'édition 2004 de TREC Novelty, des textes non pertinents ont pu être

[3]http ://www.lri.fr/ia/
[4]http ://www.lri.fr/

rajoutés dans les différentes thématiques.

Dans cet article, nous nous sommes plus spécifiquement intéressés à un sous-ensemble du corpus normalisé (Amrani et al. (2004)) rassemblant 29 articles journalistiques traitant d'une opération sur la greffe de la main. Les textes pertinents étaient ceux relatifs à la première greffe de la main effectuée sur le patient Matthew David Scott. Cependant, dans ce corpus, quatre articles se révèlent non pertinents (textes du domaine médical mais ne décrivant pas explicitement l'opération de Matthew David Scott).

Ainsi, nous avons utilisé LSA dans le but de vérifier que ces quatre textes avaient un score faible (le plus éloigné possible de 1) comparativement à l'ensemble des textes du corpus. De manière globale, ce résultat attendu n'a pas été vérifié dans nos expérimentations (voir TAB. 2). Plus précisément, les résultats montrent que deux des textes non pertinents ont un score de similarité plus faible que le score moyen (0.0933) mais les deux autres articles non pertinents ont un score plus important que ce score moyen. Ainsi, il semble globalement difficile de conclure sur l'efficacité de la méthode dans ce cas. Ceci peut s'expliquer par le fait que le vocabulaire utilisé dans les textes pertinents et non pertinents est souvent très proche. Ainsi, une classification de textes très fine peut se révéler difficile avec LSA.

Numéro des articles	Moyenne des scores	Numéro des articles	Moyenne des scores	Numéro des articles	Moyenne des scores
1 *(np)*	0.0357	11	0.0972	21	0.0970
2 *(np)*	0.0365	12	0.0922	22	0.1135
3 *(np)*	0.1083	13	0.0929	23	0.0990
4 *(np)*	0.1006	14	0.0961	24	0.1077
5	0.1068	15	0.0594	25	0.1208
6	0.1347	16	0.0951	26	0.0639
7	0.1191	17	0.0985	27	0.0395
8	0.1042	18	0.0958	28	0.1016
9	0.0952	19	0.1085	29	0.0354
10	0.1311	20	0.1197	**Moyenne des 29 textes : 0.0933**	

TAB. 2 – *Pour chaque article, la moyenne des scores de similarité avec les 28 autres articles est calculée. La notation "np" désigne les articles non pertinents.*

Dans un second temps, nous avons ajouté des textes issus d'une autre thématique pour étudier la similarité entre ces textes clairement non pertinents par rapport au sous-corpus traitant de la greffe de la main. Par exemple, TAB 3 montre que les quatre textes de thématiques singulièrement différentes[5] choisis aléatoirement ont un score de similarité globalement plus faible que les scores moyens. Ceci s'explique par le vocabulaire de ces quatre articles qui est éloigné des articles relatifs à la greffe de la main.

[5]Deux articles traitent du procès de Pinochet et deux articles sont relatifs au premier commandement d'une navette spatiale par une femme, le commandant Eileen Collins.

Sous-corpus original de "la greffe de la main" (avec les 4 premiers articles non pertinents)	
Numéro des articles	Moyenne des scores
1 *(np)*	0.0338
...	...
28	0.0966
29	0.0332
30 *(np)*	0.0482
31 *(np)*	0.0444
32 *(np)*	0.0393
33 *(np)*	0.0324
Moyenne des 33 textes : 0.0826	

Sous-corpus modifié de "la greffe de la main" (sans les 4 premiers articles non pertinents)	
Numéro des articles	Moyenne des scores
5	0.0992
...	...
28	0.0993
29	0.0318
30 *(np)*	0.0505
31 *(np)*	0.0469
32 *(np)*	0.0422
33 *(np)*	0.0353
Moyenne des 29 textes : 0.0873	

TAB. 3 – *Pour chaque article non pertinent (numéros 30 à 33) ajouté au sous-corpus relatif à la greffe de la main, la moyenne des scores de similarité avec l'ensemble des autres articles est calculée. La notation "np" désigne les articles non pertinents.*

6 Solution proposée : Ajout de connaissances syntaxiques à LSA

Afin d'améliorer la performance de LSA, Wiemer-Hastings (2000) propose d'ajouter des connaissances syntaxiques à LSA en transformant les phrases en structures syntaxiques. Pour ce faire, une segmentation syntaxique des phrases en trois groupes de mots est effectuée :

- syntagmes nominaux représentant les sujets,
- verbes en prenant en compte les adverbes et les syntagmes adverbiaux,
- syntagmes nominaux représentant les objets.

Ainsi, chaque phrase est représentée sous la forme (« verbe » « sujet » « objet »). Lorsqu'il y a deux objets (« objet1 » et « objet2 ») affectés à un même verbe, la phrase sera représentée sous la forme (« verbe » « sujet » « objet1 ») et (« verbe » « sujet » « objet2 »), de même dans le cas de la présence de deux sujets associés à un seul verbe.

Initialement, LSA ne prend pas en compte un certain nombre de mots (« stops words ») tels que « *if* », « *because* », « *have* », etc. Contrairement à la version originale de LSA, Wiemer-Hastings (2000) prend en compte de tels mots et peut les utiliser pour construire les structures de certaines phrases. Par exemple, la phrase *if the new motherboard uses the same type of RAM* sera représentée sous la forme *(« if uses » « the new motherboard » « the same type of RAM »)*.

Les résultats expérimentaux développés dans l'étude de Wiemer-Hastings (2000) restent malgré tout décevants. Ceci pourrait être dû à une analyse syntaxique pas assez fine. De plus, les données réelles utilisées (évaluation de la qualité des réponses d'étudiants) dans les travaux

de Wiemer-Hastings (2000) pouvaient se révéler trop difficiles à traiter [6].

Dans les futurs travaux que nous souhaitons mener dans l'équipe TAL du LIRMM, nous proposons également d'ajouter des connaissances syntaxiques à LSA. Pour cela, nous allons utiliser l'analyseur syntaxique SYGMART développé par Chauché (1984). À partir de textes bruts écrits en français, SYGMART fournit une analyse sous forme d'arbres morpho-syntaxiques. La qualité de l'analyseur syntaxique utilisé pour le français ainsi qu'une analyse plus fine des phrases devrait nous permettre d'améliorer les résultats obtenus par LSA.

Pour cela, nous allons décomposer chaque phrase de la manière la plus fine possible en utilisant au mieux la précision de l'analyseur SYGMART. Pour cela, nous allons nous appuyer sur la décomposition proposée par Wiemer-Hastings (2000) en trois entités : *sujet*, *verbe* et *objet*. De plus, nous pouvons ajouter toutes les informations syntaxiques supplémentaires apportées par SYGMART. Pour chacun des éléments, seuls les gouverneurs des syntagmes nominaux sont conservés. Illustrons une telle décomposition avec deux exemples donnés ci-dessous dans lesquels les termes *projet-de-recherche* et *ambitieuses-perspectives* sont déterminés par l'expert[7]. Ces deux exemples de phrases possèdent seulement deux mots en communs (*ajout* et *connaissance*). De ce fait, la méthode LSA pourrait avoir tendance à donner une similarité décevante entre les termes *projet-de-recherche* et *ambitieuses-perspectives*.

EXEMPLE 1
Phrase
L'ajout de connaissances syntaxiques à la méthode statistique LSA caractérise notre projet-de-recherche à moyen-terme
Décomposition
sujet*(ajout, connnaissance*, complément*(méthode, LSA))* **verbe***(caractériser)* **objet***(projet-de-recherche,* complément*(moyen-terme))*

EXEMPLE 2
Phrase
L'ajout de connaissances sémantiques significatives à notre approche ouvre également d'ambitieuses-perspectives
Décomposition
sujet*(ajout, connnaissance*, complément*(approche)* **verbe***(ouvrir)* **objet***(ambitieuses-perspectives)*

[6] Les textes traités dans Wiemer-Hastings (2000) ont des contextes de taille très réduite. Avec les phrases ayant un contexte moyen de 16 mots du corpus traité par Wiemer-Hastings (2000), les résultats qui sont obtenus peuvent tout de même se révéler comparables aux performances humaines. Cependant, comme le précisent Wiemer-Hastings et al. (1999), il semble difficile même pour un expert humain d'obtenir de bonnes performances avec des contextes ayant peu de mots.

[7] Un trait d'union a été placé entre les mots composant les termes pour que ces derniers soient reconnus comme des mots à part entière par les analyseurs syntaxiques.

En utilisant, les connaissances syntaxiques, les deux phrases possèdent exactement les deux mêmes mots principaux pour caractériser le sujet. Ainsi, il pourrait être intéressant de privilégier la proximité sémantique des deux termes présents comme objets des deux phrases. Une manière de prendre en compte cette information syntaxique dans la méthode LSA pourrait, par exemple, consister à ajouter une valeur numérique (notée α) au score trouvé par LSA entre les deux termes partageant le même contexte (ici, les termes *projet-de-recherche* et *ambitieuses-perspectives* partagent le même sujet). Par exemple, α pourrait prendre comme valeur le maximum ou la moyenne des scores obtenus avec LSA.

Nous avons effectué des premières expérimentations avec les deux phrases de l'exemple cité ci-dessus. Dans ces expérimentations, nous avons souhaité évaluer le taux de similarité obtenu avec LSA entre ces deux phrases comparativement à l'ensemble des phrases de l'introduction de ce présent article (section 1). Nous avons alors ajouté la deuxième phrase de notre exemple (phrase numérotée 2) aux 22 phrases représentant l'introduction (phrases numérotées de 3 à 24). Nous avons évalué le taux de similarité de la première phrase de l'exemple (phrase numérotée 1) avec les 23 phrases constituant le corpus. Ainsi, les deux meilleurs taux de similarité (avec des valeurs numériques du même ordre) avec la première phrase de l'exemple (phrase 1) ont été obtenus avec les phrases 2 et 7 (voir ci-dessous).

Phrase 1
L'ajout de connaissances syntaxiques à la méthode statistique LSA caractérise notre projet-de-recherche à moyen-terme
Phrase 2
L'ajout de connaissances sémantiques significatives à notre approche ouvre également d'ambitieuses-perspectives
Phrase 7
Dans cet article, nous ne détaillerons pas la méthode d'extraction de la terminologie qui a été utilisée (voir l'approche décrite dans les travaux de Roche).

La méthode que nous souhaitons mettre en œuvre privilégiera la similarité des deux premières phrases qui ont les mêmes mots principaux pour caractériser le sujet. Pour cela, nous donnerons une valeur de similarité plus importante au couple de phrases (1,2) comparativement au couple (1,7). Ces deux couples de phrases avaient initialement un score de similarité du même ordre obtenu avec LSA[8]. Ainsi, cette méthode permettra de privilégier la similarité entre les termes *projet-de-recherche* et *ambitieuses-perspectives* issus des phrases 1 et 2.

Les travaux de Wiemer-Hastings (2000) mais également l'approche de Faure et Nédellec (1998) s'appuient essentiellement sur les verbes pour effectuer un regroupement de termes. Par exemple, pour construire des classes sémantiques, le système ASIUM (Faure et Nédellec (1998); Faure (2000)) utilise la notion de « proximité sémantique » fondée sur le principe de distance entre les mots qui partagent le même contexte (en particulier, les verbes ayant souvent les mêmes objets). Les contextes "verbe objet" ou "sujet verbe" peuvent en effet être davantage pertinents pour effectuer un regroupement de termes comparativement à une relation "sujet objet" (sans considérer les verbes). Ainsi, nous pourrions utiliser un poids (α) à ajouter au score

[8]Dans ces expérimentations, le corpus constitué étant de taille très réduite, nous avons utilisé l'application http ://lsa.colorado.edu/cgi-bin/LSA-one2many.html avec le domaine "Français-Total".

de LSA plus important pour les termes trouvés dans les relations "verbe objet" ou "sujet verbe" partageant le même verbe comparativement à une relation "sujet objet" partageant un même contexte (sujet ou objet). Un processus d'apprentissage supervisé pourrait bien entendu être mis en œuvre pour établir les poids les plus adaptés à appliquer (voir l'ouvrage de Cornuéjols et al. (2002) qui présente un état de l'art des différentes méthodes d'apprentissage).

Nous avons présenté ici les perspectives à moyen terme que nous souhaitons mettre en place et qui, nous l'espérons, permettront d'améliorer les résultats parfois décevants obtenus en utilisant la méthode LSA de base.

7 Conclusion et perspectives

La méthode LSA qui est une méthode statistique (ou géométrique) ne prend pas en compte l'ordre des mots. Par exemple, les phrases *"le mot français est écrit dans le corpus"* et *"le corpus est écrit avec des mots français"* ont un sens très différent. Cependant, la méthode LSA conclura à une similarité parfaite entre ces deux phrases qui partagent les mêmes mots (sans considérer les mots « vides »).

Le fait d'ajouter des connaissances syntaxiques permet d'acquérir une meilleure compréhension du sens et donc de construire une classification conceptuelle de meilleure qualité. Ceci est une piste de travail particulièrement intéressante à développer. En effet, comme nous l'avons montré dans cet article, la méthode LSA possède des limites. Ainsi, l'ajout de connaissances syntaxiques pourrait pallier aux limites de LSA décrites dans cet article.

La phrase relative au deuxième exemple de la section 6 de cet article illustre une autre piste de travail... La perspective énoncée dans cet exemple propose d'ajouter des connaissances sémantiques à la méthode LSA. Ceci pourrait consister à remplacer des mots par le nom d'un concept plus général. Ceci permettrait également d'améliorer les résultats obtenus avec LSA.

Remerciements

Les auteurs remercient Yves Kodratoff (LRI) et Serge Baquedano (Société PerformanSe) pour le travail effectué sur le corpus des Ressources Humaines.

Références

Amrani, A., J. Azé, T. Heitz, Y. Kodratoff, et M. Roche (2004). From the texts to the concepts they contain : a chain of linguistic treatments. In *In Proceedings of TREC'04 (Text REtrieval Conference)*, pp. 712–722.

Assadi, H. (1997). Knowledge acquisition from texts : Using an automatic clustering method based on noun-modifier method. In *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pp. 504–509.

Assadi, H. (1998). *Construction d'ontologies à partir de textes techniques - application aux systèmes documentaires*. Ph. D. thesis, Université de Paris 6.

Aussenac-Gilles, N. et D. Bourigault (2003). Construction d'ontologies à partir de textes. In *Actes de TALN03*, Volume 2, pp. 27–47.

Bestgen, Y. (2004). Analyse sémantique latente et segmentation automatique de textes. In *Proceedings of JADT'04 (International Conference on Statistical Analysis of Textual Data)*, Volume 1, pp. 171–181.

Chauché, J. (1984). Un outil multidimensionnel de l'analyse du discours. In *In Proceedings of Coling, Standford University, California*, pp. 11–15.

Cornuéjols, A., L. Miclet, et Y. Kodratoff (2002). *Apprentissage artificiel, Concepts et algorithmes*. Eyrolles.

Faure, D. (2000). *Conception de méthode d'apprentissage symbolique et automatique pour l'acquisition de cadres de sous-catégorisation de verbes et de connaissances sémantiques à partir de textes : le système ASIUM*. Ph. D. thesis, Université Paris-Sud.

Faure, D. et C. Nédellec (1998). A corpus-based conceptual clustering method for verb frames and ontology acquisition. In P. Velardi (Ed.), *LREC workshop on Adapting lexical and corpus ressources to sublanguages and applications*, Granada Espagne, pp. 5–12.

Fontaine, L. et Y. Kodratoff (2002). Comparaison du rôle de la progression thématique et de la texture conceptuelle chez les scientifiques anglophones et francophones s'exprimant en anglais. In *Actes de la Journée de Rédactologie scientifique : L'écriture de la recherche*.

Kodratoff, Y. (2004). Induction extensionnelle : définition et application l'acquisition de concepts à partir de textes. *Revue RNTI E2, numéro spécial EGC'04 1*, 247–252.

Landauer, T. et S. Dumais (1997). A solution to plato's problem : The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review 104*(2), 211–240.

Landauer, T. K., P. W. Foltz, et D. Laham (1998). Introduction to latent semantic analysis. In *Discourse Processes*, Volume 25, pp. 259–284.

Rehder, B., M. Schreiner, M. Wolfe, D. Laham, T. Landauer, et W. Kintsch (1998). Using latent semantic analysis to assess knowledge : Some technical considerations. In *Discourse Processes*, Volume 25, pp. 337–354.

Roche, M. (2004). *Intégration de la construction de la terminologie de domaines spécialisés dans un processus global de fouille de textes*. Ph. D. thesis, Université de Paris 11.

Roche, M. et Y. Kodratoff (2003). Utilisation de LSA comme première étape pour la classification des termes d'un corpus spécialisé. In *Actes (CD-ROM) de la conférence MAJECSTIC'03 (MAnifestation des JEunes Chercheurs dans le domaine STIC)*.

Salton, G. (1991). Developments in automatic text retrieval. *Science 253*, 974–979.

Turney, P. (2001). Mining the Web for synonyms : PMI–IR versus LSA on TOEFL. In *Proceedings of ECML'01, Lecture Notes in Computer Science*, pp. 491–502.

Wiemer-Hastings, P. (2000). Adding syntactic information to LSA. In *Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society*, pp. 989–993.

Wiemer-Hastings, P., K. Wiemer-Hastings, et A. Graesser (1999). Improving an intelligent

tutor's comprehension of students with Latent Semantic Analysis. *Artificial Intelligence in Education*, 535–542.

Summary

This paper proposes a clustering method from texts which are complex data. We interested in statistical method called LSA (Latent Semantic Analysis) used to the terms and/or texts clustering. This paper shows the limits of LSA on real data. Finally, we explain how we can improve the results.

Fouille d'images IRMf : algorithme CURE

Jerzy Korczak, Aurélie Bertaux

LSIIT, Bd Sébastien Brant, 67412 Illkirch cedex France

<korczak, bertaux>@lsiit.u-strasbg.fr

Résumé. Dans cet article, nous présentons la fouille d'images IRMf à travers l'utilisation d'un algorithme hiérarchique ascendant de classification non supervisée : CURE. Nous avons adapté cet algorithme à nos contraintes pour nous permettre de traiter des volumes de données importants en temps quasi réel. Ces extensions portent sur le tirage aléatoire des signaux, leur partitionnement, leur échantillonnage, ainsi que la représentativité des clusters.

Nous exposerons les divers éléments de tests réalisés sur des données IRMf synthétiques, et nous critiquerons les resultats obtenus par CURE en dévoilant ses forces et faiblesses, en comparaison avec d'autres algorithmes implémentés préalablement et testés sur les mêmes données.

1 Introduction à la fouille d'images IRMf

L'Imagerie par Résonance Magnétique (IRM) produit des coupes virtuelles afin de visualiser l'anatomie du cerveau. Ces images d'une précision extrême permettent à des radiologues de détecter et localiser d'éventuelles lésions cérébrales chez leurs patients. L'IRM s'est spécialisée dans plusieurs branches dont récemment l'IRM fonctionnelle (IRMf), qui facilite la distinction des différentes parties du cerveau qui s'activent selon la "fonction" qui leur est associée. C'est sur ce principe que se base notamment l'imagerie neurofonctionnelle IRMf car le cerveau n'est pas une masse homogène mais elle est divisée en régions plus ou moins spécialisées. Elle repose également sur le fait que l'activation de ces régions entraîne une plus grande consommation d'oxygène et augmente le débit sanguin : effet BOLD. Ce sont ces informations qui sont détectées par l'IRMf en chaque point du cerveau lorsqu'il exécute une tâche particulière (motrice, sensorielle ou cognitive).

Une acquisition IRMf génère une série de 100 à 1000 images IRM. Chaque image d'acquisition est en général constituée de 32 ou 64 coupes de 64 pixels de côté. Les voxels de l'image volumique sont codés le plus souvent sur 16 bits ce qui produit des fichiers de 256 ou 512 Ko.

Les images normalisées par NMI (Normalized Mutual Information) qui permettent la comparaison entre plusieurs sujets ou l'utilisation d'atlas anatomiques sont plus volumineuses. Au final les séries IRMf complètes forment des volumes de données compris entre 25 Mo et 1 Go.
Jusqu'ici les protocoles expérimentaux (paradigmes) sont programmés de bout en bout. L'expérience est intégralement prévue au préalable. Elle passe par la conception des tâches à soumettre au cerveau ainsi que le modèle de réponse hémodynamique attendu pour chacune. Cette méthode guidée par le modèle (ang. model-driven) qui est la plus répandue à l'heure actuelle, mais elle ne peut conclure en dehors du modèle et le paradigme interdit toute intervention au cours de l'expérience.

Dans cet article, nous allons proposer une démarche pour découvrir le fonctionnement du cerveau en se basant sur un concept de fouille de données, déjà décrit dans nos publications précédentes : Korczak et al. (2005a) (2005b), ainsi que dans Goute et al. (1999) et Moller et al. (2001). Brièvement, ce concept peut se définir comme l'extraction de connaissances potentiellement exploitables à partir d'images IRMf. Il s'agit donc d'une démarche d'exploration et de découverte, radicalement différente de celle décrite préalablement. C'est une approche interactive qui intègre directement l'expert-médecin dans le processus de découverte et d'apprentissage de concepts pour mettre en évidence les zones fonctionnelles du cerveau et leur organisation.

La plateforme d'expérimentation de fouille d'images IRMf a été développée par Korczak et al. (2005) comprenant des algorithmes de classification de signaux IRMf qui permettent une fouille visuelle interactive en temps quasi réel : K-means, LBG (Lloyd généralisé), SOM de Kohonen et GNC (Gaz Neuronaux Croissants). En général ces algorithmes favorisent des groupes de forme sphérique et de tailles similaires. Ils sont très sensibles à la présence d'outliers (atypismes) dont la proximité induit l'algorithme en erreur en lui laissant laisser supposer qu'ils ont leur place au sein d'une classe.

Dans cet article, l'algorithme CURE sera décrit et comparé avec les algorithmes cités au-dessus.

CURE selon Guha et al. (1998) est un algorithme de classification, mais il est plus robuste face aux outliers et permet d'identifier des groupes non sphériques et d'une grande variance de taille. CURE réalise ceci en représentant chaque groupe par un nombre fixé de points qui sont générés en sélectionnant des points bien dispersés du groupe, et ensuite rapprochés du point moyen au centre du groupe en le multipliant par un coefficient. Le fait d'avoir plus d'un point représentatif permet à CURE de bien s'ajuster à la géométrie des clusters non sphériques et l'opération de rapprochement de ses points permet de diminuer les effets des outliers.

Pour manipuler de grands volumes de données, CURE propose une combinaison d'échantillonnage aléatoire et de partitionnement. Un échantillon tiré de l'ensemble des données est tout d'abord partitionné et chaque partition est partiellement mise en cluster. Chacun de ces groupes partiels sera à nouveau regroupé lors d'une seconde passe de l'algorithme pour extraire les clusters désirés.

L'article a été structuré en quatre sections principales. Dans la section suivante, l'algorithme CURE est décrit informellement en s'appuyant sur les extensions et les adaptations à la classification IRMf. La section 3 présente des possibilités développées dans le logiciel 3DSlicer pour la fouille d'images IRMf. La section 4 illustre les résultats de comparaison des algorithmes développés sur les données synthétiques. Elle discute aussi des avantages et des faiblesses de l'algorithme CURE.

2 Présentation de l'algorithme CURE

CURE est un algorithme de classification hiérarchique ascendante non supervisée, conçu pour traiter de grands volumes de données. Sa spécification détaillée est décrite en détail dans Guha et al. (1998). Brièvement, il débute en considérant chaque signal comme un cluster, et fusionne au fur et à mesure les deux clusters les plus proches (FIG. 1). Chaque cluster possède un ensemble de signaux représentatifs (R) qui le délimitent. Ils sont déterminés tout d'abord en choisissant les points les plus épars dans le cluster, et sont ensuite rapprochés du centre (X) par un coefficient. La distance entre deux clusters est la distance entre la paire de points représentatifs les plus proches, chacun appartenant à l'une des deux classes. Ainsi, seulement les signaux représentatifs d'un cluster sont utilisés pour calculer la distance aux autres clusters.

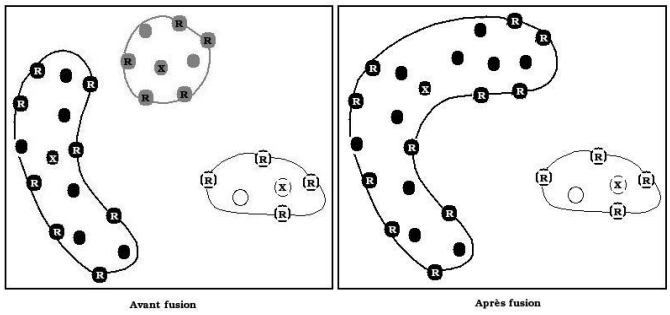


FIG. 1 - *Fusion de deux clusters*

Les signaux représentatifs tendent à s'accaparer les contours et la géométrie du cluster. En outre le rapprochement des signaux éloignés vers le centre du cluster lisse les anomalies de surface et atténue les effets des outliers qui sont généralement les signaux les plus éloignés du centre du cluster. Par conséquence, le rapprochement va obliger les outliers à se déplacer d'avantage vers le centre, pendant que les signaux représentatifs restants n'auront qu'un changement minimal à faire. Les grands mouvements des outliers devraient réduire leur capacité à indiquer les mauvais clusters à fusionner. Un traitement ultérieur pour écarter les outliers basé sur ces déplacements est envisageable. Le coefficient de rapprochement peut aussi être utilisé pour contrôler les bords du cluster, c'est à dire qu'une petite valeur rapproche peu les signaux éloignés et ainsi favorise les clusters allongés, alors qu'une grande rapproche très nettement les signaux éloignés vers le centre et les clusters tendront à être plus compacts. Comme mentionné, au départ chaque signal est considéré comme un cluster, et à chaque passe, les deux plus proches sont fusionnés en un seul. Ce processus est répété jusqu'à ce qu'il n'y ait plus que *k* clusters.

