Vers une nouvelle approche d'extraction des motifs séquentiels non-dérivables

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. L'extraction de motifs séquentiels est un défi important pour la communauté fouille de données. Même si les représentations condensées ont montré leur intérêt dans le domaine des itemsets, à l'heure actuelle peu de travaux considèrent ce type de représentation pour extraire des motifs. Cet article propose d'établir les premières bases formelles pour obtenir les bornes inférieures et supérieures du support d'une séquence S. Nous démontrons que ces bornes peuvent être dérivées à partir des sous-séquences de S et prouvons que ces règles de dérivation permettent la construction d'une nouvelle représentation condensée de l'ensembles des motifs fréquents. Les différentes expérimentations menées montrent que notre approche offre une meilleure représentation condensée que celles des motifs clos et cela sans perte d'information.

1 Introduction

Motivée par de nombreux domaines d'applications (e.g. marketing web, analyses financières, détections d'anomalies dans les réseaux, traitements de données médicales), l'extraction de motifs séquentiels fréquents est un domaine de recherche très actif Mobasher et al. (2002); Ramirez et al. (2000); Lattner et al. (2005). Les travaux menés ces dernières années ont montré que toutes les approches qui visent à extraire l'ensemble des motifs séquentiels deviennent cependant inefficaces dès que le support minimal spécifié par l'utilisateur est trop bas ou lorsque les données sont fortement corrélées. En effet, dans ce cas, et plus encore que pour les itemsets, les recherches sont pénalisées par un espace de recherche trop important. Par exemple, avec i attributs (appelés aussi *items*), il y a potentiellement $O(i^k)$ séquences fréquentes de taille k Zaki (2001). Pour essayer de gérer au mieux ces problèmes de complexités spatiale et temporelle, deux grandes tendances se distinguent à l'heure actuelle. Dans le premier cas, les propositions comme PrefixSPAN Pei et al. (2004) ou SPADE Zaki (2001) se basent sur de nouvelles structures de données et une génération de candidats efficace. Les approches de la seconde tendance considèrent l'extraction d'une représentation condensée Mannila and Toivonen (1996). Même si l'utilisation d'une représentation compacte a montré son intérêt dans le domaine de l'extraction d'itemsets, la complexité structurelle des motifs séquentiels fait qu'il existe cependant peu de travaux utilisant une représentation condensée dans ce contexte.