

# Extraction d'itemsets compacts

Bashar Saleh\* Florent Massegli\*

\*Inria Sophia-Antipolis Méditerranée  
Equipe-Projet AxIS  
2004 route des lucioles - BP 93  
FR-06902 Sophia Antipolis  
{Prénom.Nom}@sophia.inria.fr,  
<http://www-sop.inria.fr/axis>

**Résumé.** L'extraction d'itemsets fréquents est un sujet majeur de l'ECD et son but est de découvrir des corrélations entre les enregistrements d'un ensemble de données. Cependant, le support est calculé en fonction de la taille de la base dans son intégralité. Dans cet article, nous montrons qu'il est possible de prendre en compte des périodes difficiles à déceler dans l'organisation des données et qui contiennent des itemsets fréquents sur ces périodes. Nous proposons ainsi la définition des itemsets compacts, qui représentent un comportement cohérent sur une période spécifique et nous présentons l'algorithme DEICO qui permet leur découverte.

## 1 Introduction

Le problème de la recherche de règles d'association, introduit dans Agrawal et al. (1993), est basé sur l'extraction de corrélations fréquentes entre les enregistrements et connaît de nombreuses applications dans le marketing, la gestion financière ou l'analyse décisionnelle (par exemple). Au cœur de ce problème, la découverte d'itemsets fréquents représente un domaine de recherche très étudié. Dans l'analyse du panier de la ménagère, par exemple, les itemsets fréquents ont pour but de découvrir des ensembles d'items qui correspondent à un nombre significatif de clients. Si ce nombre est supérieur à un support défini (par l'utilisateur) alors cet itemset est considéré comme fréquent. Cependant, dans la définition initiale des itemsets fréquents, l'extraction est effectuée sur la base de données toute entière (*i.e.* soit  $min_{supp}$ , le support minimum donné par l'utilisateur, les itemsets extraits doivent apparaître dans au moins  $|D| \times min_{supp}$  enregistrements de  $D$ ). Toutefois, il est possible que des itemsets intéressants reste ignorés malgré des caractéristiques particulières (y compris de support). Effectivement, les itemsets intéressants sont souvent liés au moment qui correspond à leur observation. On pourrait prendre pour exemple le comportement des utilisateurs d'un site de commerce en ligne pendant une offre spéciale sur les DVD et les CD vierges pour laquelle une publicité est faite par mailing. De la même manière, le site Web d'une conférence peut voir le nombre de connexions augmenter dans une fenêtre de quelques heures avant la date limite de soumission. Une condition nécessaire à la découverte de ce type de données est liée à l'aspect temporel des données. Cet aspect a déjà été abordé pour les règles d'association dans Ale et Rossi (2000);