Découverte d'itemsets fréquents fermés sur architectures multicœurs

Benjamin Négrevergne*, Alexandre Termier*, Jean-François Méhaut*, Takeaki Uno**

*LIG - UJF-CNRS UMR 5217 - 681 rue de la Passerelle, B.P. 72, 38402 Saint Martin d'Hères, France
{Benjamin.Negrevergne, Jean-Francois.Mehaut, Alexandre.Termier}@imag.fr, http://www.liglab.fr
**National Institute of Informatics, 2-1-2, Hitotsubashi, Chiyoda-ku, 101-8430 Tokyo, Japan
uno@nii.jp, http://research.nii.ac.jp/∼uno/

Résumé. Dans ce papier nous proposons PLCM, un algorithme parallèle de découverte d'itemsets fréquents fermés basé sur l'algorithme LCM, reconnu comme l'algorithme séquentiel le plus efficace pour cette tâche. Nous présentons aussi une interface de parallélisme à la fois simple et puissante basée sur la notion de *Tuple Space*, qui permet d'avoir une bonne répartition dynamique du travail.

Grâce à une étude expérimentale détaillée, nous montrons que PLCM est le seul algorithme qui soit suffisamment générique pour calculer efficacement des itemsets fréquents fermés à la fois sur des bases creuses et sur des bases denses, améliorant ainsi l'état de l'art.

1 Introduction

La découverte de motifs fréquents est l'un des domaines majeurs de la fouille de données. A l'origine de ce domaine se trouvent les travaux d'Agrawal et Srikant (1994) sur la découverte d'itemsets fréquents. Le problème de la découverte d'itemsets fréquents consiste, étant donné une base de données où les transactions sont des listes d'*items* et un seuil de support minimal $minsup$, à découvrir tous les *itemsets* qui apparaissent plus de $minsup$ fois dans la base de de données. Ce problème est le plus simple du domaine de la découverte de motifs fréquents, comparativement à la recherche de séquences, d'arbres ou de graphes fréquents. Toutefois, il se retrouve dans de nombreuses applications, en particulier dans l'analyse de données de vente d'organisations commerciales. De plus, de part sa relative simplicité, toutes les améliorations algorithmiques significatives en découverte de motifs fréquents ont d'abord été réalisées dans le cas des itemsets fréquents avant d'être adaptées à des motifs plus complexes. C'est en particulier le cas pour la fermeture (Pasquier et al. (1999)) ou les méthodes sans génération de candidats (Han et al. (2000)).

En 2003 et 2004, le workshop FIMI (Goethals (2004)) a confronté les algorithmes de recherche d'itemsets fréquents afin de déterminer les meilleures solutions. Le gagnant de FIMI 2004, LCM2 de Uno et al. (2004b), combine des améliorations de haut niveau issues de la théorie de l'énumération, et des optimisations de bas niveau optimisant la répartition entre