

Règles graduelles et cubes de données : quand les blocs s’empilent !

Lisa Di Jorio*, Yeow Wei Choong**
Anne Laurent * Maguelonne Teisseire***

*LIRMM – Université de Montpellier 2 – CNRS
161 rue Ada, 34392 Montpellier – FRANCE
{name}@lirmm.fr,
<http://www.lirmm.fr/tatoo>

**HELP University College – Kuala Lumpur – MALAYSIA
choongyw@help.edu.my

***Cemagref - UMR Tetis
maguelonne.teisseire@teledetection.fr

Résumé. Le couplage des méthodes de fouille de données et d’entrepôts de données permet d’extraire des informations pertinentes à partir de cubes de données. Dans ce contexte, de nombreuses approches ont été proposées, permettant par exemple d’extraire des règles d’association ou des motifs séquentiels. Cependant, il n’existe pas de méthodes permettant d’extraire des règles graduelles. Dans cet article nous nous intéressons donc à la découverte de telles règles corrélant des variations sur un ensemble de dimensions ordonnées avec des variations sur la mesure du cube. Nous découvrons par exemple des règles du type “*Plus la ville est de taille importante et la catégorie socio-professionnelle de catégorie supérieure, plus le nombre de produits vendus est grand*”. Afin de découvrir ces règles de manière efficace et en prenant en compte les grandes tendances issues des cubes de données, nous nous appuyons sur des travaux précédents permettant d’extraire des blocs de données homogènes.

1 Introduction

La fouille de cube de données, initiée par Chen et Pei (2006) qui a introduit l’OLAP Mining, consiste à définir des méthodes capables d’extraire des connaissances à partir de données multidimensionnelles, agrégées, et potentiellement organisées à différents niveaux de hiérarchies. Différente de la fouille de données classique en raison des spécificités propres à cette organisation des données, elle nécessite la définition de nouvelles méthodes permettant à la fois d’extraire des connaissances intéressantes et pertinentes, mais aussi de faire face à de gros volumes de données en raison de la taille sans cesse croissante des cubes de données disponibles et des besoins grandissants d’applications en quasi temps réel des utilisateurs.

Dans ce contexte, des travaux ont été proposés ces dernières années, notamment pour extraire des règles d’association, des résumés flous, ou encore des motifs séquentiels multidimensionnels. Il est alors possible d’extraire des règles du type *la plupart des ventes de canoës à Boston sont faibles*, ou *la plupart des ventes de l’est sont effectuées à Boston* ou encore *pour la plupart des catégories socio-professionnelles, on trouve des achats de canoes dans une ville de l’Est puis des achats de combinaisons et de housses à San Francisco*. Ces différentes règles, profitant de l’organisation des cubes de données, de la présence d’une ou plusieurs mesures et de leur multidimensionnalité et organisation multi-niveaux, permettent de renseigner le décideur sur les tendances présentes dans les cubes de données. Ces tendances