# Recherche de règles d'association hiérarchiques par une approche anthropocentrée

Olivier COUTURIER *, Engelbert MEPHU NGUIFO * et Brigitte NOIRET **

* CRIL CNRS FRE 2499 – Université d'Artois
rue Jean Souvraz, SP-18
F-62307 Lens Cedex France
{couturier,mephu}@cril.univ-artois.fr
http ://www.cril.univ-artois.fr
** Caisse d'Epargne du Pas de Calais (CEPDC)
1, place de la république, B.P. 199
F-62304 Lens Cedex France
brigitte.noiret@cepdc.caisse-epargne.fr
http ://www.cepdc.caisse-epargne.fr

**Résumé.** L'Extraction de Connaissances dans les Bases de Données est devenue, pour les banques, une alternative au problème lié à la quantité de données qui sont stockées et qui ne cessent d'augmenter. Ceci aboutit à un paradoxe puisqu'il faut mieux cibler la clientèle susceptible d'être intéressée par une offre en utilisant des méthodes qui ne permettent plus de traiter le nombre croissant d'enregistrements des bases de données. Nos travaux se situent dans la continuité d'une étude que nous avons réalisée sur la recherche de règles d'association appliquée au marketing bancaire. En effet, des premiers résultats encourageants nous ont conduits à approfondir nos travaux vers une recherche de règles d'association hiérarchiques utilisant non plus une approche automatique mais une approche anthropocentrée. Il s'agit d'une approche dans laquelle l'expert fait partie intégrante du processus en jouant le rôle d'heuristique évolutive. Cet article présente les résultats de notre démarche de recherche.

**Mots clés :** Marketing bancaire, Extraction de Connaissances dans les Bases de Données (ECBD), Fouille de données, Taxinomie, Règles d'association hiérarchiques

## 1 Introduction

Actuellement, les entreprises sont confrontées à de nouveaux problèmes dus à l'augmentation croissante de la taille de leurs données [Han et Kamber, 2001]. Ces problèmes mettent en évidence les limitations des méthodes actuelles au niveau temps et espace mémoire. Notre problématique est d'établir des relations entre les différents produits proposés dans le but de mieux cibler les campagnes marketing bancaires. Ces travaux s'incrivent dans la continuité d'une étude visant à utiliser la recherche de règles d'association, afin de cibler des clients susceptibles d'être intéressés par un produit. Cependant, les premières conclusions ont montré que nous obtenions beaucoup trop

de règles qui n'étaient pas exploitables car triviales ou inutiles, et que le nombre de règles générées rend la phase de validation des connaissances beaucoup trop longue [Couturier *et al.*, 2003].

En fait, le problème n'est pas le nombre de clients mais le nombre d'attributs. Notre nouvelle problématique est de faire baisser le nombre d'attributs, en ne retenant que ceux qui sont intéressants pour l'expert. Pour ce faire, nous avons orienté nos travaux vers la recherche de règles d'association hiérarchiques en utilisant une taxinomie des produits proposés (cf Fig 1).

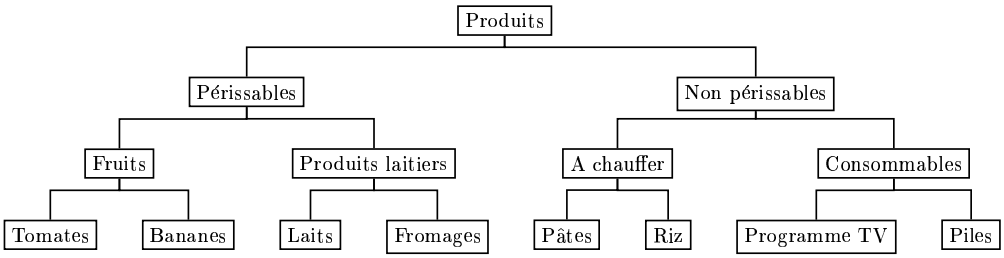


FIG. 1 – Exemple de taxinomie

L'idée générale est de faire de la recherche par niveau de hiérarchie [Han et Fu, 1995, Hipp *et al.*, 1998, Srikant et Agrawal, 1995] en allant du moins précis vers le plus précis grâce à l'intervention de l'expert qui est au centre de la recherche [Kuntz *et al.*, 2000]. On appelle cette approche *anthropocentrée* par opposition à une approche *automatique* où l'expert n'intervient qu'en début et en fin de processus. Nous ne considérons plus le processus de fouille de données de façon linéaire comme il est connu actuellement, mais de façon cyclique car l'expert juge les résultats intermédiaires et les réinjecte dans le processus pour affiner sa demande.

