Recherche anthropocentrée de règles d'association pour l'aide à la décision

User-driven association rules mining to decisions supports systems

Vincent CHEVRIN (1), Olivier COUTURIER (2), Engelbert MEPHU NGUIFO (2), José ROUILLARD (1)

(1) LIFL (Equipe NOCE) - USTL
59655 Villeneuve d'ascq cedex, France
{vincent.chevrin,jose.rouillard}@univ-lille1.fr

(2) CRIL CNRS FRE 2499 – Université d'Artois – IUT de Lens
Rue de l'université, SP-16
62307 LENS Cedex, France
{couturier,mephu}@cril.univ-artois.fr

Résumé. Cet article expose un double travail se situant à la confluence de deux domaines que sont l'Extraction de Connaissances dans les Données (ECD) d'une part et l'Interaction Homme Machine (IHM) d'autre part. Dans un premier temps, nous présentons un problème bien connu en ECD à savoir, la recherche de règles d'association dans de grandes masses de données. Nous montrons que la fusion des deux domaines, précédemment cités, converge vers un but commun qui est la conception de systèmes décisionnels orientés utilisateurs. Nous présentons ensuite notre implémentation (*Lminer*) ainsi qu'une expérimentation que nous avons réalisée dans le secteur bancaire et qui nous a permis de montrer l'intérêt de notre approche. Le nombre de règles bien que pouvant être réduit reste néanmoins élevé. Cela nous a conduit dans un second temps à aborder le problème de la visualisation de grands ensembles de règles d'association, qui est la suite logique du problème traité dans la première partie. Nous présentons une solution permettant de gérer simultanément une représentation globale et détaillée, ce qui n'est actuellement pas le cas dans les travaux existants en ECD. Ce second travail est inclus dans la plate-forme *LARM* et cette dernière est détaillée dans la seconde partie de cet article.
Mots-clés. Interaction Homme Machine, Extraction de Connaissances dans les Données (ECD), Visualisation, Système interactif d'aide à la décision (SIAD), règles d'association.

Abstract. This article presents two different works situated at the border of the areas of knowledge discovery in databases (KDD) and Human-Computer Interaction (HCI). Firstly, a well-known issue in KDD is presented: the search for association rules in large amounts of data. Then, we show that the merging between these two areas allows us to design powerful user-driven decision systems. Next, we present an implementation (*Lminer*) as well as an experiment for the banking

system. This experiment allowed us to validate our approach. The number of rules can be reduced but it remains large. So, the visualisation of large sets of association rules issue is focused. This is the logical next step of the first work presented in this paper. We then show our solution allowing to manage simultaneously both a global and a detailed representation. This is not yet the case in the recent works in ECD. This second work is included in the *LARM* platform, and this one is detailed in the second part of this article.
Keywords. Human Computer Interaction, Knowledge Discovery in Databases (KDD), Visualisation, Decision Support System (DSS), Association rules.

1 Introduction

Au vu du nombre croissant de grandes bases de données, déterminer la façon dont sont organisées les données, les interpréter et en extraire des informations utiles est un problème difficile et ouvert. À l'heure actuelle, notre capacité à collecter les données de tous types outrepasse nos possibilités d'analyse, de synthèse et d'extraction de connaissances dans les données (Han et Kamber, 2001). Une communauté s'est créée au milieu des années 90 autour de l'Extraction de Connaissances dans les Données (ECD) pour tenter d'apporter des solutions aux chercheurs et aux industriels. L'offre en matière d'outils d'ECD est en net accroissement, et certains de ces outils sont actuellement commercialisés comme par exemple *Purple Insight*[1]. Toutefois, ils sont, en règle générale, assez complexes à utiliser et difficilement adaptables à la problématique de l'utilisateur. Ils ne prennent pas toujours en compte les stratégies d'analyse envisagées par ces derniers. De plus, ces outils génèrèrent des volumes de résultats très importants qui nous amènent à une nouvelle problématique appelée la *fouille de connaissances* constituant le problème central de cet article. Pour y répondre, nous avons abordé le problème par l'utilisation de la technique de recherche de règles d'association (Agrawal *et al.*, 1996) dans le cadre d'une application réelle à cause de sa nature descriptive et exploratoire. Nous avons aussi cherché à accroître le potentiel d'interaction avec l'utilisateur-expert qui est peu présent dans les travaux existants.

Les travaux associés à cette problématique sont assez anciens puisque l'une des premières méthodes de recherche de corrélations entre valeurs booléennes est la méthode GUHA (Hajek *et al.*, 1966). L'intérêt pour les règles d'association a été relancé trois décennies plus tard suite à l'apparition de grandes bases de données contenant des transactions commerciales (Agrawal *et al.*, 1993, Houtsma et Swami, 1995). Cette application est également appelée « analyse du panier de la ménagère » et elle est à l'origine des règles d'association. Il s'agit d'obtenir des relations ou des corrélations du type « Si Condition alors Résultat ». Dans ce cas de figure, chaque panier n'est significatif que pour un client en fonction de ses besoins et de ses envies, mais si le supermarché s'intéresse à tous les paniers simultanément, des informations utiles peuvent être extraites et exploitées. Tous les clients sont différents et achètent des produits différents, en quantités différentes. Cependant, l'analyse du panier de la ménagère consiste en l'étude des comportements des clients ainsi que les facteurs qui les poussent à effectuer tel ou tel type d'achat. Elle permet d'étudier quels produits tendent à être achetés en même temps et lesquels seront les mieux adaptés à une campagne promotionnelle. Une règle d'association est une règle de la forme « Si Fumeur alors cholestérol (75%) ». Cette règle signifie qu'une

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

personne qui fume a 75% de risque d'avoir un excès de cholestérol. Bien que cette méthode soit initialement prévue pour le secteur de la grande distribution, elle peut tout à fait s'appliquer à d'autres domaines. L'approche demeure identique quel que soit le domaine étudié : proposer des modèles, outils et méthodes transdisciplinaires susceptibles de favoriser la bonne prise de décision pour l'expert[2].

Les méthodes actuelles sont limitées au niveau de l'extraction de certaines informations ainsi que dans la restitution de ces dernières. Pour y répondre, des travaux d'approche anthropocentrée appliqués à l'ECD ont été proposés. L'expert joue alors le rôle d'heuristique évolutive au sein même du processus et son analyse est exploitée afin d'affiner les résultats et les temps d'expertise. La place prépondérante de l'IHM dans notre approche est ainsi mise en avant. En effet, l'association de la connaissance explicite d'un algorithme et de la connaissance tacite d'un expert nous a permis d'obtenir des résultats concrets qui vont être présentés ici, autour d'une action marketing menée au sein d'une banque française. L'un des problèmes récurrents est le nombre de règles générées en sortie, handicapant l'expertise de l'utilisateur du fait d'une surcharge cognitive. Pour ce faire, des travaux visant à représenter cet ensemble sous forme visuelle et de manière pertinente existent dans la littérature et s'articulent autour de la fouille visuelle de données. Bien qu'étant efficaces sur de petites quantités de données, ces représentations montrent des limites dès lors que le nombre de règles augmente. Nous allons présenter dans cette contribution, nos travaux relatifs à cette problématique de recherche. Plus particulièrement, nous nous intéresserons aux facteurs humains dans un processus ECD.

Cet article est structuré comme suit. La seconde section introduit le cadre de ce travail et présente plus en détails la recherche de règles d'association qui constitue le cœur de nos travaux. La troisième section de cet article présente un travail qui a été réalisé sur une recherche interactive de règles d'association hiérarchiques. Directement liée à la troisième section, la quatrième section se focalise sur la visualisation de grands ensembles de règles d'association, en ECD afin d'améliorer la prise de décision d'un utilisateur expert pour une optimisation de la performance de ce dernier. Nous concluons cet article en rappelant les résultats obtenus et en proposant quelques pistes de recherches en guise de perspectives.

2 Problématique

Nous exposons dans cette partie de l'article notre problématique qui s'articule autour de l'ECD et de la place de l'IHM dans les processus ECD en nous focalisant sur le cas spécifique de la recherche de règles d'association.

2.1 L'extraction de connaissance dans les données

L'information qui circule aujourd'hui sur la planète est majoritairement stockée sous forme numérique. En effet, un projet à l'Université de Berkeley a estimé, il y a déjà quelques années, à un exa-octet (c'est-à-dire à 1 milliard de giga-octets) la quantité de données générées annuellement dans le monde. Parmi ces données, 99,997 % sont disponibles sous forme numérique selon (Keim, 2001) cité dans (Nigay, 2001). Dans des entreprises commerciales comme les assurances, la grande distribution, ou encore le domaine bancaire, de nombreuses données sont collectées à propos des clients et ne sont pas nécessairement exploitées par la suite. Dans ce cas, comment répondre aux entreprises qui souhaitent pouvoir employer ces

[2] Dans le cadre de la première partie de cet article, l'utilisateur final est un expert de son domaine, nous utiliserons donc indépendamment les termes « expert » et « utilisateur ».

données de manière profitable et ceci dans des temps acceptables, tout en sachant que les requêtes traditionnelles, de type SQL (Structured Query Language) ou OLAP[3] (On-Line Analytical Processing), sont limitées au niveau du type d'informations qu'elles peuvent faire ressortir des données ? Tous ces facteurs sont les éléments d'un domaine de recherche très actif actuellement : *l'Extraction de Connaissances dans les Données (ECD)* ou *Knowledge Discovery in Databases (KDD[4])*.

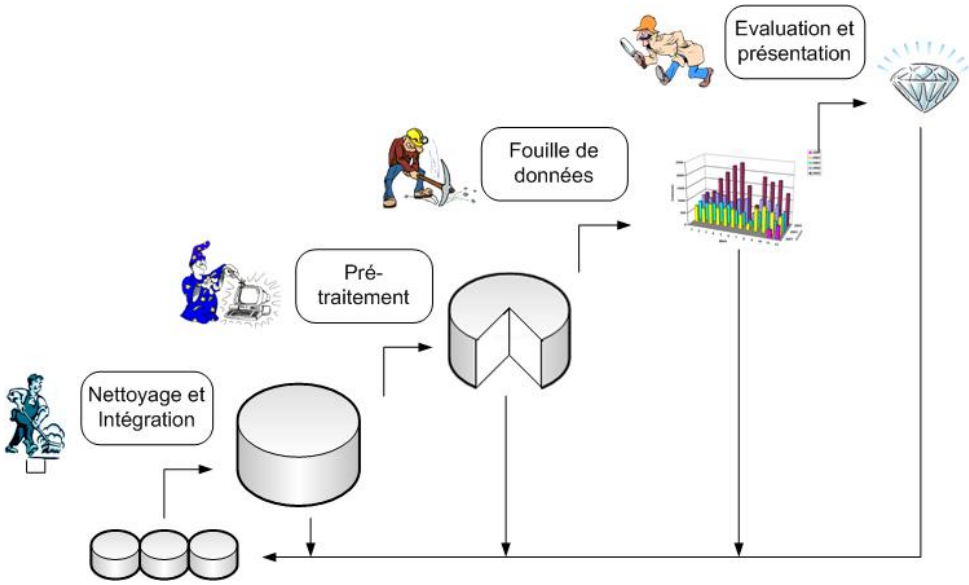

Figure 1. *Processus global en ECD*

Grâce aux techniques en ECD, les bases de données volumineuses sont devenues des sources riches et fiables pour la génération et la validation de connaissances. La *fouille de données* (ou Data Mining) constitue la phase centrale du processus, et consiste à appliquer des algorithmes d'apprentissage sur les données afin d'en extraire des *modèles* (ou *motifs*). L'ECD se situe à l'intersection de nombreuses disciplines comme l'apprentissage automatique, la reconnaissance de formes, les bases de données, les statistiques, la représentation de connaissances, l'intelligence artificielle, les systèmes experts, etc. L'ECD est un processus interactif et itératif d'analyse d'un grand ensemble de données brutes afin d'en extraire des connaissances exploitables par un utilisateur-analyste qui y joue un rôle central (Kodratoff *et al.*, 2001). L'interactivité est liée aux différents choix que l'utilisateur est amené à effectuer. L'itérativité est liée au fait que l'ECD est composée de plusieurs tâches et que l'utilisateur peut décider de revenir en arrière à tout moment si les résultats ne lui conviennent pas. La Figure 1 présente l'ECD telle qu'elle a été introduite lors de son émergence (Fayyad *et al.*, 1996). Ce processus est décomposé en quatre tâches distinctes qui sont décrites ci-après :

[3] Désigne une catégorie d'applications et de technologies permettant de collecter, stocker, traiter et restituer des données multidimensionnelles, à des fins d'analyse.

