Extraction incrémentale de séquences fréquentes dans un flux d'itemsets

Thomas Guyet[*,***], René Quiniou[**,***]

[*]AGROCAMPUS-OUEST
[**]INRIA, Centre de Rennes - Bretagne Atlantique
[***]IRISA - UMR 6074 Campus de Beaulieu, F - 35 042 Rennes Cedex
thomas.guyet@agrocampus-ouest.fr, quiniou@inria.fr

1 Introduction

De nombreuses méthodes ont été proposées pour l'extraction de motifs séquentiels d'une base de transactions. La plupart détermine la fréquence d'un motif à partir du nombre de transactions contenant ce motif, sans tenir compte des répétitions dans une même transaction. Plus rares sont les approches visant à extraire les motifs, ou épisodes, fréquents dans une séquence d'itemsets unique. Le comptage du nombre d'occurrences d'un épisode dans une séquence est plus difficile dans la mesure où il faut tenir compte des répétitions et chevauchements entre les occurrences d'un même épisode. Plusieurs méthodes de comptage ont été proposées pour résoudre cette difficulté tout en conservant des propriétés de monotonie nécessaires à l'efficacité de la recherche d'occurrences (voir Achar et al. (2010)). Par exemple, Winepi (Mannila et al. (1997)) compte toutes les occurrences d'un épisode dans la séquence et la fréquence est le nombre de fenêtres contenant cet épisode lorsque l'on fait glisser la fenêtre sur toute la séquence. Minepi, des mêmes auteurs, compte le nombre d'occurrences minimales d'un épisode.

Des solutions algorithmiques spécifiques doivent être adaptées aux flux de données pour l'extraction de motifs ou épisodes fréquents. Dans un contexte de flux de données, la fenêtre glissante sur laquelle sont extraits les épisodes fréquents est une séquence d'itemsets en perpétuelle évolution : lorsque des nouvelles données arrivent, elles rendent obsolètes celles du début de la fenêtre. Une approche naïve réitèrant l'intégralité du processus de fouille pour la séquence à chaque modification serait trop coûteuse en temps de calcul. La plupart des méthodes d'extraction de séquences fréquentes dans un flux de données traite de la gestion des séquences fréquentes extraites dans des tampons successifs du flux ou "batches" (Marascu et Masseglia (2006)). Seuls quelques algorithmes se sont intéressés au problème de la fouille dans des fenêtres glissantes.

2 Algorithme incrémental de fouille d'un flux d'itemsets

Nous présentons un algorithme incrémental, complet et correct, d'extraction de séquences d'itemsets fréquentes basé sur un dénombrement des occurrences minimales d'une séquence. L'algorithme s'appuie sur la représentation des séquences fréquentes sous la forme d'un arbre