Une approche PPC pour la fouille de données séquentielles

Jean-Philippe Métivier*, Samir Loudni*, Thierry Charnois**

*GREYC (CNRS UMR 6072) – Université de Caen
Campus II Côte de Nacre, 14000 Caen - France
** LIPN (CNRS UMR 7030) – Université PARIS 13
99, avenue Jean-Baptiste Clément 93430 Villetaneuse - France

Résumé. Nous proposons dans cet article une nouvelle approche croisant des techniques de programmation par contraintes et de fouille pour l'extraction de motifs séquentiels. Le modèle que nous proposons offre un cadre générique et déclaratif pour modéliser et résoudre des contraintes de nature hétérogène.

1 Introduction

Introduite par (Agrawal et Srikant, 1995), la fouille de données séquentielles permet de découvrir des corrélations entre des événements selon une relation d'ordre (e.g. le temps). En intégrant des connaissances sous forme d'a priori dans le processus de fouille, l'extraction de motifs sous contraintes contribue à réduire le nombre de motifs en ciblant les motifs potentiellement intéressants (Dong et Pei, 2007). De plus, elle permet souvent de concevoir des algorithmes plus efficaces en réduisant l'espace de recherche. De nombreux algorithmes sont proposés dans la littérature pour l'extraction de motifs séquentiels (Dong et Pei, 2007). Malheureusement, ces méthodes ne traitent que quelques classes particulières de contraintes (monotonicité et anti-monotonicité) avec des techniques dédiées. Ce manque de généricité est un frein à la découverte de motifs pertinents car chaque nouveau type de contraintes entraîne la conception et le développement d'une méthode ad hoc.

Pour lever ce frein, des travaux récents visent à croiser les techniques de Programmation Par Contraintes (PPC) et de fouille pour l'extraction sous contraintes de motifs d'itemsets (Guns et al., 2011; Khiari et al., 2010). Le point commun de ces travaux est de modéliser le problème de la fouille de motifs en un problème de CSP. Une telle modélisation présente l'avantage d'être flexible en permettant de définir de nouvelles contraintes sans s'occuper de leur résolution. Mais, les méthodes proposées sont conçues pour des données ensemblistes et la dimension séquentielle reste quasiment non exploitée, à l'exception des travaux de (Coquery et al., 2012) qui portent sur un cas particulier de chaîne (et non sur une base de séquences).

L'originalité de notre travail consiste à proposer une première modélisation PPC de l'extraction de motifs séquentiels sous contraintes – à partir d'une base de séquences – dans un cadre déclaratif permettant de traiter simultanément des contraintes de nature quelconque. Les contraintes traitées dans le cadre de cet article incluent les contraintes de fréquence, clôture, taille, gap et celles portant sur les items (voir (Métivier et al., 2013) pour d'autres types de contraintes traitées et pour une présentation plus détaillée de ce travail). Des expériences me-