Fouille de motifs séquentiels avec ASP

Thomas Guyet[*,**], Yves Moinard[**], René Quiniou[**], Torsten Schaub[**,***]

[*]AGROCAMPUS-OUEST/IRISA UMR 6074
65 rue de Saint Brieuc, 35042 Rennes
[**]Inria – Centre de Rennes Bretagne Atlantique
[***]Université de Potsdam, Allemagne

Résumé. Cet article présente l'utilisation de la programmation par ensembles réponses (ASP) pour répondre à une tâche de fouille de motifs séquentiels. La syntaxe de l'ASP, proche du Prolog, en fait un langage très pertinent pour représenter des connaissances de manière aisée et ses mécanismes de résolution, basés sur des solveurs efficaces, en font une solution alternative aux approches de programmation par contraintes pour la fouille déclarative de motifs. Nous proposons un premier encodage de la tâche classique d'extraction de motifs séquentiels et de ses variantes (motifs clos et maximaux). Nous comparons les performances calculatoires de ses encodages avec une approche de programmation par contraintes. Les performances obtenues sont inférieures aux approches de programmation par contraintes, mais l'encodage purement déclaratif offre plus de perspectives d'intégration de connaissances expertes.

1 Introduction

L'extraction de motifs (*pattern mining*) consiste à identifier les motifs « intéressants » dans une base de données structurée. Dans le cas de l'extraction de motifs séquentiels (Shen et al., 2014), la base de données est structurée sous la forme d'un ensemble de séquences d'itemsets (par exemple, des séquences d'achats). La tâche d'extraction de motifs consiste à identifier l'ensemble de sous-séquences qui apparaissent fréquemment dans les séquences de la base. Une fréquence suffisante, définie par un seuil f_{min} est la mesure d'intérêt classiquement utilisée pour les motifs, car elle permet de concevoir des algorithmes efficaces.

Les recherches en algorithmique ont conduit à proposer des méthodes d'extraction de motifs séquentiels de plus en plus efficaces si bien qu'elles sont en mesure d'extraire des motifs dans de très grandes bases de données et pour des seuils de fréquences très bas. Ces méthodes se heurtent maintenant à une profusion de motifs qui est devenu un problème pour les analystes qui, à la place d'être noyés sous les données, se trouvent noyés sous des motifs.

Le nouveau défi des méthodes de fouille de motifs est d'extraire moins de motifs mais plus pertinents. Il faut pour cela être en mesure d'intégrer des connaissances expertes au sein même du processus d'extraction de motifs. De telles connaissances ont été exprimées sous forme de contraintes portant, par exemple, sur la forme des motifs à extraire (*e.g.* contraintes de taille de motifs), sur la forme de leurs occurrences (*e.g.* contraintes de *max-gap* sur des séquences (Pei