[4] Le terme anglais Knowledge Discovery in Databases (KDD) a été introduit par Piatetsky-Shapiro au cours des années 90 (Fayyad *et al.*, 1996).

- Le nettoyage et l'intégration qui consiste à retravailler des données de manière à en tirer le meilleur profit.
- Le pré-traitement qui consiste à sélectionner et transformer les données utiles à une problématique de manière à les rendre exploitables par un outil de fouille de données.
- La fouille de données qui consiste à appliquer des méthodes intelligentes dans le but d'extraire des motifs.
- L'évaluation et la présentation qui consiste à mesurer l'intérêt des motifs générés et de les présenter à l'utilisateur grâce à différentes techniques de visualisation.

Ces différentes tâches, ainsi que les enchaînements en résultant sont représentés sur la Figure 1. Cette séparation n'est que théorique car en pratique, la frontière qui sépare chacune des tâches n'est pas aussi marquée. En effet, dans de nombreux systèmes, certaines de ces tâches sont regroupées. L'une des constations faites par les chercheurs du domaine lors de la conférence KDD en 2003, est que l'interaction avec les outils existants reste limitée (Fayyad *et al.*, 2003). Lors de son essor au milieu des années 90, l'ECD était censée respecter les trois « I » : Intégration, Itérativité et Interaction. Or, on observe dans la littérature que relativement peu de travaux, anciens ou récents, intègrent ces trois notions. C'est pourquoi, nous nous focalisons sur l'interaction dans un processus ECD et présentons ci-après, la place qu'elle occupe réellement en mettant en avant, un écart flagrant entre la théorie et la pratique.

2.2 La place de l'IHM dans un processus ECD

Comme vu précédemment, l'ECD est un processus itératif et interactif. La notion d'interactivité n'est pas respectée en pratique tout au long d'un processus ECD car certaines phases restent exclusivement systèmes, c'est-à-dire qu'elles ne demandent aucune intervention de l'utilisateur. Reprenons notre exemple afin de clarifier notre propos : les différents personnages représentés sur ce schéma, expriment les différents moments au cours desquels les experts sont invités à intervenir au sein du processus. Ces derniers doivent interagir avec le processus pendant son exécution mais aussi pendant le déroulement des différentes phases qui le composent. Comme nous l'avons dit, bien que l'expert soit censé avoir une place importante en ECD, ce n'est actuellement pas le cas. Nous nous situons à la frontière de deux communautés que sont l'ECD et l'interaction homme-machine (IHM). Selon nous, l'ECD doit s'inspirer des notions de systèmes orientés-tâches tels qu'ils sont définis en IHM (Diaper, 2004, Greenberg, 2004).

Dans ces systèmes, l'activité de l'utilisateur est décomposée en tâches, de manière à construire un modèle de tâches de l'application finale. Ces tâches constituent les briques de base de la conception de systèmes à base de connaissances. L'exécution réussie d'une tâche permet d'atteindre le but qui lui est associé. La tâche regroupe un ensemble de traitements et peut alors être décomposée en sous-tâches, ce qui nous permettra, entre autres, de décrire ces sous-tâches selon qu'elles sont interactives, humaines ou système. Cela nous permettra de bien distinguer le rôle de l'expert et de le faire intervenir précisément lorsque nous le souhaitons. En ECD, ce n'est pas toujours le cas. En effet, il est possible de scinder le processus en quatre phases distinctes, que nous rapprochons par la suite aux tâches telles qu'elles sont définies dans la littérature (Sutcliffe, 1997). La théorie de l'action de Norman (Norman et Draper, 1986), par exemple, se fonde sur une modélisation de l'accomplissement d'une tâche en sept étapes, à partir de

l'établissement du but jusqu'à l'évaluation de l'état du système par rapport à ce but. De manière plus générale, nous pouvons citer Kolski pour définir notre cadre de référence : « *Les systèmes d'information comportent généralement un ensemble de fonctionnalités dont certaines permettent l'accès à une ou plusieurs bases de données ; certaines fonctionnalités sont interactives et accessibles à différents types d'utilisateurs, dont le rôle et la complexité des tâches à réaliser pourront varier énormément, et qui seront amenés à travailler aussi bien de manière isolée qu'en groupe. Les tâches interactives concernent aussi bien les activités productrices de l'entreprise, que celles plus stratégiques où la composante décisionnelle sera prépondérante. L'analyse et la conception de l'interaction homme-machine est, dans ce contexte, source de nombreuses difficultés.* » (Kolski, 2001).

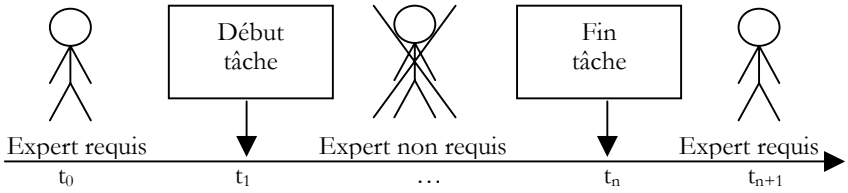

Figure 2. *Rôle actuel de l'expert lors d'une tâche dans un processus ECD*

Nos tâches en ECD ne sont actuellement pas décomposées en sous-tâches. L'intervention de l'expert est requise uniquement entre les tâches et non pas pendant ces dernières. En effet, soit une sous tâche à effectuer durant un processus ECD en un certain laps de temps entre t_0 et t_{n+1} (cf. Figure 2). L'intervention de l'expert n'est sollicitée qu'entre les tâches et non pas pendant le déroulement desdites tâches. Dans cette représentation, l'expert intervient juste avant la tâche t_1 et juste après t_n. Durant l'exécution de la tâche entre t_1 et t_n, il n'a aucune influence sur le processus ECD. Nous souhaitons affiner cette approche en redéfinissant le rôle de l'expert, de manière à ce qu'il puisse intervenir tout au long du processus afin de le guider. Pour cela, nous désirons définir des sous-tâches dont certaines seraient associées à l'utilisateur. Ainsi, nous souhaitons nous rapprocher de la définition initiale de l'ECD telle qu'elle a été rappelée précédemment. Nous nous inspirons donc des systèmes orientés-tâches afin de créer une passerelle entre l'IHM et l'ECD en la considérant comme la clé d'un système décisionnel plus orienté vers les utilisateurs.

2.3 Le cas de la recherche de règles d'association

Dans le cadre de cet article, nous nous focalisons sur une étape précise de fouille de données, à savoir, la recherche de règles d'association. Ce problème est présenté maintenant formellement. Soit $I = \{a_1, a_2, \ldots, a_m\}$, un ensemble de m attributs binaires distincts, appelés *items*. Dans la suite, *item* sera employé pour désigner un attribut. L'espace de recherche pour l'énumération de tous les ensembles possibles de $|I| = m$ est de 2^m, et donc exponentiel en m (Agrawal *et al.*, 1996). Soit $T = \{t_1, t_2, \ldots, t_n\}$, une base de données contenant n transactions, où chaque transaction t_i est constituée d'un sous-ensemble $X \subseteq I$ d'items possèdant un identifiant unique (*TID*). Un ensemble d'items $X \subseteq I$ est appelé *itemset*. Un sous-ensemble d'items $X \subseteq I$ de taille k est également appelé *k-itemset*. Une transaction t_i contient un *itemset* X si et seulement si $X \subseteq t_i$. Le support d'un *itemset* X est le pourcentage de transactions de T dans lequel X est un sous-ensemble :

$$SUPPORT(X) = \frac{\left|\left\{ t \in T \mid X \subseteq t \right\}\right|}{\left|t \in T\right|}$$

Un *itemset X* est dit fréquent, si SUPPORT(X) $\geq$ *minsup*, où *minsup* est spécifié par l'utilisateur et fixe la borne inférieure du support. Une règle d'association est une implication de la forme $A \mapsto B$, où A et B sont des *itemsets*, tels que $A, B \subseteq I$ et $A \bigcap B = \emptyset$. La partie gauche de la règle A est appelée *prémisse* et la partie droite *conclusion*. Le support d'une règle d'association r : $A \mapsto B$ est égal au support de l'union des *itemsets* qui la constituent :

$$SUPPORT(r) = SUPPORT(A \cup B)$$

Le support permet de restreindre le nombre d'*itemsets* fréquents mais plus ce support est faible, plus le nombre de règles générées est important. Pour ce faire, la mesure de confiance associée à une règle d'association a été proposée. Cette mesure est la probabilité conditionnelle que la transaction contienne B sachant A :

$$CONFIANCE(r) = \frac{SUPPORT(A \cup B)}{SUPPORT(A)}$$

Une règle d'association est fréquente, si l'*itemset* $A \bigcup B$ est fréquent. Une règle d'association est dite de confiance, si CONFIANCE(r) $\geq$ *minconf*, où *minconf* est spécifié par l'utilisateur et fixe la borne inférieure de confiance. La confiance est typiquement une mesure objective qui est orientée données permettant de juger de la qualité d'une règle. D'autres mesures objectives, également appelées indices de qualité, existent dans la littérature et des états de l'art ainsi que des études comparatives sont disponibles (Ohsaki *et al.*, 2004, Lenca *et al.*, 2007). Malgré tous ces travaux, le nombre de règles générées demeure élevé et difficilement gérable pour l'analyse d'un expert. Par ailleurs, le savoir de l'expert n'est pas exploité dans ce type d'approche. Pour y répondre, les mesures subjectives ont été proposées et elles sont quant à elles orientées utilisateurs. Un état de l'art est fait dans (Couturier, 2005). Pour y répondre de manière optimale, il faut se focaliser sur deux phases distinctes du processus global en ECD. À savoir, en amont de la fouille de connaissances, en utilisant l'expert telle une heuristique évolutive de manière à restreindre le nombre de règles. Puis, en aval, en proposant de nouvelles solutions permettant de traiter des volumes de connaissances plus importants tout en réduisant la charge cognitive de l'expert. Dans la section suivante, nous présentons une recherche de règles d'association centrée sur l'expert.

3 Recherche interactive de règles d'association hiérarchique centrée sur l'expert

La recherche de règles d'association telle qu'elle a été présentée précédemment est une approche automatique. En effet, l'utilisateur intervient à deux moments distincts du processus. Tout d'abord, il fixe les seuils de support et de confiance qui vont être utilisés durant la recherche. Une fois celle-ci terminée, il est seul juge des résultats qui sont présentés. Le principal inconvénient est le nombre de règles générées. En effet, la conséquence directe est que certains items ne sont pas pris en compte s'ils sont peu représentés dans la base de données. Pour y remédier, il suffit de baisser le support au détriment des temps de calcul. La définition initiale du seuil de recherche est donc un travail délicat. Par opposition à une approche automatique, l'approche anthropocentrée a été proposée. Cette nouvelle approche permet de regrouper les étapes de fouille de données et de post-traitement de l'information comme un tout, où l'expert est au cœur de la recherche en la guidant tout au long du processus. L'utilisateur n'intervient plus juste en début de processus en fournissant les données à étudier et en fin de processus, en jugeant de la

pertinence des résultats, cette fois-ci il fait partie intégrante du processus mais il pourra ainsi rétroagir avec le processus en fonction de ses choix lors de l'étape de post-traitement. Des travaux existent en IHM pour la conception de SIAD centrés sur le décideur (Lepreux, 2005). Nous traitons le problème du point de vue ECD.

3.1 Motivations

Un des éléments qui n'entre pas en compte dans les études est le temps de validation des résultats qui est non négligeable dans le cadre d'application au niveau industriel. En effet, plus le support est bas, plus le nombre d'*itemsets* fréquents sera important et donc plus le nombre de règles générées sera, lui aussi, important. Cette dernière étape est une problématique à part entière. Un expert pourra juger de la pertinence des résultats seulement si ses capacités le lui permettent. Afin de diminuer la charge cognitive des experts vis-à-vis des règles d'association à analyser, différentes approches ont été proposées comme la recherche de règles d'association par contraintes (Srikant et Agrawal, 1997, Goethals et Vandenbusshe 2000, Tseng, 1998, Soulet et Crémilleux, 2005). Préalablement à la recherche, un ensemble de contraintes fixées par l'utilisateur, est imposé à l'algorithme de manière à ne prendre en compte que les connaissances n'ayant pas de contraintes particulières. Le principal inconvénient de cette approche est que dans un processus de fouille de données, il n'est pas possible de formaliser la connaissance personnelle d'un expert. Cette connaissance est appelée *connaissance tacite*. De plus, il est possible que l'expert ne sache pas initialement, quel type de règles il souhaite obtenir. Dans l'approche que nous proposons, ce problème va être solutionné puisque l'expert va pouvoir juger au fur et à mesure de la pertinence de différents sous-ensembles de connaissances et faire les choix adéquats.

