

Une structure basée sur les hiérarchies pour synthétiser les itemsets fréquents extraits dans des fenêtres temporelles

Yoann Pitarch, Anne Laurent
Pascal Poncelet

LIRMM - Université Montpellier 2, CNRS
161 rue Ada, 34392 Montpellier, France
nom.prénom@lirmm.fr,
<http://www.lirmm.fr/~nom>

Résumé. Le paradigme des flots de données rend impossible la conservation de l'intégralité de l'historique d'un flot qu'il faut alors résumer. L'extraction d'itemsets fréquents sur des fenêtres temporelles semble tout à fait adaptée mais l'amoncellement des résultats indépendants rend impossible l'exploitation de ces résultats. Nous proposons une structure basée sur les hiérarchies des données afin d'unifier ces résultats. De plus, puisque la plupart des données d'un flot présentent un caractère multidimensionnel, nous intégrons la prise en compte d'itemsets multidimensionnels. Enfin, nous pallions une faiblesse majeure des Tilted Time Windows (TTW) en prenant en compte la distribution des données.

1 Contexte

Un flot de données peut être défini comme une séquence potentiellement infinie de données précises, changeantes et rapides. Dès lors, il est impossible de lire ces données plus d'une fois Aggarwal (2007) tout comme il n'est pas réaliste de conserver l'intégralité de l'historique d'un flot. Proposer des techniques pour résumer au mieux des flots permet alors de guider le décideur dans son analyse.

Parmi les approches de résumé existantes, l'extraction d'itemsets fréquents exhibe les relations fréquentes entre items. Cette méthode est largement utilisée dans un contexte statique pour découvrir les comportements fréquents d'une population. De plus, cette technique a l'avantage de s'accorder avec une théorie actuelle sur le fonctionnement de la mémoire humaine. En effet, en psychologie cognitive, une théorie connue sous le nom de *mémoire constructive* affirme que les souvenirs sont reliés entre eux et ne sont pas la simple accumulation d'événements indépendants (Schacter et Addis (2007)). Par exemple, l'évocation d'une personne familière entraîne une multitude de souvenirs associés tels que sa dernière rencontre, son physique, ... L'extraction d'itemsets fréquents sur des fenêtres temporelles d'un flot permet de mesurer l'évolution des tendances d'un flot mais souffre d'un inconvénient de taille : il faut considérer manuellement les différents résultats d'extraction. Cette tâche devient rapidement impossible à réaliser.

De plus, la plupart des flots de données présentent un caractère multidimensionnel. Ces dimensions peuvent être considérées à différents niveaux de granularité tels que *la ville, le*