Extraction de Séquences Multidimensionnelles Convergentes et Divergentes

Marc Plantevit, Anne Laurent, Maguelonne Teisseire

LIRMM, Université Montpellier 2, CNRS, 161 Rue Ada 34392 Montpellier, France
prenom.nom@lirmm.fr, http://www.lirmm.fr

Résumé. Les motifs séquentiels sont un domaine de la fouille de données très étudié depuis leur introduction par Agrawal et Srikant. Même s'il existe de nombreux travaux (algorithmes, domaines d'application), peu d'entre eux se situent dans un contexte multidimensionnel avec la prise en compte de ses spécificités : plusieurs dimensions, relations hiérarchiques entre les éléments de chaque dimension, etc. Dans cet article, nous proposons une méthode originale pour extraire des connaissances multidimensionnelles définies sur plusieurs niveaux de hiérarchies mais selon un certain point de vue : du général au particulier ou vice et versa. Nous définissons ainsi le concept de séquences multidimensionnelles convergentes ou divergentes ainsi que l'algorithme associé, *M2S_CD*, basé sur le paradigme "pattern growth". Des expérimentations, sur des jeux de données synthétiques et réelles, montrent l'intérêt de notre approche aussi bien en terme de robustesse des algorithmes que de pertinence des motifs extraits.

1 Introduction

Les motifs séquentiels sont étudiés depuis plus de dix ans (Agrawal et Srikant (1995)), ils permettent de mettre en exergue des corrélations entre événements suivant leur chronologie d'apparition. Les motifs séquentiels ont été récemment étendus dans un contexte multidimensionnel par Pinto et al. (2001), Plantevit et al. (2005) et Yu et Chen (2005). Ils permettent ainsi de découvrir des motifs définis sur plusieurs dimensions et ordonnés par une relation d'ordre (*e.g.* temporelle). Par exemple, dans Plantevit et al. (2005), des motifs de la forme *"La plupart des consommateurs achètent une planche de surf et un sac à N.Y., puis ensuite une combinaison à SF"* sont découverts. Les motifs séquentiels multidimensionnels sont bien adaptés aux contextes de stockage et de gestion des données actuels (entrepôts de données). En effet, les motifs ou règles obtenus permettent une autre appréhension des données sources. Cependant leur découverte nécessite certains paramètres dont en particulier le support minimal. Celui-ci correspond à la fréquence minimale d'apparition des motifs au sein de la base considérée. Si le support minimal choisi est trop élevé, le nombre de règles découvertes est faible mais si le support est trop bas, le nombre de règles obtenues est très important et rend difficile l'analyse de celles-ci. Un autre problème est la longueur des motifs extraits. Comment ajuster au mieux le support afin d'obtenir des séquences suffisamment longues pour être réellement utilisables ? L'utilisateur est alors confronté au problème suivant : comment baisser le support minimal sans