La seconde approche est la recherche de règles d'association hiérarchiques proposée par (Hipp *et al.*, 1998, Srikant et Agrawal, 1997). Cette approche permet de proposer à l'expert des règles à différents niveaux de granularité. L'avantage de cette approche réside dans le fait que le nombre de règles est plus ou moins grand en fonction du degré de généralisation. Malgré cela, cette approche n'inclut toujours pas la connaissance tacite de l'expert. Elle ne permet pas de limiter des calculs aboutissant à des connaissances explicites qui vont se révéler par la suite inexploitables. Le nombre de transactions n'est pas un problème en soi car les machines actuelles ont la possibilité de les traiter. Le véritable problème à résoudre est de pouvoir faire baisser le nombre d'items utiles sans pour autant perdre trop d'informations. Il est également possible d'utiliser initialement la sélection d'attributs dans le cas où l'expert sait à l'avance quelle typologie de règles il souhaite obtenir, ce qui n'est pas notre cas. L'utilisation de bases génériques permet d'éliminer les règles redondantes mais ce nombre reste malgré tout élevé (Gasmi *et al.*, 2005).

Dans les différentes approches existantes, il n'y a qu'une sorte de connaissance utilisée. Il s'agit de la connaissance explicite, celle fournie par un algorithme par exemple. Un expert doit analyser cette connaissance afin de la juger en fonction de sa connaissance tacite. Pour répondre à ce problème, une approche anthropocentrée, appliquée à la recherche de règles d'association, a été proposée. Il s'agit de donner à l'utilisateur, un rôle d'heuristique évolutive (Kuntz *et al.*, 2000). Cette approche permet à l'utilisateur de faire des choix intermédiaires qui vont orienter la suite du processus. De cette manière, une fois le processus terminé, le nombre d'items pourra baisser et les motifs ainsi extraits seront directement liés aux buts de l'expert. Le problème principal de cette approche est de proposer des résultats clairs et rapides, afin que l'utilisateur ne perde pas son temps à les analyser.

Nous proposons de coupler cette approche anthopocentrée avec une recherche de règles d'association hiérarchiques présentée précédemment. L'avantage

de cette solution hybride réside dans le fait que l'utilisateur devra valider les règles d'associations jugées intéressantes, niveau par niveau, afin qu'elles soient détaillées au niveau suivant. De cette manière, le nombre de règles à évaluer sera moins important, et les règles seront plus intéressantes pour l'expert puisque c'est lui qui va guider la recherche du début à la fin du processus. Nous redéfinissons ainsi le rôle de l'expert, de manière à ce qu'il puisse intervenir à tout moment du processus afin d'en diriger la suite. De cette façon, nous nous rapprochons de la définition de l'ECD telle qu'elle a été donnée initialement. Pour ce faire, nous nous appuyons sur les travaux définissant les systèmes orientés-tâches (Diaper, 2004, Greenberg, 2004). De plus, l'interaction de l'expert avec un outil intuitif, simple, ergonomique et adapté est nécessaire. En effet, il ne faut pas oublier que l'utilisateur final est spécialiste du domaine qu'il étudie, mais pas de l'outil informatique en lui-même.

3.2 Activité cognitive de l'expert dans plusieurs approches de règles d'association

Nous nous intéressons ici, à l'analyse des tâches au sein de différentes approches de recherche de règles d'association : la recherche « classique » (cf. Figure 3 (a)), la recherche par contraintes (cf. Figure 3 (b)) et la recherche hiérarchique (cf. Figure 3 (c)).

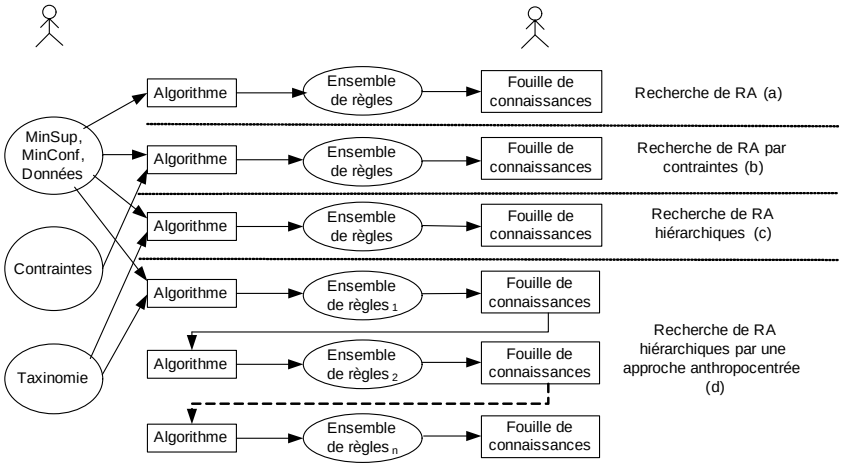

Figure 3. *Activité cognitive de l'expert au sein de différentes approches de recherche de règles d'association*

Dans ces trois approches, le rôle de l'expert est identique. Il sélectionne les données en entrée (minsup, minconf, contrainte et/ou taxinomie) et il attend la fin du processus pour intervenir de nouveau durant la fouille de connaissances. Dans ces approches, l'expert n'a pas d'activité cognitive durant l'exécution de l'algorithme de fouille de données. C'est pour cette raison, que nos travaux s'appuient sur une recherche de règles d'association hiérarchique mais centrée sur l'expert (Couturier *et al.*, 2005a). Dans cette approche, l'expert est inclus dans le processus et intervient durant son exécution (cf. Figure 3 (d)).

3.3 Rôle de l'expert dans notre approche

Nous allons illustrer la suite de cet article en utilisant un exemple factice simplifié, transposé au domaine de la consommation (cf. Tableau 1). Dans cet

exemple, le client 1 a acheté des tomates, du lait, des pâtes, du riz et des piles. Un même client peut apparaître plusieurs fois dans la liste ce qui correspond à des achats effectués à différents moments.

TID	*Tomates*	*Bananes*	*Laits*	*Fromages*	*Pâtes*	*Riz*	*Prog TV*	*Piles*
$client_1$	1	0	1	0	1	1	0	1
$client_2$	0	0	0	1	0	0	1	0
$client_3$	1	0	1	0	0	1	1	0
$client_4$	1	1	0	0	0	0	0	1
…	…	…	…	…	…	…	…	…
$client_n$	0	0	1	1	1	1	1	0

Tableau 1. *Exemple de contexte d'extraction de règles d'association*

La première étape consiste à générer le contexte d'extraction et fait intervenir l'expert pour la première fois. La base de travail pour la suite du processus va être construite à partir de la réflexion de ce dernier. Il doit réfléchir aux items qu'il souhaite faire figurer dans son étude. Il doit également se poser la question des objets à sélectionner. Doit-il prendre la totalité des objets qui lui sont proposés ou faire une partition d'objets plus ciblée sur une sous population particulière ? Ceci dépend de la problématique à traiter. Sur l'exemple, le client 2 est en relation avec les items *Fromage* et *Programme TV*, c'est-à-dire que les items sont présents sur son ticket de caisse (cf. Tableau 1).

Il convient à ce moment de créer une taxinomie des items. Ce travail doit être réalisé par un expert du domaine d'application (cf. Figure 4). Il s'agit de regrouper certains items par famille d'items ayant des racines communes et de constituer ainsi la taxinomie correspondante. Le dernier niveau correspond à une recherche classique dans lequel l'expert intervient en amont et en aval. La réalisation de cette taxinomie est le fruit d'une réflexion d'un ou de plusieurs experts. Il s'agit de la seconde intervention obligatoire de l'expert et constitue l'une des étapes incontournables de notre approche. Les résultats finaux dépendent directement de la qualité de cette taxinomie. La Figure 4 schématise notre exemple.

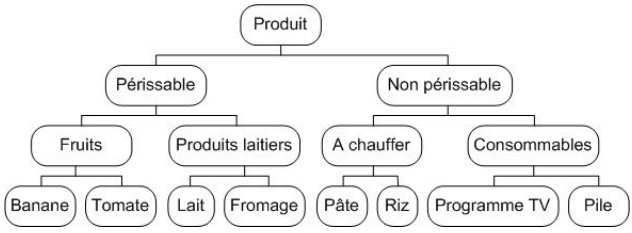

Figure 4. *Taxinomie d'items associés au contexte d'extraction*

À partir de cette taxinomie, la recherche va être réalisée niveau par niveau, guidée par les choix faits par l'expert. Pour chaque niveau, deux possibilités s'offrent à lui. Premièrement, si aucune règle n'est générée sur un niveau, la recherche est relancée au niveau suivant. Dans tous les autres cas, des *n*-règles (règles de niveau *n*) sont générées et soumises à expertise. Les règles les plus intéressantes vont être sélectionnées et la taxinomie va être élaguée en fonction de cette sélection. Les items correspondants sont conservés et les autres, ainsi que tous leurs descendants sont élagués. Les règles sont réinjectées dans le processus pour les calculs suivants et ainsi de suite. Pour illustrer notre exemple, considérons la

problématique suivante : relancer la vente de pâtes d'un grand magasin. La première étape est de lancer une recherche sur le niveau 1 de la taxinomie. Or, à ce stade, seul l'itemset *Produits* correspond à :

Tomate ∨ *Banane* ∨ *Lait* ∨ *Fromage* ∨ *Pâte* ∨ *Riz* ∨ *Programme TV* ∨ *Pile*

Il est donc impossible de construire des règles d'association avec cet unique itemset. En pratique, la recherche commencera au niveau suivant. Cette étape constitue réellement l'entrée en lice de l'expert. Ce n'est qu'à partir de ce moment que ses choix vont s'avérer cruciaux pour la suite du processus. L'algorithme génère des règles d'association à partir du niveau 2 et les soumet à l'expert (cf. Tableau 2). Une seule règle a été générée et elle est proposée à l'expert :

Ensemble de 2-règles
r1:Périssables → Non périssables

Tableau 2. *Règles d'association générées au niveau 2*

L'expert doit analyser cette règle. Toutefois, elle ne lui donne pas encore énormément d'informations. Il va poursuivre la recherche sans l'orienter pour l'instant. Pour cela, il interagit directement avec le processus en sélectionnant cette règle pour le niveau suivant. De la même façon qu'au niveau précédent, des règles sont calculées à partir des *itemsets* fréquents de niveau 3 pour que l'expert approfondisse l'analyse qu'il a entamée précédemment. Une fois les règles générées, elles sont soumises à l'expert (cf. Tableau 3) :

Ensemble de 3-règles
r1:Fruits ∨ Consommables → À chauffer
r2:Produits laitiers → Fruits
r3:À chauffer ∨ Produits laitiers → Consommables
r4:Consommables ∨ Produits laitiers → Fruits

Tableau 3. *Règles d'association générées au niveau 3*

L'expert juge les règles qui viennent de lui être proposées. Au vue des règles, l'expert souhaite se focaliser, par exemple, sur la vente de pâtes. Il va alors s'intéresser aux règles dont la conclusion contient l'*itemset* « Pâte », donc a fortiori « A chauffer » puisque nous sommes au niveau 3 de la taxinomie. D'après les règles qui lui sont proposées, seule la règle r1 est potentiellement intéressante pour lui. Il la sélectionne pour la développer au niveau suivant et ne retient pas les autres. Les *itemsets* ainsi retenus sont : « Fruits, Consommables » et « A chauffer ». L'*itemset* « Produits laitiers » n'y figure pas et peut être éliminé de la taxinomie. Cet élagage a pour conséquence de supprimer également *Lait* et *Fromage* qui ne seront plus pris en compte au niveau suivant. La recherche continue en fonction du dernier niveau, telle une recherche classique. Les règles ainsi obtenues ne contiendront pas les informations que l'expert a jugées inutiles lors des niveaux précédents.

