SPAMS, une nouvelle approche incrémentale pour l'extraction de motifs séquentiels fréquents dans les *Data streams*

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

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

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

Résumé. L'extraction de motifs séquentiels fréquents dans les data streams est un enjeu important traité par la communauté des chercheurs en fouille de données. Plus encore que pour les bases de données, de nombreuses contraintes supplémentaires sont à considérer de par la nature intrinsèque des streams. Dans cet article, nous proposons un nouvel algorithme en une passe : SPAMS, basé sur la construction incrémentale, avec une granularité très fine par transaction, d'un automate appelé SPA, permettant l'extraction des motifs séquentiels dans les streams. L'information du stream est apprise à la volée, au fur et à mesure de l'insertion de nouvelles transactions, sans pré-traitement a priori. Les résultats expérimentaux obtenus montrent la pertinence de la structure utilisée ainsi que l'efficience de notre algorithme appliqué à différents jeux de données.

1 Introduction

Concerné par de nombreux domaines d'application (e.g. le traitement des données médicales, le marketing, la sécurité et l'analyse financière), l'extraction de motifs séquentiels fréquents est un domaine de recherche actif qui intéresse la communauté des chercheurs en fouille de données. Initialement, les premiers travaux présentés traitent du cas des bases de données statiques et proposent des méthodes dites exactes d'extraction de motifs séquentiels. On peut citer, à titre d'exemple, les algorithmes GSP, SPADE, PrefixSpan et SPAM, respectivement proposés par Srikant et Agrawal (1996); Zaki (2001); Pei et al. (2001); Ayres et al. (2002). Plus récemment ces dernières années, de nouvelles applications émergentes, telles que l'analyse de traffic dans les réseaux, la fouille de données "clickstream"[1] ou encore la détection de fraudes et d'intrusions, induisent de nouvelles problématiques qui impactent les méthodes de fouilles. En

[1] clickstream : flot de requêtes d'utilisateurs sur des sites web.