L'organisation de cet article se fera comme suit. Nous commencerons par définir les bases de la recherche de règles d'association hiérarchiques. Nous verrons ensuite notre application de cette méthode en marketing bancaire et enfin nous présenterons nos conclusions et perspectives pour la suite de nos travaux.

## 2 La recherche de règles d'association hiérarchiques

Trouver des règles d'association provenant de données est un problème ouvert [Agrawal *et al.*, 1996]. Il se décompose en deux sous-problèmes, la génération du treillis des ensembles fréquents et la génération des règles d'association qui sont des implications du type Si $A$ alors $B$ ou $A \mapsto B$.

Soit $\mathcal{I} = \{a_1, a_2, ..., a_m\}$ un ensemble de $m$ attributs distincts, appelés *items*. Dans la suite, *item* sera employé pour désigner un attribut et *itemset* pour désigner un ensemble

d'attributs. Soit $\mathcal{T} = \{t_1, t_2, ..., t_n\}$ une base de données contenant $n$ transactions, où chaque transaction $t_i$ est constituée d'un sous-ensemble $X \subseteq \mathcal{I}$ d'items possédant un identifiant unique ($TID$). Un ensemble d'items $X \subseteq I$ est appelé *itemset*. Un sous-ensemble d'items $X \subseteq \mathcal{I}$ de taille $k$ est également appelé *k-itemset*. Une transaction $t_i$ "contient" un *itemset* $X$ si et seulement si $X \subseteq t_i$. Soit $\mathcal{G}$, une taxinomie ou un arbre hiérarchique. Une taxinomie est un graphe acyclique dirigé où les attributs de $\mathcal{I}$ sont les feuilles de l'arbre, et où chaque arête est une relation d'héritage (cf Fig 1). La recherche va utiliser cette taxinomie pour générer des règles d'association sur différents niveaux.

# 3 Apport pour le marketing bancaire

## 3.1 Contexte de l'application

Le ciblage marketing est directement impliqué dans le problème de l'augmentation de la taille de données. La limitation des méthodes actuelles a été prouvée par la mise en situation de nos jeux de données appliqués à un gros volume de données [Couturier *et al.*, 2003]. En partant de cette base, nous avons pu tester l'algorithme Apriori sur des données fortement corrélées (le cas critique) et faiblement corrélées (le meilleur cas), en modifiant le nombre d'attributs (cf Fig 2). Nos données fortement corrélées sont représentées par une matrice binaire pleine et inversement, nos données faiblement corrélées, par une matrice binaire creuse. Ces résultats représentent le temps de calcul du treillis d'ensembles fréquents par rapport au nombre d'attributs, dans le but de découvrir à quel moment les résultats commencent à se dégrader.

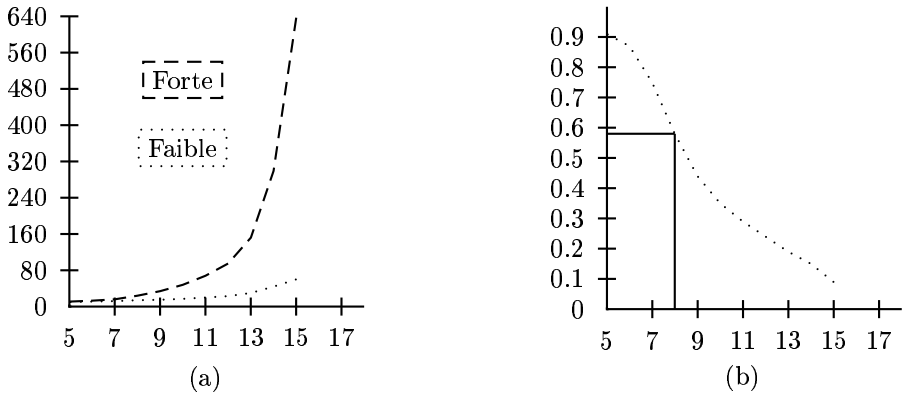


FIG. 2 – Résultats d'une étude menée sur des données fortement et faiblement corrélées