Suite au travail réalisé en amont, il n'y a plus 2^8 *itemsets*, mais 2^6, grâce à l'élagage de la taxinomie. L'exemple utilisé est assez simple mais ceci n'est pas négligeable quand la taille de la base de données est importante comme pour celle d'une banque comme nous le verrons dans la section suivante. Comme précédemment, l'algorithme génère des règles d'association et les soumet à l'expert (cf. Tableau 4).

Ensemble de 4-règles
r1:Tomate → Pâte
r2:Pile ∨ Riz → Programme TV ∨ Banane
r3:Banane → Tomate
r4:Pile ∨ Banane → Pâte
r5:Riz → Pâte ∨ Tomate

Tableau 4. *Règles d'association générées au niveau 4*

À la vue de ces règles, l'expert va retenir les règles qu'il juge intéressantes et les analyser. Pour cet exemple, les règles contenant *pâte* en conclusion sont r1, r4 et r5. Les règles r1 et r5 sont correctes mais ne fournissent pas beaucoup d'informations nouvelles à l'expert. En effet, tous ces produits sont déjà regroupés en magasin. Par contre, l'expert va pouvoir s'attarder sur la règle r4 qui elle n'est pas courante. Il va devoir analyser cette règle et définir si cette règle est exploitable. S'il juge qu'elle l'est, la conséquence directe pourrait être par exemple de rapprocher le rayon des piles de celui des pâtes.

L'avantage de cette méthode est de faire varier les complexités en espace et temps en fonction du niveau sur lequel la recherche est effectuée. Plus le niveau de recherche est haut, c'est-à-dire vers le niveau 1, plus les temps de calcul sont bas. Les règles d'association, qui ne sont pas sélectionnées par l'expert, ne sont pas développées dans les niveaux suivants de la taxinomie. Le nombre de règles généralisées sera ainsi diminué. Cette méthodologie peut être appliquée dans beaucoup de domaines exploitant des données similaires.

3.4 Mise en œuvre sur une application réelle

Nous avons développé *Lminer* (cf. Figure 5), un système informatique d'aide à la décision (SIAD) dans lequel nous avons inclus l'algorithme *Apriori* (Agrawal *et al.*, 1996) ainsi que l'algorithme SHARK (Search Hierarchic Association Rules for Knowledge) que nous avons développé (Couturier *et al.*, 2005b). Cette méthodologie repose sur l'algorithme *Apriori* et reprend les idées citées précédemment. L'utilisation de cet algorithme n'est qu'un paramètre. Il est possible d'utiliser d'autres outils cités dans l'état de l'art. Pour une meilleure interaction avec l'utilisateur qui doit orienter la recherche, une interface conviviale, facile d'utilisation a été développée. Cet outil est un support obligatoire pour aider l'expert dans son analyse. De plus, la convivialité et l'ergonomie s'avèrent des points très importants qui font intervenir des notions d'IHM (Shneiderman, 1996). En effet, un logiciel peut être programmé avec les meilleures structures de données et les algorithmes les plus performants, et être en même temps inutilisable car trop complexe au niveau du mode opératoire. Il faut donc travailler en gardant à l'esprit que le logiciel final doit être adapté à ses utilisateurs finaux.

Nous avons utilisé une interface graphique spécifique pour pouvoir implémenter notre algorithme fondé sur une présentation textuelle des règles. Nous avons fait ce choix pour une mise en œuvre rapide de notre approche. Une analyse liée au choix de la représentation sera faite dans la seconde partie. La Figure 5 présente cette interface qui est scindée en quatre parties : informations sur les données (fenêtre I), la taxinomie (fenêtre II), la liste des règles générées (fenêtre III) et la liste des règles retenues par l'expert (fenêtre IV) (Couturier, 2005).

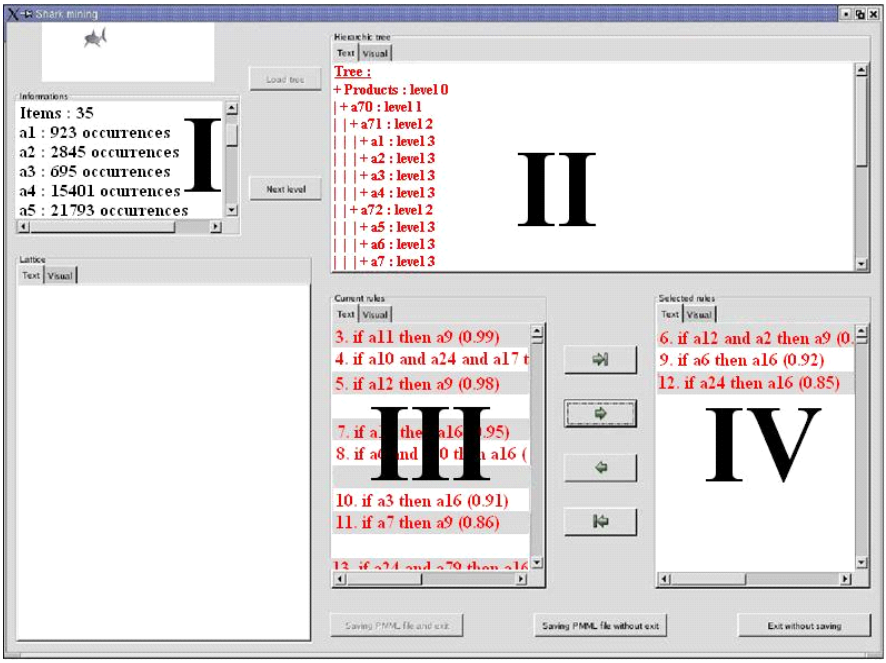

Figure 5. *Interface graphique de SHARK*

Domaine d'application

Pour illustrer cet article, nous présentons plus particulièrement une application relative au domaine bancaire. Nous travaillons en collaboration avec le service marketing d'une grande banque française qui entreprend des démarches de gestion de la connaissance. Un aspect important de cette démarche est le renforcement du lien avec les clients. En effet, investir uniquement sur de nouvelles techniques, outils, etc., ne suffit pas pour maximiser l'efficacité. Selon (Lefébure et Venturi, 2001) il faut, en amont, mesurer la qualité de la relation et de la fidélité des clients. La problématique initiale est liée au marketing qui consiste en l'identification plus efficace pour chaque client, du ou des offres susceptibles de l'intéresser. Typiquement, si un client n'est jamais contacté par sa banque, il a de grandes chances de se sentir délaissé à un moment ou un autre.

Selon les experts, la fidélisation d'un client doit donc passer par une phase de renouvellement au niveau de la diversité des produits et services qui lui sont proposés. Ces services peuvent être des services de maintenance sur un produit, ou la vente de produits de long terme comme un prêt immobilier, un prêt de consommation, un plan d'épargne logement, etc. Nos travaux consistent à anticiper le choix des clients, et à prédire ainsi quels types de produits vont pouvoir les intéresser avec un certain facteur temporel. Selon (Kotler et Dubois, 2000), il est possible de dénombrer deux catégories distinctes de clients. Tout d'abord, les clients que l'entreprise essaye de conserver et sur lesquels un effort de fidélisation doit être réalisé. Selon les termes de ces auteurs, il sera alors question de marketing défensif. Ensuite, les clients de la concurrence que l'entreprise tente de démarcher. C'est ce que ces mêmes auteurs appellent le marketing offensif. La problématique sur laquelle nous avons travaillé se situe dans le cadre du marketing défensif mais la solution que nous préconisons peut être utilisée dans les deux cas.

Prétraitement du jeu de données

La recherche de règles d'association s'est effectuée sur les différents produits proposés. Pour extraire et travailler les données, nous avons utilisé des outils d'analyse multidimensionnelle comme *Business Object*[5]. À partir de ce dernier, une simple requête a permis de récupérer les différents attributs. Mais à l'exception de certains d'entre eux qui étaient déjà sous forme binaire, tous les autres ont dû être retravaillés sous *SPSS*[6] pour respecter la contrainte qui nous était imposée, à savoir le regroupement de certains attributs en un seul, nous permettant ainsi, de prendre en compte les besoins des utilisateurs dans la construction des données.

Les seules obligations fixées par les experts étaient de ne considérer que des clients dont nous étions sûrs qu'ils possédaient au minimum deux produits. En effet, initialement, 800 000 clients étaient stockés dans la base mais un grand nombre d'entre eux ne possédaient qu'un seul produit. Nous les avons supprimés de notre sélection car aucune corrélation ne pouvait être trouvée. Ensuite, nous n'avons retenu que des clients significatifs.

Par conséquent, nous n'avons pas sélectionné les clients peu ou très peu actifs, c'est-à-dire qui n'ont pas réalisé d'opérations depuis un certain temps. Ces deux critères nous ont permis de diviser par deux le nombre de clients. Pour cette recherche, une liste d'environ 57 attributs a été définie par les experts. Notre contexte d'extraction final est donc composé d'environ 400 000 clients pour 57 attributs.

Résultats

Nous avons appliqué l'algorithme *SHARK* à ce jeu de données de manière à diminuer les temps d'expertise en ciblant mieux la connaissance utile. La première étape de notre méthode consiste à créer la taxinomie avec les experts du domaine. Dans l'exemple factice que nous donnons, l'arbre est équilibré. Ce n'est pas le cas pour notre expérimentation mais ceci n'a pas d'influence sur le calcul de l'algorithme. Notre taxinomie est composée de 114 nœuds et de 57 feuilles. Afin d'estimer les règles plus rapidement, un outil graphique appelé *LARM* (Large Association Rules Mining) a été développé dans le but de permettre à l'expert de faire varier plusieurs critères pour sélectionner des règles en fonction du volume global (Couturier *et al.*, 2005c).

En effet, même si l'espace de recherche diminue, les quantités de règles peuvent rester importantes en fonction du sous-espace de recherche à exploiter. *LARM* a permis de sélectionner rapidement un ensemble de règles en faisant varier le support (relatif ou absolu) ainsi que la confiance. Le nombre d'items à retenir en prémisse et en conclusion a pu être réduit grâce à l'utilisation de *LARM*. Cela s'est fait par la diminution du nombre de règles à exploiter à certains niveaux et par la sélection d'items particuliers à d'autres niveaux.

Grâce à cet outil, le nombre de règles est passé de 2445 à quelques dizaines en fonction des choix de l'expert. C'est la raison pour laquelle les temps d'expertise présentés ci-après sont faibles. Cette expertise a été réalisée sur les règles sélectionnées (cf. Tableau 5, Colonne *Sélection*).

[5] http://www.businessobjects.com/

[6] http://www.spss.com/

	Nb règles	*Temps exécution*	*Sélection*	*Expertise*	*Sélection finale*
SHARK (Niveau 1)	ø	ø	ø	ø	ø
SHARK (Niveau 2)	5	1s	2	30s	2
SHARK (Niveau 3)	546	70s	6	60s	6
SHARK (Niveau 4)	2113	1080s	36	300s	8
SHARK (Total)	2113	1151s	36	**390s**	**8**
Apriori	2445	1140s	56	**600s**	**10**

Tableau 5. *Résultats expérimentaux avec Apriori et SHARK*

Pour ces expérimentations, la recherche a été réalisée sur le jeu de données dans son intégralité. Les résultats obtenus fournissent un ensemble de connaissances qui peuvent être utiles pour différents sous problèmes. C'est dans le cadre de l'un de ces derniers que nous avons été amenés à travailler. Nous avons donc opté pour l'utilisation de *LARM* de manière à extraire la connaissance adéquate par rapport à l'ensemble des règles générées. Cette sélection se résume souvent à limiter le nombre d'items contenus dans les *itemsets*. Nous avons testé notre jeu de données avec une approche automatique (algorithme *Apriori)* et une approche anthropocentrée (algorithme *SHARK*). Les résultats sont synthétisés dans le Tableau 5. La colonne *Nb règles* indique le nombre de règles générées, la colonne *Temps exécution* représente le temps qu'il a fallu pour générer ces règles, la colonne *Sélection* indique le nombre de règles restant après la sélection effectuée sur les règles générées, la colonne *Expertise* indique le temps nécessaire aux experts pour valider les règles sélectionnées. Il ne s'agit pas d'une moyenne car il n'y a qu'une seule passe. Pour le mesurer, une fonction permettant de calculer le temps d'expertise a été utilisée. Une première mesure de temps est effectuée dès que les règles sont proposées à l'expert. Une seconde mesure est prise au moment où l'expert valide sa sélection. Le temps d'expertise est calculé en réalisant la différence entre ces deux mesures. Enfin, la colonne *Sélection finale* représente les règles qui ont été validées pendant l'expertise. Les deux principaux critères de comparaison sont le temps d'expertise et le nombre de règles car notre objectif initial est de diminuer la charge cognitive liée à l'analyse des règles.

