Découverte des soft-skypatterns avec une approche PPC

Willy Ugarte*, Patrice Boizumault*
Samir Loudni*, Bruno Crémilleux*, Alban Lepailleur**

* GREYC (CNRS UMR 6072)-Université de Caen Basse-Normandie
Campus Côte de Nacre, Bd du Maréchal Juin BP 5186 - 14032 Caen CEDEX - France
** CERMN (UPRES EA 4258 - FR CNRS 3038 INC3M)
Université de Caen Basse-Normandie - Bd Becquerel, 14032 CAEN Cedex -France
{prénom.nom}@unicaen.fr

Résumé. Les skypatterns sont des motifs traduisant des préférences de l'utilisateur selon une relation de dominance. Dans cet article, nous introduisons la notion de souplesse dans la problématique des skypatterns et nous montrons comment celle-ci permet de découvrir des motifs intéressants qui seraient manqués autrement. Nous proposons une méthode efficace d'extraction de skypatterns ainsi que de *soft*-skypatterns, méthode fondée sur la programmation par contraintes. La pertinence de notre approche est illustrée à travers une étude de cas en chémoinformatique pour la découverte de toxicophores.

1 Introduction

La découverte de motifs est une tâche centrale en fouille de données et est utilisée avec succès dans un grand nombre d'applications. Une limite bien connue des processus de fouille de données est la production d'un grand nombre de motifs qu'il n'est pas possible d'examiner manuellement et parmi lesquels l'information utile est diluée. L'extraction de motifs sous contraintes permet de cibler l'information recherchée selon les centres d'intérêt de l'utilisateur. Un prolongement récent de cette voie de recherche est la prise en compte de l'intérêt d'un motif en fonction des autres motifs extraits, afin de produire des ensembles de motifs qui satisfont des propriétés sur l'ensemble des motifs considérés conjointement (Raedt et Zimmermann, 2007; Khiari et al., 2010). Notre travail se situe dans cette lignée et porte sur la notion de requêtes *skylines* (Börzsönyi et al., 2001). Notre originalité est d'introduire la souplesse dans la relation de dominance caractérisant les skylines dans le contexte de la fouille de données et de montrer l'apport de la Programmation Par Contraintes (PPC) pour cela.

La notion de skylines a été récemment étendue à la fouille de données pour extraire des *motifs skylines* (appelés *skypatterns*) (Soulet et al., 2011). Les skypatterns traduisent les préférences d'un utilisateur selon une relation de *dominance*. Dans un espace multidimensionnel où chaque dimension définit une préférence, un point p_1 domine un autre point p_2 ssi p_1 est meilleur ou égal à p_2 sur toutes les dimensions, et est strictement meilleur sur au moins une dimension. Par exemple, un utilisateur peut préférer les motifs ayant une fréquence peu élevée, une petite taille et une confiance élevée. Dans ce cas, un motif p_1 domine un autre motif p_2 ssi : $freq(p_1) \leq freq(p_2) \wedge taille(p_1) \leq taille(p_2) \wedge confiance(p_1) \geq confiance(p_2)$, où au moins une