SALINES : un automate au service de l'extraction de motifs séquentiels multidimensionnels

Yoann PITARCH*, Lionel VINCESLAS**
Anne LAURENT*, Pascal PONCELET*, Jean-Emile SYMPHOR**

*LIRMM - Université Montpellier 2, CNRS
{pitarch,laurent,poncelet}@lirmm.fr
**CEREGMIA, Université des Antilles et de la Guyane, Martinique, France
{lionel.vinceslas,je.symphor}@martinique.univ-ag.fr

Résumé. Les entrepôts de données occupent aujourd'hui une place centrale dans le processus décisionnel. Outre leur consultation, une des finalités des entrepôts est de servir de socle aux techniques de fouilles de données. Malheureusement, les approches existantes exploitent peu les particularités des entrepôts (multidimensionnalité, hiérarchies et données historiques). Parmi ces méthodes, l'extraction de motifs séquentiels multidimensionnels a récemment été étudiée. Nous montrons dans cet article que ces dernières ne tirent pas pleinement profit des hiérarchies et ne découvrent par conséquent qu'une partie seulement des motifs qualitativement intéressants. Nous proposons alors une méthode d'extraction de motifs séquentiels multidimensionnels basée sur un automate et extrayant de nouveaux motifs. Les différentes expérimentations menées sur des jeux de données synthétiques attestent des bonnes performances de notre proposition.

1 Introduction [1]

Initiallement introduites pour faciliter le processus de prises de décisions, les bases de données multidimensionnelles (Codd et al. (1993)) sont de plus en plus utilisées comme support à la fouille de données Han (1998). Dans la mesure où les entrepôts de données stockent des donnés historisées, il apparaît pertinent d'y rechercher des corrélations temporelles. Les motifs séquentiels (Agrawal et Srikant (1995)) sont une des principales techniques utilisées dans cet objectif. Ces motifs, de la forme $\langle(pain, lait)(beurre)(pain, vin)\rangle$, ont été étudiés depuis une dizaine d'années et sont aujourd'hui appliqués dans de nombreux domaines (e.g., profilage de client, détection de fraudes). Cependant, très peu d'approches se sont intéressées à extraire de tels motifs dans un contexte multidimensionnel (Plantevit et al. (2006, 2008)). Par exemple, la découverte du motif $\langle\{(pain, epicerie)(lait, supermarche)\}\{(beurre, epicerie)\}\{(pain, supermarche)(vin, supermarche)\}\rangle$ signifierait que de nombreux clients ont acheté du pain à l'épicerie et dans la même période du lait au supermarché puis du pain au supermarché et enfin du vin au supermarché. Ici, considérer deux dimensions (le produit et le type de magasin) rend les motifs extraits plus intéressants car contenant plus d'informations. De même, il est également possible de prendre en compte les hiérarchies associées aux dimensions (Plantevit et al. (2006, 2008)). Dans HYPE (Plantevit et al. (2006)), il n'est pas possible d'obtenir des motifs où plusieurs niveaux de granularité d'une même dimension apparaissent. Par

1. Ce projet fait parti du projet ANR MIDAS (ANR-07-MDCO-008)