Au cours de la classification, chaque cluster est renseigné par son centre et l'ensemble de ses signaux représentatifs. Pour une paire de signaux *p* et *q*, *dist(p, q)* renvoie la distance entre ces deux signaux, qui peut être une distance de Manhattan ou bien une distance euclidienne. D'autres types de distance pourraient aussi être employés. La distance entre deux clusters *u* et *v* peut être définie comme suit :

$$dist(u,v) = min\ dist(p, q) \mid p \in u\ ,\ q \in v$$

Dans notre projet, l'algorithme CURE a été adapté d'une part aux images IRMf, et d'autre part aux structures de données préexistantes dans la plateforme développée. Il se base sur deux grandes étapes que sont celle du *clustering* à proprement parler et celle de la *fusion* des dits clusters.

L'étape de *clustering* est le processus même de classification. Au départ, chaque cluster n'est constitué que d'un unique signal, qui fait donc office de centre et de signal représentatif. CURE, après avoir déterminé pour chacun quel était le cluster qui lui était le plus proche, va tous les ranger dans le tableau de clusters dans l'ordre croissant des distances à leur cluster le plus proche. Suite à cette initialisation, à chaque itération de la boucle principale, jusqu'à obtention du nombre souhaité de clusters, les deux plus proches clusters sont fusionnés par la méthode du même nom : celui qui se trouve en tête du tableau de cluster, et celui qui lui a été assigné comme plus proche. Le cluster émergeant de la fusion, se voit donc pourvu d'un jeu approprié de signaux représentatifs. L'algorithme détermine alors à nouveau pour chaque cluster, celui qui lui est le plus proche. Notamment pour ceux qui étaient en relation avec les clusters absorbés par la fusion, ceux qui ont le nouveau cluster comme plus proche voisin et enfin et surtout, pour ce nouveau cluster. Une fois cette opération terminée, les clusters, y compris nouveau, se voient rangés à leur juste place dans le tableau de clusters.

L'étape de *fusion*, consiste à regrouper les deux clusters les plus proches. Sa fonction principale est de déterminer les signaux représentatifs pour le cluster issu de la fusion. Par rapport à la version originale qui recherche les signaux représentatifs du nouveau cluster parmi tous les signaux qui le constituent, nous avons implémenté une recherche restreinte aux signaux représentatifs des clusters ayant été fusionnés.

De plus, nous déterminons de manière dynamique le nombre de signaux représentatifs pour ce nouveau cluster. Contrairement à l'algorithme original qui propose un nombre fixe de signaux représentatifs quelque soit les dimensions du cluster, nous avons opté pour un nombre variable proportionné à la dispersion de ses signaux.

Une force de CURE selon les auteurs est de pouvoir s'adapter à de grandes bases de données pour un algorithme hiérarchique. L'implémentation de la version originale a démontré certaines faiblesses de performances de la classification de signaux IRMf qui est très lourde car il s'agit de voxels auxquels s'ajoute la quatrième dimension de leur évolution dans le temps. Pour réduire le temps de classification, nous avons appliqué quelques améliorations.

– *Tirage aléatoire.* Un tirage aléatoire des données est utilisé ayant pour vertu d'améliorer la qualité de la classification. En effet, les signaux sont enregistrés selon l'ordre dans lequel l'IRM les balayent, ce qui fait que deux signaux qui sont issus de zones voisines dans le cerveau, peuvent être séparés lors de l'enregistrement car une couche est balayée dans un sens avant de passer à la couche inférieure.

– *Echantillonnage.* Cela permet de déterminer les classes, avec moins de signaux. La taille de l'échantillon est paramétrable et pourrait donc être modifiable par le médecin. Ce cas est extrêment important car CURE fonctionnant de manière hiérarchique, plus le nombre de signaux est important, plus il génère de classes au départ et plus les calculs entre toutes les classes prennent du temps et des ressources.
– *Partitionnement.* Sur cette même constation, un système de rechargement en signaux a été réalisé. Ici encore la taille du partitionnement pourrait être laissé à la discrétion du médecin. CURE classant les clusters par ordre croissant de leur distance au cluster qui leur est le plus proche, impose donc un calcul de distance entre chaque paire de clusters, et pour chaque paire, leur distance est la distance minimale entre tous les couples de signaux représentatifs des deux classes. Nous avons déterminé expérimentalement un nombre de clusters seuil au delà duquel l'algorithme est trop ralenti. Pas à pas l'algorithme fusionne deux à deux les clusters jusqu'à atteindre une valeur plancher à partir de laquelle nous effectuons un rechargement en nouveaux clusters pour réatteindre le nombre maximal fixé. Ce procédé est répété jusqu'à épuisement du nombre de signaux.

Les outliers sont des sources de perturbation importante des algorithmes de classification. Dans notre cas, les clusters de petite cardinalité peuvent se trouver être des clusters d'un intérêt important, alors que d'autres qui sont très nombreux comme ceux comportant des signaux émis par la matière blanche ou le liquide céphalo-rachidien sont d'une inutilité flagrante. La présence du médecin expert et sa faculté à pouvoir intervenir lors du processus, permet de s'abstenir pour l'instant de ce genre de traitement de suppression de ces signaux intrus, puisqu'il peut le faire lui-même en pleine connaissance de cause.

L'implémentation dans la plateforme de telles modifications rendent CURE plus efficace et robuste à la classification des données qui sont lourdes, et nombreuses. Nous avons pour conséquence, des temps de calcul amoindris et grâce au tirage aléatoire une meilleure classification.

3 Fouille visuelle d'images IRMf

L'algorithme CURE a été inclus dans un outil de visualisation 3DSlicer (www.slicer.org), développé par l'école de médecine d'Havard en collaboration avec le MIT. 3DSlicer a été conçu principalement pour la visualisation temps réel au cours d'une intervention chirurgicale sur le cerveau. Il peut présenter cinq vues différentes du cerveau, trois coupes orthogonales, une vue 3D d'ensemble et une vue 3D dite endoscopique pour naviguer à l'intérieur du volume (cette dernière n'apparaît pas sur l'exemple et n'a de toute façon pas d'intérêt pour l'IRMf).

Elles peuvent combiner et afficher simultanément 3 volumes qui n'ont pas nécessairement la même résolution. La vue 2D affiche les trois coupes et les labels modélisés sous forme de surface. En dehors de l'algorithme CURE, nous avons implémenté les fonctions de 3DSlicer permettant

l'affichage des résultats dont notamment la partie gauche qui permet de renseigner les paramètres de calcul et d'afficher des résultats tels que la forme du signal du cluster en cours. Sur la figure 2 les clusters apparaissent de couleurs différentes et cohérentes avec la liste déroulante de la fenêtre de gauche, ainsi que les statistiques affichées en temps réel durant le processus de classification.

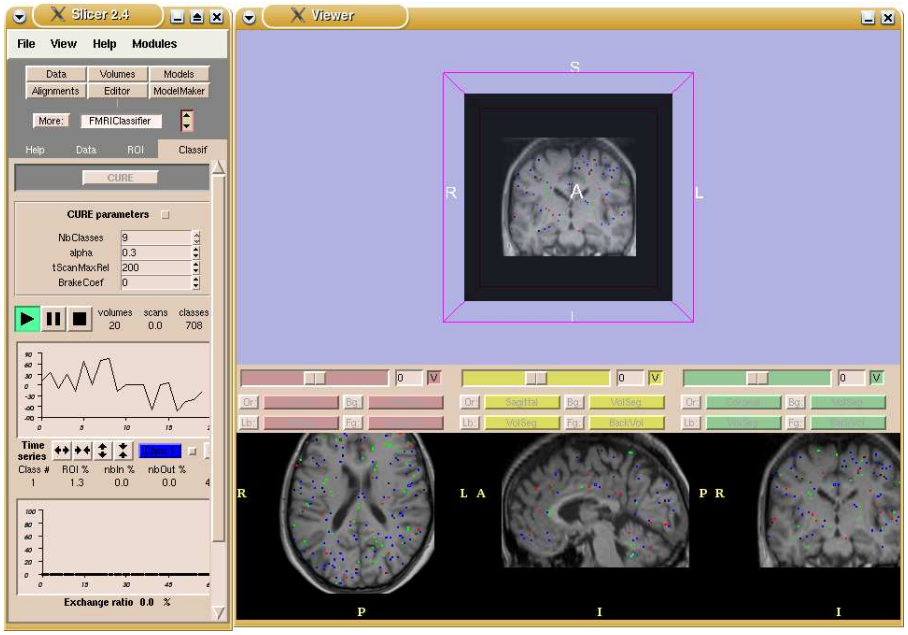


FIG. 2 - *Interface de contrôle FMRIClassifier dans 3DSlicer*

Durant la classification, le médecin peut changer facilement de point de vue avec la souris et modifier le contenu de l'affichage. Pour l'aspect interactif, il est pourvu de deux points forts. Le premier est qu'il est utilisé pour visualiser par IRM en temps réel, ce qui garantit une très bonne réactivité. Deuxièmement, 3DSlicer affiche les résultats des algorithmes à mesure de leur découverte. C'est par cette interface permettant la visualisation 3D que l'expert médecin peut également intervenir. Il peut mettre la procédure en pause pour modifier certains paramètres, focaliser sur son centre d'intérêt, éliminer les classes indésirables comme les yeux par exemple. L'interface du module de classification inclus dans 3DSlicer a été détaillée par Korczak (2005b).

4 Résultats d'expérimentation

Dans cette étude nous avons voulu comparer CURE aux autres algorithmes déjà implémentés. Ceux-ci avaient déjà subit des protocoles comparatifs par Hommet (2005). Nous avons donc repris

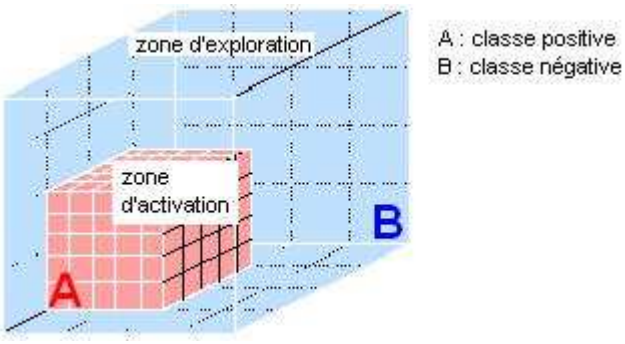


FIG. 3 - *Schéma simplifié des classes d'activation positives et négatives*

les mêmes démarches permettant de comparer leurs performances de détection des zones d'activation. Travailler sur des données réelles impliquerait d'avoir un outil permettant de nous indiquer que l'algorithme a tort ou raison. La station d'acquistion et de fouille d'images IRMf en temps reel n'existe pas, donc dans nos évaluations des algorithmes nous avons reformulé les expérimentations effectuées sur les benchmarks de SPM (Statistical Parametric Mapping, http ://www.fil.ion.ucl.ac.uk /spm), qui sont connus comme la référence dans le domaine. Il s'agit d'un assemblage de deux parties, les activations simulées purement artificielles et le fond constitué de données réelles. La série d'IRMf du cerveau utilisée comme fond a été composée à partir d'images issues du test auditif SPM. Le test auditif est un enchaînement de deux conditions : "silence" et "parole". Toutes les images de la condition "silence" ont été rassemblées et enchaînées pour former le fond de notre jeu de test qui se rapproche ainsi au mieux de données réelles.

Sur cette série de 40 images, viennent s'ajouter les activations synthétiques formées par des séries temporelles en créneau (signal carré) simulant un paradigme de type bloc. La zone d'activation est un volume cubique de 5 voxels de côté. Le niveau de bruit moyen de ces 125 voxels du fond a été mesuré en prenant le double de la variance des intensités de ces voxels dans le temps. A partir de cette mesure, il est ensuite possible de contrôler le rapport signal sur bruit de nos activations en ajoutant des créneaux d'intensités voulues aux voxels considérés. La figure 3 illustre de manière simplifiée comment sont déclarées les classes positives en fonction de leur recouvrement d'une zone d'activation.

En utilisant des données synthétiques, les algorithmes ont été comparés en variant différents critères qui avaient été pris en considération sur les précédents tests, à savoir : le nombre de classes, le rapport signal sur bruit et le rapport de dillution de la zone d'activation dans la zone d'exploration.

La qualité de la classification est proportionnelle à la faculté de l'algorithme à retrouver les bonnes classes.

Ainsi pour être déclarée positive, une classe doit avoir tous ses signaux qui appartiennent à la zone activée. Une classe négative ne possède pas de signaux dans cette zone. Le point critique intervient lorsqu'une classe se trouve à cheval sur ces deux définitions.

Un critère nous permet de déterminer dans quelle catégorie ranger la classe :

$$C_a \text{ positive} \equiv (inC_a / outC_a) > (2nZA / ((nZE - nZA) / (nbClasses - 1)))$$

où *nZE* est le nombre de voxels dans la zone explorée ; *nZA* est le nombre de voxels dans la zone activée ; *nbClasses*, le nombre total de classes ; C_a, une classe ; *et* inC_a, $outC_a$, le nombre de voxels de C_a à l'intérieur et à l'extérieur de la zone activée. Ce critère a été repris des tests effectués sur les algorithmes préimplémentés sur la plateforme afin de pouvoir les comparer sur les mêmes bases. Les tests consistent à soumettre aux algorithmes les séries constituées et à comparer les résultats de classification avec les résultats attendus. Ils ont été effectués de manière automatique et répétés 25 fois pour estimer d'une part l'erreur réelle et d'autre part pour voir la convergeance des algorithmes. Les autres valeurs ont été choisies dans les limites (intervalles) définies par des experts médecins. La moyenne des essais constitue alors le résultat définitif d'un test et renseigne sur la fréquence de détection de la zone activée.

Les tableaux 1, 2 et 3 présentent les résultats obtenus aux côtés des évaluations précédentes des autres algorithmes. Les classifications ont été réalisées par variation respective des paramètres que sont le nombre de classes (*nC*), le rapport de dilution des voxels activés (*nZA/nZE*) et le rapport signal sur bruit (*S/B*). Les résultats sont exprimés en pourcentage de détection de la zone d'activation.

Un algorithme est d'autant meilleur qu'il permet de retrouver des zones actives même si elles sont peu nombreuses en rapport du nombre total de voxels. Et aussi qu'il ne se laisse pas trop influencer par les bruits parasitants les signaux.

nC	4	6	8	9	10	12	14	15	16	18	20	21	24
GNC	10	0	95	100	100	100	100	100	100	100	100	100	100
SOM	4	24	56	40	52	70	100	98	100	100	100	100	100
LBG	2	26	36	30	36	54	64	72	88	96	100	96	100
K-Means	9	16	36	40	44	40	60	58	64	84	68	70	84
CURE	56	60	64	68	76	85	80	80	68	72	76	80	84

TAB. 1 - *Fréquence de détection (%) en fonction du nombre de classes imposé nC*

nZA/nZE	1/343	1/216	1/125	1/64	1/27	1/8
GNC	0	85	100	100	100	100
SOM	0	15	40	100	100	100
LBG	10	34	30	95	100	100
K-Means	10	35	40	45	95	100
CURE	52	72	68	68	72	72

TAB. 2 - *Fréquence de détection (%) en fonction du rapport nZA/nZE*

S/B	1,0	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2,0	3,0	4,0
GNC	0	0	55	95	100	100	100	100	100	100	100	100	100
SOM	0	0	0	0	0	40	70	85	100	100	100	100	100
LBG	0	5	0	15	15	30	55	40	65	75	69	100	100
K-Means	6	1	15	35	15	40	44	70	50	55	70	82	90
CURE	52	64	52	68	76	68	72	80	72	68	72	88	84

TAB. 3 - *Fréquence de détection (%) en fonction du rapport signal sur bruit S/B*

Le choix de ces paramètres a été discuté avec les experts médecins. Le nombre de classes qui est généralement la condition d'arrêt des algorithmes permet de juger d'une bonne valeur par défaut à fournir au médecin, et par la suite savoir dans quelle fourchette de valeur, l'algorithme s'égare. La dillution des voxels activés dans les voxels explorés permet de savoir si l'algorithme peut retrouver des classes de faible cardinalité. En effet, notre cerveau n'utilise qu'une petite partie de ses neuronnes pour certaines tâches. Si le médecin désire trouver ces petites classes il est bon de pouvoir lui recommander un algorithme plutôt qu'un autre. Enfin le rapport signal sur bruit est capital à cause de la présence de signaux parasites comme le mouvement des yeux ou les bruits émis par l'IRM.

Il apparaît que CURE présente des résultats meilleurs que les autres algorithmes pour les tests avec un nombre de classes inférieur à 8. Ces performances augmentent avec le nombre de classes mais de manière plus faible que les autres algorithmes. CURE présente une alternative intéressante dans les cas de forte dillution (faible pourcentage de zones actives *nZA/nZE*) et de signal sur bruit très faible. En comparaison avec le meilleur algorithme testé, CURE offre des performances médiocres. Il est à noter cependant que ces données synthétiques n'exploitent pas les qualités d'adaptation de CURE aux classes non sphériques et de forte variance.

Notons que ces résultats sont issus des algorithmes tels qu'ils sont implémentés à l'heure actuelle, autrement dit en mode batch. Le déroulement de l'algorithme est pausé régulièrement pour lui permettre une communication avec l'interface visuelle 3DSlicer, mais cela ne modifie pas la classification. Une prochaine étape qui s'inscrit dans le cadre de l'interactivité avec l'expert médecin serait de remodeler les algorithmes pour leur permettre un déroulement autre que batch. Le concept serait de pouvoir réserver une classe qui nous semble pertinente à un moment donné de la classification. En la figeant, l'algorithme ne pourrait plus interagir avec elle pour lui ajouter des signaux ou la fusionner à d'autres. Cela améliorerait grandement les résultats de classification, car le médecin sachant identifier la sémantique des classes, pourra empécher qu'une classe intéressante soit altérée d'une part, mais également que d'autres classes puissent être perturbées par une fusion malencontreuse avec celle-ci. De plus, des règles de classification pourront ainsi progressivement se dégager pour faire tendre la classification elle-même à la détection de ces classes.

5 Conclusion

Dans cet article, un algorithme hiérarchique de classification non supervisée a été présenté et validé sur des images IRMf. L'algorithme développé est une extension d'algorithme CURE proposé par Guha et al. (1998) pour traiter des volumes de grande taille en temps quasi réel.

Les extensions effectuées se portent sur les domaines suivants que sont le tirage aléatoire des signaux, leur partitionnement, leur échantillonnage ainsi que la représentativité des clusters. En tant qu'algorithme hiérarchique, CURE est extrêment gourmand en ressources. Nos améliorations ont réduit la complexité de l'algorithme notamment en limitant le nombre de calculs de distance et en conséquence ont réduit les temps de calculs. Selon la simulation on peut envisager une utilisation d'algorithme CURE étendue avec des contraintes de temps réel. L'intégration de cet algorithme dans la plateforme 3DSlicer facilite l'interaction d'un expert médecin dans le processus d'analyse d'images IRMf. L'algorithme CURE a été testé sur des données simples bi-dimmensionnelles et sur des données synthétiques. Si sur les premières, CURE obtient une très bonne performance, cependant il s'avère que sur les secondes, il présente des performances moyennes, mais reste de bonne robustesse. Cette constatation ne concerne que des données synthétiques ne lui permettant pas de mettre en avant ses qualités d'adaptation à des clusters d'une morphologie non sphérique.

La plateforme de fouille d'images IRMf et l'algorithme CURE dans la version actuelle nécessitent des validations de grande envergure par des médecins spécialistes. Pour l'instant les algorithmes classent les signaux, mais n'ont aucune mémoire de leurs précédents classements, ils ne possèdent donc aucune expérience. Ce travail est d'une très grande ampleur et nécessitera de longs mois ou années de travail pour aboutir. En effet, l'objectif futur sera l'intégration de connaissances médicales préalables dans l'algorithme, ainsi que l'acquisition des connaissances issues de la classification, dans une base de connaissances au niveau du patient pour suivre sa propre évolution, mais aussi au niveau de tous les sujets traités pour les comparer et tirer des enseignements médicaux, et des conseils lors des prochaines fouilles.

Remerciements

Les auteurs remercient Christian Scheiber (IHC, Lyon) pour les données expérimentales et Jean Hommet, Nicolas Lachiche, LSIIT, Illkirch, pour les conseils et aides dans la réalisation de ce projet.

Références

C. Goute, P. Toft, E. Rostrup, F.A. Nielsen, et A.K. Hansen (1999). *On clustering FRMI time series.* NeuroImage, 9, pages 2398-3100.

S. Guha, R. Rastogi, K. Shim (1998). *CURE : An Efficient Clustering Algorithm for Large Databases*. SIGMOD 1998, pages 73-84.

J. Hommet (2005). *Fouille interactive de séquences d'images 3D d'IRMf*. Rapport de LSIIT, CNRS, Illkirch.

J. Korczak, C. Scheiber, J. Hommet, N. Lachiche (2005a). *Fouille interactive en temps réel de séquences d'images IRMf.* Numéro Spécial RNTI.

J. Korczak, C. Scheiber, J. Hommet, N. Lachiche (2005b). *Exploration visuelle d'images IRMf basée sur des Gaz Neuronaux Croissants.* Atelier sur la fouille de données complexes, EGC 2005, Paris.
U. Moller, M. Ligges, C. Grunling, P. Georgiewa, W.A. Kaiser, H. Witte et B. Blanz (2001). *Pitfalls in clustering of neuroimage data and improvements by global optimization strategies*. NeuroImage, 14, pages 206-218.
SPM. Statistical Parametric Mapping. *Welcome Department of Imaging Neuroscience*.
http ://www.fil.ion.ucl.ac.uk/spm.

Summary

The human brain provides typical complex data for data mining. Extracting active voxels from brain images is often very difficult due to a very high level of various noises. First experiments of current data mining algorithms in this domain showed their low performances and recognition abilities. In this article, an extended unsupervised data mining algorithm CURE is described and evaluated. CURE is compared with several unsupervised algorithms on fMRI images, reporting results with respect to the number of classes, the noise level, and the ratio of the activated/observed areas.

Estimation et fusion des Temps de parcours routiers par la théorie de l'évidence

Eric Lefevre*, Nour-Eddin El Faouzi**

*Laboratoire d'Informatique et d'Automatique de l'Artois (LGI2A), EA 3926
Université d'Artois, Faculté des Sciences Appliquées
Technoparc Futura, 62400 Béthune
eric.lefevre@iut-geii.univ-artois.fr,
**Laboratoire d'Ingénierie Circulation Transports (LICIT)
INRETS-ENTPE
25, Avenue François Mitterand Case 24
69675 Bron Cedex
nour-eddin.elfaouzi@inrets.fr

Résumé. Nous abordons ici le problème de l'estimation du temps de parcours sur un axe urbain par des méthodes de classification basées sur la théorie de l'évidence. Les informations issues des différentes sources de recueil de temps de parcours (capteurs au sol, boucle magnétique, vidéo, véhicules traceurs,...) sont complémentaires et redondantes. Il est alors nécessaire de mettre en œuvre des stratégies de fusion multi-capteurs. L'approche de fusion que nous avons retenue, est basée sur la théorie des fonctions de croyance. Cette théorie permet de prendre en compte de manière plus naturelle les imprécisions ainsi que les incertitudes liées aux différentes informations traitées. Deux stratégies sont mises en œuvre. Tout d'abord une approche fusion de classifieurs dans laquelle les sources d'information sont considérées comme des classifieurs. La seconde approche est une approche de classification basée sur une approche distance pour la modélisation des fonctions de croyance. Des résultats de ces approches sur des données recueillies à l'issue d'une campagne de mesures réalisée sur un axe urbain de Toulouse montre les avantages de la fusion dans le cadre de cette application.

1 Introduction

Dans cet article nous abordons le problème de l'estimation du temps de parcours sur un axe urbain par des méthodes de classification basées sur la théorie de l'évidence. La notion de temps de parcours est utilisée pour répondre aux préoccupations de la gestion du trafic, qui consistent à apporter la réponse la plus satisfaisante aux besoins de déplacements, mais aussi comme information routière pour l'usager.