Nous montrons sur le premier graphique (a), la complexité en temps des algorithmes existants, avec une croissance plus rapide pour les données fortement corrélées. D'après nos jeux de données, nous avons montré que ces algorithmes sont très nettement limités pour la construction du treillis. D'après le second graphique (b) qui calcule le rapport entre les temps de calcul des deux types de données, les résultats commencent à se

dégrader à partir de 8 attributs. Il faut cependant noter que ces résultats varient en fonction du matériel utilisé. Nous travaillons actuellement sur une liste de 57 attributs. Les temps de réponse ne peuvent pas être validés dans un service marketing qui a besoin de trouver des connaissances le plus rapidement possible.

De plus, un des éléments qui n'entre pas en compte dans les études est le temps de validation des résultats qui représente un temps non négligeable au niveau industriel. Cette dernière étape est une problématique à part entière. En effet, un expert ne pourra juger de la pertinence des résultats que si ses capacités le lui permettent. Pour ces différentes raisons, nous avons conçu une stratégie d'analyse sous le contrôle de l'expert répondant aux objectifs, basée sur la même approche que l'algorithme Apriori en incluant l'expert dans le processus .

**Algorithme 1** Algorithme de recherche de règles d'association hiérarchiques
**fonction** *AssociationRuleGenerate(Liste d'attributs $\mathcal{I}$, Base de données $\mathcal{T}$, Taxinomie $\mathcal{G}$)*
Données: La liste contenant tous les attributs qui seront étudiés,
la base de données sur laquelle la recherche va être exécutée,
la taxinomie des différents attributs
Résultat: Liste de règles d'association $F_{ra}$
**début**
  Level=1;
  **tant que** *NextLevel==OK et user!=FIN* **faire**
    // Incrémentation du niveau
    Level++;
    // Modification de la liste d'attributs par rapport au niveau de la recherche
    $\mathcal{I} \leftarrow$ Update($\mathcal{I}$,Level);
    // Recherche des itemsets fréquents
    $F_g \leftarrow$ FrequentSearch($\mathcal{I}$,$\mathcal{T}$,$\mathcal{G}$,minsup);
    // Génération des règles d'association
    $F_{ra} \leftarrow$ RuleGeneration($F_g$,$\mathcal{G}$,minconf);
    // Règles soumises à l'utilisateur
    $F_{raFinal} \leftarrow$ ChooseRule($F_{ra}$);
    // Elagage de la taxinomie
    PruneTaxinomy($\mathcal{G}$,$F_{raFinal}$);
    // Test de fin
    **si** *Level==MaxLevel* **alors**
      NextLevel==NotOK;
  **retourner** $F_{raFinal}$;
**fin**

Nous proposons de structurer nos données sous la forme d'une taxinomie et d'effectuer des recherches d'informations par niveau de hiérarchie avec validation de l'expert à chaque étape, pour permettre ainsi de ne calculer que les informations intéressantes et innovatrices, au détriment de celles triviales et inutiles. La section suivante présente

les expérimentations que nous avons réalisées.

### 3.2 Expérimentations et résultats obtenus

Nous avons implémenté l'algorithme SHARK (Search Hierarchic Association Rules for Knowledge) pour pouvoir valider notre approche. Nous avons également développé une interface graphique adaptée pour permettre une meilleure interaction avec l'expert car il devient très vite difficile d'extraire de la connaissance lorsque les informations pertinentes sont cachées dans une masse importante de données.

Un arbre d'attributs est construit et introduit dans l'algorithme (cf. Fig 1). Cet arbre est un arbre n-aire non équilibré. Il a été réalisé en collaboration avec les experts du domaine et respecte un certain formalisme. Notre expérimentation porte sur une population constituée de 61 000 clients pour 35 attributs particuliers.

Les résultats des expérimentations montrent que l'algorithme SHARK permet de diminuer les temps d'expertise des règles d'association. De plus, l'avantage de cette méthode est qu'elle fournit des règles généralisées au premier niveau et plus la recherche va progresser, plus les règles vont être spécialisées en fonction des choix de l'expert. Les règles généralisées sont plus rapides à calculer et fournissent déjà des informations exploitables en fonction du problème à traiter.