L'algorithme *Apriori* génère 2445 règles en 19 minutes sur une machine PC cadencée à 1 Ghz, avec 128 Mo de RAM. Il faut alors ajouter le temps d'expertise par rapport aux 56 règles sélectionnées qui est de 10 minutes, soit un total de 29 minutes. L'algorithme *SHARK* travaille sur des données plus ou moins agrégées. Au niveau 1, aucune règle ne peut être générée. La recherche commence donc au niveau 2 et donne 5 règles en 1 seconde, avec 30 secondes d'expertise sur les 2 règles sélectionnées. Au niveau 3, 546 règles sont générées en 1 minute et 10 secondes avec 1 minute d'expertise pour les 6 règles sélectionnées. Enfin au dernier niveau, c'est-à-dire le même que pour *Apriori*, 2113 règles sont générées en 18 minutes avec 5 minutes d'expertise pour les 36 règles sélectionnées. Les règles des niveaux précédents ne sont pas prises en compte dans le total de règles. En effet, ces règles constituaient de la connaissance intermédiaire permettant d'obtenir les 36 règles du dernier niveau. Au niveau qualitatif, toutes les règles générées avec *SHARK* se retrouvent dans les règles générées avec *Apriori*. Ceci s'explique facilement puisque *Apriori* est un paramètre de notre approche. La seule différence notable est que les règles qui n'ont pas été générées avec *SHARK* n'ont pas engendré de temps d'expertise inutile.

Le temps total d'exécution de l'algorithme *SHARK* est de 1151 secondes, il est légèrement plus long que pour *Apriori* (1140 secondes). Ceci est lié au caractère interactif de l'interface graphique utilisée. En effet, l'expert valide des règles et c'est ce dernier qui lance la recherche au niveau suivant. Donc, tant qu'il n'a pas effectué cette action, la recherche est bloquée. C'est la raison pour laquelle en ayant moins de règles à calculer, les temps peuvent être plus longs pour *SHARK*. Notre but était de diminuer les temps d'expertise en diminuant le nombre de règles à traiter par intervention de l'expert, ceci ayant pour conséquence de ne pas calculer, par la suite, des règles que l'expert ne retiendra pas et ainsi faire baisser les temps de calcul et d'analyse. Le temps total d'expertise de *SHARK* (390s) est plus rapide que celui d'*Apriori* (600s).

Retour des experts

Notre approche a permis d'extraire des règles nouvelles qui n'étaient pas connues des experts. Ces résultats ont été mis en œuvre sur une application bancaire réelle dans le cadre du lancement d'un nouveau produit. Deux cibles ont été réalisées : une avec les critères actuels qui sont définis par les experts de la banque et la seconde, à partir des règles d'association qui ont été extraites. Ces deux cibles ont généré deux listes de clients potentiels à contacter et sont parties en exploitation au niveau des commerciaux pour contacter ces clients. Les résultats ont montré que les deux cibles ont obtenu des performances similaires mais sur deux segments de clientèle différents. En effet, nous avons proposé un segment clientèle non exploité jusqu'à présent ce qui a permis, en complément de la cible des experts, d'augmenter le taux de retour global concernant les contacts. Le retour en gain de productivité n'a, quant à lui, pas pu être mesuré par les experts par manque d'informations.

Dans la continuité de cette première partie, nous nous intéressons à l'aspect interaction avec cet expert. En effet, ceci est étroitement lié au travail qui vient d'être présenté car pour qu'un système décisionnel soit performant, il est nécessaire que l'interaction entre l'expert et le processus soit dans la mesure du possible optimale. En effet, un algorithme même efficace, devient complètement inutile si l'utilisateur final est incapable de l'utiliser. Comme nous l'avons dit, nous avons opté pour une présentation textuelle des règles ce qui nous a permis de valider notre approche mais qui présente toutefois des limites en terme de charge cognitive. Pour ces expérimentations, nous avons dû utiliser un module externe (*LARM*) pour sélectionner nos règles. Cet outil de visualisation utilise un format XML en entrée. Il est donc tout à fait envisageable d'utiliser d'autres outils adaptés à notre problématique (Couturier, 2005). Néanmoins, afin d'améliorer davantage l'interaction avec l'expert, nous présentons dans la section suivante, une approche graphique permettant de compléter ce premier travail en proposant le choix à l'expert quant à la représentation qu'il souhaite en sortie. Pour ce faire, nous reprenons les modules graphiques de *LARM*.

4 Visualisation de grands ensembles de règles d'association

L'essor de l'ECD a fait apparaître de nouveaux problèmes comme la fouille de connaissances que nous avons déjà abordée dans la section précédente. Ces masses d'information doivent être soumises à expertise pour validation mais cela demande parfois des efforts cognitifs importants impliquant, entre autres, une perte de temps dont ne peuvent pas se permettre les industriels en terme de rentabilité. En effet, extraire de la connaissance devient vite difficile lorsque les informations pertinentes sont cachées dans une grande masse de données. Cette nouvelle problématique a eu pour conséquence l'apparition de la fouille visuelle de données dont le but est de

proposer des outils de visualisation adaptés à différentes tâches bien connues en ECD (Do et Poulet, 2003). Ces outils contribuent à l'efficacité des processus mis en oeuvre au niveau de l'extraction de connaissances en offrant aux utilisateurs des représentations intelligibles tout en facilitant l'interaction avec ces derniers. L'ECD et la visualisation présentent des similitudes qui sont présentées ci-après.

4.1 Lien entre l'ECD et la visualisation

Une étude comparative a été réalisée (Buttenfield, 2003) et propose un croisement entre les techniques de visualisation et certaines tâches en ECD comme par exemple *trouver* des motifs, *représenter* ce qui a été trouvé, *valider* leurs significations et *optimiser* les performances de calculs permettant d'aboutir à ces motifs. Cette étude est assez générale et ne fournit que les pistes à exploiter mais permet cependant de dire qu'une technique est dépendante de la tâche à effectuer. Malgré cela, extraire de l'information et de la connaissance à partir d'un corpus de données constitue l'objectif commun de l'ECD et de la visualisation. Les enjeux similaires sont détaillés ci-après :

- *Bases de données conséquentes* : le besoin d'adapter et développer des méthodes pour des volumes de données de plus en plus importants est véritablement l'un des points communs majeur entre les deux disciplines. Au sein de la visualisation, les méthodes sont testées sur des problèmes de taille relativement proches, et le résultat qui revient généralement, est qu'il n'y a pas assez de pixels disponibles sur un écran pour l'affichage des informations. Les deux disciplines travaillent actuellement dans ce sens.
- *Données multi-variables* : un autre point commun est l'exploration de données avec de nombreuses variables. La plupart des techniques performantes sur peu de variables se révèlent souvent inadéquates sur des corpus de données exploitant des centaines de variables. Le problème commun qui en résulte est qu'il faut savoir initialement quelles sont les variables importantes à la problématique, et *a fortiori*, quelles sont celles qui peuvent être considérées comme inutiles à l'analyse.
- *Données hétérogènes* : l'exploitation de données provenant d'une table unique est une chose usuelle. Dès que les tables multiples relationnelles apparaissent, ou que les formats de données sont peu courants, comme les données multimédias ou spatiales, etc., les approches utilisées se révèlent une nouvelle fois, inadaptées.
- *Accessibilité pour les experts du domaine :* dans le cas d'applications en situation réelle, il est indispensable que les techniques développées soient accessibles à des experts qui ont peu de formation en visualisation, fouille de données, informatique ou statistiques et qui au contraire ont une très bonne connaissance des données à étudier.
- *Intégration :* comme pour toutes jeunes disciplines, l'ECD et la visualisation, ont débuté par le développement d'outils en fonction d'une situation bien précise. Puisque ces disciplines se sont énormément développées, il faut maintenant pouvoir proposer des outils et des méthodes qui peuvent s'intégrer dans un environnement polyvalent et qui puissent s'adapter à tous types de problèmes.

Malgré les enjeux communs qui viennent d'être cités, il y a une différence fondamentale dans l'approche de l'ECD et de la visualisation. En effet, bien que devant respecter en théorie la notion des trois « I », en pratique les méthodes actuelles d'ECD tendent à analyser les données le plus automatiquement possible, tout en minimisant l'interaction avec les experts. Dans le même temps, la

visualisation maximise l'interaction avec ces derniers de sorte que leurs capacités analytiques puissent être exploitées. Ce constat va à l'encontre de la définition initiale de l'ECD qui stipule que l'utilisateur-analyste doit faire partie intégrante du processus. La fusion entre l'ECD et la visualisation va permettre de construire des modèles prédictifs qui vont inclure la réflexion de l'expert. Pour que ceci soit réellement applicable, il s'avère indispensable que l'expert puisse interagir de manière optimale.

4.2 Motivations

Nous nous intéressons à l'amélioration de la visualisation de grands ensembles de règles d'association pour permettre d'augmenter les performances du système tout en diminuant la charge cognitive de l'utilisateur. Ce travail se situe une nouvelle fois à l'intersection de deux domaines de recherche distincts que sont l'IHM d'une part, et l'ECD d'autre part. Il existe peu de travaux dans la littérature allant dans ce sens. Un premier travail propose une approche qui vise à intégrer les étapes du processus ECD dans un modèle de développement enrichi sous l'angle des interactions homme-machine appelé le modèle en U (Lajnef *et al.*, 2005). Dans ce papier, les auteurs notent des points de convergence entre les étapes d'un processus ECD et celles d'un processus d'IHM. De plus, ils indiquent que ces deux processus doivent être réalisés en parallèle pour une conception et une utilisation optimale de systèmes s'appuyant sur les facteurs humains. En ce qui nous concerne, nous considérons ces facteurs uniquement lors de l'utilisation du système. Lors d'une génération de règles, il est possible d'en dénombrer plusieurs milliers voire plusieurs millions si le jeu de données est conséquent. Toute la problématique est de pouvoir présenter ces règles à l'expert de manière optimale. Des travaux en visualisation existent mais ils ne sont plus adaptés dès lors que les quantités de règles augmentent.

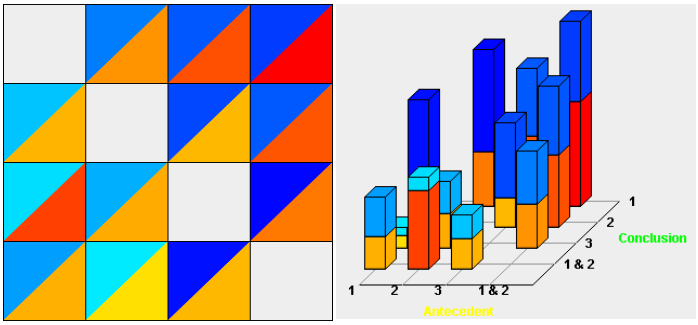

Figure 6. *Visualisation 2D et 3D de règles d'association avec LARM*

La Figure 6 montre des règles représentées dans un environnement (2D ou 3D) à l'intersection de leur prémisse et de leur conclusion. Des couleurs graduelles permettent de représenter la valeur des métriques (*support* et *confiance*) associées aux règles. De plus, la 3D permet de renforcer cette information en empilant ces métriques dont la taille est proportionnelle à leurs valeurs. Ces représentations sont limitées car elles sont, soit globales, soit détaillées. En vue d'améliorer la performance et la satisfaction de l'utilisateur dans cette tâche particulièrement cognitive, nous proposons une approche hybride basée sur une représentation de type matrice 2D colorée (Couturier, 2005) et sur une FEV (FishEyeView) (Furnas, 2006). Ainsi, la représentation est simultanément globale et détaillée ce qui n'est pas le cas actuellement pour la visualisation de règles.

4.3 Fouille visuelle de données

Depuis la fin des années 90, les techniques reposant sur la visualisation sont devenues de plus en plus importantes en ECD pour l'exploration d'ensembles de données multidimensionnelles de grandes tailles. La visualisation intervient à différentes étapes de la chaîne de traitement : dans les phases amont pour appréhender les données et effectuer les premières sélections, lors du processus de fouille puis dans la phase aval pour évaluer les résultats obtenus et les communiquer. Du fait de l'importance croissante accordée au rôle de l'utilisateur en fouille de données, les outils de visualisation sont devenus des composantes majeures des logiciels qui s'utilisent, de plus en plus, en coopération étroite avec des méthodes automatiques, à la fois en pré et post-traitement. La fouille visuelle de données intègre par essence des concepts issus de disciplines diverses : perception visuelle, psychologie cognitive, métaphores de visualisation, visualisation d'information, etc.

