# Recherche de règles non redondantes par vecteurs de bits dans des grandes bases de motifs [1]

François Jacquenet, Christine Largeron, Cédric Udréa

Laboratoire EURISE – Université Jean Monnet
23 rue du Docteur Michelon – 42023 Saint-Etienne Cedex 2 – France
{Francois.Jacquenet,Christine.Largeron,Cedric.Udrea}@univ-st-etienne.fr

Deux voies sont envisageables pour limiter le nombre de motifs extraits dans un processus de fouille de données. La première s'efforce, lors de la génération des motifs, de ne conserver que les seuls motifs semblant présenter un intérêt immédiat pour l'utilisateur (Boulicaut, 2005), tandis que la seconde voie consiste à stocker tous les motifs extraits par les algorithmes de fouille de données dans des structures de données efficaces et à développer des outils d'interrogation et de manipulation permettant de les traiter (Grossman et al., 1999; Tuzhilin et Liu, 2002; Zaki et al., 2005). C'est en suivant cette démarche que nous nous sommes intéressés à la recherche de règles d'association non redondantes alors que la plupart des travaux antérieurs consacrés à ce problème se sont plutôt attachés à l'extraction de règles non redondantes directement à partir des données (Zaki, 2000; Bastide et al., 2000; Li et al., 2004; Li et Hamilton, 2004; Goethals et al., 2005).

Dans la suite, en nous inspirant d'une définition de (Bastide et al., 2000), nous considérons qu'une règle d'association $B \rightarrow H$ est non redondante si et seulement si il n'existe pas de règle de la forme $B' \rightarrow H'$ telle que $B' \subseteq B$ et $H \subseteq H'$. Chaque partie de la règle d'association peut être représentée par un vecteur qui possède autant de bits qu'il existe d'items dans la base de transactions (Morzy et Zakrzewicz, 1998). Chaque bit est alors associé à un item particulier et la valeur du bit est de '1' si et seulement si l'item correspondant est présent dans la partie de la règle associée au vecteur de bits.

En utilisant ce codage, nous proposons de déterminer la redondance d'une règle $R = B \rightarrow H$ vis-à-vis d'une autre règle $R' = B' \rightarrow H'$, en exploitant la propriété suivante :

Etant donné $IB_X = \{IB_1^X,...,IB_k^X\}$ (respectivement $IH_X = \{H_1^X,...,IH_k^X\}$) le vecteur de bits correspondant à la partie gauche (respectivement droite) de la règle $X$ où $IB_i^X$ (respectivement $IH_i^X$) est égal à 1 si l'item $i$ est présent dans la partie gauche (respectivement droite) de la règle $X$, 0 sinon. Nous démontrons alors que la règle $R$ est redondante par rapport à la règle $R'$ si et seulement si $N_b(R \text{ AND } R')$=$N_b(R')$ et $N_h(R \text{ AND } R')$=$N_h(R)$ où $N_b(X)$ désigne le nombre de '1' dans $IB_X$, $N_h(X)$ le nombre de '1' dans $IH_X$ et ($R$ AND $R'$) désigne la règle ayant en partie gauche l'intersection des parties gauches des règles $R$ et $R'$ et en partie droite l'intersection des parties droites des règles $R$ et $R'$.

Nous avons développé un algorithme, basé sur cette propriété, et réalisé plusieurs tests pour comparer les temps nécessaires pour extraire les règles non redondantes d'un ensemble

1. Ce travail a été partiellement soutenu par le projet BINGO de l'ACI Masses de Données 2004-2007, financé par le Ministère de la Recherche

de règles en utilisant des vecteurs de bits par rapport à l'approche faisant appel à un mode de stockage plus classique nécessitant une table comportant trois attributs : l'identifiant de la règle, l'identifiant de la partie concernée (gauche ou droite) et l'identifiant de l'item. Ces expérimentations, menées en faisant varier le nombre global de règles ainsi que le nombre de règles redondantes, ont confirmé l'intérêt de l'approche par vecteurs de bits.

## Références

Bastide, Y., N. Pasquier, R. Taouil, G. Stumme, et L. Lakhal (2000). Mining minimal non-redundant association rules using frequent closed itemsets. In *Proceedings of the first International Conference on Computational Logic*, LNCS 1861, pp. 972–986.

Boulicaut, J. F. (2005). Condensed representations for data mining. In *Encyclopedia of Data Warehousing and Mining*, pp. 207–211. Idea Group Reference.

Goethals, B., J. Muhonen, et H. Toivonen (2005). Mining non-derivable association rules. In *Proceedings of the fifth International Conference on Data Mining*.

Grossman, R. L., S. Bailey, A. Ramu, B. Malhi, P. Hallstrom, I. Pulleyn, et X. Qin (1999). The management and mining of multiple predictive models using the predictive model markup language (pmml). In *Information and Software Technology*, Volume 41, pp. 589–595.

Li, G. et H. Hamilton (2004). Basic association rules. In *Proceedings of the fourth SIAM International Conference on Data Mining*. SIAM.

Li, Y., Z. T. Liu, L. Chen, W. Cheng, et C. H. Xie (2004). Extracting minimal non-redundant association rules from QCIL. In *International Conference on Computer and Information Technology*, pp. 986–991. IEEE Computer Society.

Morzy, T. et M. Zakrzewicz (1998). Group bitmap index: A structure for association rules retrieval. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, pp. 284–288. AAAI Press.

Tuzhilin, A. et B. Liu (2002). Querying multiple sets of discovered rules. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 52–60. ACM.

Zaki, M. J. (2000). Generating non-redundant association rules. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pp. 34–43.

Zaki, M. J., N. Parimi, N. De, F. Gao, B. Phoophakdee, J. Urban, V. Chaoji, M. A. Hasan, et S. Salem (2005). Towards generic pattern mining. In *Proceedings of the Third International Conference on Formal Concept Analysis*, pp. 1–20.

## Summary

The management of large pattern databases rapidly becomes untractable. This paper presents the way we have efficiently implemented the search for non redundant rules, in post treatment, thanks to a representation of rules in the form of bit strings.