Echantillonnage pour l'extraction de motifs séquentiels : des bases de données statiques aux flots de données

Chedy Raïssi[*,**], Pascal Poncelet[**]

[*]LIRMM, 161 rue Ada,34392 Montpellier Cedex 5, France
raissi@lirmm.fr,
[**]EMA-LGI2P, Parc Scientifique Georges Besse, 30035 Nîmes Cedex, France
prénom.nom@ema.fr

Résumé. Depuis quelques années, la communauté fouille de données s'est intéressée à la problématique de l'extraction de motifs séquentiels à partir de grandes bases de données en considérant comme hypothèse que les données pouvaient être chargées en mémoire centrale. Cependant, cette hypothèse est mise en défaut lorsque les bases manipulées sont trop volumineuses. Dans cet article, nous étudions une technique d'échantillonnage basée sur des réservoirs et montrons comment cette dernière est particulièrement bien adaptée pour résumer de gros volumes de données. Nous nous intéressons ensuite à la problématique plus récente de la fouille sur des données disponibles sous la forme d'un flot continu et éventuellement infini ("*data stream*"). Nous étendons l'approche d'échantillonnage à ce nouveau contexte et montrons que nous sommes à même d'extraire des motifs séquentiels de flots tout en garantissant les taux d'erreurs sur les résultats. Les différentes expérimentations menées confirment nos résultats théoriques.

1 Introduction

La problématique de l'extraction de motifs séquentiels dans de grandes bases de données intéresse la communauté fouille de données depuis une dizaine d'années et différentes méthodes ont été développées pour extraire des séquences fréquentes. L'extraction de tels motifs est toutefois une tâche difficile car l'espace de recherche considéré est très grand. Afin de gérer au mieux cet espace de recherche, différentes stratégies ont été proposées. Les plus traditionnelles utilisent une approche à la *Apriori* Srikant et Agrawal (1996) et diffèrent principalement par les structures de données utilisées (vecteurs de bits, arbres préfixés, ...). Les approches les plus récentes considèrent, quant à elles, des projections multiples de la base de données selon le principe de *pattern-growth* proposé dans Pei et al. (2001) et évitent ainsi de générer des candidats. Outre ces différentes stratégies, les propositions les plus efficaces considèrent comme hypothèse que la base de données peut être chargée directement en mémoire centrale. Cependant, avec le développement des nouvelles technologies, ces dernières se trouvent de plus en plus mises en défaut dans la mesure où la quantité de données manipulées est trop volumineuse et qu'il devient irréaliste de stocker l'intégralité de la base en mémoire centrale.

Le développement des nouvelles technologies permet également de générer de très grands volumes de données issues de différentes sources : trafic TCP/IP, transactions financières, en-