Nous avons vu précédemment que la principale limitation est le nombre élevé de règles générées en sortie ce qui représente une charge cognitive assez lourde pour l'utilisateur. Celui-ci doit pouvoir interagir facilement avec un environnement de fouille de données de manière à comprendre plus aisément les résultats qui lui sont proposés, ce qui est essentiel à la performance globale du système (Couturier et Mephu Nguifo, 2006). Des outils de visualisation de règles d'association ont été proposés de manière à réduire cette charge cognitive mais ils restent limités (Couturier, 2005).

Visualisation de règles d'association

Les outils que nous venons d'évoquer s'appuient sur différentes techniques de représentation en 2D (Klemettinen *et al.*, 1996, Lehn, 2000, Ben Yahia et Mephu Nguifo, 2004), ou en 3D (Liu *et al.*, 1999, Wong *et al.*, 1999, Blanchard *et al.*, 2003, Couturier *et al.*, 2005c). Le choix de l'une de ces visualisations est une tâche complexe et constitue une problématique à part entière. De plus, leurs interprétations peuvent varier d'un expert à l'autre. Chacune de ces techniques possède des avantages et des limites qu'il faut prendre en compte au moment du choix initial de la représentation. Pour plus de détails, un état de l'art et une étude comparative sont faits dans (Couturier et Mephu Nguifo, 2007). L'efficacité de ces approches est dépendante des jeux de données initiaux. Certaines représentations sont assez claires pour de petites quantités de données mais deviennent complexes lorsque ces quantités augmentent. En effet, il est probable qu'une information, qui devrait apparaître de manière saillante à l'expert, ne soit pas suffisamment perceptible dans la masse. La limitation commune à toutes les représentations est que si elles sont globales, elles deviennent vite illisibles (taille des objets en 2D, occlusions en 3D) et que si elles sont détaillées, elles ne fournissent plus à l'expert une vue d'ensemble sur ces données.

Nous avons donc entrepris une expérimentation préliminaire sur un corpus de 10 utilisateurs. Pour ce faire, nous avons successivement proposé plusieurs représentations (résumé textuel de règles, visualisations 2D, puis 3D colorées, (cf. Figure 6)) représentant le même jeu de données constitué de 62 règles. Le but des utilisateurs étaient de classer « à la volée » les trois représentations selon leur ordre de préférence sans leur donner aucune explication. Les résultats sont présentés ci-après (cf. Tableau 6). Une ligne du tableau représente un individu du corpus par son identifiant (*ID*). Par exemple, l'individu 1 a classé la représentation 2D en première position, la représentation textuelle en deuxième position et la représentation 3D en troisième position (cf. Tableau 6).

ID	Textuelle	2D	3D
1	2	1	3
2	3	1	2
3	3	1	2
4	1	2	3
5	1	2	3
6	3	2	1
7	3	1	2
8	2	1	3
9	3	2	1
10	3	2	1

Tableau 6. *Expérimentation sur le classement des représentations*

L'objet de cette expérimentation était de mettre en avant une tendance. Celle-ci a pu être dégagée malgré la taille réduite de l'échantillon. La moitié du corpus classe la représentation 2D en premier. Les classements moyens pour les représentations textuelles et 3D sont proches. Cette interprétation pourrait être spécifique à notre échantillon mais les arguments avancés par les utilisateurs nous laissent penser le contraire car ils vont tous dans le même sens. En effet, les remarques générales concernant les représentations textuelles et 3D sont que s'il y a beaucoup de données, il est assez difficile de s'y repérer. Dans le premier cas, il faut utiliser la barre de défilement pour rechercher une information. Dans le second cas, même si toutes les informations sont sur le même espace écran, les informations se chevauchent. Pour le cas de la 2D, les utilisateurs l'ayant classée en premier, estiment que même si les données sont nombreuses, cette représentation n'est pas limitée comme les deux autres. Néanmoins, ils constatent que la représentation devient vite illisible du fait de la taille des objets qu'ils voient.

Mieux visualiser pour mieux décider : les outils de visualisation en IHM

Du point de vue des IHM, selon Shneiderman (Shneiderman, 1996), le mantra de la recherche visuelle d'informations est : « fournir tout d'abord une vue d'ensemble, puis zoomer, filtrer et enfin détailler à la demande ». Il convient donc de fournir en premier lieu une vue d'ensemble du système dans le but d'en déduire rapidement les caractéristiques principales. Cette vue doit permettre à l'utilisateur d'identifier les informations importantes afin qu'il puisse décider où commence son analyse. Durant celle-ci, il doit pouvoir explorer certaines parties et en obtenir des informations détaillées s'il le souhaite. En effet, il est inutile pour l'utilisateur, que tous les détails du système soient affichés en même temps. Selon Le Grand (Le Grand, 2001), « ces deux types de besoins, qui consistent à disposer d'une vue synthétique et de détails précis, ne sont pas réalisables simultanément ». Ceci a constitué le point de départ de notre travail : comment obtenir une représentation adaptée à la visualisation de grands ensembles de règles d'association en présentant conjointement une vue générale (le global) et une vue ciblée sur un ou plusieurs éléments particuliers (le détail) ?

Nous avons vu que la fouille de données n'est efficace que si l'on dispose d'outils logiciels capables de présenter de grandes masses de résultats obtenus. Nous savons également que depuis plusieurs années, dans le domaine des IHM, différentes fonctions d'interaction ont été proposées (vues d'ensemble, zooms, filtrages de données, visualisations de relations entre objets graphiques affichés) afin de faciliter la tâche de l'utilisateur qui doit prendre connaissance de ces informations et décider du niveau de pertinence d'un élément parmi d'autres. Les travaux

présentés dans (Couturier, 2005) ont montré que, même experts d'un domaine, les utilisateurs ont souvent besoin d'outils facilitant la recherche et la navigation dans de grands ensembles d'informations.

Nous avons concentré nos efforts sur la représentation 2D puisque d'après notre évaluation, cette représentation se distingue des autres. Parmi différentes représentations connues, permettant d'appréhender une grande masse d'information, telles que les lentilles ou les arbres hyperboliques (Lamping *et al.*, 1995), les murs en perspective (MacKinlay *et al.*, 1991), les vues en oeil de poisson (FishEyeView ou FEV) (Furnas, 2006) ou encore les superpositions de vues transparentes (Harrison et Vicente, 1996), la FEV est la plus facilement adaptable à une matrice 2D.

4.4 Visualisation de règles d'association exploitant une FEV

Nous présentons ici plus en détail notre étude qui se situe dans la continuité des travaux engagés par (Vernier et Nigay, 1997) et (Rouillard, 1999) en vue d'interpréter des résultats de manière visuelle tout en préservant le contexte (cf. Figure 8b). Nous avons également testé InfoVis (Fekete, 2004) dans le cadre de nos recherches. A l'usage, il apparaît que certains éléments ne sont pas réellement adaptés à nos besoins : par exemple, dans InfoVis, une règle est représentée à l'intersection de ses métriques ; ceci limitant l'utilisation de plus de deux métriques. Or, il existe dans la littérature un grand nombre de mesures statistiques associées aux règles d'association (Couturier, 2005) et dans la pratique, l'utilisateur ne se limitera pas uniquement à deux métriques. De ce fait, plusieurs règles peuvent se chevaucher, comme on peut le voir sur la Figure 8c, ce qui en limite la lisibilité. Nous nous focalisons donc sur une représentation dans laquelle les règles seraient dessinées dans un espace alloué à l'intersection de leur prémisse et de leur conclusion (cf. Figure 8d) afin de remédier à ce problème.

Ce type de visualisation permet de représenter différentes métriques dans cet espace, grâce à différentes palettes de couleurs. Dans notre exemple, nous utilisons les deux métriques usuelles (*support* et *confiance*) pour caractériser nos règles et faciliter la compréhension de cet article. L'espace alloué est ainsi divisé en deux. Avec N métriques, l'espace serait alors divisé en N parties. Notre hypothèse consiste à supposer que nous aurons de meilleurs résultats en couplant une vue colorée sémantiquement (Fekete et Plaisant, 2002) et une FEV. L'utilisateur pourra ainsi directement pointer le polygone (coloré de manière graduelle en fonction de la valeur d'une ou plusieurs métriques associées à la règle) de la FEV qui lui paraît le plus intéressant par rapport à sa tâche.

Outils implémentant les FEV

Nous avons alors cherché des outils, permettant d'entrer des jeux de données et offrant la possibilité de visualiser les règles à l'aide d'une FEV. D'après la littérature, il n'existe pas vraiment de système totalement adapté à notre problématique. Néanmoins, nous avons étudié le cas d'IDL (Interactive Data Language)[7] qui est dédié au traitement et à la visualisation de données (séries temporelles, images, cubes, ...). Ce logiciel propriétaire s'imposa très rapidement dans les laboratoires d'astronomie, notamment, dès le début des années 90, parce qu'il répondait à certaines des contraintes de ces chercheurs : portabilité du code, existence de librairies graphiques[8], simplicité de lecture/écriture/échange de

[7] http:// www.rsinc.com

[8] Les physiciens ou astronomes ne sont pas des programmeurs, mais ils ont besoin d'outils pour manipuler des nombres et faire des statistiques, afficher des images, des tracés, etc.

fichiers. IDL est donc un langage interactif apprécié des scientifiques ayant besoin de manipuler et d'afficher des données sans pour autant avoir besoin de maîtriser un langage de programmation particulier. Ensuite, nous avons testé aiSee[9], basé sur GDL (Graph Description Language) (proche d'IDL, mais libre). Le logiciel aiSee permet de lire un jeu de données, (cf. Figure 8) issu d'un fichier (.gdl) et de visualiser ces données sous différentes formes, notamment en FEV (cf. Figure 8a). GDL décrit alors un graphe en terme de nœuds, d'arcs, de sous graphes et d'attributs. Ces attributs peuvent être la couleur, la taille, etc.

```
graph: {
        node: { title:  "A"  color: blue }
        node: { title:  "B"  color: red  }
        edge: { source: "A"  target: "B" }
}
```

Figure 7. *Format d'entrée interprété par le langage GDL*

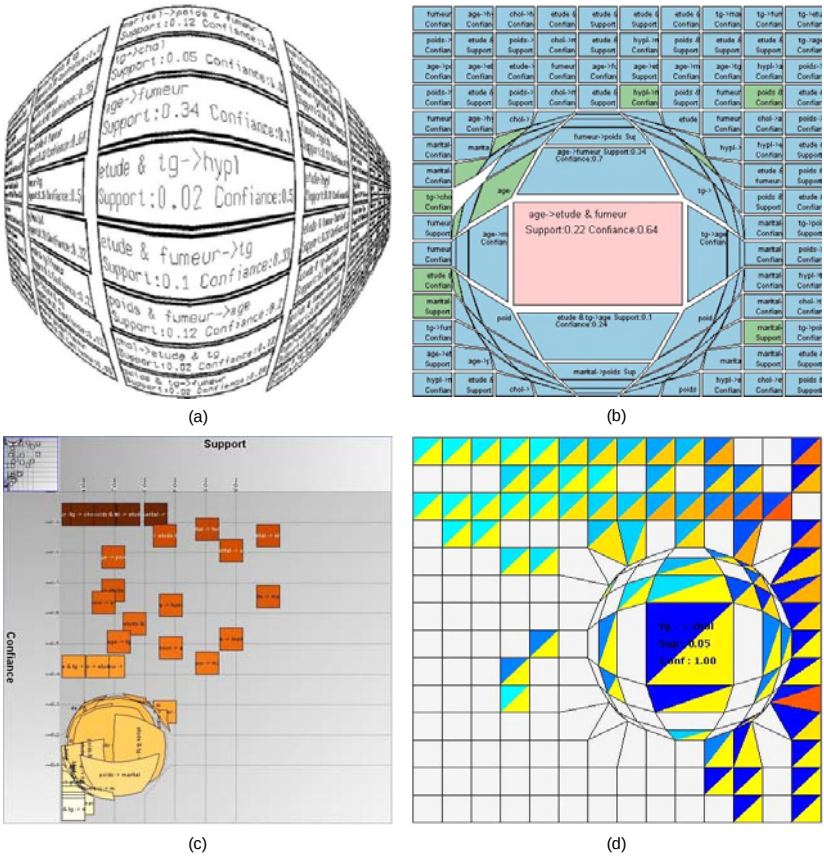

Figure 8. *Visualisations avec (a) aiSee, (b) applet JAVA, (c) InfoVis, et (d) LARM*

[9] http://www.aisee.com

Ces deux solutions paraissent, de prime abord, adaptées à notre problématique (cf. Figure 8a), mais à l'usage, des limitations apparaissent, du fait de leur trop grande généricité. Certes, on peut aussi bien exploiter ces logiciels en généalogie qu'en bioinformatique, en passant par le domaine du management, mais, de manière assez standard, et peu évoluée, selon nous. Par exemple, dans le cadre de la visualisation de règles d'association, il est nécessaire d'afficher plusieurs couleurs sur un même nœud, ce que ces logiciels ne savent pas faire de manière classique. Ces langages nous paraissent plus adaptés aux scientifiques désirant manipuler et visualiser des données de manière traditionnelle, sans nécessité de produire du code.

