

HYPE : Prise en compte des hiérarchies lors de l'extraction de motifs séquentiels multidimensionnels

Marc Plantevit, Anne Laurent, Maguelonne Teisseire

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

<http://www.lirmm.fr/~plantevi>

<http://www.lirmm.fr/laurent>

<http://www.lirmm.fr/teisseir>

Résumé. Les entrepôts de données contiennent de grosses masses de données historisées stockées à des fins d'analyse. Si les méthodes et outils d'analyse sont maintenant bien connus (OLAP), il reste difficile de fournir aux utilisateurs des outils de fouille de données permettant la prise en compte des spécificités de ces contextes (e.g. multidimensionnalité, hiérarchies, données historisées). Dans cet article, nous proposons une méthode originale d'extraction de motifs séquentiels prenant en compte les hiérarchies. Cette méthode extrait des connaissances plus précises et étend notre approche précédente M^2SP . Nous définissons les concepts liés à notre problématique ainsi que les algorithmes associés. Les expérimentations que nous avons menés montrent l'intérêt de notre proposition.

1 Introduction

La technologie OLAP et la fouille de données ne sont pas incompatibles (Han (1997)). Les techniques d'extraction de connaissances peuvent apporter une aide non négligeable dans le contexte OLAP où l'utilisateur doit désormais prendre les décisions les mieux adaptées en un minimum de temps. De façon plus précise, la fouille de données constitue une étape clef dans le processus de décision face à de gros volumes de données multidimensionnelles. En effet, les motifs ou règles obtenues 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. Le décideur est alors confronté au problème suivant : comment baisser le support minimal sans générer la découverte de règles non pertinentes ? Ou comment augmenter le support minimal sans perdre les règles utiles ? Est-il alors nécessaire de faire un compromis entre qualité des connaissances extraites et support ?

L'utilisation des hiérarchies dans l'extraction de connaissances représente un excellent moyen de résoudre ce dilemme. Elle permet de découvrir des règles au sein de plusieurs niveaux de hiérarchies. Ainsi, même si un support élevé est utilisé, les connaissances importantes