# Extraction de Motifs Séquentiels Multidimensionnels Clos sans Gestion d'Ensemble de Candidats

Marc Plantevit, Anne Laurent, Maguelonne Teisseire

LIRMM, Université Montpellier 2, CNRS,
prenom.nom@lirmm.fr

**Résumé.** L'extraction de motifs séquentiels permet de découvrir des corrélations entre événements au cours du temps. Introduisant plusieurs dimensions d'analyse, les motifs séquentiels multidimensionnels permettent de découvrir des motifs plus pertinents. Mais le nombre de motifs obtenus peut devenir très important. C'est pourquoi nous proposons, dans cet article, de définir une représentation condensée garantie sans perte d'information : les motifs séquentiels multidimensionnels clos extraits ici sans gestion d'ensemble de candidats.

## 1 Introduction

Les motifs séquentiels sont étudiés depuis plus de 10 ans (Agrawal et Srikant (1995)). Ils ont donné lieu à de nombreuses applications.Des algorithmes ont été proposés, basés sur le principe d'Apriori (Masseglia et al. (1998); Zaki (2001); Ayres et al. (2002)) ou sur d'autres propositions (Pei et al. (2004)). Récemment, les motifs séquentiels ont été étendus aux motifs séquentiels multidimensionnels par Pinto et al. (2001), Plantevit et al. (2005), et Yu et Chen (2005) dans l'objectif de prendre en compte plusieurs dimensions d'analyse. Par exemple, dans Plantevit et al. (2005), les règles telles que *Un client qui achète une planche de surf avec un sac à NY achète plus tard une combinaison à SF* sont découvertes. Toutefois, le nombre de motifs extraits dans une base de données peut être très important. C'est pourquoi des représentations condensées telles que les motifs *clos* ont été proposées pour l'extraction des itemsets (Pasquier et al. (1999); Pei et al. (2000); Zaki et Hsiao (2002); El-Hajj et Zaïane (2005)) et des séquences (Yan et al. (2003); Wang et Han (2004)). Les clos permettent de disposer à la fois d'une représentation condensée des connaissances extraites et d'un mécanisme d'extraction plus efficace afin d'élaguer significativement l'espace de recherche. Néanmoins, ces propositions ne peuvent pas être directement appliquées aux motifs séquentiels multidimensionnels pour la raison suivante : une super séquence peut être obtenue de deux façons (1) une plus longue séquence (plus d'items) ou (2) une séquence plus générale (plus de valeurs non spécifiées) ce qui modifie les définitions des méthodes précédemment introduites.

Notre contribution majeure est la définition d'un cadre théorique pour l'extraction de motifs séquentiels multidimensionnels clos ainsi qu'un algorithme permettant de rechercher de tels motifs. Nous adoptons une méthode basée sur le paradigme "pattern growth" (Pei et al. (2004)) afin de proposer une solution d'extraction de motifs séquentiels multidimensionnels clos efficace. De plus, nous souhaitons définir un algorithme qui se dispense de gérer un ensemble de clos candidats, seules les séquences closes étant ajoutées à l'ensembles des clos.