C'est pourquoi, dans le but d'amorcer une passerelle entre ECD et IHM, nous avons conçu et développé notre propre système de visualisation exploitant une FEV (cf. Figure 8d). Cette implémentation est réalisée en JAVA et elle est complètement intégrée au système LARM (Large Association Rules Mining) qui est développé dans (Couturier *et al*., 2005c). Nous avons donc proposé un système de visualisation de règles d'association qui permet d'alléger la charge cognitive associée à une analyse. Selon nos investigations, il semble que notre approche soit la seule à proposer simultanément une vue détaillée et générale. En effet, ce dernier se trouve devant une vue générale des règles colorées graduellement et grâce à la FEV, il peut obtenir une information détaillée sur une règle l'ayant attiré visuellement. Nous avons réalisé une évaluation de notre prototype implémentant une FEV.

Evaluation

Cette étude avait pour but de déterminer si des utilisateurs estimaient que notre approche présentait un intérêt ou pas. En complément, nous voulions aussi évaluer le caractère intuitif de notre prototype. Par exemple, l'utilisateur découvre t-il par lui-même notre problématique de départ, à savoir la limitation résultante du nombre de règles à visualiser ? Afin d'éviter tout biais qui pourraient résulter de la connaissance préliminaire d'un domaine applicatif particulier par les utilisateurs du test, nous avons choisi de visualiser un jeu de données complètement abstrait. Le fichier chargé dans le prototype est composé de 62 règles pour 10 itemsets. Nous sommes bien conscients qu'il s'agit d'un jeu de règles réduit. Nous l'avons cependant choisi dans cette première étape de notre démarche car notre but était d'évaluer en premier lieu la compréhension du couplage entre une matrice 2D et une FEV. Par la suite, il est prévu d'estimer les limites en termes de capacité de représentation de grands ensembles de données. Notre hypothèse est que notre approche sera réellement intéressante sur de grands ensembles, si et seulement si elle l'est déjà sur des jeux restreints. Chaque test concerne un évaluateur et un utilisateur. Le test est composé en trois parties : une exploration libre (EL), une démonstration (DM) et une seconde utilisation (SU). La durée totale du test est de 15 minutes. Au cours de la première partie du test (EL), l'utilisateur est placé face à un ordinateur sur lequel le prototype est démarré. Il ne reçoit aucune explication, ni à propos du logiciel ni à propos de ses fonctionnalités. L'évaluateur lui demande d'explorer à sa guise, de commenter les fonctionnalités qu'il découvre et d'expliquer ce que la visualisation représente pour lui. Cette partie dure 5 minutes durant laquelle l'évaluateur n'aide pas l'utilisateur. Le formulaire d'évaluation stipule plusieurs points à relever : est-ce que l'utilisateur a compris que la FEV permet d'affiner une information ? Quel sens donne t-il à ce qu'il voit ? Quelles sont les critiques qu'il exprime de prime abord ?

La seconde phase du test (DM) dure également 5 minutes pendant lesquelles l'évaluateur prend la main sur le prototype et réalise une démonstration en expliquant sommairement ce qui est représenté à l'écran. Ensuite, l'utilisateur est invité à utiliser une seconde fois le prototype (SU) encore pendant 5 minutes.

L'évaluateur lui demande alors ce qu'il pense des différentes fonctionnalités, notamment en termes de facilité d'utilisation et de potentiel d'usage dans des contextes applicatifs. Cette partie du test est construite sous la forme d'une discussion informelle. Le formulaire utilisé par l'évaluateur ne contient plus de directives sur des observations formelles à renseigner. L'évaluateur transcrit simplement les commentaires du sujet et son comportement vis-à-vis du logiciel.

Nous avons repris le même échantillon que précédemment qui contient 10 individus. Il ne s'agit pas d'un choix volontaire mais nous avons sélectionné des étudiants d'une promotion de DUT Informatique constituée exclusivement d'hommes. Nous avons pris ces étudiants car nous avions besoin de personnes qui soient familiers avec l'usage de l'ordinateur mais qui en même temps, ne soient pas des spécialistes de fouille visuelle de données.

Résultat de l'évaluation

La taille réduite de notre échantillon ne permet pas de tirer des conclusions statistiquement significatives. Néanmoins, la méthode de test utilisée (discount usability testing) (Shneiderman et Plaisant, 2005) fournit une évaluation relativement correcte de la perception initiale de notre prototype par des utilisateurs réels. Une telle approche est souvent considérée comme efficace pour mettre en évidence les avantages principaux et les défauts majeurs d'une interface utilisateur. Notre commentaire se focalisera sur deux types de résultats : des données quantitatives sur la reconnaissance par l'utilisateur de certaines fonctionnalités de base du prototype (cf. partie EL du test) et d'un résumé des commentaires formulés par les utilisateurs du test (cf. parties DM et SU du test).

Pour débuter, il est intéressant de remarquer que tous les sujets ont compris très rapidement qu'il s'agissait d'un ensemble de données. Sans connaître exactement le contenu, ils ont su exprimer avec leurs mots ce qu'ils voyaient : groupe, plateau ou encore ensemble. 8 étudiants ont découvert le but des curseurs (sliders permettant une sélection d'un sous ensembles de règles ayant une métrique supérieure à celle du curseur) et ont fait le lien avec les couleurs de la représentation une nouvelle fois avec leurs mots puisqu'ils ne savaient ce qui était représenté par les couleurs : palette de couleurs, une couleur est une information particulière. Tous les étudiants ont réussit à activer la FEV en cliquant sur la représentation. A ce moment, l'évaluateur leur a demandé de décrire ce qu'il voyait et de donner des avantages et des inconvénients. Ne connaissant pas le concept de la FEV, les descriptions ont été assez vastes : zoom, boule disco ou encore distorsion. 7 des étudiants y trouvent alors un intérêt pour faire ressortir une information. Les 3 étudiants restants trouvent qu'ils ne voient pas assez bien sur les côtés de la FEV et ne se focalisent alors pas sur son centre d'intérêt. Par conséquent, ils conservent une vue générale de la représentation.

Nous avons tenté de comprendre pourquoi il n'y avait pas plus d'étudiants qui étaient réceptifs à la FEV. Nous avons croisé nos résultats avec notre première expérimentation de classement des représentations (cf. Tableau 6). Nous avons découvert un point important : 5 des 7 étudiants qui trouvent un intérêt à la FEV sont des étudiants qui avaient classé la représentation 2D en première position. En dernier lieu, nous leur avons demandé si l'apport de la FEV leur ferait changer leur classement des représentations. Un seul a souhaité modifier son classement en choisissant la représentation 2D alors qu'il avait classé la 3D en première position. D'un point de vue méthodologique, nous sommes raisonnablement confiants sur le fait que les sujets n'aient senti aucune pression qui les incite à formuler des avis particulièrement favorables. Ces premiers résultats nous montrent donc que la FEV

apporte une réelle valeur ajoutée à la représentation 2D et nous conforte dans notre choix.

5 Conclusions et perspectives

Depuis quelques temps, l'importance accordée aux facteurs humains dans la conception et l'utilisation des systèmes informatique s'accroît. En effet, le facteur humain n'est plus seulement vu comme un paramètre dans un système, mais il fait bel et bien partie du système. Son degré d'implication dans le système va faire fluctuer les performances qui vont en résulter. Sa prise en compte s'avère donc plus que jamais incontournable. Ces facteurs ont leur importance dans différents domaines comme l'ergonomie, l'informatique ou encore l'ingénierie système. Nous avons situé les facteurs humains à la confluence de toutes ces problématiques, et avons présenté dans cet article leurs rôles prépondérants dans la réussite d'un processus ECD. En effet, bien qu'en théorie l'expert doit jouer un rôle central en ECD, en pratique son potentiel n'est pas exploité et ceci au détriment d'approches automatiques. Nos travaux tendent à rendre sa place à l'expert tout en montrant que cette action va permettre d'améliorer les performances des systèmes décisionnels. C'est dans ce cadre que se situe cet article.

Nous avons présenté deux travaux concrets à la confluence de l'ECD et de l'IHM. Dans le premier travail, nous avons décrit l'activité cognitive de l'expert au sein de différents processus de recherche de règles d'association. L'idée qui en ressort est que la connaissance tacite de l'expert n'est pas exploitée durant les différentes tâches du processus de l'ECD. Il intervient en amont du processus, entre chaque tâche et en fin de processus pour juger des résultats. Nous avons donc proposé un algorithme interactif de recherche de règles d'association hiérarchique permettant de pallier ce problème. Dans notre approche, l'expert fait partie du processus ce qui permet de coupler ses connaissances tacites et les connaissances explicites de l'algorithme durant la tâche de fouille de données. Cette solution a été testée avec succès sur un problème de marketing bancaire en collaboration avec des experts du domaine. Elle est tout à fait généralisable à d'autres domaines. Il s'avère également nécessaire d'étudier l'impact de nos experts sur nos résultats mais cette tâche se révèle assez complexe. Pour cette tâche, nous avons travaillé avec trois experts. Notre question initiale était de savoir si nous prenions nos experts séparément ou ensemble. Le premier cas n'était, selon nous, pas judicieux, puisque les trois experts possédaient des connaissances différentes ce qui aurait pu aboutir sur des sous ensembles différents. Ces derniers n'auraient donc pas pu être comparés. De plus, nous ne pouvions pas non plus enchaîner les deux types d'analyses, car nous travaillons sur une connaissance qui n'est pas connue de ces experts. Le fait qu'ils analysent plusieurs fois la même question, aurait influé irrémédiablement sur leurs secondes analyses. Pour ces raisons, nous avons choisi de regrouper les experts tout en sachant qu'il devenait impossible de mesurer le degré d'intervention de ces experts sur notre recherche. Une piste possible pour la suite de nos travaux, est de valider notre approche sur d'autres jeux de données de manière à obtenir un gain d'expertise moyen.

Le second travail, qui est imbriqué au premier, a mis l'accent sur l'importance d'une bonne interaction entre l'expert et le processus. L'esprit humain peut traiter un plus grand nombre de données visuellement et en extraire de l'information plus rapidement (Aupetit *et al.*, 2003). Les techniques de visualisation contribuent à proposer des outils permettant de faciliter l'acquisition de connaissances par l'expert (Shneiderman, 1996). Du fait du nombre important de connaissances générées, il

s'avère nécessaire de travailler sur des techniques performantes pour la représentation de règles. L'expert doit pouvoir aller au plus vite à l'essentiel lors d'une expertise. Nous avons donc traité dans la seconde partie de l'article, de ce problème délicat en ECD, à savoir, la visualisation de grands ensembles de données, et plus particulièrement de règles d'association. Les techniques actuelles de visualisation de tels ensembles sont peu performantes si le nombre de règles est important. Ceci est dû au fait qu'elles ne sont pas simultanément globales et détaillées. Nous avons proposé une approche hybride couplant une matrice 2D colorée et une FEV afin de résoudre ce problème et ainsi diminuer la charge cognitive de l'utilisateur en améliorant sa perception visuelle des résultats.

