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'unifiant 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 fenêtre temporeles 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 résutats 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*

département ou *la région* pour la dimension *lieu d'achat*. Il semble alors intéressant d'exploiter ces spécificités pour guider la synthèse des itemsets fréquents.

À notre connaissance, aucune approche n'aborde cette problématique. Cependant, deux approches de résumé de flots de données multidimensionnelles exploitent les hiérarchies associées aux données pour construire un cube de données alimenté par un flot (Han et al. (2005); Pitarch et al. (2008). Aucune d'entre elles ne remet en cause l'utilisation des TTW qui peut pourtant entraîner un important manque de précision. En effet, l'agrégation à intervalle fixe des données rend imprécis le résumé de données dont la distribution est changeante. Par exemple, si un élément apparaît rarement dans le flot, conserver ses apparitions exactes est certainement utile. Avec une TTW, cette précision est automatiquement perdue dès la première agrégation.

2 Apperçu de l'approche

Nous exploitons l'aspect multidimensionnel et multiniveaux des données et proposons une structure pour synthétiser en temps réel les résultats indépendants d'extraction d'itemsets fréquents dans des fenêtres temporelles. Nous facilitons ainsi la tâche d'un utilisateur en lui offrant la possibilité d'observer (sans synthèse manuelle préalable) l'évolution des itemsets fréquents d'un flot sur différents niveaux de granularité. De plus, nous pallions un défaut majeur des TTW en proposant une structure de liste dynamique pour stocker les apparitions des itemsets dans le flot. Une agrégation aura lieu uniquement si un itemset est fréquent dans plusieurs fenêtres proches. À l'inverse, les apparitions des itemsets rarement ou périodiquement fréquents seront conservées pendant une période de temps significative.

Références

Aggarwal, C. C. (2007). *Data Streams : Models and Algorithms*. Advances in Database Systems.

Han, J., Y. Chen, G. Dong, J. Pei, B. W. Wah, J. Wang, et Y. D. Cai (2005). Stream cube : An architecture for multi-dimensional analysis of data streams. *Distrib. Parallel Databases 18*(2), 173–197.

Pitarch, Y., A. Laurent, M. Plantevit, et P. Poncelet (2008). Fenêtres sur cube. In *Bases de Données Avancées*, pp. 1–20.

Schacter, D. L. et D. R. Addis (2007). The cognitive neuroscience of constructive memory : remembering the past and imagining the future. *Philos Trans R Soc Lond B Biol Sci 362*(1481), 773–786.

Summary

In the data stream context, storing the whole data stream history is unfeasible and providing a high-quality summary is required for decision makers. A practical and consistent summarization method is the extraction of the frequent itemsets over temporal windows. Nevertheless, this method suffers from a critical drawback: results pile up quickly making the analysis either uncomfortable or impossible for users. We propose to unify these results thanks to a synthesis method for multidimensional frequent itemsets based on a graph structure and taking advantage of the data hierarchies. We overcome a major drawback of the Tilted Time Window (TTW) standard framework by taking into account the data distribution.