A Galois connection semantics-based approach for deriving generic bases of association rules

S. Ben Yahia, N. Doggaz Y. Slimani, J. Rezgui
Département des Sciences de l'Informatique
Faculté des Sciences de Tunis
Campus Universitaire, 1060 Tunis, Tunisie.
sadok.benyahia;yahya.slimani;narjes.doggaz@fst.rnu.tn;jihen_rezgui@yahoo.fr

Résumé. L'augmentation vertigineuse de la taille des données (textuelles ou transactionnelles) est un défi constant pour la "scalabilité" des techniques d'extraction des connaissances. Dans ce papier, on présente une approche pour la dérivation des bases génériques de règles associatives. Les principales caractéristiques de cette approches sont les suivantes. D'une part, l'introduction d'une structure de données appelée "Trie-itemset" pour le stockage de la relation en entrée. D'autre part, on utilise une méthode "Diviser pour régner" pour réduire le coût de construction de structures partiellement ordonnées, à partir desquelles les bases génériques de règles sont directement extraites.

1 Introduction

Much research in data mining from large databases has focused on the discovery of association rules [Agrawal et Skirant, 1994, Brin *et al.*, 1997, Manilla *et al.*, 1994]. Association rule generation is achieved from a set F of frequent itemsets in an extraction context $\mathcal{D}$, for a minimal support *minsupp*. An association rule r is a relation between itemsets of the form $r: X \Rightarrow (Y - X)$, in which X and Y are frequent itemsets, and $X \subset Y$. Itemsets X and $(Y - X)$ are called, respectively, *antecedent* and *conclusion* of the rule r. The valid association rules are those of which the measure of confidence $Conf(r) = \frac{support(Y)}{support(X)}$ [1] is greater than or equal to the minimal threshold of confidence, named *minconf*. If $Conf(r) = 1$ then r is called *exact association rule (ER)*, otherwise it is called *approximative association rule (AR)*. Exploiting and visualizing association rules is far from being a trivial task, mostly because of the huge number of potentially interesting rules that can be drawn from a dataset. Various techniques are used to limit the number of reported rules, starting by basic pruning techniques based on thresholds for both the frequency of the represented pattern (called the *support*) and the strength of the dependency between antecedent and conclusion (called the *confidence*). More advanced techniques that produce only a limited number of the entire set of rules rely on closures and Galois connections [Bastide *et al.*, 2000, Stumme *et al.*, 2001, Zaki, 2000], which are in turn derived from Galois lattice theory and formal concept analysis (FCA) [Ganter et Wille, 1999]. Finally, works on FCA have yielded a row of results on compact representations of closed set families, also called *bases*, whose impact on association rule reduction is currently under intensive investigation within the community [Bastide *et al.*, 2000, Stumme *et al.*, 2001].

In this paper, we propose a trie-based new data structure called "**Itemset-trie**" tree. Itemset-trie tree extends the idea claimed by the authors of FPTree [Han *et al.*, 2000] and CATS [Cheung et Zaiane, 2003], aiming to improve storage compression and to allow (closed) frequent pattern mining without "explicit" candidate itemsets generation. Next, we propose an algorithm, falling in the characterization "Divide and Conquer" to extract the frequent closed itemsets with their associated minimal generators. It is noteworthy that the derivation of Luxemburger base is based on the exploration of such closed itemsets organized upon their natural partial order (also called *precedence relation*). That's why we construct on the