Une nouvelle fois, il est difficile de mesurer la performance de notre approche. En effet, comment déterminer qu'une représentation est plus performante qu'une autre. Les conclusions d'une étude, qui a été réalisée sur les différentes représentations existantes (Couturier, 2005), étaient qu'il n'y pas de représentation optimale sur un jeu de données quelconque. Le choix 2D ou 3D est déjà discutable. En effet, leurs interprétations peuvent varier en fonction des différents utilisateurs comme nous l'avons vu dans les expérimentations. Le seul véritable critère de comparaison est le nombre de règles affichées qui restent toutefois compréhensibles dans leur ensemble. Notre approche permet d'exploiter des quantités plus importantes qu'avec les représentations existantes puisque les représentations les plus illisibles en 2D, du fait d'un nombre important d'informations sur chacun des axes, vont pouvoir être traitées grâce à la FEV sans utiliser de zoom, tout en conservant une vue globale. Une question pertinente et ouverte peut alors être émise : à partir de quel moment une représentation devient cognitivement lourde pour un utilisateur ?

Tous les problèmes liés à la mesure de performance associée à des facteurs humains constituent des questions ouvertes sur lesquelles nous nous focalisons pour la suite de nos travaux. Cependant, il s'avère indispensable de prendre en compte dans notre approche l'erreur humaine ainsi que les conséquences qui en résultent. Néanmoins, nous sommes convaincus que la prise en compte des facteurs humains est indispensable à la réussite d'un système décisionnel et que dans un futur proche, cet axe de recherche est amené à se développer.

Remerciements

Ce travail est soutenu par le CNRS et l'ANRT. Nous souhaitons également remercier le service marketing de la caisse d'épargne du Pas-de-Calais et plus particulièrement Mme Brigitte Noiret, également le programme MIAOU du contrat de plan État Région Nord Pas-de-Calais et le FEDER pour leurs supports financiers partiels. Par ailleurs, les auteurs souhaitent remercier les étudiants du DUT Informatique (2007) de Lens pour leurs participations lors des évaluations.

6 Bibliographie

Agrawal, R., Imielinski, T., Swami, A. (1993). 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., USA, 207-216.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A.I. (1996). Fast discovery of association rules. In *Advances in knowledge discovery and data mining*, American Association for Artificial Intelligence, 307-328.

Aupetit, S., Monmarché, N., Guinot, C., Venturini, G., Slimane, M. (2003). Exploration de données multimédias par réalité virtuelle. *Actes de la 3ème conférence en Extraction et Gestion de Connaissances (EGC'03)*, volume 17, 71-82, Lyon, France.

Ben Yahia, S., Mephu Nguifo, E. (2004). Emulating a cooperative behavior in a generic association rule visualization tool. In *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'04)*, Boca Raton, Florida, USA.

Blanchard, J., Guillet, F., Briand, H. (2003). Exploratory Visualization for Association Rule Rummaging. In *Proceedings of the 4th International Workshop on Multimedia Data Mining MDM/KDD2003*, Washington, D.C., USA., 107-114.

Buttenfield, B. (2003). Representing information for knowledge discovery: pattern detection and database structure. In *UCGIS workshop Knowledge Discovery and Visualisation*, Boulder, Colorado, USA.

Couturier, O. (2005). *Contribution à la fouille de données : règles d'association et interactivité au sein d'un processus d'extraction de connaissances dans les données*. Thèse d'Université, Université d'Artois, CRIL, Lens, France, Décembre.

Couturier, O., Mephu Nguifo, E. (2006). Une approche anthropocentrée interactive pour l'aide à la décision en marketing bancaire. *Actes de la 18ème conférence Francophone en Interaction Homme-Machine (IHM'06)*, 253-256, Montréal, Québec, Canada.

Couturier, O., Mephu Nguifo, E. (2007). Visualisation de règles d'association en 3D par réduction des occlusions. *Revue Information, Interaction, Intelligence (I3)*, à paraître.

Couturier, O., Mephu Nguifo, E., Noiret, B. (2005a). Interactive datamining process based on human-centered system for banking marketing applications. In *Proceedings of the 7th International Conference on Enterprise Information Systems (ICEIS'05)*, 104-109, Miami, USA.

Couturier, O., Mephu Nguifo, E., Noiret, B. (2005b). A hierarchical user-driven method for association rules mining. In *Proceedings of the 11th International Conference on Human-Computer Interaction (HCI'05)*, Las Vegas, USA.

Couturier, O., Mephu Nguifo, E., Noiret, B. (2005c). A formal approach to occlusion and optimization in association rules visualization. In *Proceedings of International Symposium of Visual Data Mining (VDM) of IEEE 9th International Conference on Information Visualization (IV@VDM'05)*, Poster, London, UK.

Diaper, D. (2004). Understanding task analysis for Human-Computer Interaction. In *The handbook of task analysis for human-computer interaction*, D. Diaper, N. Stanton (Eds.), 5-47, LEA Pub.

Do, T.N., Poulet, F. (2003). Interactive visualization tools for visual data-mining. In *Proceedings of Human Centered Processes (HCP'03)*, 299-304, Luxembourg.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. (1996). From Data Mining to Knowledge Discovery: An Overview. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining*, 1-43, AAAI Press/ The MIT Press.

Fayyad, U.M., Piatetsky-Shapiro, G., Uthurusamy, R. (2003). Summary from the KDD-03 panel - Data Mining: The next 10 years. *SIGKDD Explor. Newsl.*, vol. 5, num. 2, 191-196, ACM Press.

Fekete, J.D. (2004). The InfoVis Toolkit. In *Proceedings of the 10th IEEE Symposium on Information Visualization (InfoVis'04),* IEEE Press, 167-174.

Fekete, J.D., Plaisant, C. (2002). Interactive Information Visualization of a Million Items. *IEEE Symposium on Information Visualization,*, 117-124, Boston.

Furnas, G.W. (2006). A fisheye follow-up: further reflections on focus + context. In *Proceedings of the SIGCHI conference on Human Factors in computing systems (CHI'06)*, 999-1008, Montréal, Québec, Canada.

Gasmi, G., Ben Yahia, S., Mephu Nguifo, E., Slimani, Y. (2005). IGB: A New Informative Generic Base of Association Rules. In *Proceedings of the 9th Pacific-Asia Conference on Knowledge Discovery and Datamining (PAKDD'05)*, 81-90, Hanoi, Vietnam.

Goethals, B., Van Den Bussche, J. (2000). On supporting interactive association rules mining. In *Proceedings of the 2nd International Conference on Data Warehousing and Knowledge Discovery (DWKD'00)*, London, UK.

Greenberg, S. (2004). Working through task-centered system design. In *The handbook of task analysis for human-computer interaction*, D. Diaper, N. Stanton (Eds.), 49-65, LEA Pub.

Hajek, P., Havel, I., Chytil, M. (1966). The GUHA method of automatic hypotheses determination. *Computing*, (1), 293-308.

Han, J., Kamber, M. (2001). *Data Mining; concepts and techniques*. Morgan Kauffman.

Harisson, B.L., Vicente, K.J. (1996). An experimental evaluation of transparent menu usage. In *Proc. of ACM Conference CHI'96*, New York: ACM Press, 391-398.

Hipp, J., Myka, R., Güntzer, U. (1998). A new algorithm for faster mining of generalized association rules. In *Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery in Database (PKDD'98)*, Nantes, France.

Houtsma, M.A.W., Swami, A.N. (1995). Set-oriented mining for association rules in relational databases. In *Proceedings of the 11th International Conference on Data Engineering (ICDE'95)*, pages 25-33, 1995.

Keim, D. (2001). *Visual Exploration of large data Sets*. Communications of the ACM, Vol. 44., N. 8, 39-44.

Klemettinen, M., Mannila, H., Toivonen, H. (1996). *Interactive Exploration of Discovered Knowledge : A Methodology for Interaction and Usability Studies*. Technical report, University of Helsinki.

Kodratoff, Y., Napoli, A., Zighed, D. (2001). L'extraction de connaissances, fouille de données et intelligence artificielle. *Bulletin AFIA,* numéro spécial ECD, numéro 46/47, 30-45.

Kolski, C. (2001). *Analyse et conception de l'IHM*, *Interaction pour les systèmes d'information*, volume 1. Editions HERMES, Paris.

Kotler, P., Dubois, B. (2000). *Marketing management*. Publis-union.

Kuntz, P., Guillet, F., Lehn, R., Briand, H. (2000). A user-driven process for mining association rules. In *Proceedings of the 4th European conference on Principles and Practice of Knowledge Discovery in Database (PKDD'00)*, Lyon, France, 160-168.

Lampin, G J., Rao, R., Pirolli, P. (1995). A Focus+Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies. In *Proceedings ACM Conference on Human Factors in Computing Systems (CHI'95)*, Vancouver, Canada, 401-408.

Lehn, R. (2000). *Un système interactif de visualisation et de fouille de règles pour l'extraction de connaissances dans les bases de données.* Thèse d'Université, École Polytechnique de Nantes, France.

Lajnef, M-A., Ben Ayed, M., Kolski, C. (2005). Convergence possible des processus du data mining et de conception-évaluation d'IHM : adaptation du modèle en U. *In Proceedings of IHM 2005*, ACM Press, Toulouse, 243-246.

Lefébure, R., Venturi, G. (2001). *Data mining: gestion de la relation client - Personnalisation de sites Web*. Eyrolles, Paris.

Le Grand, B. (2001). *Extraction d'information et visualisation de systèmes complexes sémantiquement structurés*. Thèse de doctorat, Université Paris VI.

Lenca, P., Vaillant, B., Meyer, P., Lallich, S. (2007). Association rule interestingness measures: experimental and theoretical studies. In *Quality Measures in Data Mining, Studies in Computational Intelligence (SCI) 43,* Guillet F., Hamilton H. J. (eds.), Springer-Verlag Berlin Heidelberg, 51-76.

Lepreux, S. (2005). *Approche de développement centré décideur et à l'aide de patrons de Systèmes Interactifs d'Aide à la Décision, application à l'investissement dans le domaine ferroviaire*. Thèse de doctorat, Université de Valenciennes et du Hainaut-Cambrésis.

Liu, B., Hsu, W., Wang, K., Chen, S. (1999). Visually aided exploration of interesting association rules. In *Proceedings of the 3rd Pacific-Asia Conference on Knowledge Discovery and Datamining (PAKDD'99)*, Beijing, China, 380-389.

Mackinlay, J.D., Robertson, G.G., Card, S.K. (1991). Perspective Wall: Detail and Context Smoothly Integrated. In *Proc. ACM Conference CHI'91*, ACM Press, New York, 173-179.

Nigay, L. (2001). *Modalité d'Interaction et Multimodalité*. Habilitation à Diriger des Recherches, spécialité Informatique de l'Université Joseph Fourier - Grenoble I.

Norman, D.A., Draper, S. (1986). *User Centered System Design : New Perspectives on Human-Computer Interaction*. Lawrence Erlbaum Associates.

Ohsaki, M., Kitaguchi, S., Okamoto, K., Yokoi, H., Yamaguchi, T. (2004). Evaluation of rule interestingness measures with a clinical dataset on hepatitis. In *Proceedings of the 8th European conference on Principles and Practices on Knowledge Discovery in Database (PKDD'04)*, Pisa, Italia.

Rouillard, J. (1999). *Navigation versus dialogue sur le web,* Une étude des préférences. *Actes d'IHM'99*, Montpellier, Novembre.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualization. In *Proceedings of IEEE Symposium on Visual Languages*, 336-343, Boulder, Colorado, USA.

Shneiderman, B., Plaisant, C. (2005). *Designing the user interface, 4th Edition*. Boston: Addison-Wesley.

Soulet, A, Crémilleux, B. (2005). An efficient framework for mining flexible constraints. In *Proceedings of the 9th Pacific-Asia Conference on Knowledge Discovery and Datamining (PAKDD'05)*, Hanoï, Vietnam, 661-671

Srikant, R., Agrawal, R. (1997). Mining generalized association rules. *Future Generation Computer Systems*, 13(2-3), 161-180.

Sutcliffe, A. (1997). Task-Related Information Analysis. *International Journal of Human-Computer Studies*, 47, 223-257.

Tseng S.M. (1998). Efficient Mining of Association Rules with Item Constraints. In *Proc. 1998 International Conference on Discovery Science*, Springer-Verlag Lecture Notes in Artificial Intelligence, vol. 1532, 423-424, Fukuoka, Japan, Dec.

Vernier, F., Nigay, L. (1997). Représentations multiples d'une grande quantité d'information. *Actes d'IHM'97*, Futuroscope de Poitiers, France.

Wong, P.C., Whitney, P., Thomas, J. (1999). Visualizing Association Rules for Text Mining. In *Proceedings of the 1999 IEEE Symposium on Information Visualization (INFOVIS'99)*, Salt Lake City, Utah, USA, 120-128.