Les techniques de recueil de l'information temps de parcours peuvent être classées en deux familles. La première famille repose sur des moyens liés à l'infrastructure routière. Ces moyens sont généralement des détecteurs (capteurs au sol, boucle magnétique, vidéo, ...) et/ou des

observateurs (enquête minéralogique relevée manuellement, ...). La seconde famille concerne les moyens de mesures embarqués à bord de véhicules. On parle alors de *véhicules tests*. Il existe une complémentarité entre les données issues de capteurs au sols et celles mesurées par des capteurs embarqués à bord de véhicules empruntant les différents tronçons d'un réseau.

Les données fournies par les capteurs au sols sont des mesures quasi exhaustives, c'est-à-dire couvrant l'ensemble des véhicules ayant empruntés le tronçon, avec un échantillonnage et une résolution temporelle excellente. Cependant, ces mesures sont très imprécises (principalement le taux d'occupation qui est fortement bruit,...) avec un échantillonnage spatial qui dépend de la densité et de l'emplacement des capteurs. En effet, ces mesures ne représentent l'état du trafic qu'à l'endroit où le capteur est placé et non sur l'ensemble du tronçon.

A l'inverse, les données fournies par des véhicules munis de capteurs embarqués (véhicules tests), sont, quant à elles, des mesures très précises avec une excellente couverture spatiale. Elles expriment l'état du trafic sur l'ensemble du tronçon. Cependant, elles sont non exhaustives, car ne couvrant qu'une partie des véhicules ayant emprunté le réseau pendant une période temporelle donnée.

Les propriétés de complémentarité et de redondance de ces deux sources de données peuvent donc être mises à profit en élaborant une solution fusion de données multi-capteurs pour le problème d'estimation du temps de parcours en milieu urbain. L'objectif de cette approche multi-capteurs sera d'exploiter au mieux les avantages de chacune des sources d'information, tout en essayant de pallier leurs limitations individuelles respectives et de faire face, par la même, aux imperfections des mesures (données manquantes, aberrantes. ..). Ceci afin de fournir une meilleure image (globale et complète) de l'état du trafic.

Parmi les techniques de fusion de données multi-capteurs connues dans la littérature, nous avons retenu les approches basées sur la théorie des fonctions de croyance. En effet, il apparaît que cette théorie permet de prendre en compte de manière plus naturelle les imprécisions ainsi que les incertitudes liées aux informations. Dans cet article, nous présentons deux approches de fusion utilisant cette théorie. Tout d'abord une approche fusion de classifieurs dans laquelle les sources d'information sont considérées comme des classifieurs. La seconde approche est une approche de classification basée sur une approche distance pour la modélisation des fonctions de croyance.

Cet article est décomposé de la manière suivante. Nous présentons dans la section 2, les bases mathématiques de la théorie des fonctions de croyance nécessaires à la compréhension de la suite de l'article. Les deux méthodes de classification employées dans l'estimation du temps de parcours sont présentées dans la section 3. Enfin, dans la section 4, des résultats de ces approches sur des données recueillies à l'issue d'une campagne de mesures réalisée sur un axe urbain de Toulouse montre les avantages de la fusion dans le cadre de cette application.

2 Théorie des Fonctions de Croyances

Dans cette section, nous rappelons brièvement quelques concepts de base de la théorie des fonctions de croyance. Le point de vue du modèle des croyances transférables, proposé par Smets et Kennes (1994), est adopté dans cet article. Celui-ci distingue deux niveaux de traitement de l'information : le niveau crédal où les croyances sont modélisées et révisées et le niveau pignistique dans lequel les fonctions de croyance sont transformées en fonctions de probabilités pour la prise de décision.

2.1 Niveau crédal

Soit $\Omega = \{\omega_1, \cdots, \omega_k, \cdots, \omega_K\}$ un ensemble fini, généralement appelé cadre de discernement. Une fonction de croyance bel est une mesure floue non additive de 2^{Ω} dans $[0, 1]$ définie par :

$$bel(A) \triangleq \sum_{\emptyset \neq B \subseteq A} m(B) \qquad \forall\, A \subseteq \Omega \tag{1}$$

où m, appelé généralement jeu de masses, est une fonction de 2^{Ω} dans $[0, 1]$ qui vérifie :

$$\sum_{A \subseteq \Omega} m(A) = 1.$$

Chaque sous-ensemble $A \subseteq \Omega$ tel que $m(A) > 0$ est appelé élément focal de m. Ainsi, la masse $m(A)$ représente le degré de croyance attribué à la proposition A et qui n'a pas pu, compte tenu de l'état de la connaissance, être affectée à un sous-ensemble plus spécifique que A. Tirés des travaux de Dempster (1968) et de Shafer (1976), les fonctions de croyance sont de nos jours reconnues pour la modélisation des informations incertaines (de l'ignorance totale à la connaissance complète) et imprécises. Nous verrons plus en détail, dans la section 3, les méthodes employées dans notre application afin de construire les fonctions de croyances.

La seconde étape au niveau credal correspond à la révision des croyances. Parmi les outils de la théorie de l'évidence, il en est un qui concerne la combinaison de deux fonctions de croyance. A partir des fonctions de masse m_1 et m_2, la combinaison conjonctive de ces deux sources d'information ($m_\cap = m_1 \cap m_2$) peut être calculée $\forall\, A \subseteq \Omega$ par :

$$m_\cap(A) \triangleq \sum_{A = B \cap C} m_1(B) m_2(C). \tag{2}$$

Il est à noter que cette règle, généralement appelée règle de Dempster non normalisée, permet de combiner des informations incertaines extraites sous forme de fonctions de croyance. Si nécessaire, la condition $m(\emptyset) = 0$ peut être retrouvée en divisant chaque masse par un coefficient de normalisation. L'opération résultante, appelée règle de Dempster et notée $\oplus$ est définie $\forall\, A \subseteq \Omega$ par :

$$(m_1 \oplus m_2)(A) \triangleq \frac{(m_1 \cap m_2)(A)}{1 - m(\emptyset)} \tag{3}$$

où la quantité $m(\emptyset)$ est appelée degré de conflit entre les fonctions m_1 et m_2 et peut être calculé en utilisant l'équation suivante :

$$(m_1 \cap m_2)(\emptyset) = \sum_{B \cap C = \emptyset} m_1(B) m_2(C). \tag{4}$$

L'utilisation de la règle de Dempster est possible si et seulement si m_1 et m_2 ne sont pas en conflit total, c'est-à-dire s'il existe deux éléments focaux B et C de m_1 et m_2 qui satisfassent $B \cap C \neq \emptyset$.

2.2 Niveau pignistique

L'étape d'aggrégation précédemment définie permet ainsi d'obtenir un résumé exhaustif de l'information sous forme d'une fonction de croyance unique m qui est utilisée pour la prise

de décision. En basant leur raisonnement sur des arguments de rationnalité développés dans le modèle des croyances transférables, Smets et Kennes (1994) proposent de transformer m en une fonction de probabilité $BetP$ définie sur Ω (appelée fonction de probabilité *pignistique*) qui se formalise pour tout $\omega_k \in \Omega$ par :

$$BetP(\omega_k) = \frac{1}{1 - m(\emptyset)} \sum_{A \ni \omega_k} \frac{m(A)}{|A|} \tag{5}$$

où $|A|$ représente la cardinalité de $A \subseteq \Omega$. Dans cette transformation, la masse de croyance $m(A)$ est uniformément distribuée parmi les éléments de A.

3 Application à l'estimation de Temps de Parcours

Afin de réaliser l'estimation de temps de parcours à l'aide de la théorie de l'évidence, nous allons mettre en œuvre 2 approches différentes.

Pour la première approche, nous définissons un espace à 2 dimensions où chaque composante correspond à l'estimation d'un capteur (véhicule traceur ou boucle magnétique). Dans cet espace, nous calculons des dissimilarités entre un nouveau couple de mesures et les mesures d'apprentissage. Ces dissimilarités vont nous permettre de construire des fonctions de croyance et ainsi attribuer une classe de temps de parcours au nouveau couple de mesures.

La seconde approche correspond à une approche de fusion de classifieurs. Dans ce cas, les mesures obtenues par les capteurs sont considérées comme des classes. Elle peuvent donc être assimilées à des sorties de classifieurs qu'il convient alors de fusionner afin de prendre une décision globale plus fiable.

Dans les sections suivantes, nous détaillons ces deux approches.

3.1 Approche distance

Dans cette section, nous présentons le travail original de Denœux (Denoeux (1995); Zouhal et Denoeux (1998)) qui introduit l'approche distance de la modélisation des fonctions de croyance de la manière suivante.

Considérons un nouvel individu de vecteur forme x connu et de vecteur d'appartenance u inconnu. Si un élément de vecteur forme $x^{(i)}$ et d'étiquette $u_n^i = 1$ de l'ensemble d'apprentissage $\mathcal{L}$ est proche de x dans l'espace des caractéristiques, alors une partie de la croyance sera affectée à ω_n et le reste à l'ensemble des hypothèses du cadre de discernement. Ainsi, nous obtenons alors à partir de l'élément i une masse de croyance m_i. Jusqu'à présent, nous n'avons considéré pour l'appartenance de x qu'un seul élément de $\mathcal{L}$. Si la même opération est répétée pour l'ensemble des I exemples d'apprentissage, on obtient alors I fonctions de croyance qui peuvent être combinées à l'aide de l'opérateur de Dempster. En pratique, les éléments éloignés de x ont peu d'influence et peuvent être négligés. Deux techniques peuvent alors être mises en œuvre. Dans la première approche, on ne prend en compte que les k plus proches voisins de x, méthode que nous appellerons KNN-DS (Zouhal et Denoeux (1998)). La seconde approche repose sur la caractérisation des données d'apprentissage à l'aide de prototypes (Denoeux (2000)), méthode que nous nommerons ProDS. Chacun des prototypes i, initialisés par

un algorithme de type C-means, permet la construction d'une fonction de croyance ayant l'expression suivante :

$$\begin{cases} m_i(\{\omega_n\}) & = \alpha_i \phi_i(d_i) \\ m_i(\Omega) & = 1 - \alpha_i \phi_i(d_i) \end{cases} \tag{6}$$

où $0 < \alpha_i < 1$ est une constante, $\phi_i(.)$ est une fonction décroissante monotone vérifiant $\phi_i(0) = 1$ et $\lim_{d \to \infty} \phi_i(d) = 0$, d_i est la distance euclidienne entre le vecteur x et le $i^{\text{ème}}$ prototype. La fonction ϕ_i peut être une fonction exponentielle de la forme :

$$\phi_i(d_i) = exp^{-\gamma_i (d_i)^2} \tag{7}$$

où γ_i est un paramètre associé au $i^{\text{ème}}$ prototype. Le paramètre α_i empêche l'affectation de toute la masse de croyance à l'hypothèse ω_n lorsque x et le $i^{\text{ème}}$ prototype sont égaux. Il traduit l'incertitude relative à la caractérisation du prototype i. En outre, la contrainte $\alpha_i < 1$ garantit la possibilité de combiner m_i avec n'importe quelle autre fonction de croyance puisque quelque soit d_i, on aura toujours $m_i(\Omega) > 0$ (la certitude de ω_n pourrait entraîner un conflit total avec une autre source de croyance incompatible). Le paramètre γ_i, quant à lui, permet de spécifier la vitesse de décroissance de la masse avec la distance selon le prototype.

Les fonctions de croyance m_i, obtenues pour chacun des prototypes, sont ensuite fusionnées avec la règle de combinaison de Dempster.

Nous pouvons remarquer que cette méthode repose sur l'estimation de plusieurs paramètres : nombre de voisins k (ou nombre de prototypes), position des prototypes, valeurs de γ_i et les valeurs de α_i. Au cours d'une première étude (Zouhal (1997)), il a été montré que l'approche présentée ici est très peu sensible au choix du paramètre k, et donc *a priori* au nombre de prototypes.

Pour les autres paramètres, une méthode d'optimisation, fondée sur l'utilisation de l'information contenue dans l'ensemble d'apprentissage $\mathcal{L}$, a été introduite par Zouhal et Denoeux (1998). Cette optimisation est basée sur la minimisation d'un critère d'erreur quadratique moyenne E_{MS} entre les probabilités pignistiques et les vecteurs d'appartenance aux classes :

$$E_{MS} = \sum_{i=1}^{I} \sum_{n=1}^{N} [BetP^{(i)}(\omega_n) - u_n^i]^2 \tag{8}$$

où $BetP^{(i)}$ représente la probabilité pignistique d'un vecteur $x^{(i)}$ de la base d'apprentissage.

3.2 Approche combinaison de classifieurs

Dans la section précédente, nous avons présenté un classifieur basé sur la théorie des fonctions de croyance. Dans le cas où plusieurs classifieurs sont disponibles, l'agrégation de ces classifieurs permet le plus souvent d'améliorer la qualité de la classification en terme de taux de bien classés. Cette propension à l'amélioration est d'autant plus importante que les classifieurs soient complémentaires.

Les schémas d'agrégation de ces classifieurs varient selon le type de leurs sorties. Dans le cas de notre application, les sorties des capteurs (qui peuvent être assimiler aux classifieurs) sont des classes (des estimations de temps de parcours). Certaines techniques d'agrégation de

classifieurs de type classe sont basées sur les théories de l'incertain (approche bayésienne et réseaux du même nom) et de l'imprécis (approches crédibiliste et possibiliste).

Les schémas de fusion de classifieurs que nous allons introduire reposent essentiellement sur la prise en compte des erreurs des classifieurs individuels. Les erreurs de chaque classifieur sont usuellement consignées dans la matrice de confusion donnée par :

$$\mathcal{M}^j = \begin{pmatrix} n_{11} & n_{12} & \dots & n_{1j} & \dots & n_{1N} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots \\ n_{i1} & n_{i2} & \dots & n_{ij} & \dots & n_{iN} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots \\ n_{N1} & n_{N2} & \dots & n_{Nj} & \dots & n_{NN} \end{pmatrix}$$

où N représente le nombre de classes dans le problème. La ligne i correspond à la classe ω_i de l'apprentissage et la colonne j correspond à la classe décidée par la classifieur j, i.e. $\mathcal{C}^j$. Cette matrice est obtenue sur un échantillon d'apprentissage et peut être considérée comme la connaissance *a priori* sur les performances du classifieur. Les éléments diagonaux sont les pourcentages de concordances entre les classes reconstituées par le classifieur et les classes de référence. En d'autres termes, ceci représente le nombre de fois où la classe de référence et la classe reconstituée par le classifieur coïncident. Les éléments hors diagonale donnent quant à eux les pourcentages de discordances (i.e. de confusion). On définit alors le taux de reconnaissance τ^j_{rec} et le taux de confusion τ^j_{con} par :

$$\tau^j_{\text{rec}} = \frac{\sum_{i=1}^{N} n_{ii}}{\sum_{i,j}^{N} n_{ij}} \qquad \text{et} \qquad \tau^j_{\text{con}} = 1 - \tau^j_{\text{rec}} \tag{9}$$

Nous supposons que l'on dispose d'une matrice de confusion par classifieur et nous présentons dans les sections suivantes trois approches utilisées pour l'élaboration des masses de croyance à partir de ces matrices de confusion. Ces différentes approches de fusion de classifieurs sont présentées plus en détail par Aissa et al. (2004).

3.2.1 Méthode de Xu

L'une des premières approches de fusion de classifieurs de type classe basée sur la théorie des fonctions de croyance a été proposée par Xu et al. (1992). Les fonctions de croyances sont alors définies de la façon suivante :

$$\begin{cases} m_j(\{\omega_n\}) & = \tau^j_{\text{rec}} \\ m_j(\overline{\omega_n}) & = \tau^j_{\text{con}} \end{cases} \tag{10}$$

dans le cas où le classifieur $\mathcal{C}^j$ a sélectionné la classe ω_n. Les méthodes que nous avons développées et qui sont décrites par la suite découlent de cette construction de jeu de masses mais nous apparaissent plus précises.

3.2.2 Méthode n°2

Les fonctions de croyance construites dans la méthode précédente sont indépendantes de la classe sélectionnée par le classifieur. Les jeux de masses sont alors identiques quelque soit la classe retenue par le classifieur. Ainsi un classifieur ayant une taux de reconnaissance global important produira une fonction de croyance importante même pour une classe qu'il reconnaîtra très peu. Il apparaît alors plus judicieux de prendre en compte des taux de reconnaissance par classe. C'est cette approche qui est proposée avec cette méthode. Nous obtenons alors les fonctions de croyance de la manière suivante :

$$\begin{cases} m_j(\{\omega_k\}) & = \frac{n_{kk}}{\sum\limits_{i=1}^{N} n_{ki}} \\ m_j(\overline{\omega_k}) & = 1 - \frac{n_{kk}}{\sum\limits_{i=1}^{N} n_{ki}} \end{cases} \tag{11}$$

lorsque le classifieur $\mathcal{C}^j$ sélectionne la classe ω_k.

3.2.3 Méthode n°3

Comme pour les méthodes précédentes, celle-ci permet de construire des fonctions de croyance à partir de la matrice de confusion. Au contraire de la méthode n°2, elle permet de placer de la masse de croyance non seulement sur la classe retenue par le classifieur, son complémentaire mais aussi sur le cadre de discernement Ω. Les croyances se répartissent alors de la manière suivante :

$$\begin{cases} m'_j(\{\omega_i\}) & = \frac{n_{ki}}{\sum\limits_{j=1}^{N} n_{jk}} \qquad \forall i = 1, \ldots, N \\ m'_j(\Omega) & = \frac{\sum\limits_{j=1}^{N} n_{kj} - n_{kk}}{\sum\limits_{j=1}^{N} n_{kj}} \end{cases} \tag{12}$$

lorsque le classifieur $\mathcal{C}^j$ sélectionne la classe ω_k. Les fonctions de masse m'_j ne sont pas normalisées, il faut alors passer par l'étape de normalisation suivante :

$$m_j(A) = \frac{m'_j(A)}{\sum\limits_{B \subseteq \Omega} m'_j(B)} \qquad \forall A \subseteq \Omega \tag{13}$$

4 Résultats

4.1 Les données traitées

Les données utilisées, dans cette étude, ont été recueillies à l'issue de la campagne de mesures réalisée à la ZELT (Zone Expérimentale, Laboratoire de Trafic) de Toulouse. L'enquête de mesure a consisté en un recueil de données issues de plusieurs sources d'information :

- capteurs à boucles magnétiques,
- véhicules traceurs,

	Véhicule Traceur	Boucle Magnétique
Pourcentage de Classification	26.57	27.27

TAB. 1 – *Pourcentage de bonne classification sans le processus de fusion pour les véhicules traceurs et les boucles magnétiques.*

– une enquête de reconnaissance des véhicules par les plaques minéralogiques.

Les 2 premières sources d'information seront exploitées afin de définir une estimation du temps de parcours alors que la dernière source d'information sera considérée comme le temps de parcours de référence.

Les boucles magnétiques nous donnant des informations trafic (débit, taux d'occupation) alors que les véhicules traceurs nous donne un temps de parcours, il est nécessaire de transformer les informations fournies par les boucles magnétiques en une estimation du temps de parcours. Pour cela, nous utilisons la méthode proposée par Bonvalet et Robin-Prévallée (1987) qui est aussi l'approche retenue dans les premiers travaux sur l'estimation du temps de parcours par les fonctions de croyance (Faouzi (2000)).

Le recueil d'information a permis d'obtenir 230 données en ce qui concerne les boucles magnétiques et l'enquête minéralogique (temps de parcours de référence). Les véhicules traceurs ont fourni eux 156 observations. En réalisant, le croisement de ces deux fichiers, il apparaît que 143 observations sont communes. Pour la suite des essais, nous ne travaillerons que sur ces données.

Par ailleurs, pour les tests nous avons choisi de discrétiser notre temps de parcours en 6 classes. Ces classes sont définies à partir des temps de parcours de référence (enquête minéralogique) de manière à avoir le même nombre d'observations dans chaque classe (méthode des quantiles).

Afin d'évaluer les performances des différentes stratégies de fusion mises en œuvre dans cet article, nous présentons dans la tableau TAB. 1 les résultats de classification sans ces approches c'est-à-dire en prenant de manière individuel les véhicules traceurs et les boucles magnetiques. On peut constater sur ce tableau que l'estimation issue des boucles magnétiques semble donner de meilleurs résultats que ceux issus des véhicules traceurs.

4.2 Tests sur la constitution de la base d'apprentissage

Pour vérifier l'influence de la constitution de la base d'apprentissage nous avons fait varier le nombre d'observations dans la base d'apprentissage de 5 à 95% de la population initiale. Cela consiste alors à prendre une partie des observations initiales qui sont employées dans la base d'apprentissage et le restant des observations constituant alors la base de test. Afin de moyenner les résultats, 300 tirages de bases d'apprentissage ont été réalisés pour une proportion donnée. Ainsi sur les différentes courbes, nous avons représenté la moyenne, la valeur minimale et maximale ainsi que l'écart type du pourcentage de bonne classification.

La figure FIG. 1 montre l'évolution du pourcentage de bonne classification obtenu par la méthode de fusion de classifieurs proposée par Xu et al. (1992), en fonction du pourcentage de points dans la base d'apprentissage.

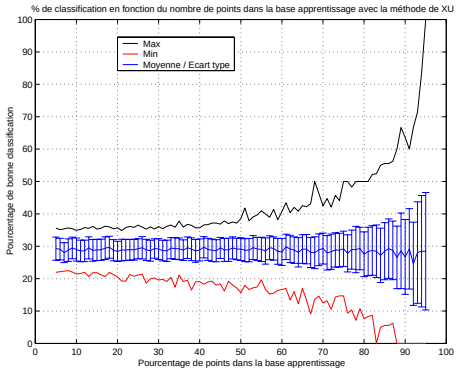


FIG. 1 – *Pourcentage de bonne classification en fonction du nombre de points dans la base d'apprentissage pour la méthode de XU.*

La figure FIG. 2 montre l'évolution du pourcentage de bonne classification en fonction du pourcentage de points dans la base d'apprentissage en utilisant l'amélioration, proposée dans Aissa et al. (2004), de la méthode de Xu .

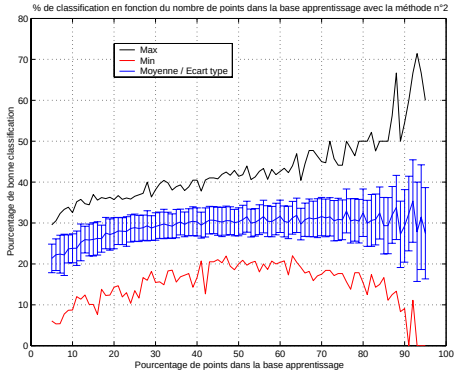


FIG. 2 – *Pourcentage de bonne classification en fonction du nombre de points dans la base d'apprentissage pour la méthode n°2.*

La figure FIG. 3 montre l'évolution du pourcentage de bonne classification en fonction du pourcentage de points dans la base d'apprentissage pour la méthode n°3 Aissa et al. (2004).

La figure FIG. 4 récapitule les pourcentages de bonne classification pour les 3 méthodes de fusion de classifieurs.

On peut constater sur ces figures que les performances obtenues par la méthode de Xu sont moins sensibles aux nombres d'observations constituants la base d'apprentissage. Toutefois, ces performances restent globalement moins bonnes que celles obtenues par les méthodes n°2 et n°3 lorsque 20% des observations initiales constituent la base d'apprentissage. Par ailleurs,

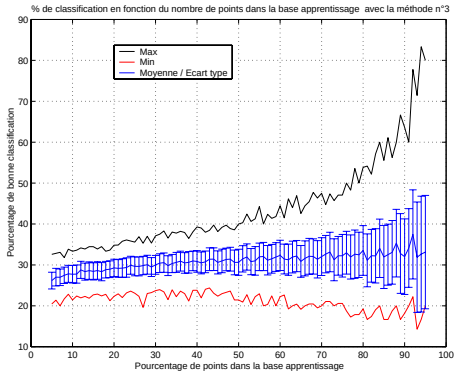


FIG. 3 – *Pourcentage de bonne classification en fonction du nombre de points dans la base d'apprentissage pour la méthode n°3.*

la méthodes n°3 donne les meilleurs résultats, en terme de pourcentage de bonne classification, une fois que la base d'apprentissage est constituée de plus de 20% des observations initiales.

La figure FIG. 5 représente respectivement l'évolution du pourcentage de classification en fonction de la constitution de la base d'apprentissage en travaillant avec des méthodes de classification. La méthode KNN-DS donne des résultats de bonne classification inférieurs à 20% quelque soit la constitution de la base d'apprentissage. La méthode ProDS donne elle des résultats similaires à ceux obtenus par la méthode n°3 de fusion de classifieurs.

5 Conclusion

L'objectif du travail effectué dans cet article est de proposer un cadre méthodologique ainsi que des solutions au problème d'estimation de temps de parcours en présence de données issues de sources hétérogènes. Deux sources ont été considérées ici : des boucles électromagnétiques, qui permettent d'obtenir une estimation de temps de parcours moyen, et un échantillon réduit de véhicules traceurs, qui recueillent les temps de parcours qu'ils ont réalisés. Dans cet article, nous abordons le problème de l'estimation du temps de parcours comme un problème typique de fusion de classifieurs en utilisant la théorie des croyances pour sa finesse de modélisation de la connaissance. Dans le cadre de cette théorie, nous avons présenté deux approches afin de modéliser les informations boucles magnétiques et véhicules traceurs en fonctions de croyance.

