Motifs Séquentiels δ-Libres

Marc Plantevit*, Chedy Raïssi**, Bruno Crémilleux***

* Université de Lyon, CNRS, INRIA
Université Lyon 1, LIRIS Combining, UMR5205, F-69622, France
Marc.Plantevit@liris.cnrs.fr,
**INRIA Nancy Grand-Est
Chedy.Raissi@loria.fr
*** Université de Caen Basse-Normandie
GREYC, UMR6072, F-14032, France
Bruno.Cremilleux@info.unicaen.fr

Résumé. Bien que largement étudiée, l'extraction de motifs séquentiels reste une tâche très difficile et pose aussi le défi du grand nombre de motifs produits. Dans cet article, nous proposons une nouvelle approche extrayant les motifs séquentiels les plus généraux à fréquence similaire. Nous montrons en quoi l'extension de cette notion, déjà connue pour les motifs ensemblistes, est un problème particulièrement difficile pour les séquences. Les motifs δ-libres ainsi produits sont en nombre réduit et facilitent les usages d'un processus de fouille et nous montrons leur apport comme descripteurs dans un contexte de classification de séquences.

1 Introduction

La fouille de données temporelles est une problématique rencontrant un vif succès notamment parce qu'elle est motivée par de nombreuses applications telles que l'analyse de données clients ou financières, le web usage mining ou encore l'analyse de séquences biologiques. Pour un aperçu des méthodes de fouilles de séries temporelles ou de flots de données, voir Dong et Pei (2007). Les données représentées par des séquences d'événements discrets sont un cas particulier très étudié de données temporelles. Une tâche importante dans de telles données est de découvrir les régularités présentes en extrayant des *motifs séquentiels*, tâche introduite par Agrawal et Srikant (1995). Un motif séquentiel est un motif local représentant une séquence d'itemsets régulièrement observée dans les données. Au-delà de leur intérêt intrinsèque, ceux-ci sont aussi exploités à des fins de clustering ou de classification.

Il est bien connu que l'extraction de motifs pose le problème de l'abondance des résultats produits : l'utilisateur a peu de guide pour découvrir les motifs qui lui seront utiles parmi la masse de motifs proposés. Cela a conduit la communauté à mener de nombreux travaux pour offrir une vision plus synthétique des motifs extraits (représentations condensées Mannila et Toivonen (1996), compression des données van Leeuwen et al. (2009)) ou se focaliser sur les plus pertinents suivant des critères a priori (telles que l'extraction sous contraintes de motifs locaux ou d'ensembles de motifs, voir Bonchi et Lucchese (2007)).