Enfin, un problème de confidentialité des données lié à notre étude ne nous permet pas de faire ressortir l'ensemble des règles d'association extraites. Cependant ces résultats sont en phase d'exploitation.

## 4 Conclusions

Nous venons de présenter dans cet article, nos travaux d'hybridation de la recherche hiérarchique de règles d'association et de l'approche anthropocentrée. L'idée est de pouvoir diminuer les temps de calcul et le nombre de règles d'association générées par les méthodes exactes. Dans le but de diminuer les temps de calcul, la recherche hiérarchique est utilisée pour faire des recherches par niveaux afin de proposer rapidement des règles généralisées à un utilisateur qui jugera de leur pertinence. Ses choix guideront le processus dans les niveaux suivants pour spécialiser les règles. De cette façon, le nombre de règles générées est moindre et plus ciblé puisque c'est l'utilisateur qui oriente la recherche du début à la fin du processus.

Une méthodologie de fouille de données visuelle, couplant les concepts de recherche hiérarchique et d'approche anthropocentrée, a été développée. Elle est en cours de validation. Cependant, l'avantage de cette hybridation a été démontrée puisque les temps d'expertise ont diminué, ainsi que les temps de calcul. En effet, l'elagage se déroulant à chaque étape a permis d'éliminer un certain nombre d'attributs inutiles selon l'expert. Notre approche présente des limites. Le nombre de règles générées reste malgré tout important. Le problème a été réduit mais il peut encore être amélioré.

De plus, nous définissons une taxinomie avec des notions d'héritage simple. Nous ne traitons pas le cas d'héritage multiple.

### Remerciements

Ce travail est soutenu par le Centre National de la Recherche Scientifique (CNRS), par l'Association Nationale de la Recherche Technique (ANRT), par l'IUT de Lens et par l'université d'Artois. Nous souhaitons remercier tout le service marketing de la Caisse d'Epargne du Pas de Calais ainsi que les relecteurs anonymes pour leurs remarques.

## Références

[Agrawal *et al.*, 1996] R. Agrawal, H. Manilla, R. Srikant, H. Toivonen, et A.I. Verkamo. Fast discovery of association rules. *In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, Advances in Knowlegde Discovery and Data Mining*, pages 307–328, 1996.

[Couturier *et al.*, 2003] O. Couturier, E. Mephu Nguifo, et B. Noiret. Recherche de règles d'association dans une base de données bancaires. *In proceedings of EGC'03 (Journée entreprise)*, pages 71–82, 2003.

[Han et Fu, 1995] J. Han et Y. Fu. Discovery of multiple-level association rules from large databases. *In proceedings of VLDB'95*, 1995.

[Han et Kamber, 2001] J. Han et M. Kamber. Data mining : Concepts and techniques. 2001.

[Hipp *et al.*, 1998] J. Hipp, A. Myka, R. Wirth, et U. Guentzer. A new algorithm for faster mining of generalized association rules. *In proceedings of PKDD'98*, 1998.

[Kuntz *et al.*, 2000] P. Kuntz, F. Guillet, R. Lehn, et H. Briand. A user-driven process for mining association rules. *In Proceedings of PKDD'00*, pages 160–168, 2000.

[Srikant et Agrawal, 1995] R. Srikant et R. Agrawal. Mining generalized association rules. *In Proceedings of VLDB'95*, pages 1–12, 1995.

## Summary

Knowledge Discovery in Databases (KDD) is the new hope for banking marketing due to increase of the collection of large databases. There is a paradox because the bank must improve the development's policy of customer loyalty by using methods that do not allow to treat big quantities of data. Our current works are the results of a study that we led on a association rules search in banking marketing. Our first encouraging results steered our works towards a hierarchic association rules search, using an user-driven approach rather than an automatic approach. The user is the heart of the process by playing a role of evolutionnary heuristic. This paper presents the results of our new method for finding association rules.

**Keywords :** Banking marketing, Knowledge Discovery in Databases (KDD), Datamining, Taxinomy, hierarchic association rules