La première technique repose sur le concept de fusion de classifieurs. Dans ce cas, chaque source d'information est considérée comme un classifieur. Pour chacun de ces classifieurs, nous définissons une matrice de confusion qui reflète le pouvoir de discrimination des sources. Cette matrice nous permet par la suite de construire les fonctions de croyance.

La seconde approche emploie une technique classique de classification basée sur le calcul de distance pour construire les fonctions de croyance.

L'utilisation de ces deux approches dans le cadre de notre application réelle s'est avérée efficace comparativement à des approches mono-capteur. Cela est d'autant plus vrai pour une technique original de fusion de classifieurs que nous avons mise en oeuvre (Méthode n°3)

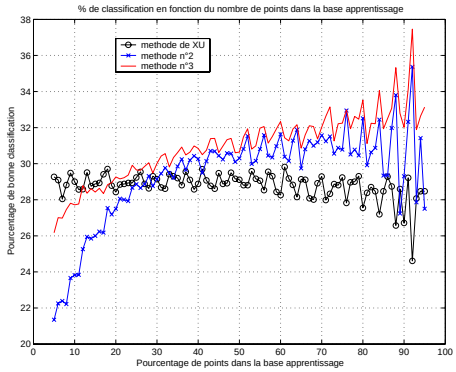


FIG. 4 – *Comparaison du pourcentage de bonne classification en fonction du nombre de points dans la base d'apprentissage pour les 3 méthodes présentées.*

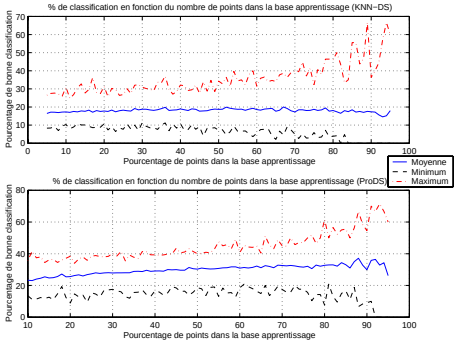


FIG. 5 – *Pourcentage de bonne classification en fonction du nombre de points dans la base d'apprentissage pour les KNN-DS et ProDS.*

et pour une approche classique de classification par la théorie de l'évidence (ProDS). Il est évident que l'on ne peut pas généraliser ces résultats car ils ont été obtenus dans le cadre d'une application spécifique. Toutefois, ces premiers travaux nous permettent de penser que l'emploi de techniques de fusion de données basées sur la théorie de l'évidence tendrai à améliorer l'estimation du temps de parcours.

Les perspectives liées à ce travail concernent tout d'abord l'apprentissage adaptatif. Pour cette perspective, il s'agira de construire une base d'apprentissage pour chaque créneau horaire de la journée (l'amplitude du créneau horaire restant à définir). Ensuite, selon l'heure de la journée où l'on désire connaître le temps de parcours, on utilise la base d'apprentissage adéquate. Ce principe permettra de prendre en compte l'aspect dynamique et donc évolutif de l'application.

La dernière perspective concerne l'emploi des différents développements récents sur la prise en compte des aspects continus dans le cadre de la théorie de l'évidence. Ces travaux

permettraient d'estimer les temps de parcours non plus de manière discrète comme cela est le cas actuellement mais de façon continu permettant ainsi une estimation plus précise.

Références

Aissa, A. B., N. E. E. Faouzi, et E. Lefevre (2004). Classification multisource via la théorie des fonctions de croyance. In *Workshop Fouille de Données Complexes Dans un Processus D'extraction de Connaissances, 4ème Journée d'Extraction Des Connaissances, EGC'2004*, pp. 31–44.

Bonvalet, F. et Y. Robin-Prévallée (1987). Mise au point d'un indicateur permanent des conditions de circulation en ile-de-france. *Transport Environnement Circulation* (84-85).

Dempster, A. (1968). A generalization of bayesian inference. *Journal of Royal Statistical Society, Serie B 30*, 205–247.

Denoeux, T. (1995). A k-nearest neighbour classification rule based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man and Cybernetics 25*(5), 804–813.

Denoeux, T. (2000). A neural network classifier based on Demspter-Shafer theory. *IEEE Transactions on Systems, Man and Cybernetics Part A 30*(2), 131–150.

Faouzi, N. E. E. (2000). Fusion de données pour l'estimation des temps de parcours via la théorie de l'évidence. *Recherche Transport Sécurité* (68), 15–30.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton, New Jersey : Princeton University Press.

Smets, P. et R. Kennes (1994). The transferable belief model. *Artificial Intelligence 66*(2), 191–234.

Xu, L., A. Krzyzak, et C. Y. Suen (1992). Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition. *IEEE Transactions on Systems, Man and Cybernetics 22*(3), 418–435.

Zouhal, L. M. (1997). *Contribution à l'application de la théorie des fonctions de croyance en reconnaissance des formes*. Ph. D. thesis, Université de Technologie de Compiègne.

Zouhal, L. M. et T. Denoeux (1998). An evidence-theoretic k-NN rule with parameter optimization. *IEEE Transactions on Systems, Man and Cybernetics - Part C 28*(2), 263–271.

Summary

The paper address the road travel time on an urban axis by classification method based on evidence theory. The information used to estimate the travel time are complementary and redundancy. It is then necessary to implement a strategy of multi-sensors data fusion. The selected approach is the evidence theory. This theory takes into account more naturally the imprecision and uncertainty of the used data. Two strategies are implemented. The first one is a approach by classifier fusion where each information source are considered as a classifier. The second approach is a classification approach based on distance to modeling the belief function. Results of these approaches, on data collected on an urban axis of Toulouse, show the advantages of fusion in the framework of this application.

Classification de flots de séquences basée sur une approche centroïde

Alice Marascu, Florent Masseglia

INRIA Sophia Antipolis,
Equipe AxIS,
2004 route des Lucioles - BP 93,
06902 Sophia Antipolis, France
{Alice.Marascu,Florent.Masseglia}@sophia.inria.fr

Résumé. Les flots de données séquentielles (data streams) se trouvent impliqués dans des domaines de plus en plus nombreux. Dans un processus de fouille appliqué sur un data stream, l'utilisation de la mémoire est limitée, de nouveaux éléments sont générés en permanence et doivent être traités le plus rapidement possible, aucun opérateur bloquant ne peut être appliqué sur les données et celles-ci ne peuvent être observées qu'une seule fois. A l'heure actuelle, il existe très peu de méthode de classification non supervisée des séquences dans un data stream. Notre objectif, dans cet article, est de montrer que cette classification est possible grâce à une approche centroïde, qui permettra de maintenir un représentant de chaque classe. Une attention particulière sera portée à une mesure de la qualité des classes produites en cours de traitement. Cette mesure permettra de redistribuer certaines séquences si besoin. Nos expérimentations montrent que notre approche permet d'obtenir les classes de façon très rapide.

1 Introduction

Depuis peu, des applications émergentes comme (entre autres) l'analyse du trafic réseaux, la détection de fraude ou d'intrusion, la fouille de clickstream[1] ou encore l'analyse des données issues de capteurs ont introduits de nouveaux types de contraintes pour les méthodes de fouille. Ces applications ont donné lieu à une forme de données connues sous le nom de "data streams". Dans le contexte des data streams l'utilisation de la mémoire doit être réduite, les données sont générées de manière continue et très rapide, les opérations bloquantes ne sont pas envisageables et, enfin, les nouvelles données doivent être prises en compte aussi vite que possible. Dans ce domaine, l'approximation a rapidement été reconnue comme un facteur clé pour fournir des motifs à la vitesse imposée par l'application (Garofalakis et al. (2002), Teng et al. (2003), Giannella et al. (2003)). Dans Marascu et Masseglia (2006) nous avons proposé SMDS, une méthode d'extraction de motifs séquentiels dans les data streams. Parmi les différentes fonctionnalités de cette méthode, la segmentation des séquences est utilisée comme préalable à l'extraction des motifs fréquents. Nous y avons montré l'efficacité d'une approche

[1] clickstream : flot de requêtes d'un utilisateur sur un site Web

heuristique gloutonne pour la classification des séquences.
Dans cet article, nous proposons l'algorithme SCDS (Sequence Clustering in Data Streams) qui a pour but d'améliorer les performances de l'algorithme de classification utilisé dans SMDS. Le principe de SCDS est de comparer chaque séquence à un représentant de chaque cluster. La méthode calcul de ce représentant est basée sur des techniques d'alignement de séquence. De plus, ce représentant sera surveillé par SCDS en permanence afin de garantir la qualité du cluster. En effet nous proposons une mesure du déséquilibre de la séquence alignée qui permet de décider si un cluster doit être divisé. SCDS est implémenté et a été testé sur des données réelles issues du serveur Web de l'Inria Sophia Antipolis. Ces données collectent les informations sur les usages qui sont faits d'un site Web. Les techniques d'analyse de ces usages (WUM ou Web Usage Mining) fournissent des informations sur le comportement des utilisateurs du site. Notre objectif est d'extraire des classes de comportement à partir des flots de données d'usage d'un site Web. Nous montrerons que SCDS satisfait les contraintes liés à la rapidité du data stream et peut être inclus dans un environnement temps réel.
Cet article est organisé de la manière suivante : tout d'abord nous présentons les concepts d'extraction de motifs séquentiels dans la section 2. Ensuite la section 3 expose la technique que nous avons développée dans ce travail afin de proposer une classification des séquences issues d'un data stream. Nous proposons des expérimentations en section 4 avant de conclure.

2 Définitions

2.1 Motifs séquentiels

Ce paragraphe expose et illustre la problématique liée à l'extraction de motifs séquentiels dans de grandes bases de données. Il reprend les différentes définitions proposées dans Agrawal et Srikant (1995), Srikant et Agrawal (1996) et Masseglia et al. (1998). La notion de séquence est définie de la manière suivante :

Définition 1 *Une transaction constitue, pour un client C, l'ensemble des items achetés par C à une même date. Dans une base de données client, une transaction s'écrit sous forme d'un triplet : <id-client, id-date, itemset>. Un itemset est un ensemble non vide d'items noté $(i_1 i_2 \dots i_k)$ où i_j est un item (il s'agit de la représentation d'une transaction non datée). Une séquence est une liste ordonnée, non vide, d'itemsets notée $< s_1 s_2 \dots s_n >$ où s_j est un itemset (une séquence est donc une suite de transactions avec une relation d'ordre entre les transactions). Une séquence de données est une séquence représentant les achats d'un client. Soit $T_1,\ T_2, \dots,\ T_n$ les transactions d'un client, ordonnées par date d'achat croissante et soit itemset(T_i) l'ensemble des items correspondants à T_i, alors la séquence de données de ce client est :*
$< itemset(T_1)\ itemset(T_2) \dots\ itemset(T_n) >$.

Exemple 1 *Soit C un client et S=$< (3)\ (4\ 5)\ (8) >$, la séquence de données représentant les achats de ce client. S peut être interprétée par "C a acheté l'item 3, puis en même temps les items 4 et 5 et enfin l'item 8".*

Dans cet article nous considérons des flots de données séquentielles. Il s'agit donc à chaque instant i d'entrées de la forme $[C,\ T_1,\ T_2, \dots,\ T_n]$ qui représentent les transactions T_1 à T_n du client C à l'instant i.

3 SCDS : principe général

Notre méthode est basée sur un environnement de découpage du data stream en "batches" de taille fixe. Soient $B_1, B_2, ...B_n$, les batches et B_n, le batch courant. Le principe de SCDS est de segmenter les séquences contenues dans chaque batch b de $[B_1..B_n]$.

Dans le but d'obtenir une classification des navigations aussi rapidement que possible, notre approche fonctionne de la manière suivante : l'algorithme est initialisé avec une seule classe, qui contient la première navigation. Ensuite, pour chaque navigation n dans le batch, n est comparée avec chaque cluster c. Soit c le cluster dont le centroïde est le plus proche de n, alors n est insérée dans c. Si n n'est insérée dans aucun cluster, alors un nouveau cluster est crée et n est insérée dans ce nouveau cluster. Trois étapes sont donc essentielles dans ce processus. La première est le calcul du centroïde ς_c du cluster c. Ce calcul est détaillé dans la section 3.1. Ensuite pour comparer la séquence de navigation n avec le cluster c, nous proposons de définir la similitude entre n et le centroïde de c. Cette étape est expliquée dans la section 3.2. Enfin, l'ajout d'une séquence dans un cluster implique de mettre à jour son centroïde. Nous mesurons la qualité du cluster grâce à ce centroïde et nous proposons des techniques de détection de la dégradation du cluster dans la section 3.3.

3.1 Calcul du centroïde

Le centroïde du cluster est déterminé par une technique d'alignement appliquée sur le cluster (comme Kum et al. (2003) et Hay et al. (2002) l'ont déjà utilisée pour la fouille de bases de données statiques). A l'initialisation d'un cluster son centroïde est la séquence unique qu'il contient.
L'alignement des séquences renvoie une séquence alignée du type :

$$SA =< I_1 : n_1, I_2 : n_2, ..., I_r, n_r >: m$$

Dans cette représentation, m représente le nombre total de séquences impliquées dans l'alignement. I_p ($1 \leq p \leq r$) est un itemset représentée sous la forme $(x_{i_1} : m_{i_1}, ...x_{i_t} : m_{i_t})$, où m_{i_t} est le nombre de séquences qui contiennent l'item x_i à la p^{eme} position dans la séquence alignée. Enfin, n_p est le nombre d'occurrences de l'itemset I_p dans l'alignement. L'exemple 2 décrit le processus d'alignement de quatre séquences. À partir de deux séquences, l'alignement commence par insérer des itemsets vides (au début, au milieu ou à la fin des séquences) jusqu'à ce que les deux séquences contiennent le même nombre d'itemsets.

Exemple 2 *considérons les séquences suivantes : $S_1 =<$ (a,c) (e) (m,n) $>$, $S_2 =<$ (a,d) (e) (h) (m,n) $>$, $S_3 =<$ (a,b) (e) (i,j) (m) $>$ et $S_4 =<$ (b) (e) (h,i) (m) $>$. Les étapes conduisant à l'alignement de ces séquences sont détaillées dans la figure 1. Tout d'abord, un itemset vide est inséré dans S_1. Ensuite S_1 et S_2 sont alignées dans le but de produire SA_{12}. Le processus d'alignement est alors appliqué entre SA_{12} et S_3. La méthode d'alignement continue à traiter les séquences deux par deux jusqu'à la dernière séquence.*

À la fin du processus d'alignement, la séquence alignée (SA_{14} dans la figure 1) est considéré comme le centroïde du cluster. Dans SCDS, l'alignement se fait de manière incrémentale à chaque ajout d'une séquence dans le cluster. Pour cela nous maintenons une matrice de comptage des items dans chaque séquence et un tableau des distances entre chaque séquence

Etape 1 :				
S_1 :	<(a,c)	(e)	()	(m,n)>
S_2 :	<(a,d)	(e)	(h)	(m,n)>
SA_{12} :	(a :2, c :1, d :1) :2	(e :2) :2	(h :1) :1	(m :2, n :2) :2
Etape 2 :				
SA_{12} :	(a :2, c :1, d :1) :2	(e :2) :2	(h :1) :1	(m :2, n :2) :2
S_3 :	<(a,b)	(e)	(i,j)	(m)>
SA_{13} :	(a :3, b :1, c :1, d :1) :3	(e :3) :3	(h :1, i :1, j :1) :2	(m :3, n :2) :3
Etape 3 :				
SA_{13} :	(a :3, b :1, c :1, d :1) :3	(e :3) :3	(h :1, i :1, j :1) :2	(m :3, n :2) :3
S_4 :	<(b)	(e)	(h,i)	(m)>
SA_{14} :	(a :3, b :2, c :1, d :1) :4	(e :4) :4	(h :2, i :2, j :1) :3	(m :4, n :2) :4

FIG. 1 – *Etapes de l'alignement de séquences*

et les autres. Ces éléments sont illustrés par la figure 2. La matrice (à gauche) stocke pour chaque séquence le nombre d'apparition de chaque item dans cette séquence. Par exemple la séquence s_1 contient deux fois l'item a. Le tableau des distances stocke la somme des similitudes ($similMatrice$) entre chaque séquence et les autres séquences du cluster. Soit s_{1_i} le nombre d'apparitions de l'item i dans la séquence s_1 et m le nombre total d'items. $similMatrice$ est calculé grâce à la matrice de la manière suivante :
$similMatrice(s_1, s_2) = \sum_{i=1}^{m} min(s_{1_i}, s_{2_i})$.

Par exemple, avec les séquences s_1 et s_2 de la matrice donnée à la figure 2 cette somme vaut $s_{1_a} + s_{2_b} + s_{2_c} = 1 + 0 + 1 = 2$.

Cet alignement n'est cependant pas toujours calculé de manière incrémentale. Considérons l'ajout d'une séquence s_n. Tout d'abord s_n est ajoutée à la matrice et sa distance aux autres séquences est calculée ($\sum_{i=1}^{n} similMatrice(s_n, s_i)$). s_n est alors insérée dans le tableau de distances, en gardant l'ordre décroissant des valeurs de distances. Par exemple, dans la figure 2, s_n est insérée après s_2. Soit r le rang auquel s_n est insérée (dans notre exemple, $r = 2$) dans c. Il y a alors deux possibilités après l'insertion de s_n :

1. $r > 0.5 \times |c|$. Dans ce cas, l'alignement est calculé de manière incrémentale et $\varsigma_c = alignement(\varsigma_c, s_n)$.
2. $r \leq 0.5 \times |c|$. Dans ce cas il faut rafraîchir le centroïde du cluster et l'alignement est recalculé pour toutes les séquences du cluster.

3.2 Comparaison séquence/centroïde

Soit s la séquence à affecter dans un cluster et C l'ensemble des clusters. SCDS parcours l'ensemble des clusters de C et pour chaque cluster $c \in C$, effectue une comparaison entre s et ς_c (le centroïde de c, qui est donc un alignement). Cette comparaison est basée sur la plus longue sous-séquence commune (PLSC) entre s et ς_c. Ensuite, la longueur de la séquence est

Séq	a	b	c
s_1	2	0	1
s_2	1	0	1
$\vdots$			

Séq	$\sum_{i=1}^{n} similMatrice(s, s_i)$
s_1	16
s_2	14
s_n	13
s_3	11
$\vdots$	
s_{n-1}	1

FIG. 2 – *Distances entre les séquences*

également prise en compte car elle doit être comprise entre 80% et 120% de la longueur de la séquence alignée.

Définition 2 Soient s_1 et s_2 deux motifs séquentiels. Soit $|PLSC(s_1, s_2)|$ la longueur de la plus longue sous-séquence commune entre s_1 et s_2. La distance $dist(s_1, s_2)$ entre s_1 et s_2 est définie de la manière suivante : $dist = 1 - \frac{2 \times |PLSC(s_1,s_2)|}{longueur(s_1)+longueur(s_2)}$.

Soit t la longueur de la première séquence insérée dans c. Les conditions pour que s soit affectée à c sont donc les suivantes :

- $\forall d \in C/d \neq c, dist(s, \varsigma_c) \leq dist(s, \varsigma_d)$
- $0.8 \times t \leq |s| \leq 1.2 \times t$
- $dist(s, \varsigma_c) < 0.3$

La première condition assure que s est affectée dans le cluster dont le centroïde est le plus similaire à s. La deuxième condition assure que les clusters contiendront des séquences de taille homogène et que la taille moyenne des séquences d'un cluster variera peu. Enfin la troisième condition assure que si aucun centroïde ayant un degré de similitude supérieur à 70% avec s n'est trouvé, alors s n'est affecté à aucun cluster. Dans ce dernier cas, un nouveau cluster est crée et s y est affectée.

3.3 Détection des clusters dégradés

Lors de nos expérimentations, nous avons pu détecter la formation de clusters non optimaux. Ces clusters contiennent souvent une séquence répétée qui forme la majorité du cluster et quelques séquences "satellites" qui sont légèrement différentes. Notre objectif est de détecter ces clusters sans augmenter les temps de calcul afin de redistribuer leur contenu dans plusieurs sous-clusters. L'un de ces sous-clusters contiendrait alors les répétitions de la séquences "noyau" (séquence majoritaire du cluster) et un (ou plusieurs) cluster de taille moindre contenant les séquence satellites.

Nous avons déjà vu dans la section 3.1 que selon le niveau d'insertion d'une nouvelle séquence (son rang) l'alignement était recalculé pour tout le cluster. Il s'agit d'une première technique d'optimisation permettant de garder un centroïde le plus représentatif possible du cluster.

La deuxième technique que nous avons mise en place est une mesure permettant de définir le déséquilibre d'un alignement. Il s'agit de détecter d'après les valeurs (nombre d'oc-

currences) de la séquence alignée si il existe une séquence noyau et des séquences satellites. Prenons l'exemple du cluster suivant :

< (57) (68)>, < (57) (68)>, < (57) (563)>, < (68) (68)>, < (68)>, < (57) (68)>, < (57) (68)>, < (68)>, < (563)>

La séquence alignée de ce cluster est : < (57 :5, 68 :3) :8 (68 :5, 563 :2) :7 >. De notre point de vue, il serait préférable d'obtenir à partir de ce cluster les deux sous-clusters suivants :

Sous-cluster 1	Sous-cluster 2
< (57) (68)>, < (57) (68)>, < (68) (68)>, < (68)>, < (57) (68)>, < (57) (68)>, < (68)>	< (57) (563)>, < (563)>

Dans la séquence alignée du cluster d'origine il est possible de détecter un déséquilibre en comptant le nombre d'apparition des items. Par exemple l'item 68 apparaît 8 fois et fait donc partie de la séquence noyau. Par contre l'item 563 n'apparaît que 2 fois. Il est donc nécessaire de reconstruire ce cluster en séparant les séquences qui contiennent l'item 563 des autres séquences. Cela permet d'obtenir les deux sous-clusters décrits dans le tableau ci-dessus.

Plus formellement, soit a le nombre maximum d'occurrences d'un item et b le nombre minimum. Soit t la longueur minimum d'une séquence de ce cluster s. Si $\frac{a-b}{|c|} \geq x$ (avec x un pourcentage défini par l'utilisateur) alors le cluster sera divisé. Dans notre exemple $t = 2$, $a = 8$, $b = 2$ et $|c| = 9$. Pour diviser ce cluster, le paramètre spécifié par l'utilisateur doit donc être $x = 66\%$.

4 Expérimentations

SCDS a été implémenté en Java sur un Pentium (2,1 Ghz) exploité par un système Linux Fedora. Nous avons évalué notre proposition sur des réelles issues des usages du Web de l'Inria Sophia Antipolis.

4.1 Temps de réponse et robustesse de SCDS

Dans le but de montrer l'efficacité de SCDS, nous reportons à la figure 3 le temps nécessaire pour classer les séquences sur chaque batch correspondant à des données d'usage sur le site Web de l'Inria. Les données ont été collectées sur une période de 14 mois et représentent 14 Go. Le nombre total de navigations est de 3,5 millions pour 300000 navigations. Nous avons découpé le fichier log en batches de 4500 transactions (soit environ 1500 séquences en moyenne). Nous avons comparé ce temps de réponse à celui de SMDS Marascu et Masseglia (2006) qui n'optimise pas la phase de clustering. En effet dans SMDS, la séquence à classer s est comparée à toutes les séquences de tous les clusters, jusqu'à ce que l'un des clusters présente une séquence compatible avec s. Nous pouvons observer que le temps de réponse de SCDS varie de 1000 ms à 2000 ms alors que le temps d'exécution de SMDS varie de 2500 ms à 4000 ms. Nous avons ajouté à la figure 3 le nombre de séquences de chaque batch pour expliquer les différences de temps d'exécution d'un batch à un autre. On peut observer, par exemple, que le batch 1 contient 1750 séquences et que SCDS demande 1400 ms pour en extraire les motifs séquentiels.

