Le FIA: un nouvel automate permettant l'extraction efficace d'*itemsets* fréquents dans les *flots de données*

Jean-Émile SYMPHOR*, Alban MANCHERON*, Lionel VINCESLAS* et Pascal PONCELET**

*GRIMAAG, Université des Antilles et de la Guyane, Martinique, France.
{je.symphor;alban.mancheron;lionel.vinceslas}@martinique.univ-ag.fr

** EMA-LG2IP/site EERIE, Parc Scientifique Georges Besse, 30035 Nîmes Cedex, France.
pascal.poncelet@ema.fr

Résumé. Le FIA (*Frequent Itemset Automaton*) est un nouvel automate qui permet de traiter de façon efficace la problématique de l'extraction des itemsets fréquents dans les *flots de données*. Cette structure de données est très compacte et informative, et elle présente également des propriétés incrémentales intéressantes pour les mises à jour avec une granularité très fine. L'algorithme développé pour la mise à jour du FIA effectue un unique passage sur les données qui sont prises en compte tout d'abord par *batch* (*i.e.*, itemset par itemset), puis pour chaque itemset, item par item. Nous montrons que dans le cadre d'une approche prédictive et par l'intermédiaire de la bordure statistique, le FIA permet d'indexer les itemsets véritablement fréquents du *flot* en maximisant le rappel et en fournissant à tout moment une information sur la pertinence statistique des itemsets indexés avec la P-valeur.

1 Introduction

L'extraction d'itemsets fréquents est une problématique de recherche qui intéresse la communauté fouille de données depuis plus d'une dizaine d'années et intervient pour la recherche de règles d'association, de motifs séquentiels ou encore d'itemsets maximaux. Les premiers à traiter cette question furent Agrawal et Srikant (1994), ils ont été suivis en ce sens par Han et al. (2000). Traditionnellement, les différents algorithmes proposés dans la littérature reposent sur des structures de données de type arbre ou encore treillis (*e.g.* : `A-priori` (Agrawal et Srikant, 1994), `FP-growth` (Han et al., 2000), ...). La problématique de recherche de motifs (*i.e.*, une généralisation des itemsets) apparaît dans des domaines aussi variés que la bioinformatique ou la fouille de textes. En ce qui concerne ce dernier, de nouvelles structures de données, basées sur des automates sont apparues afin d'extraire les sous-séquences communes à une ensemble de textes (Troníček, 2002). Par exemple, Hoshino et al. (2000) ont introduit, un nouvel automate déterministe et acyclique : le SA (Subsequence Automaton) qui permet de reconnaître toutes les sous-séquences d'un ensemble de textes. L'un des problèmes principaux auxquels doit faire face une approche d'extraction de motifs est de disposer de structures qui soient suffisamment compactes et informatives afin de minimiser l'explosion combinatoire liée à d'importants espaces de recherche. En effet, l'applicabilité des algorithmes