Extraction des Top-k Motifs par Approximer-et-Pousser

Arnaud Soulet et Bruno Crémilleux

GREYC, CNRS - UMR 6072, Université de Caen
Campus Côte de Nacre
14032 Caen Cedex France
{Prenom.Nom}@info.unicaen.fr

Résumé. Cet article porte sur l'extraction de motifs sous contraintes *globales*. Contrairement aux contraintes usuelles comme celle de fréquence minimale, leur vérification est problématique car elle entraine de multiples comparaisons entre les motifs. Typiquement, la localisation des k motifs maximisant une mesure d'intérêt, i.e. satisfaisant la contrainte top-k, est difficile. Pourtant, cette contrainte globale se révèle très utile pour trouver les motifs les plus significatifs au regard d'un critère choisi par l'utilisateur. Dans cet article, nous proposons une méthode générale d'extraction de motifs sous contraintes globales, appelée Approximer-et-Pousser. Cette méthode peut être vue comme une méthode de relaxation d'une contrainte globale en une contrainte locale évolutive. Nous appliquons alors cette approche à l'extraction des top-k motifs selon une mesure d'intérêt. Les expérimentations montrent l'efficacité de l'approche Approximer-et-Pousser.

Mots clés : extraction de motifs, contraintes.

1 Introduction

L'extraction de motifs contraints est un champ significatif de l'Extraction de Connaissances dans les Bases de Données, notamment pour dériver des règles d'association. L'intérêt des motifs extraits est garanti par le point de vue de l'analyste exprimé à travers la sémantique de la contrainte. Par ailleurs, la complétude de l'extraction assure qu'aucun motif jugé pertinent par l'utilisateur ne sera manqué. La contrainte la plus populaire est certainement celle de fréquence minimale (Agrawal et al., 1993) qui permet de rechercher des régularités au sein d'une base de données. Malheureusement, le nombre de motifs fréquents est souvent prohibitif. Les motifs les plus pertinents sont alors noyés au milieu d'informations triviales ou redondantes que même d'autres contraintes d'agrégats (Ng et al., 1998) n'arrivent pas davantage à isoler.

Dans ces conditions, plusieurs approches proposent de comparer les motifs entre eux pour ne sélectionner que les meilleurs (Fu et al., 2000) ou une couverture (Mannila et Toivonen, 1997; Pasquier et al., 1999). De tels motifs révèlent alors une structure globale au sein des données. Le critère d'appartenance ou non à cette structure s'apparente à une contrainte *globale*. L'extraction de motifs satisfaisant une contrainte globale présente donc une finalité importante pour les utilisateurs. Cependant, leur extraction s'avère souvent ardue car leur localisation dans l'espace de recherche est loin d'être triviale. En particulier, trouver les k motifs maximisant