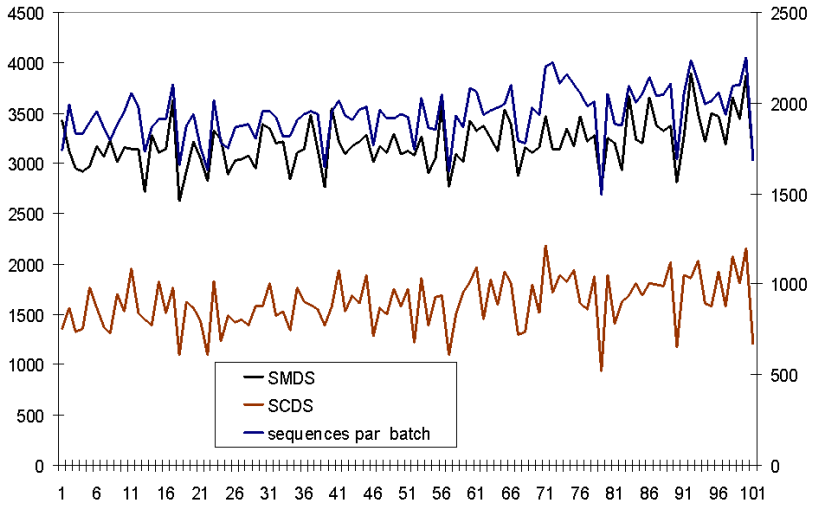


FIG. 3 – *Temps d'exécution de SCDS.*

4.2 Analyse de la qualité des clusters

Afin de mesurer la qualité des classes produites par SCDS, notre principal outil sera la distance entre deux séquences. Soit s_1 et s_2, deux séquences, la distance $dist(s_1, s_2)$ entre s_1 et s_2 est basée sur $sim(s_1, s_2)$, la mesure de similitude donnée par la définition 2 et telle que $dist(s_1, s_2) = 1 - sim(s_1, s_2)$. On a donc $dist(s_1, s_2) \in [0..1]$ et $dist(s_1, s_2)$ proche de 0 signifie que les séquences sont proches (similaire si cette valeur est nulle) alors que $dist(s_1, s_2)$ proche de 1 signifie que les séquences sont éloignées (ne partagent aucun item si cette valeur est 1). Nous reportons dans la figure 4 la double moyenne DBM après avoir traité chaque batch. DBM est calculée de la manière suivante : soit C l'ensemble des classes, $DBM = \frac{\sum_{i \in C} \sum_{x \in C_i} dist(x, c_i)}{|C|}$ avec c_i le centre de C_i (la i^{eme} classe). La valeur finale de DBM à la fin du batch est donnée par la figure 4. On peut y observer que DBM est comprise entre 15% et 45%. A la fin du processus, la valeur moyenne de DBM est de 28% (une qualité moyenne des classes de 72%). Ce résultat reste à améliorer et nous explorons actuellement les pistes qui permettront d'obtenir une valeur de DBM en fin de batch la plus faible possible.

5 Conclusion

Dans ce papier, nous avons proposé la méthode SCDS pour classer les séquences dans les data streams. Notre algorithme repose sur une technique d'alignement des séquences et une détection du déséquilibre de cet alignement afin de proposer de rafraîchir les clusters. Nos expérimentations ont montré que SCDS traite le data stream assez rapidement pour être intégré dans un contexte temps réel. En effet les temps de réponse de SCDS montrent que ce dernier

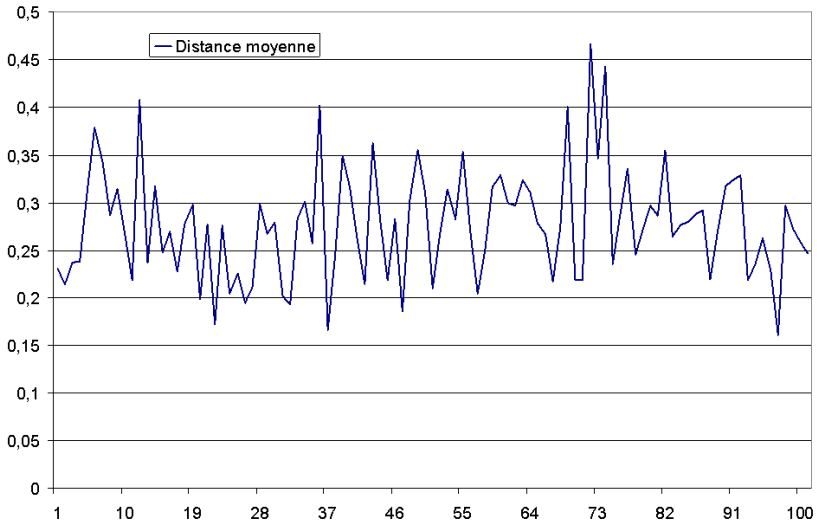


FIG. 4 – *Distance globale, batch par batch*

est bien plus performant que SMDS, proposé dans une communication précédente. Cependant, la qualité des clusters, lors de nos analyses, montre qu'il reste des efforts à fournir afin de proposer des clusters plus cohérents. En effet, la distance moyenne entre le centre du cluster et les autres séquences du cluster montre que les clusters pourraient être plus homogènes. Nous explorons des solutions possibles à ce problème à l'heure actuelle.

Références

Agrawal, R. et R. Srikant (1995). Mining Sequential Patterns. In *Proceedings of the 11th International Conference on Data Engineering (ICDE'95)*, Taiwan.

Garofalakis, M., J. Gehrke, et R. Rastogi (2002). Querying and mining data streams : you only get one look a tutorial. In *SIGMOD '02 : Proceedings of the 2002 ACM SIGMOD international conference on Management of data*.

Giannella, C., J. Han, J. Pei, X. Yan, et P. Yu (2003). *Mining Frequent Patterns in Data Streams at Multiple Time Granularities*. In H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yesha (eds.), Next Generation Data Mining. AAAI/MIT.

Hay, B., G. Wets, et K. Vanhoof (2002). Web Usage Mining by Means of Multidimensional Sequence Alignment Method. In *WEBKDD*, pp. 50–65.

Kum, H., J. Pei, W. Wang, et D. Duncan (2003). ApproxMAP : Approximate mining of consensus sequential patterns. In *Proceedings of SIAM Int. Conf. on Data Mining*, San Francisco, CA.

Marascu, A. et F. Masseglia (2006). Extraction de motifs séquentiels dans les flots de données d'usage du Web. In *Actes des 6èmes journées "extraction et gestion des connaissances" (EGC'06)*, Lille, France.

Masseglia, F., F. Cathala, et P. Poncelet (1998). The PSP Approach for Mining Sequential Patterns. In *Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery*, Nantes, France.

Srikant, R. et R. Agrawal (1996). Mining Sequential Patterns : Generalizations and Performance Improvements. In *Proceedings of the 5th International Conference on Extending Database Technology (EDBT'96)*, Avignon, France, pp. 3–17.

Teng, W.-G., M.-S. Chen, et P. S. Yu (2003). A Regression-Based Temporal Pattern Mining Scheme for Data Streams. In *VLDB*, pp. 93–104.

Summary

Data streams are often involved in domains getting more and more numerous. In a data mining process applied to streaming data, the usage of memory is limited, new elements are generated continuously and have to be considered as fast as possible, no blocking operation can be performed on the data and those data can be observed only once. At this time, there is only a few studies dedicated to clustering sequences in a data stream. Our goal, in this paper, is to provide a clustering method that will be based on a centroid approach. This approach will allow to maintain the representative of each cluster. A particular attention will be given to the quality of the clusters throughout the process. This measure will allow us to redistribute some sequences if needed. Our experiments show that our approach allow to discover the clusters in a very efficient way.

Techniques de généralisation des URLs pour l'analyse des usages du Web

Yves Lechevallier*, Florent Masseglia**, Doru Tanasa**, Brigitte Trousse**

*Projet AxIS, INRIA Rocquencourt,
Domaine de Voluceau-Rocquencourt - BP 105
78153 Le Chesnay Cedex, France
**Projet AxIS, INRIA Sophia Antipolis,
2004 route des Lucioles - BP 93
06902 Sophia Antipolis, France
{Prénom.Nom}@inria.fr
http://www-sop.inria.fr/axis/

Résumé. L'analyse des usages d'un site Web à partir d'une extraction de motifs est souvent limitée par le faible support de ces motifs. Cela est dû principalement à la grande diversité des pages et des comportements. Il est pourtant possible de regrouper la plupart des pages dans différentes catégories lors d'un pré-traitement. Travailler sur ces catégories, plutôt que sur les URLs, peut permettre de faire émerger certains comportements de manière "générique". Cet article présente une méthodologie originale d'analyse des usages du Web à partir d'une généralisation des URLs. Cette généralisation est réalisée à partir d'une catégorisation des URLs à l'aide d'informations extraites soient des pages elles-mêmes soient à partir de l'accès à ces pages par les internautes. Nous présentons ensuite une expérimentation relative à une généralisation des URLs basée sur des informations relatives aux accès aux pages : celle-ci permet de mettre en avant les changements de support des motifs extraits selon qu'ils sont obtenus avec ou sans généralisation. Pour conclure, nous présentons quelques pistes de prolongement de ces premiers résultats

1 Introduction

Le Web Usage Mining désigne l'ensemble des techniques basées sur le Data Mining (ou Fouille de Données) pour analyser le comportement des utilisateurs d'un site web (Cooley et al. (1999); Masseglia et al. (2000); Mobasher et al. (2002); Spiliopoulou et al. (1999)). Reposant généralement sur la quantité de données enregistrées dans les fichiers de type access log, ces méthodes permettent de mettre en évidence des comportements fréquents. En particulier, l'extraction de motifs séquentiels (Agrawal et Srikant (1995), Masseglia et al. (1998), Pei et al. (2001), Wang et Han (2004), Kum et al. (2003), Tanasa (2005)) est adaptée au contexte de l'analyse des logs, étant donnée la nature temporelle de leurs enregistrements. Sur un portail d'ordre général, on pourrait découvrir par exemple que « 25% des utilisateurs naviguent sur le

site dans un ordre particulier, en consultant la page d'accueil puis la page concernant la guerre en Irak, puis le CAC40 puis reviennent sur la page d'accueil avant de consulter leur mail en tant qu'abonné ».
En théorie cette analyse permet de mettre en évidence des comportements fréquents assez facilement. Cependant, la réalité montre que la diversité des pages et des comportements rend cette approche délicate. En effet il faut souvent chercher des seuils de fréquence de l'ordre de 1% ou 2% avant de révéler des comportements. De tels supports combinés à des caractéristiques de fichiers importantes (nombre d'enregistrements dans les logs) sont le plus souvent la source d'échecs ou de limitations des techniques existentes d'analyse des usages.
Une solution à ce phénomène consiste à regrouper les pages par thème, sous forme de taxonomie par exemple, afin d'obtenir un comportement plus global. En reprenant l'exemple du début, on aurait pu obtenir : « 70% des utilisateurs naviguent sur le site dans un ordre particulier, en consultant la page d'accueil puis une page de news, puis une page sur le marché financier, puis reviennent sur la page d'accueil avant de consulter un service de communication offert par le portail ». Une page sur le marché financier pouvant concerner aussi bien le CAC40 que le DOW JONES ou le NIKKEI (et de manière similaire : le mail ou le chat sont des services de communication, la guerre en Irak fait partie des news, etc.) et le fait de grouper ces pages sous le terme "marché financier" augmente directement le seuil de fréquence des comportements et donc leur lisibilité, leur pertinence et leur signification.
Le problème de la taxonomie vient du temps et de l'énergie nécessaires à sa mise en place et à sa maintenance. Dans cet article, nous proposons des solutions pour faciliter (ou guider le plus possible) la création automatique de cette taxonomie afin de rendre un processus de Web Usage Mining plus efficace et pertinent (catégoriser les pages selon l'organisation logique du site, selon leur contenu, par la façon dont elles sont accédées, par classification, etc...). Nous montrerons lors d'une expérimentation la pertinence de notre approche en termes d'efficacité de l'extraction des résultats.

L'article est structuré de la manière suivante : la section 2 présente les notions liées à l'extraction de motifs séquentiels et à l'analyse des usages du Web. La section 3 présente les travaux relatifs à ces thèmes en gardant un point de vue attentif à la temporalité des données. Nous détaillons les aspects de notre approche dans la section 4 et l'expérimentation menée est présentée en section 5. Enfin nous présentons nos conclusions en section 6.

2 Définitions

2.1 Motifs séquentiels

Ce paragraphe expose et illustre la problématique liée à l'extraction de motifs séquentiels dans de grandes bases de données. Il reprend les différentes définitions proposées dans Agrawal et al. (1993) et Agrawal et Srikant (1995).

Dans Agrawal et al. (1993), le problème de la recherche de règles d'association dans de grandes bases de données est défini de la manière suivante.

Définition 1 Soit $I = \{i_1, i_2, ..., i_m\}$, un ensemble de m achats (*items*). Soit $D = \{t_1, t_2, ...t_n\}$, un ensemble de n transactions ; chacune possède un unique identificateur appelé TID et porte

sur un ensemble d'items (*itemset*) I. I est appelé un *k-itemset* où k représente le nombre d'éléments de I. Une transaction $t \in D$ contient un itemset I si et seulement si $I \subseteq t$. Le *support* d'un itemset I est le pourcentage de transactions dans D contenant I : $supp(I) = \|\{t \in D \mid I \subseteq t\}\|/\|\{t \in D\}\|$. Une règle d'association est une implication conditionnelle entre les itemsets, $I_1 \Rightarrow I_2$ où les itemsets I_1, $I_2 \subset I$ et $I_1 \cap I_2 = \emptyset$. La *confiance* d'une règle d'association $r : I_1 \Rightarrow I_2$ est la probabilité conditionnelle qu'une transaction contienne I_2 étant donné qu'elle contient I_1. Le support d'une règle d'association est défini par $supp(r) = supp(I_1 \cup I_2)$. Etant donné deux paramètres spécifiés par l'utilisateur, *minsupp* et *minconfiance*, le problème de la recherche de règles d'association dans une base de données D consiste à rechercher l'ensemble des itemsets fréquents dans D, *i.e.* tous les itemsets dont le support est supérieur ou égal à *minsupp*. Puis, à partir de cet ensemble, de générer toutes les règles d'association dont la confiance est supérieure à *minconfiance*.

Pour étendre la problématique précédente à la prise en compte du temps des transactions, les mêmes auteurs ont proposé dans Agrawal et Srikant (1995) la notion de séquence définie de la manière suivante :

Définition 2 Une *transaction* constitue, pour un client C, l'ensemble des items achetés par C à une même date. Dans une base de données client, une transaction s'écrit sous forme d'un triplet : <id-client, id-date, itemset>. Un *itemset* est un ensemble non vide d'items noté $(i_1 i_2 \dots i_k)$ où i_j est un *item* (il s'agit de la représentation d'une transaction non datée). Une *séquence* est une liste ordonnée, non vide, d'itemsets notée $< s_1 s_2 \dots s_n >$ où s_j est un itemset (une séquence est donc une suite de transactions avec une relation d'ordre entre les transactions). Une *séquence de données* est une séquence représentant les achats d'un client. Soit $T_1, T_2, \dots, T_n$ les transactions d'un client, ordonnées par date d'achat croissante et soit itemset(T_i) l'ensemble des items correspondants à T_i, alors la séquence de données de ce client est $< itemset(T_1)\ itemset(T_2) \dots\ itemset(T_n) >$.

Exemple 1 *Soit C un client et S=$< (a)\ (d\ e)\ (h) >$, la séquence de données représentant les achats de ce client. S peut être interprétée par "C a acheté l'item a, puis en même temps les items d et e et enfin l'item h".*

Définition 3 Le $support$ de s, noté $supp(s)$, est le pourcentage de toutes les séquences dans D qui supportent (contiennent) s. Si $supp(s) \geq minsupp$, avec une valeur de support minimum $minsupp$ fixée par l'utilisateur, la séquence s est dite *fréquente*.

2.2 Adapter la problématique des Motifs Séquentiels aux logs d'accès Web

Ce paragraphe propose de reprendre les concepts essentiels d'un processus de Web Usage Mining, afin de présenter de façon synthétique, les procédés mis en œuvre lors de l'analyse du comportement des utilisateurs d'un site Web. Les principes généraux sont similaires à ceux du processus d'extraction de connaissances exposé dans Fayad et al. (1996). La démarche se décompose en trois phases principales. Tout d'abord, à partir d'un fichier de données brutes, un prétraitement décrit en Tanasa et Trousse (2004) est nécessaire pour éliminer les informations inutiles et organiser les données. Dans la deuxième phase, à partir des données transformées, des algorithmes de fouille de données sont utilisés pour extraire les itemsets ou les séquences

fréquentes. Enfin, l'exploitation par l'utilisateur des résultats obtenus est facilitée par un outil de requête et de visualisation.
Les données brutes sont collectées dans des fichiers access log des serveurs Web. Une entrée dans le fichier access log est automatiquement ajoutée chaque fois qu'une requête pour une ressource atteint le serveur Web (*demon http*). Les fichiers access log peuvent varier selon les systèmes qui hébergent le serveur, mais présentent tous en commun trois champs : l'adresse du demandeur, l'URL demandée et la date à laquelle cette demande a eu lieu. Parmi ces différents types de fichiers, nous avons retenu dans cet article le format spécifié par le CERN et la NCSA Consortium (1998), une entrée contient des enregistrements formés de 7 champs séparés par des espaces : `host user authuser [date:time] ''request'' status bytes`

```
138.96.69.8 - - [03/Mar/2003 :18 :42 :14 +0100] "GET /axis/presentation HTTP/1.1" - -
138.96.69.8 - - [03/Mar/2003 :18 :48 :00 +0100] "GET /aid/people/ HTTP/1.1" - -
138.96.69.8 - - [03/Mar/2003 :19 :03 :14 +0100] "GET /axis/ra/ HTTP/1.0" - -
```

FIG. 1 – *Exemple de fichier access log*

La figure 1 illustre un extrait de fichier access log du serveur Web de l'Inria Sophia Antipolis. Deux types de traitements sont effectués sur les entrées du log. Tout d'abord, le fichier access log est trié par adresse et par transaction.

Définition 4 Soit Log un ensemble d'entrées dans le fichier access log. Une entrée g, $g \in Log$, est un tuple $g =< ip_g, \{(l_1^g.URL, l_1^g.time), ..., (l_m^g.URL, l_m^g.time)\} >$ tel que pour $1 \leq k \leq m$, $l_k^g.URL$ représente l'objet demandé par le client g à la date $l_k^g.time$, et pour tout $1 \leq j < k$, $l_k^g.time > l_j^g.time$.

Toutes les dates sont également traduites en temps relatif par rapport à la plus petite date du fichier. Ensuite une étape d'élimination des données "non intéressantes" pour l'analyse (phase de sélection) est réalisée. La figure 2 illustre un exemple de fichier obtenu après la phase de pré-traitement. A chaque client correspond une suite de "dates" (événements) et la traduction de l'URL demandée par ce client à cette date.

Client	d1	d2	d3	d4	d5
1	a	c	d	b	c
2	a	c	f	b	c
3	a	g	c	b	c

FIG. 2 – *Exemple de fichier résultat issu de la phase de pré-traitement*

L'objectif est alors d'extraire de ce jeu de données les séquences fréquentes selon la définition 3. Les résultats obtenus sont du type `< ( a ) ( c ) ( b ) ( c ) >` (ici avec un support minimum de 100% et en appliquant les algorithmes de fouille de données sur le fichier représenté par la figure 2). Ce dernier résultat, une fois re-traduit en termes d'URL, confirme la découverte d'un comportement commun à $minsup$ utilisateurs et fournit l'enchaînement des pages qui constituent ce comportement fréquent.

3 Etat de l'art

Différentes techniques d'extraction de motifs séquentiels ont été appliquées aux logs d'accès Web (Masseglia et al. (2000); Spiliopoulou et al. (1999); Bonchi et al. (2001); Hay et al. (2004); Zhu et al. (2002); Nakagawa et Mobasher (2003); Masseglia et al. (2003); Tanasa (2005)). Le principal intérêt d'utiliser cette technique pour les données d'usage Web est la prise en compte de la temporalité comme dans les deux travaux que nous décrivons par la suite (Spiliopoulou et al. (1999); Masseglia et al. (2000)).

L'outil WUM (Web Utilisation Miner) proposé dans Spiliopoulou et al. (1999) permet la découverte de patrons de navigation qui sont intéressants du point de vue statistique ou de leur structure. L'extraction de motifs séquentiels proposée par WUM repose sur la fréquence (support minimum) des motifs considérés. On peut aussi spécifier un autre critère subjectif pour les patrons de navigations comme par exemple le fait de passer par des pages avec certaines propriétés ou que la confiance soit élevée entre deux ou plusieurs pages du patron de navigation.

Dans Masseglia et al. (2000), les auteurs proposent la plateforme WebTool. L'extraction des motifs dans WebTool repose sur PSP, un algorithme développé par les auteurs, dont l'originalité est de proposer un arbre préfixé pour gérer à la fois les candidats et les fréquents.

Malheureusement, la dimensionalité de ces données (tant en nombre d'items - pages - différents, qu'en nombre de séquences) pose des problèmes aux techniques d'extraction de motifs séquentiels. Plus précisément, à cause du nombre important d'items différents le nombre de résultats obtenus est très faible. La solution consisterait dans une baisse du support utilisé mais dans ce cas les algorithmes ne parviennent pas à finir et donc à fournir des résultats.
Une proposition a été faite pas les auteurs de Masseglia et al. (2003) qui se sont penchés sur la prise en compte des motifs séquentiels de support très faible en partant du constat que les supports élevés génèrent souvent des motifs évidents. Pour palier les difficultés qu'ont les méthodes d'extraction de motifs lors de la baisse du support les auteurs ont proposé de diviser le problème de manière récursive afin de procéder à une phase de fouille de données sur chaque sous-problème. Les sous-problèmes correspondent à des objectifs de navigations communs (sous-log contenant uniquement les utilisateurs qui sont passés par des pages similaires). Les motifs obtenus vont de la navigation sur une démonstration d'une équipe de recherche, jusqu'à un ensemble d'attaques pirates qui utilisent les mêmes techniques d'intrusion. Une autre proposition a été faite par Tanasa dans sa thèse (Tanasa (2005)), inscrivant ces deux propositions dans une méthodologie plus générale d'extraction de motifs séquentiels de faible support.

Une autre solution est de réduire le nombre d'items en utilisant une généralisation des URLs. Dans Fu et al. (2000) les auteurs utilisent une généralisation syntaxique des URLs avec un type diffèrent d'analyse (classification). Avant d'appliquer une classification par BIRCH (Zhang et al. (1996)), les rubriques syntaxiques de niveau supérieur à deux sont remplacées par leurs rubriques syntaxiques de niveau inférieur. Par exemple, au lieu de *http ://www-sop.inria.fr/axis/ Publications/2005/all.html* ils utiliseront *http ://www-sop.inria.fr/axis/* ou bien *http ://www-sop.inria.fr/axis/Publications/*. Cependant cette généralisation syntaxique, même si elle est automatique, est naïve car elle s'appuie trop sur l'organisation qui a été donnée aux pages du site Web. Une mauvaise organisation va implicitement générer une mauvaise classification et donc des résultats de faible qualité.
Dans Tanasa et Trousse (2004), une généralisation basée sur des rubriques sémantiques est

faite lors du pré-traitement de logs Web. Ces rubriques (ou catégories) sont données a priori par un expert du domaine relatif au site Web considéré. Cependant ceci est une tâche couteuse en temps aussi bien pour la définition que pour la mise à jour de telles catégories.

Enfin notons qu'il existe d'autres méthodes (comme Srikant et Agrawal (1996)) pour extraire des motifs séquentiels en tenant compte d'une généralisation mais dans d'autres domaines que le Web. Aussi la construction automatique de généralisations (sous forme de classes) dans le domaine du Web (généralisation des URLs) et à des fins d'analyse des usages n'a pas encore été étudiée. Nous proposons dans la section suivante une telle methode qui s'appuie sur les caracteristiques des données d'usage du Web.

4 GWUM : principe de généralisation

Dans la section 4.2 nous présentons notre principe de généralisation des Urls. L'idée consiste à extraire des informations relatives aux pages Web référencées par ces URls dans toutes les sources qui sont mises à notre disposition. La section 4.1 présente un exemple motivant notre travail et montre comment ces informations peuvent être extraites puis exploitées à des fins classificatoires. Ensuite, dans la section 4.3, nous indiquons comment instancier cette méthodologie à partir d'informations extraites des "referers" associés aux pages Web dans le fichier log et celles extraites de l'URL comme le contenu de ces pages.

4.1 Motivation illustrée sur un exemple

Comme nous l'avons indiqué dans l'introduction, la généralisation des items est un facteur clé lors de l'extraction de motifs séquentiels. Pour comprendre l'enjeu de nos travaux, nous proposons l'exemple suivant.

Client	Date1	Date2	Date3
C1	accueil_DT	publications_DT	accueil_Inria
C2	accueil_SC	publications_SC	logiciels_AxIS
C3	accueil_DT	publications_AxIS	publications_DT
C4	accueil_AxIS	accueil_SC	publications_SC

TAB. 1 – *Accès au site regroupés par client*

Considérons les enregistrements du log de l'Inria Sophia-Antipolis reportés dans le tableau 1. On peut y lire que le client 1 à la date 1 a fait une requête sur l'URL "accueil_DT" qui est la page d'accueil de Doru Tanasa, puis à la date 2 une requête sur la page des publications de Doru Tanasa et enfin une requête sur la page d'accueil de l'Inria. De la même manière, le client 2 a fait une requête sur la page d'accueil de Sergiu Chelcea, et ainsi de suite...
Avec une analyse basée sur les motifs séquentiels et un support de 100%, aucun motif ne sera trouvé dans ce log (aucun item n'est partagé par 100% des enregistrements). Pour trouver un comportement fréquent, il faudra descendre le support jusqu'à un seuil de 50%, ce qui permet d'extraire les comportement suivants :

1. < (accueil_DT) (publications_DT) > (vérifié pour les clients C1 et C3)
2. < (accueil_SC) (publications_SC) > (vérifié pour les clients C2 et C4)

Malheureusement, le fait de baisser le support :

1. Est un facteur de ralentissement, voir de bloquage, du processus d'extraction.
2. Retourne généralement des résultats difficiles à interpréter car nombreux et similaires (ou redondants).

Considérons maintenant que nous soyons en mesure de classer les URLs de ce log dans différentes catégories. Par exemple la catégorie "Pub" contiendrait les pages relatives aux publications de chercheurs (dans notre cas : "publications_DT" et "publications_SC"). La catégorie "Mining" contiendrait les pages relatives au data mining (dans notre cas les pages d'accueil de Doru Tanasa et Sergiu Chelcea qui font leur travail de recherche sur ce thème). Avec de telles informations, nous serions en mesure d'extraire un motif avec un support de 100% qui serait : < (Mining) (Pub) >. L'interprétation de ce motif est que 100% des utilisateurs consultent une page relative au data mining puis une page relative à des publications de chercheurs. On peut en effet vérifier ce comportement sur les enregistrements du tableau 1.

4.2 Méthodologie

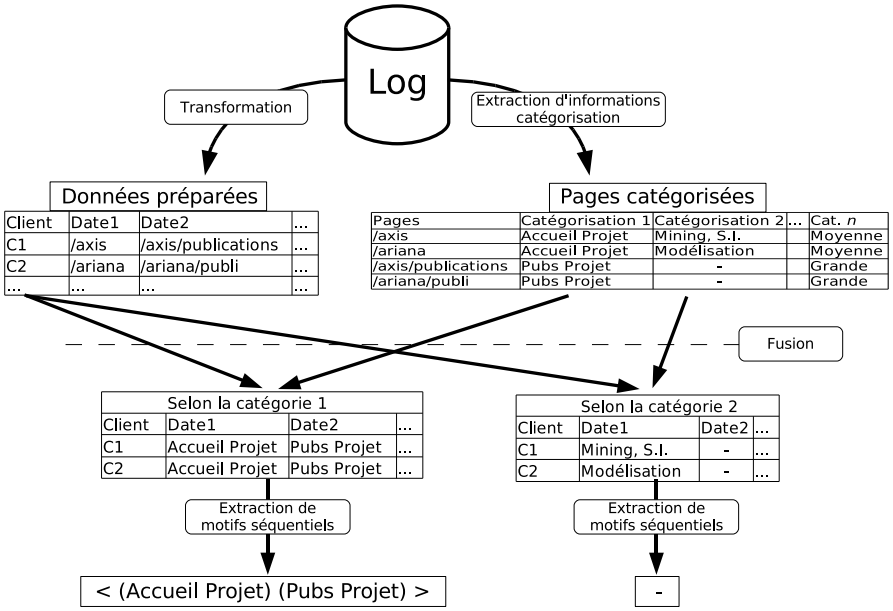


FIG. 3 – *Méthodologie de catégorisation et d'analyse des usages*

Le principe général de notre méthodologie est illustré par la figure 3. A partir du fichier log, notre objectif est d'obtenir le plus possible d'informations sur les pages Web, afin de les catégoriser. Une ligne d'enregistrement dans le fichier log se présente de la manière suivante :

```
111.222.111.222 - - [01/Oct/2005:10:45:05 0200]
"GET /axis/Publications/ HTTP/1.1" 200 3754
"/axis" "Mozilla/4.0"
```

Dans cet enregistrement, on peut trouver les informations suivantes (entre autres) :

- la machine d'IP `111.222.111.222`
- à la date du `[01/Oct/2005:10:45:05]`
- a demandé la page `/axis/Publications/`
- `3754` octets ont été transférés
- la page précédente était `/axis`
- et le navigateur `Mozilla/4.0`

Ces informations constituent une première source d'information sur la page (façon dont elle est accédée, navigateur utilisé, etc.) qui nous renseignent sur les usages qui en sont fait. Un premier ensemble de catégorie est alors envisageable à partir de ces informations. Il peut s'agir par exemple de faire une catégorisation selon le pays d'origine des IP qui accèdent aux différentes pages. Cela permettrait de classer les pages selon qu'elles sont plus accédées à partir de la France, des USA, du Japon, etc. Il peut également s'agir de travailler sur la taille de la page. Par exemple la page d'accueil du projet AxIS ("/axis") est catégorisée comme étant de "taille moyenne" dans la figure 3 (tableau "Pages catégorisées").

Une deuxième catégorisation peut être réalisée à partir des informations concernant la page elle-même. Par exemple, une analyse du contenu de la page d'accueil du projet AxIS peut montrer que ses mots clés sont "Bienvenue", "Projet" et "Inria". Encore une fois, l'application d'une méthode de classification sur l'ensemble des pages Web accédées dans le log à l'aide de ces mots clés permettra d'obtenir une catégorisation des pages. Dans le cas de la page d'accueil du projet AxIS la figure 3 montre que cette page serait catégorisée comme "Accueil Projet" (une page d'accueil d'un projet Inria).

Enfin, une fois ces catégories extraites, notre objectif est de les exploiter lors de l'analyse des usages. Par exemple, avec les motifs séquentiels, il s'agirait de préparer les données dans le format approprié à l'extraction de motifs séquentiels avec en tant qu'item une catégorie de l'URL plutôt que l'URL elle-même. Dans le cas de la figure 3 la page d'accueil du projet AxIS aurait pour item "/axis" (*e.g.* l'item du client 1 à la date 1). Si l'on veut exploiter la catégorisation 1, alors l'item "/axis" sera remplacé par sa catégorie ("Accueil Projet"). L'extraction de motifs séquentiels sur de telles données peut alors aboutir à des motifs comme < (Accueil Projet) (Pubs Projet)> qui serait interprété comme « x% des utilisateurs consultent une page d'accueil de projet suivie d'une page de publications de projet ».

4.3 Extraction d'informations en vue d'une classification automatique des pages

Pour instancier notre principe de généralisation des pages, nous proposons une extraction d'informations soient relatives à l'accès à ces pages (utilisation du champ '*referer*') soient aux

pages elle-mêmes (contenu, taille, rubrique syntaxique, etc.). Deux types d'informations pouvant être utilisés en vue d'une telle généralisation sont données en exemple ci-dessous.

Informations du referer : Pour l'extraction des informations concernant l'accès par les utilisateurs aux pages elle-mêmes, nous avons utilisé le champ referer d'un ligne HTTP quand celui-ci contenait une requête issue d'un moteur de recherche. Nous en avons extrait les mots clés. Les mots clés utilisés dans une requête sont ensuite passés dans l'outil TreeTagger développé à l'Institut de Linguistique Computationnelle de l'Université de Stuttgart Schmidt (1994). TreeTagger marque les mots d'un texte avec des annotations grammaticales (nom, verbe, article, etc.) et transforme les mots en leur racine syntaxique (lemmatisation).

Contenu de la page : Concernant l'analyse des pages référencées par les URLs, on peut s'intéresser à leur contenu c'est à dire à la partie textuelle des pages en tenant compte ou non de l'aspect structurel en vue d'une catégorisation. Une sélection des mots représentatifs de chacune des pages est retenue à partir d'algorithmes classiques en recherche d'informations comme ceux de Korfhage (1997) ou de Porter (1980) en vue d'une catégorisation.

Dans Sellah (2005), nous avons identifié des critères, inspiré de TF/IDF, afin de proposer des mots clés caractéristiques des pages Web à partir de leur contenu. Ces mots clés peuvent ensuite être utilisés dans l'étape de catégorisation comme indiqué dans la figure 3 qui décrit notre méthodologie.

5 Expérimentation

GWUM (Sellah (2005)) a été implémenté en Java et s'appuie sur PSP de Masseglia et al. (1998). L'expérimentation a été menée sur un Pentium avec des données issues des logs HTTP du site www-sop.inria.fr du mois d'octobre 2005.

5.1 Pré-traitement des données et catégorisation

Sur les 845 208 requêtes HTTP du log, 164 000 requêtes (presque 20%) ont été accédées par une requête dans un moteur de recherche. A partir de ces requêtes, sur les 35 907 mots différents extraits, seuls les mots reconnus par l'outil TreeTagger ont été gardés en vue d'une catégorisation. Malgré ce filtre nous avons obtenu un nombre très important de mots (3134 mots différents) ce qui rend difficile l'utilisation d'une méthode de type K-means sans faire avant une réduction de la taille de ce vocabulaire. Aussi nous avons appliqué une méthode de classification croisée sur des tables de contingence développée par Govaert (1977) et adaptée à ce problème par Lechevallier et Verde (2004). L'objectif de cette méthode est de réduire de manière simultanée les lignes et les colonnes d'un tableau de contingence en maximisant le critère du $\chi 2$ de ce tableau de contingence réduit. Nous avons choisi de partitionner l'ensemble des 17 671 Urls en 20 classes et l'ensemble des mots de ce vocabulaire en 10 classes. Actuellement la partition du vocabulaire n'est pas utilisée dans la phase d'extraction de motifs séquentiels. Nous pensons que cette approche peut etre encore opérationnelle si le nombre de classes de l'ensemble des Urls est inférieur à 100. Au delà nous devrons mettre en oeuvre une autre méthode de classification.

	Nombre	Urls diffe'rents
Requêtes	845 208	62 721
Requêtes avec referer non vide	593 564	53 573
Requêtes avec referer MR	164 000	17671

TAB. 2 – *Données du log considéré*

5.2 Extraction de motifs séquentiels

L'objectif de cette section est de montrer l'intérêt de l'approche par catégorisation des URLs dans un processus de Web Usage Mining en terme de support minimum et de nombre de motifs extraits. En effet, notre but est de montrer que le support minimum peut être singulièrement augmenté avec une généralisation des URLs. Dans cette expérimentation, la généralisation est basée sur la technique du referer, décrite en section 4.3. La figure 4 montre en effet le nombre de motifs extraits à différents supports pour les données décrites dans la section 5.1. On peut y observer, par exemple, que pour un support de départ fixé à 11,3%, le nombre de motifs généralisés extraits est de 2 alors qu'aucun motif basé sur les URLs uniquement ne peut être trouvé. Il faut atteindre un support minimum de 3,3% avant qu'apparaissent deux motifs basés sur les URLs uniquement. Pour ce même support, on trouve 19 motifs généralisés.

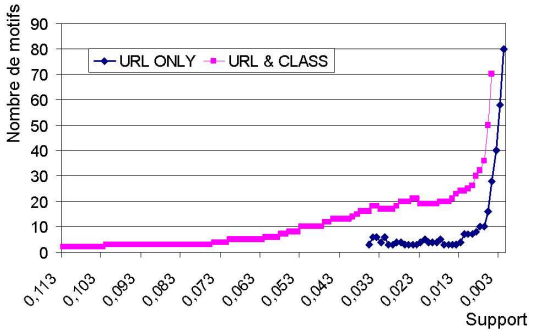


FIG. 4 – *Nombre de motifs extraits à différents supports avec et sans généralisation des URLs*

L'interprétation des motifs extraits est encore à l'étude car le nombre d'URLs, le nombre de requêtes et le nombre de classes ne permet pas de caractériser facilement chaque classe.

6 Conclusion

Nous avons présenté une méthodologie d'analyse des usages basée sur une généralisation des URLs via une catégorisation des pages que ces URLs référencent. Les informations extraites de ces pages en vue d'une catégorisation concernent soit la page elle-même (cf. son contenu, ses propriétés, sa structure) soit son accès par les utilisateurs (informations obtenues par une analyse du champ referer dans les logs HTTP). L'expérimentation que nous avons menée a permis d'illustrer notre méthodologie et de montrer le gain obtenu d'une telle approche

découvrant des motifs séquentiels fréquents en plus grand nombre et avec un support plus élévé. Notre objectif est désormais d'explorer d'autres critères permettant la catégorisation des pages, en poussant l'analyse de l'accès à ces pages. Nous prévoyons également d'intégrer plusieurs catégories dans l'extraction des motifs séquentiels au cours d'un même processus. Cela permettra d'obtenir des motifs présentant plusieurs catégories simultanément au lieu d'une seule.

Remerciements : les auteurs tiennent à remercier Sofiane Sellah pour son aide dans l'implémentation de la méthodologie présentée et Sergiu Chelcea pour sa participation dans l'extraction des mots clés à partir du referer.

Références

Agrawal, R., T. Imielinski, et A. Swami (1993). Mining Association Rules between Sets of Items in Large Databases. In *Proceedings of the 1993 ACM SIGMOD Conf.*, Washington DC, USA, pp. 207–216.

Agrawal, R. et R. Srikant (1995). Mining Sequential Patterns. In *Proceedings of the 11th Int. Conf. on Data Engineering (ICDE'95)*, Tapei, Taiwan.

Bonchi, F., F. Giannotti, C. Gozzi, G. Manco, M. Nanni, D. Pedreschi, C. Renso, et S. Ruggieri (2001). Web log data warehousing and mining for intelligent web caching. *Data Knowledge Engineering 39*(2), 165–189.

Consortium, W. W. W. (1998). httpd-log files. In *http ://lists.w3.org/Archives*.

Cooley, R., B. Mobasher, et J. Srivastava (1999). Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems 1*(1), 5–32.

Fayad, U., G. Piatetsky-Shapiro, P. Smyth, et R. Uthurusamy (Eds.) (1996). *Advances in Knowledge Discovery and Data Mining*. Menlo Park, CA : AAAI Press.

Fu, Y., K. Sandhu, et M. Shih (2000). A generalization-based approach to clustering of web usage sessions. In *Proceedings of the 1999 KDD Workshop on Web Mining, San Diego, CA. Springer-Verlag*, Volume 1836 of *LNAI*, pp. 21–38. Springer.

Govaert, G. (1977). Algorithme de classification d'un tableau de contingence. In *Proc. of first international symposium on Data Analysis and Informatics, INRIA, Versailles*, pp. 487–500.

Hay, B., G. Wets, et K. Vanhoof (2004). Mining Navigation Patterns Using a Sequence Alignment Method. *Knowl. Inf. Syst. 6*(2), 150–163.

Kum, H., J. Pei, W. Wang, et D. Duncan (2003). ApproxMAP : Approximate mining of consensus sequential patterns. In *Proceedings of SIAM Int. Conf. on Data Mining*, San Francisco, CA.

Lechevallier, Y. et R. Verde (2004). Crossed clustering method : An efficient clustering method for web usage mining. In *Complex Data Analysis*, Pekin, Chine. CDA'2004.

Masseglia, F., F. Cathala, et P. Poncelet (1998). The PSP Approach for Mining Sequential Patterns. In *Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery*, Nantes, France.

Masseglia, F., P. Poncelet, et R. Cicchetti (April 2000). An efficient algorithm for web usage mining. *Networking and Information Systems Journal (NIS)*.

Masseglia, F., D. Tanasa, et B. Trousse (2003). Diviser pour découvrir : une méthode d'analyse du comportement de tous les utilisateurs d'un site web. In *Les 19èmes Journées de Bases de Données Avancées*, Lyon, France.

Mobasher, B., H. Dai, T. Luo, et M. Nakagawa (2002). Discovery and evaluation of aggregate usage profiles for web personalization. *Data Mining and Knowledge Discovery 6*(1), 61–82.

Nakagawa, M. et B. Mobasher (2003). Impact of Site Characteristics on Recommendation Models Based On Association Rules and Sequential Patterns. In *Proceedings of the IJCAI'03 Workshop on Intelligent Techniques for Web Personalization*, Acapulco, Mexico.

Pei, J., J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, et M. Hsu (2001). PrefixSpan : Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. In *17th International Conference on Data Engineering (ICDE)*.

Schmidt, H. (1994). Probabilistic part-of-speech tagging using decision trees, revised version, original work. In *the International Conference on New Methods in Language Processing*, Manchester, UK, pp. 44–49. http ://www.ims.uni-stuttgart.de/ftp/pub/corpora/treetagger1.ps.gz.

Sellah, S. (2005). Web usage mining : Extraction de motifs séquentiels selon plusieurs points de vue. Master's thesis, Université Lumière Lyon2.

Spiliopoulou, M., L. C. Faulstich, et K. Winkler (1999). A data miner analyzing the navigational behaviour of web users. In *Proceedings of the Workshop on Machine Learning in User Modelling of the ACAI'99 Int. Conf.*, Creta, Greece.

Srikant, R. et R. Agrawal (1996). Mining Sequential Patterns : Generalizations and Performance Improvements. In *Proceedings of the 5th Int. Conf. on Extending Database Technology (EDBT'96)*, Avignon, France, pp. 3–17.

Tanasa, D. (2005). *Web Usage Mining : Contributions to Intersites Logs Preprocessing and Sequential Pattern Extraction with Low Support*. Ph. D. thesis, University of Nice Sophia Antipolis.

Tanasa, D. et B. Trousse (2004). Advanced Data Preprocessing for Intersites Web Usage Mining. *IEEE Intelligent Systems 19*(2), 59–65. ISSN 1094-7167.

Wang, J. et J. Han (2004). BIDE : Efficient Mining of Frequent Closed Sequences. In *Proceedings of the International Conference on Data Engineering (ICDE'04)*, Boston, M.A.

Zhang, T., R. Ramakrishnan, et M. Livny (1996). Birch : An efficient data clustering method for very large databases. In H. V. Jagadish et I. S. Mumick (Eds.), *Proceedings of the 1996 ACM SIGMOD Int. Conf. on Management of Data, Montreal, Quebec, Canada, June 4-6, 1996*, pp. 103–114. ACM Press.

Zhu, J., J. Hong, et J. G. Hughes (2002). Using Markov Chains for Link Prediction in Adaptive Web Sites. In *Proceedings of Soft-Ware 2002 : First Int. Conf. on Computing in an Imperfect World*, Belfast, UK, pp. 60–73.

Summary

The usage analysis of a Web site based on the extracted sequential patterns is often limited by the low support of these patterns. That is mainly due to the great diversity of the pages and behaviors. However, it is possible to group the majority of these pages in various categories during a preprocessing. Then, using these categories, rather than the URLs, will allow us to discover "generic" behaviors. This article presents a methodology for Web usage mining that uses such a generalization of the URLs. This generalization is based on a categorization of the URLs using the information extracted either from the pages or from the Web users' accesses to these pages. Then, we present an experiment which shows how the support of the extracted sequential patterns changes according to whether the patterns are obtained with or without this generalization. To conclude, we give some future directions for our work.

Vers un algorithme RASMA : RAS basé multi-agent

Radhia Ben Hamed * Hajer Baazaoui ** Sami Faiz***
* Unité des Stratégies d'Optimisation des Informations et de la connaissancE (SOIE) – Institut Supérieur de Gestion – Tunis - Tunisie
Radhia.benhamed@isg.rnu.tn
**Laboratoire Riadi – GDL Ecole Nationale des Sciences de l'Informatique – Campus Universitaire la Manouba – Tunisie
hajer.baazaouizghal@riadi.rnu.tn
***Institut National des Sciences Appliquées et de Technologie
BP. 676-1080-Tunis-Tunisie
Sami.faiz@insat.rnu.tn

Résumé. Le premier algorithme d'extraction des règles d'association spatiales (RAS), a été proposé dans (Koperski, 1995). Des algorithmes ayant pour but l'amélioration des performances de cet algorithme ont été présentés dans la littérature. Ceci de point de vue type ou nombre de règles extraites, mais l'algorithme d'extraction de règles d'association spatiales nécessite toujours un temps d'exécution important. Notre objectif est d'améliorer les performances de l'algorithme RAS en terme du temps d'exécution. Nous présentons dans cet article la nouvelle écriture de l'algorithme des Règles d'Association Spatiales basé Multi-Agent (RASMA). Une description détaillée des agents et de leurs rôles est donnée. Les résultats retournés par chaque agent et les messages échangés entre les agents sont exposés. Les résultats expérimentaux, les tests effectués sur RAS et RASMA et l'évaluation de l'expérimentation sont ensuite présentés.

1 Introduction

La Fouille de Données (FD) ou le data mining (Fayyad et al.(1996)) permet d'extraire des connaissances à partir des données qui ont tendances à devenir de plus en plus volumineuses et ceci dans plusieurs domaines la gestion, l'administration, les systèmes d'informations géographiques.
Les techniques de Fouille de Données ont été étendus aux données spatiales pour donner naissances à la Fouille de Données Spatiales (FDS). La FDS est l'extraction des connaissances implicites, les rapports spatiaux et d'autres modèles qui ne sont pas explicitement stockés dans une base de données géographiques (Han et al. (1997)).

L'algorithme de règles d'association spatiales (SAR) a été présenté par Koperski et al (1995). Cet algorithme étant basé sur l'algorithme proposé par Agrawal et al (1994) *Apriori.* L'algorithme SAR est complexe et a besoin d'un temps d'exécution important.

Les règles d'association spatiales sont des règles de la forme :

$$P1 \wedge ... \wedge Pm \rightarrow Q1 \wedge ... \wedge Qm(S\%, C\%)$$

Où au moins un des prédicats est spatial. C% représente le taux de confiance qui indique que C% des objets satisfont l'antécédent satisfont aussi la conséquence de la règle. S% représente le taux de support qui indique que S% des objets satisfont l'antécédent et la conséquence. Donc deux notions sont introduites : le support minimum et la confiance minimale, ces deux notions sont proposés pour chaque niveau de généralisation.

Plusieurs types de prédicats spatiaux sont utilisés dans les règles d'association. Des prédicats qui représentent des relations topologiques entre des objets spatiaux, d'autres qui présentent des relations d'orientation ou d'ordre ou bien ils contiennent des informations sur la distance proche_de, loin_de, etc.

L'extraction des règles d'association spatiales se déroule en cinq étapes. La première étape c'est l'extraction des données suivant la requête spécifiée. La deuxième étant la transformation des données en prédicats et en calculant leurs fréquences. Les prédicats générés sont de type spatial. La deuxième étape consiste à supprimer les 1_predicats dont le support est inférieur au support minimum et nous obtenons un ensemble des prédicats fréquents. Dans la troisième étape, on applique une méthode très fine de calcul, qui calcule les k_prédicats à partir de l'ensemble des prédicats fréquents. La quatrième étape permet de supprimer les k_prédicats dont le support est inférieur au minimum support et nous obtenons un ensemble des k_prédicats fréquents. La dernière étape c'est la génération des règles spatiales à partir de l'ensemble des k_prédicats fréquents pour chaque niveau de généralisation selon le seuil de généralisation.

D'autres extensions de l'algorithme RAS ont été faites et se sont intéressés à l'amélioration du type ou le nombre de règles extraites. L'algorithme ARGIS proposé par A.Salleb et C.Vrain (2000) est un algorithme itératif basé sur la table de lien. Un autre algorithme a été proposé par D.Malerba et F.A.Lisi (2001). Cet algorithme se base sur la programmation logique inductive. Son inconvénient est que son utilisation nécessite la transformation coûteuse des données relationnelles en un ensemble de faits exprimés en logique du premier ordre.

Nous pouvons noter l'évolution d'un nouveau domaine qui est le data mining haute performance (Philipe Market 2002), qui a pour but d'obtenir des méthodes de fouille de données qui fournissent des résultats dans un temps trop court et donc un calcul très rapide. Le data mining haute performance se base sur plusieurs techniques qui sont :

- Echantillonnage aléatoire ou Echantillonnage stratifié qui permettent de réduire les données
- Echantillonnage incrémentale qui permet de construire un modèle sur un échantillon et de tester le modèle sur un sur-ensemble de l'échantillon
- Boosting qui permet de découvrir un modèle sur les données et de redsitribuer les données de l'échantillon selon une loi de distribution
- Bagging permet la partition des données par la suite de découvrir des modèles indépendants sur ces données et de fournir un modèle résultat
- Méta apprentissage permet la partition des données, de découvrir des modèles indépendants et par la suite de faire la combinaison des prédictions.
- Utilisation d'heuristiques permet d'obtenir rapidement une solution approchée
- Informatique parallèle et distribué

La dernière technique a été adopté par Agrawal et Shafer 1996 pour augmenter les performances de l'algorithme Apriori. La parallélisation d'Apriori est une parallélisation des

données. Cette nouvelle technique a permis d'obtenir une méthode d'extraction des règles d'association incrémentales et des règles interactives.

Nous proposons d'améliorer les performances de l'algorithme RAS en se basant sur le point évoqué précédemment. Pour cela nous proposons une approche d'extraction des Règles d'Association Spatiale basée Multi-Agent, nous baptisons notre algorithme RASMA, qui se base sur la parallélisation des traitements.

Cet article propose une amélioration du temps d'exécution de l'algorithme SAR en employant la technologie Agent. En effet, les Systèmes Multi-Agent permettent de résoudre des problèmes complexes en se basant sur les principes de base : la collaboration entre les agents et le parallélisme.

Cet article s'articule comme suit, nous présentons d'abord, dans la section 2, l'algorithme des règles d'association spatiales. Dans la section 3, nous présentons notre nouvel algorithme RASMA. La section 4 est consacrée à la présentation des résultats de l'expérimentation et son évaluation. Nous terminons par une conclusion et les perspectives de ce travail.

2 Une approche pour l'extraction des Règles d'Association Spatiales basée Multi-Agent : RASMA

La figure FIG. 1 présente les étapes de l'extraction des règles d'association spatiales dans RAS. Nous avons détaillé les étapes de RAS de manière à distinguer les étapes qui peuvent s'exécuter en même temps et les différentes interactions entre ces étapes. Nous partons du fait que l'extraction des règles d'association spatiales se divise en deux étapes. L'étape de la recherche des données et l'étape de la génération des règles.
Nous avons décomposé la première étape en trois sous étapes :

1. La recherche des données sous conditions non spatiales ;
2. La recherche des données sous conditions spatiales ;
3. La jointure entre les données recherchées.

La deuxième étape, se fait pour chaque niveau de généralisation. Cette étape est composée de deux sous étapes :

4. La génération des k_prédicats fréquents ;
5. La génération des règles.

Notons que les étapes 1 et 2 et les étapes 4 et 5 s'exécutent en même temps. Nous pouvons également remarquer que l'étape 1 peut se divise en n sous étapes pour des données différentes et dont le traitement est le même.

A partir de cette décomposition, nous avons dégagé les traitements qui peuvent s'exécuter en même temps et les différentes interactions entre les étapes, nous proposons de profiter de ce point pour améliorer le temps d'exécution de l'algorithme RAS. La technique qui nous offre la possibilité de réaliser ces tâches proposées par l'approche est le système multi-agent.

Le système multi-agent parmi ces caractéristiques est qu'il permet de résoudre des problèmes complexes à travers la coopération entre les agents, la distribution et l'exécution en même temps de plusieurs tâches (Ferber 1995). Nous proposons une approche d'extraction des Règles d'Association Spatiales basée Multi-Agent (RASMA) (Baazaoui et al. 2005).

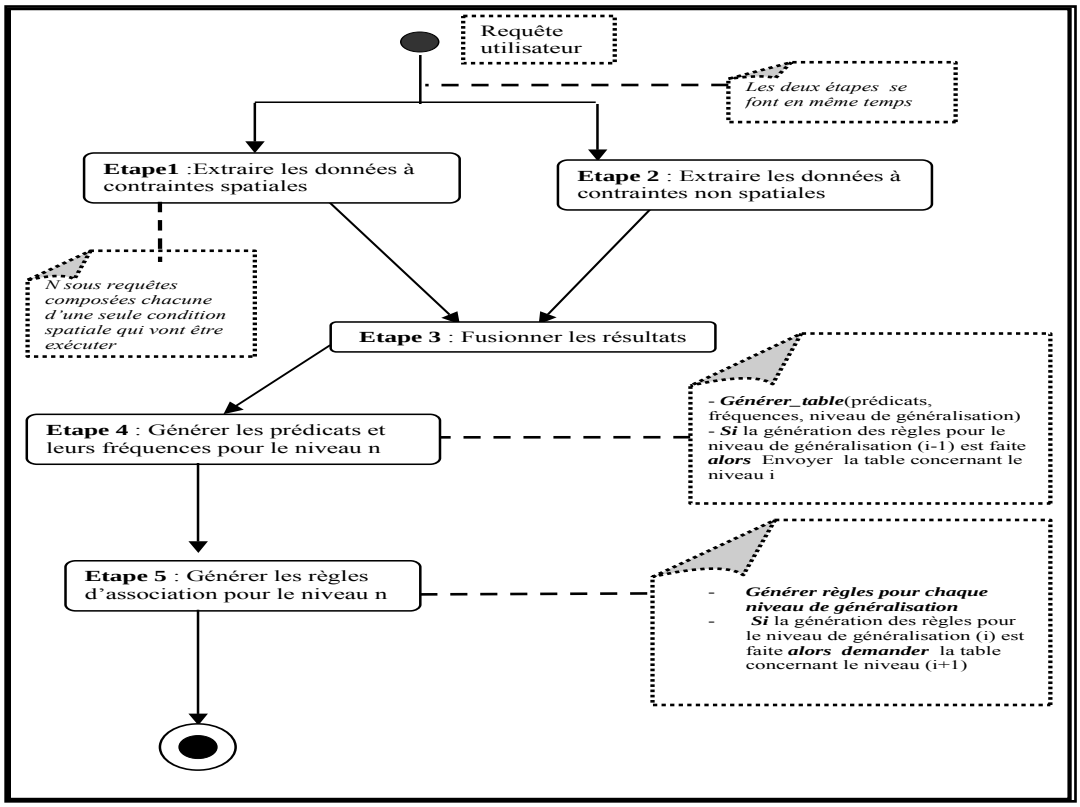


FIG. 1 – Algorithme d'extraction des règles d'association spatiales

2.1 Architecture de RASMA

L'architecture de RASMA englobe huit agents, comme illustré par la figure 2.

RASMA nous offre la possibilité de profiter de la propriété du clonage d'un agent (O.Shehory et al (1998)) et donc de réaliser le même traitement n fois et en même temps, mais sur des données différentes. Il nous offre la possibilité de créer des agents qui coopèrent ensemble pour diminuer le temps d'exécution de RAS.

Les agents constituants notre système RASMA sont les suivants:

- Un agent_coordinateur ;
- Un agent prédicats spatiaux (Agent_PS) ;
- Un agent prédicats non spatiaux(Agent_PNS) ;
- Des agents clones (Agent_CL1…Agent_CLn) ;
- Un agent extraction prédicats spatiaux (Agent_EPS) ;
- Un agent extraction prédicats non spatiaux (Agent_EPNS) ;
- Un agent fusion ;
- Un agent construction table prédicats fréquents (Agent_CTPF) ;
- Un agent génération règles association spatiales (Agent_GRAS).

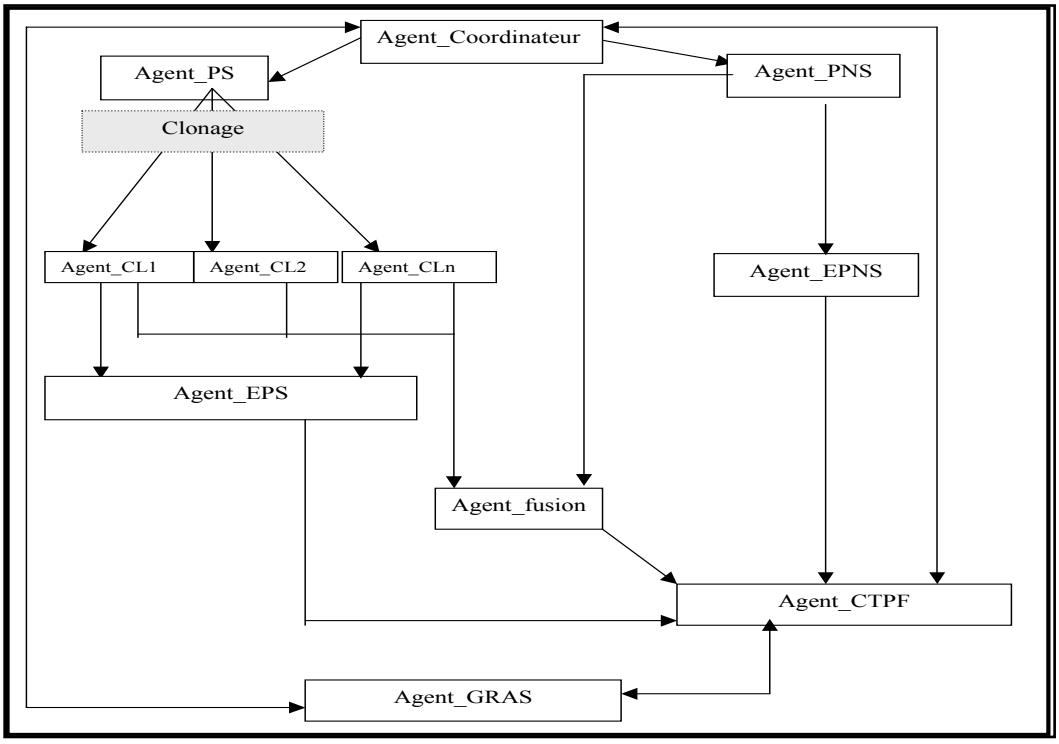


FIG. 2 – *Architecture du système multi-agent RASMA*

2.2 Présentation des agents

Dans cette section, nous présentons les rôles de chaque agent ainsi que les résultats retournés par chacun de ces agents à travers un exemple.

La base de données exemple :

Soit la base de données qui décrit les accidents, elle est composée d'un ensemble de données spatiales et non spatiales :

- Accident (Acc_ID, cond_ID, date_ID, Gravité …SDO_GID),
- Ecoles (Ecole_ID, …, SDO_GID),
- Espace-vert (Espace_vert_ID, Type, …., SDO_GID),
- Condition (Cond_ID, Luminosité, Etat_de_Surface_Route),
- Route (R_ID,…, SDO_GID)

Le jeu de données utilisé présente les accidents dans la commune urbaine de Lille. Ce jeu de données a été aimablement fourni par l'équipe BDG du laboratoire Prism. Ce jeu est réalisé sous le SIG ArcView de la firme ESRI. Les données sont relatives à une région d'une grande variété aussi bien au niveau de la morphologie urbaine que du risque routier. On dispose de données urbaines et suburbaines assez riches en matière de qualité et de contenu :

- *Données d'accidents* : Les données descriptives sont fournies par les services de gendarmerie. Ces données sont enrichies par l'information spatiale représentant les localisations des accidents en système de coordonnées universelles Lambert II. On travaille sur un fichier de 29810 accidents ayant eu lieu entre 1984 et le premier trimestre de 1998.
- Données de voirie : Le domaine routier (communal, départemental …) sur le territoire, est divisé en tronçons élémentaires compris entre deux carrefours et identifiés par un numéro et caractérisé par un point de début et un point de fin et donc par une direction et une longueur.
- Données sur le tissu urbain (Ecoles, espace_vert…).

a) L'Agent Coordinateur. Il assure le dialogue entre l'utilisateur et le système. Il offre à l'utilisateur une interface graphique. Cette interface a plusieurs rôles. Le premier étant d'offrir à l'utilisateur un assistant qui l'aide à spécifier la requête. Le deuxième rôle, est de permettre à l'utilisateur de modifier les configurations du système (la valeur de confiance minimale et la valeur du support minimum). Le troisième rôle, est l'affichage des règles. Le quatrième rôle est la décomposition de la requête spécifiée par l'utilisateur en deux sous requêtes. Une sous requête qui contient les conditions spatiales, elle sera envoyée à l'Agent_Prédicat_Spatial (Agent_PS). La deuxième sous requête contient les conditions non spatiales et sera envoyée à l'Agent_ Prédicats_Non_Spatiaux (Agent_PNS). Le dernier rôle, consiste à fournir le seuil de généralisation à l'Agent_Constructeur_Table_Prédicats_Fréquents (Agent_CTPF) pour être capable de généraliser les attributs selon le niveau de hiérarchie, ainsi que les deux autres seuils : le support minimum et la confiance minimale modifiés par l'utilisateur.

b) L'Agent Prédicats Spatiaux (Agent_PS), il reçoit la sous requête spatiale de la part de l'Agent_ Coordinateur. Si cette sous requête contient n conditions spatiales, alors, il divise la sous requête en n sous requêtes dont chacune contient une seule condition spatiale. Par la suite, il se clone en n agents clone. Chaque agent clone reçoit une sous requête à exécuter. A la fin, chaque agent clone retourne un tableau (cf. Tableau 1). Le résultat est envoyé à l'Agent_Extraction_ Prédicats_Spatiaux (Agent_EPS) et à l'Agent_fusion.

Acc_ID	Relation _Spatial	Ecole.libelle
1	Proche_de	Emc A. Comte,
1	Proche_de	Emc A. Daudet,
2	Proche_de	Emc Application Jean Aicard
7	Proche_de	Emc B. Desrousseaux
64	Proche_de	Emc A. Franck
……..	……..	……..

TAB. 1 – *Résultat de l'Agent clone_PS*

c) Agent Prédicats Non Spatiaux (Agent_PNS), il reçoit la sous requête non spatiale de la part de l'Agent_ Coordinateur. Il exécute la sous requête et fournit le résultat sous forme d'un tableau (cf. Tableau 2). Le résultat est envoyé à l'Agent_Extraction_ Prédicats_Non_Spatiaux (Agent_EPNS) et à l'Agent_fusion.

Acc_ID	Gravité	Luminosité	Etat_de_surface
1	Léger	Nuit éclairée	Sec normal
2	Grave	Jour	Sec normal
5	Léger	Jour	Sec normal
7	Grave	Jour	Humide
15	Léger	Nuit éclairée	Verglacée
24	Léger	Jour	Mouillée

TAB. 2 – *Résultat de l'Agent_PNS*

d) Agent Extraction Prédicats Spatiaux (Agent_EPS), il reçoit les tableaux envoyés par l'Agent_PS. Par la suite, il génère les prédicats spatiaux et leurs fréquences (cf. Tableau 3). Le résultat est envoyé à l'Agent_Construction_Table_Prédicats_Fréquents (Agent_CTPF).

K	Prédicats	Fréquence
1	Proche_de (Emc A. Conte)	89
1	Proche_de (Emc A. Franck)	60
1	Proche_de (Emc B. Desrousseaux)	60
1	Proche_de (EV)	65

TAB. 3 – *Résultat de l'Agent_ EPS*

e) Agent Extraction Prédicats Non Spatiaux (Agent_EPNS), il reçoit le tableau envoyé par l'Agent_PNS. Par la suite, il génère les prédicats non-spatiaux et leurs fréquences. Le résultat est envoyé à l'Agent_Construction_Table_Prédicats_Fréquents (Agent_CTPF)

k	Prédicats	Fréquence
1	Gravité (sans gravité)	38
1	Gravité (léger)	23634
1	Gravité (grave)	5313
1	Gravité (mortel)	825
1	Luminosité (jour)	20444
1	Luminosité (demi-jour)	1483
1	Luminosité (Nuit éclairée)	6999
1	Luminosité (Nuit éclairée insuffisant)	188
1	Luminosité (Nuit sans éclairage)	696
1	Etat_de_surface (Humide)	4159
1	Etat_de_surface (Mouillée)	2699
1	Etat_de_surface (Enneigée)	99
1	Etat_de_surface (Verglacée)	221
1	Etat_de_surface (Gras boueux)	51
1	Etat_de_surface (Gravillons)	34
1	Etat_de_surface (Sec normal)	22547

TAB. 4 – *Résultat de l'Agent_ EPNS*

f) Agent fusion, il reçoit les tableaux (*TAB. 1 et TAB. 2*) envoyés par les deux agents : Agent_PS et Agent_PNS. Il vérifie les identificateurs de chaque table et nous obtenons un tableau comme illustré par le tableau *5*. Par la suite l'agent fusion génère les séquences des prédicats et leurs fréquences et nous obtenons le deuxième tableau (cf. Tableau *6*). Le résultat est envoyé à l'Agent_Construction_Table_Prédicats_Fréquents (Agent_CTPF).

A	L	G	EDS	Ecole	EV
1	E-N	Léger	Sec Normal	Emc A. Comte, Emc A. Daudet,	EV
7	Jour	grave	Humide	Emc B. Desrousseaux	EV
11	Jour	Léger	Sec normal	Emc A. Franck, Emc Application Jean Aicard ,	EV
18	Jour	Léger	Mouillée	Emc A. Comte, Emc A. Daudet,	Null
….	…	….	……	……	……

Où A: Acc_ID, L: Luminosité, G : Gravité, EDS : Etat_de_surface, EV : Espace Vert; N-E: NUIT ECLAIREE

TAB. 5 – *Résultat intermédiaire de la fusion*

K	Prédicats	Fréquence
2	Proche_de (Emc B. Desrousseaux) et Proche_de (EV)	36
2	Proche_de (Emc A. Franck) et Luminosité (demi-jour)	1
3	Proche_de (EMC A. FRANCK) et Gravité(léger) et Luminosité(Jour)	26
4	Proche_de (Emc A. Comte) et Proche_de (EV) et Gravité (Léger) et Luminosité(Jour)	32
5	Proche_de (Emc A. Comte) et Proche_de (EV) et Gravité (Léger) et Luminosité(Jour) et Etat_de_surface (sec normal)	32
5	Proche_de (Emc Application Jean Aicard) et Gravité (Léger) et Luminosité(Jour) et Etat_de_surface (sec normal) et Proche_de (EV)	20

TAB. 6 – *Résultat de l'agent fusion*

g) Agent Construction Table Prédicats Fréquents (Agent_CTPF), il reçoit les trois tableaux (*TAB. 3, TAB. 4 et TAB. 6)*, il fusionne les tableaux reçus et génère les prédicats fréquents selon le support minimum dans un tableau (cf. Tableau *7*). Ensuite, il envoie le tableau *7* à l'Agent_Génération_des_Règles_d'Association_Spatiales (Agent_GRAS). Ce processus se répète pour chaque niveau de hiérarchie en fonction des seuils fournis par l'utilisateur.

K	Prédicats	Fréquence
1	Proche_de (Emc A. Conte)	89
1	Proche_de (Emc A. Daudet)	99
1	Proche_de (EV)	65
1	Gravité (Léger)	69
1	Luminosité (Jour)	67
1	Etat_de_surface (SEC NORMAL)	81
2	Proche_de (Emc A. Conte) et Gravité (Léger)	69

3	Proche_de (Emc A. Conte) et Proche_de (EV) et Gravité (Léger)	42
4	Proche_de (Emc A. Conte) et Proche_de (EV) et Gravité (Léger) et Luminosité (Jour)	32
5	Proche_de (Emc A. Conte) et Proche_de (EV) et Gravité (Léger) et Luminosité (Jour) et Etat_de_surface (SEC NORMAL)	32
…	……..	….

TAB. 7 – *Résultat de l'Agent_CTPF*

h) Agent Génération Règles Association Spatiales (Agent_GRAS), il reçoit le tableau des prédicats fréquents, ensuite il génère les règles (cf. Tableau *7*) avec les deux mesures, le support et la confiance. Il est nécessaire de noter que l'utilisateur pour chaque niveau de généralisation introduit ces deux mesures. En fait, l'Agent_GRAS à chaque fois qu'il génère les règles pour un niveau donné de la hiérarchie, il demande à l'Agent_CTPF de lui envoyer le tableau des prédicats fréquents pour le niveau suivant.

Prémisse	Conclusion	S%	C%
Accident (X) et Proche_de (Emc A. Conte)	Gravité (Léger)	69	70.4
Accident (X) et Proche_de (EV)	Etat_de_surface (Humide)	9	14
Accident (X) et Proche_de (Emc A. Conte) et Proche_de (EV) et Gravité (Léger)	Luminosité (Jour)	32	76

TAB. 7 – *Règles d'association spatiales*

3 Expérimentation, résultats et performances

L'implémentation a été réalisée sous l'environnement Oracle et la plate-forme multi-agent JADE. Les algorithmes 1 et 2 correspondent respectivement à l'approche multi-agent et l'approche classique Koperski et al.(1995). Le tableau 8 résume les coûts d'exécution en milliseconde de chacun des deux algorithmes : l'algorithme multi-agent RASMA et l'algorithme classique RAS.

Les tests visent à comparer les performances de chacun des deux approches. La figure 3 donne le temps d'exécution de deux algorithmes en fonction de la taille de la table cible. Nous pouvons remarquer l'énorme différence entre le temps d'exécution des deux algorithmes. Par exemple, pour le cas où nous avons la taille de la table cible égale à 825, le temps d'exécution réalisé par RASMA est de 254485ms par contre le temps d'exécution réalisé par l'algorithme RAS est de 854687ms.

Nombre des objets cibles	Nombre des objets en relation	RASMA	RAS (Koperski et al.(1995))
		Temps d'exécution (ms)	Temps d'exécution (ms)
29	3490	244656	785172
38	58	250172	854188
825	1084	254485	854687
5313	7342	254250	855531

23634	31765	182235	856891

TAB. 8 – *Les temps d'exécution*

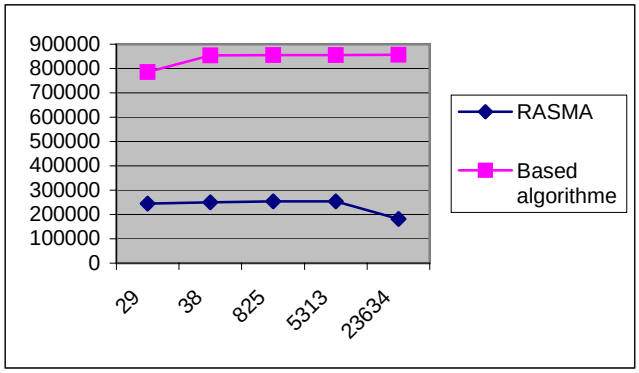


FIG. 3 – *temps d'exécution en fonction de la taille cible*

4 Evaluation

A partir des résultats de l'expérimentation, nous avons remarqué une nette amélioration du temps d'exécution de RASMA par rapport à RAS. Le gain en temps d'exécution est du à la distribution des taches entre les agents et au parallélisme.

Notre approche offre aussi une interface graphique qui permet à l'utilisateur de configurer le système en introduisant la requête et les différents seuils d'une manière interactive. Concernant les règles extraites, RASMA permet de générer des règles qui englobent en même temps des prédicats spatiaux et des prédicats non spatiaux (cf. Tableau 7) qui n'est pas le cas pour RAS. RASMA filtre les prédicats une seule fois par l'Agent-CTPF et génère un tableau des k_prédicats fréquents au lieu de filtrer à deux reprises, une fois pour générer les 1_prédicats fréquents et une deuxième fois pour générer les k_prédicats fréquents, pour l'algorithme RAS.

5 Conclusion

Dans cet article, nous avons présenté l'algorithme d'éxtraction des règles d'association spatiale proposé par Koperski et al . (1995). Nous avons remarqué que l'algorithme est composé de plusieurs étapes dont plusieurs d'entres elles peuvent s'exécuter en même temps. D'autres part, nous avons identifié des traitements semblables.

L'idée proposée est d'utiliser Système Multi-Agent (SMA) pour l'extraction des règles d'association spatiales. Nous avons implémenté notre système en utilisant la plate-forme JADE. Pour tester notre approche, nous avons utilisé des données sur les accidents de la

commune de Lille. L'expérimentation a montré une nette amélioration du temps d'exécution obtenu par RASMA par rapport à l'algorithme de base.
Le travail mené dans ce mémoire ouvre la voie à plusieurs perspectives de recherche. Une première perspective concerne la sauvegarde des relations spatiales extraites durant l'exécution d'une requête, elles peuvent être utiles pour d'autres demandes. En plus, les résultats extraits sont stockés pour d'éventuelles mises à jour. Une deuxième perspective concerne le travail dans un environnement distribué et profiter du parallélisme offert par le système multi-agent.

Références

R.Agrawal, T.Imielinski and A. N.Swami, “Mining association rules between sets of items in large databases”, in proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C., 1993, pp. 207-216.

R.Agrawal and R.Srikant, “Fast Algorithms for Mining Association Rules”, in proceedings 20th International Conference Very Large Data Bases, VLDB, Morgan Kaufmann, 1994, pp. 487-499.

R.Agrawal et J.C. Shafer: "Parallel Mining of Association Rules", IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, December 1996.

Baazaoui H., Ben Hamed R., Faiz S., Ben Ghézala H. ,

D.Malerba et F.A.Lisi, An ilp method for spatial association rule mining, proceedings workshop multi-relational data mining MRDM 2001 (Freiburg, Germany), pp. 60– 66, Septembre 2001.

J.Ferber, “Les Systèmes multi-agents vers une intelligence collective”, interEditions, France, 1995.

J.Han et M.Kamber, “Data mining: Concepts and techniques”, Morgan Kaufman Publisher, 2000.

K.Koperski and J.Han, “Discovery of spatial association rules in geographic information databases”, 4th International Symposium Advances in Spatial Databases, SSD, Springer-Verlag, 1995, vol.951, pp. 47-66.

K.Zeitouni and N.Chelghoum, “A decision tree for multilayered spatial data”, IJoint International Symposium on Geospatial Theory, Processing and Applications, Ottawa Canada, Juillet 2002, pp. 45–63.

M.Ester, H.P.Kriegel et J. Sander, “Spatial data mining: a database approach”, proceedings 5th International Symposium on Spatial Databases (SSD), 1997, pp. 25–36.

O.Shehory, K.Sycara, P.Chalasami, et S.jha, Agent cloning: an approach to agent mobility and resource allocation. IEEE Communications, Vol. 36, No. 7, Juillet, 1998, pp. 58-67.

O.Teytaud et S.Lallich, “Evaluation et validation des règles d’association”, Revue des Nouvelles Technologies de l’Information, RNTI-E-1, 2004, numéro spécial Mesures de qualité pour la fouille des données.28, pp. 193–218.

P.Marquet, "data mining haute performance", DEA d'informatique, USTL, novembre 2000 révision 2001/2002.

R.Laurini et D.Thompson, "*Fundamentals of Spatial Information Systems"*, Academic Press, London, 1994.

U.Fayyad, G.Piatetsky-Shapiro, and P. Smytn, "From data mining to knowledge discovery in databases", in the Second International Conference on knowledge discovery and data mining (KDD-96), AAAI Press Portland, Oregan, Août 2-4 1996.

A.Salleb and C.Vrain, "An application of association rule discovery to geographic information systems", In PKDD'2000, 4th European Conference on Principles and Practice of Knowledge Discovery in Databases. Lyon, France, Septembre 2000.

Summary

Koperski proposed the first algorithm for mining spatial association rules. The purpose of other algorithms presented in the literature is to improve the performances of the algorithm : the type or the number of extracted rules, but the algorithm for spatial data mining association rules requires a very important execution time. Our goal is to improve the performances of this algorithm in term of the execution time. We present in this article the new algorithm based Multi-Agent and this through a detailed description of the agents, which describes the roles, the results turned over by each agent and the messages exchanged between the agents. We expose the results of the tests carried out on Multi-Agent system for mining Spatial Association Rules compared to the first algorithm and the evaluation of the experimentation.

Optimisation de la technique de RBC pour la classification dans un processus de data mining

Mounir Ben Ayed, Issam Feki, Adel Alimi

Research Group in Intelligent Machines, National school of engineering, Sfax, Tunisia
Computer science department, Faculty of science of Sfax, Tunisia
{mounir.benayed, adel.alimi}@ieee.org

Résumé : L'utilisation de la technique « raisonnement à base de cas » pour la classification dans un processus de data mining présente un inconvénient majeur à savoir la longue durée de l'accomplissement de cette tâche, surtout sur des bases de cas de très grandes taille. Notre travail a pour objectif l'optimisation de ce processus de classification en diminuant le temps nécessaire pour accomplir cette tâche. Notre approche d'optimisation consiste à réduire les bases de cas utilisées selon les poids affectés aux attributs de ces bases. Le temps d'exécution de l'algorithme de notre approche est de 43 à 88 % inférieur au temps mis pour obtenir le même résultat avec l'algorithme classique de Kppv.
Mots clés: Data Mining, Raisonnement à Base de Cas, classification, optimisation.

1 Introduction

Les outils automatisés de collecte de données et l'évolution de la technologie des systèmes d'informations mènent à des quantités énormes de données stockées dans de grandes bases de données. Pour la prise de décision, on n'a pas besoin de toutes les données stockées mais juste des connaissances extraites à partir de ces données. Le Data Mining est le processus capable d'extraire ces connaissances cachées dans de grands volumes de données.

Une connaissance peut être obtenue sous la forme d'une : classification, segmentation, prédiction, estimation, etc. Ces actions peuvent être obtenues par une ou plusieurs techniques d'Extraction de Connaissances à partir des Données (les arbres de décision, les réseaux de neurones, les réseaux bayésiens, le raisonnement à base de cas, les algorithmes génétiques, etc.).

Nous nous intéressons dans notre travail à la tâche de classification réalisée par la technique ‘’Raisonnement à Base de Cas’’ (RBC) à base de l'algorithme standard K ppv (cas de plus proche voisins).

Le processus de classification de RBC utilisant l'algorithme Kppv est un processus dont le temps d'exécution dépend de la taille de la base de données utilisée : plus la taille de la base est importante plus le processus de classification par K ppv est long.

Notre étude vise à proposer une démarche pour améliorer le processus du RBC par la réduction de la taille de la base de données.

Dans ce travail nous présentons : le principe de la technique du RBC ; notre approche proposée qui est fondée sur la réduction des bases de données et son évaluation en la comparant à la méthode classique.

2 Le Raisonnement à Base de Cas

Traditionnellement, le RBC s'appuie sur des expériences décrites dans des formats complètement structurés tels que des objets ou des enregistrements de base de données.

Les fondements du raisonnement à base de cas proviennent des travaux en sciences cognitives menés par Roger Schank et son équipe de recherche durant les années 80 [1]. Leurs travaux ont mené à la théorie de la mémoire dynamique selon laquelle les processus cognitifs de compréhension, de mémorisation et d'apprentissage utilisent une même structure de mémoire. Cette structure, les "memory organization packets" (MOP), est représentée à l'aide de schémas de représentation de connaissance tels que des graphes conceptuels et des scripts.

Au début de la dernière décennie, on a assisté à un regain de popularité de cette technique du domaine, avec de nouvelles tendances [2].

2.1 Principe :

Le RBC est une approche de résolution de problèmes qui utilise des expériences passées pour résoudre de nouveaux problèmes [3]. L'ensemble des expériences forme une base de cas. Typiquement, un cas contient au moins deux parties : une description de situation représentant un "problème" et une "solution" utilisée pour remédier à cette situation. Parfois, le cas décrit également les conséquences résultant de l'application de la solution (exemple succès ou échec). Les techniques de RBC permettent de décrire (ou classer) de nouveaux cas en extrapolant les situations similaires à ce nouveau cas. Dans le K ppv, l'expérience (basée sur les cas déjà résolus) guide la compréhension des nouvelles situations.

L'algorithme générique de classification d'un nouvel exemple par la technique PPV est [4] :

Algorithme de classification par k-PPV
paramètre : le nombre k de voisins
donnée : un échantillon de m enregistrements classés $(\vec{x},c(\vec{x}))$
entrée : un enregistrement $\vec{y}$
1. déterminer les k plus proches enregistrements de $\vec{y}$
2. combiner les classes de ces k exemples en une classe c
sortie : la classe de $\vec{y}$ est $c(\vec{y})=c$

Le choix de la distance est primordial au bon fonctionnement de la technique. Quoique les distances les plus simples permettent d'obtenir des résultats satisfaisants (lorsque c'est possible).

Dans la figure 1, les points représentent les enregistrements d'une base de données. On peut noter qu'un point *A* peut avoir un plus proche voisin *B* tandis que *B* possède de nombreux voisins plus proches que *A* (figure 1).

***Figure 1** : À un plus proche voisin B, B a de nombreux voisins proches autres que A*

Pour définir la fonction de distance, on définit d'abord une distance sur chacun des champs, puis on combine ces distances pour définir la distance globale entre enregistrements. La distance est calculée selon le type du champ. Pour les champs numériques la distance entre deux valeurs numériques *x* et *y* peut être choisie égale à : $d(x,y) = |x-y|$,

ou bien : $d(x,y) = |x-y|/d_{max}$ (d_{max} : distance maximale dans le domaine considéré).

La seconde des deux distances normalise les distances dans l'intervalle réel [0,1].

Pour les champs dont les données sont discrètes la définition dépend des valeurs possibles :

- données binaires : 0 ou 1. On choisit $d(0,0)=d(1,1)=0$ et $d(0,1)=d(1,0)=1$.
- données énumératives : La distance vaut 0 si les valeurs sont égales et 1 sinon.
- données énumératives ordonnées : Elles peuvent être considérées comme des valeurs énumératives mais on peut également définir une distance utilisant la relation d'ordre. Par exemple, si un champ prend les valeurs 1, 2, 3, 4 et 5, on peut définir la distance en considérant 5 points de l'intervalle [0,1] avec une distance de 0,2 entre deux points successifs, on a alors $d(1,2)=0,2$; $d(1,3)=0,4$; ... ; $d(4,5)=0,2$.

L'idée de la technique est la recherche de cas similaires au cas à résoudre et d'utiliser les décisions des cas proches déjà résolus pour choisir une décision, telle que la détermination de la classe à la quelle appartient le nouveau cas.

2.2 Améliorations apportées à la technique

Le RBC est une technique qui ne nécessite pas de phase d'apprentissage. Ainsi le temps d'apprentissage est inexistant. Par contre la classification d'un nouveau cas est coûteuse en temps car il faut comparer le nouveau cas à tous les cas déjà classés.

Pour remédier à cet inconvénient, Iwayama a proposé une autre méthode appelée "Category-Based Search" [5] qui consiste à représenter tous les cas rangés dans une classe par un cas unique (par exemple la moyenne des données associées à une classe). Pour classer un nouveau cas, on cherche le représentant le plus proche du cas à classer.

Quant à Salton, il a proposé une méthode, dite "Cluster-based search" [6]. Il utilise un algorithme de classification non supervisée (clustering) qui regroupe les cas par similarité. Pour classer un nouveau cas, il le compare aux représentants de chacune des classes automatiquement découvertes. L'intérêt par rapport à la méthode de Iwayama (Category-based Search) est qu'un cas peut être rangé dans plusieurs classes. On peut ainsi donner le résultat sous forme des fréquences décroissantes d'appartenance.

3 Optimisation du processus de classification par K ppv

3.1 Procédure de classification proposée

A fin de remédier au problème de la durée nécessaire pour la classification d'un nouveau cas, nous avons proposé une modification du processus classique en ajoutant une phase de réduction de la base de cas. Cette réduction consiste à éliminer des cas (enregistrements) qui n'ont pas de similitude avec le nouveau cas.

On obtient ainsi une nouvelle base de cas de taille plus réduite que la première. La recherche des cas similaires dans la nouvelle base de données (résultat de la réduction) sera plus rapide que si cette recherche a était effectuée dans la base de cas originale (Figure 2).

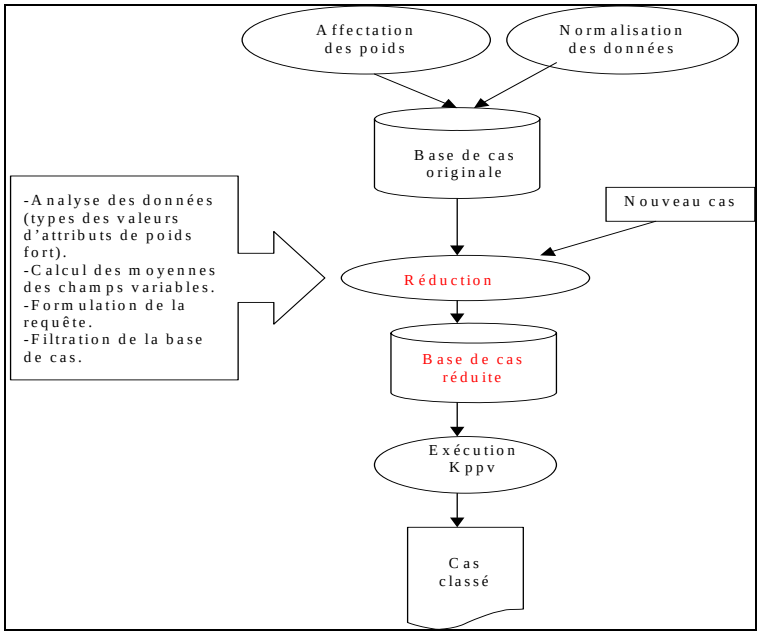


Figure 2 : processus de classification par Kppv proposé

3.2 Démarche de réduction des bases de données

Pour évaluer notre approche nous avons utilisé des bases de données (Benchmark) disponibles sur le net à l'adresse :
http://ida.first.fraunhofer.de/projects/bench/benchmarks.htm.

3.2.1 Affectation des poids

Les poids sont des valeurs affectées par un expert aux attributs de la base de données. Ces poids, appelés aussi des scores d'attributs ont des valeurs comprises entre 0 et 10.

3.2.2 Types des données

Les attributs de plus fort poids seront utilisés pour la réduction de la base de données. Dans les bases étudiées, on trouve en général deux catégories de données : (1) des valeurs discrètes tels que : sexe (0,1), type de douleur (1, 2, 3, 4), etc. ou continues tels que : âge (60, 48, 21,…), revenue (900, 180, 500….), etc.

3.2.3 Réduction de la base des cas

Pour réduire la base de données, nous sommes passés par 2 sous étapes qui sont : la construction de la requête et la génération de la nouvelle base de données. (1) Pour la construction de la requête nous avons utilisé ‘’DB Commander Pro’’. Il s’agit d’un toolbox permettant de sélectionner les données d’une base de cas en exécutant des requêtes SQL que nous avons formulées. (2) Après l’exécution de la requête, une nouvelle base, présentant seulement les cas qui pourraient être parmi les cas les plus proches voisins du nouveau cas à classer, est générée. C’est dans cette nouvelle base de données que nous allons chercher les Kppv du nouveau cas.

4 Résultats

Une étude comparative avec l’approche standard a été réalisée.

4.1 Influence du contenu de la base de cas sur les performances

Afin de mettre en évidence l’influence du contenu de la base de cas sur les performances de l’approche, nous montrons ici les résultats obtenus sur 2 bases de cas : la base de cas « heart » dont les attributs de plus grande pertinence (poids) sont qualitatifs et la base de cas « breast » dont les attributs de plus fort poids sont quantitatifs.

La base de données ‘’heart’’ comprend 270 enregistrements décrits par 12 attributs, le dernier étant le résultat de la classification, validée. Le tableau suivant présente un cas (enregistrement) de la base. Dans la 1ère ligne sont indiqués les poids affectés par l’expert.

8	0	10	9	7	9	10	8	1	4	1	4	
Age	sexe	type de douleur	Tension	sérum	sucre	électro	Fréqu.	Angine	dépression	pente	nombre navire	Résultat
60	0	4	0,1	0,21	0	2	0,132	0,1	0,2	4,2	0,2	0,7

***Tableau 1** : un enregistrement (parmi les 270) de la base de cas ‘’heart ‘’ de ‘’Bench mark’’.*

Le nouveau cas à classer se présente comme suit :

8	0	10	9	7	9	10	8	1	4	1	4

age	Sexe	t. doul.	Tension	Sérum	sucre	électro	Fréq.	angine	dépression	pente	nb nav.	Rés.
44	0	4	0,1	0,32	0	0	0,112	0,1	0	6.1	0,1	**???**

***Tableau 2** : Le nouveau cas à classe*

La requête de réduction de la base de données donne les enregistrements ayant les mêmes valeurs, des attributs de poids fort, que le nouveaux cas à classer :

Select **all** From **Heart** Where **(type de douleur = 4)** And **(électro = 0)**

Le résultat d'exécution de la requête ci-dessus est une base de données de 40 enregistrements.

La base de données « breast » (du même benchmark) comprend 701 enregistrements. Le tableau suivant représente un exemple (un cas) :

10	10	9	8	4	4	0	0	1	
Epaisseur en mm	Taille en mm	forme	Adhesion	Taille S	N Noyaux	Chromatin	Nucleoli	Mitoses	Classe
1	1	2	1	2	1	2	1	1	Bénin

***Tableau 3**: le 1er enregistrement de la base de cas ''Breast ''.*

Le nouveau cas à classer se présentent comme suit :

10	10	9	8	4	4	0	0	1	
épaisseur	Taille	forme	Adhésion	Taille S	N Noyaux	Chromatin	Nucleoli	Mitoses	Classe
4	3	2	1	2	1	3	1	1	???

***Tableau 4** : le nouveau cas à classer*

La requête de réduction de la base de données est la suivante :

Select **all** From **BREAST** Where **(epaisseur = 4)** And **(taille = 3)**

Le résultat d'exécution de la requête est une table vide. Dans cette situation nous paramétrons la requête de sélection avec la valeur du champ de poids fort suivant. Dans notre exemple, l'attribut ''forme'' présente une valeur de type ''champ fixe et dont le poids = 9.

La nouvelle requête de réduction de la base de données est la suivante :

Select all **From** BREAST **Where** (forme= 2)

Le résultat d'exécution de cette requête est une table de 61 enregistrements.

4.2 Influence de la taille de base de cas sur les performances

Pour étudier l'influence de la taille de la base sur le temps de réduction, nous avons calculé ce temps avec les requêtes formulées pour quatre bases de données du même benchmark.

Nous avons noté que la taille de la base de cas (nombre d'enregistrements ou nombre d'attribut) ne modifie pas d'une manière significative le temps d'exécution de avec notre approche. Par contre, nous avons noté que plus le nombre d'attributs de poids fort est élevé, plus la période d'optimisation (surtout de réduction) augmente.

4.3 Etude comparative entre les deux approches : Classification Kppv par optimisation de base de cas vs classification K ppv standard

Il faut remarquer que, du moins pour les bases de cas étudiées, la nouvelle approche permet de donner la même qualité de classification avec un temps d'exécution réduit.

Les temps de classement par Kppv que nous avons obtenu avec l'approche classique sont in iqués dans le tableau suivant :

Configurations	Bases de Données			
	Cancer	Cœur	Véhicule	Hépatite
Nbr Tot Attr	10	12	17	19
Nbr Attr P Fort	2	2	4	6
Nbr Enreg	699	270	846	155
Temps de classement par K ppv classique	**571s**	**220s**	**691s**	**126s**
Nbr Enreg après réduction	61	40	112	65
Temps de classement par K ppv optimisé	**52s**	**33s**	**92s**	**54s**

***Tableau 5** : Résultats de la classification avec les 2 approches (classique et optimisée)*

Le tableau suivant résume les résultats obtenus :

Bases de données	Temps Approche standards	(A) Temps de Réduction	(B) Temps d'exécution	Avec notre approche (A+B)	Pourcentage de réduction
Cancer	571	12	52	64	88%
Cœur	220	10	33	43	80%
Véhicule	691	19	92	107	84%

Hépatite	126	27	54	71	43%

***Tableau 6** : comparaison des temps de classification avec l'approche standard et notre approche.*

5 Conclusion

Dans ce papier nous avons présenté le principe du RBC comme technique de classification basée sur l'algorithme de Kppv. Cette technique permet de chercher des cas similaires dans une base de données à un nouveau cas à fin de le classer. Pour traiter le problème du temps d'exécution de l'algorithme de classification par Kppv, nous avons proposé son optimisation en réduisant la base de cas. Cette réduction est réalisée selon les poids (déterminés par un expert)affectés aux différents attributs selon leurs importances.

La nouvelle approche a permis de baisser le temps de classification de 43 à 88 % par rapport au temps mis pour obtenir le même résultat avec l'algorithme classique.

6 Référence :

[1] Riesbeck C., Shank R., Inside Case-Based Reasoning, Lawrence Erlbaum Associates, 1989.
[2] Watson I., Applying Case-Based Reasoning: Techniques for Enterprise Systems, Morgan Kaufmann Publishers Inc., 1997.
[3] R. Lefébure, G.Venturi Data mining. Ed Eyrolles, 2001.
[4] Fukunaga, K., Flick, T.. An optimal global nearest neighbor metric. IEEE Trans. Pattern Anal. PAMI-6 (3), 314–318.1984
[5] M. Iwayama and T. Tokunaga , Cluster-Based Text Categorization: A Comparison of Category Search Strategies, Proc. of ACM SIGIR'95. 1995
[6] Salton, G. and McGill, M. J. Introduction to Modern Information Retrieval. McGraw-Hill, New York, NY 1983
[7] C. Apte and S. Weiss. Data mining with decision trees and decision rules. Future Generation Computer Systems, 13, 1997.
[8] U. M. Fayyad. Branching on attribute values in decision tree generation. In Proc. 1994 AAAI Conf., pages 601-606, AAAIPress, 1994.

Summary

The use of "Case based reasoning" technique for classification in a Data mining process presents a major disadvantage which is the long period necessary for this task, specially when big data bases are used. The aim of our work is the optimisation of the classification process. Our approach consists in the reduction of the cases base according to the scores (or weight) given to the attributes by an expert. With our approach the execution time is reduced by 43 to 88% in comparison with the execution time needed by the classical approach.

Key words: Data mining, Case base reasoning, classification, optimisation

Fouille de données en appui des analyses métabonomiques

Institut National de la Recherche Agronomique

F. Jourdan[a], F. Vinson[a], C. Canlet[a], N. Priymenko[a], G. Gottardi[a], J. Molina[a] et A. Paris[a]

[a]*Laboratoire des Xénobiotiques, UMR 1089, INRA-ENVT, 180 chemin de Tournefeuille, 31931 Toulouse cedex 9, FRANCE.*

METABONOMIQUE

La méthode

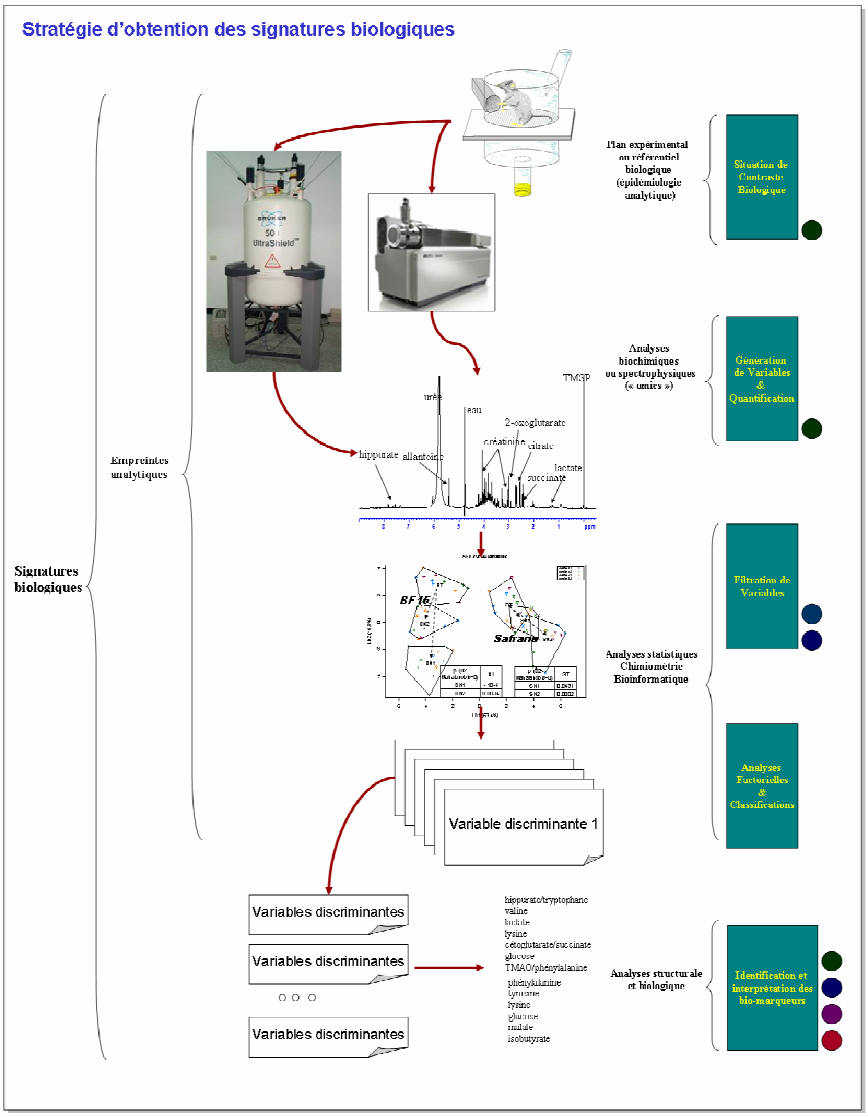
Les disruptions biologiques dues à des perturbations d'origine pathologique, génétique, nutritionnelle, hormonale ou toxicologique se traduisent par des variations des processus métaboliques. La mise en évidence de ces phénomènes nécessite une méthode chimiométrique associant l'analyse spectrale des fluides biologiques ou des tissus à des outils statistiques multidimensionnels.

Les moyens

L'équipe « Physiologie, Métabolisme et Signatures biologiques » est une équipe pluridisciplinaire regroupant des savoir-faire en biologie, chimie analytique, statistique et bioinformatique.

PROCESSUS D'ANALYSE ET FOUILLE DE DONNEES

Stratégie d'obtention des signatures biologiques



Entrepôts de données :

Méta données (e.g. plan expérimental)
Données de spectrométrie
Données RMN
Résultats d'analyses précédentes
Bibliographie

Sélection, nettoyage des données :

Extraction de motifs répétés sur des séquences de variables

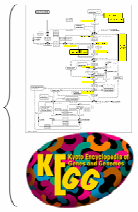
Datamining :

Intégrer les divers sources de l'entrepôt
Plusieurs générateurs de variables (spectroscopie, RMN)

Traiter les données complexes :

Text Mining sur des données bibliographiques

Extraction de connaissances via l'étude des réseaux métaboliques :
- Modélisation
- Simulation
- Visualisation



PERSPECTIVES

Entrepôts de données :
Les données traitées lors d'une analyse métabonomique sont de nature variée : chimiques, métaboliques, bibliographiques. Une approche type « entrepôts de données » semble adaptée pour leur stockage.

Sélection, nettoyage des données :
Pour chaque condition, les variables générées par les outils de chimie analytique sont associées à des valeur quantitatives. Cette information peut permettre un nettoyage automatique des données, notamment en identifiant des groupes de variables corrélées.

Datamining :
Le filtrage des variables d'un jeu de données correspond au repérage les variables fortement associées à la disruption biologique étudiée. Actuellement les méthodes utilisées sont celles de l'analyse statistique multidimensionnelle. Ces méthodes tiennent compte uniquement des variables produites par un générateur de variables dans un plan expérimental. Une approche type datamining permettrait d'intégrer à ces données, celles produites par d'autres générateurs de variables. Avec le datamining il sera également possible d'exploiter lors du filtrage les données contenues dans l'entrepôt de données.

Traiter les données complexes :
Pour l'identification des variables les chimistes analystes ont recourt à des bases de connaissances telles que les bases de données bibliographiques. Cette recherche pourrait être facilitée par des méthodes de text mining. Ces approches pourront aussi être utilisées dans la dernière étape de l'analyse qui consiste à valider les hypothèse biologiques émises lors de la conception du plan expérimental.

Les phénomènes biologiques étudiés résultent de l'évolution du réseau métabolique. La modélisation et la simulation de ce réseau, en intégrant les données métabonomiques, permettraient d'appréhender différemment les phénomènes étudiés. Enfin, la visualisation du réseau offrirait une vue globale de la déformation métabolique.