

Fouille de flots de données multidimensionnelles

A. Laurent, P. Poncelet et M. Plantevit

laurent@lirmm.fr pascal.poncelet@ema.fr marc.plantevit@lirmm.fr

Même si les supports de stockage (e.g. disques durs) ont connu un développement considérable, de plus en plus de données apparaissent sous la forme d'un flot (data stream) de manière continue, à un rythme rapide et éventuellement de manière infinie et il n'est alors plus possible de les stocker dans leur intégralité (e.g. trafic TCP/IP, transactions financières en ligne, enregistrements médicaux, capteurs, ...). L'interrogation ou l'extraction de connaissances à partir de tels flots a récemment donné lieu à de nombreux travaux de recherche [Aggarwal 2007]. L'un des problèmes auquel la communauté fouille de données se trouve confrontée aujourd'hui est la recherche d'un équilibre entre l'efficacité (il ne faut pas bloquer les données du flot) et la précision des résultats (il faut tenir compte de ce qui s'est passé avant dans le flot). En d'autres termes, il devient acceptable d'obtenir des réponses avec des approximations.

Ces dernières années de nombreuses structures de résumé ont été proposées et elles possèdent les mêmes propriétés d'applicabilité, d'efficacité mémoire, et de robustesse [Muthukrishnan 2005, Raïssi et al 2007]. Dans le cadre de ce stage, nous souhaitons étudier les méthodes permettant de traiter ces flots de données multidimensionnelles en les agrégeant sous la forme de cubes de flots de données.

De manière plus précise, il s'agira d'étendre l'approche proposée par [Han et al. 2005] utilisant les 'tilted time' windows pour le maintien des cubes de flots de données. Alors que la majorité des approches existantes se contentent de décrire l'état d'un cube à un instant t , nous proposons de décrire à l'utilisateur l'évolution du cube au cours du temps. Ainsi, en ne gommant pas l'historique, nous nous retrouvons dans un contexte plus proche des paradigmes OLAP et des buts de la fouille de données historisées pour la détection des tendances et exceptions.

Ce travail entre dans le cadre du projet d'ANR MIDAS en collaboration étroite LIRMM/EMA-LGI2P et en collaboration avec l'ENST Paris et EDF (fournisseur de données de flots de consommations électriques).

Références bibliographiques :

- [Aggarwal 2007] C. Aggarwal. « Data Streams : Models and Algorithms ». Springer Verlag, 2007.
- [Chen et al 2002]. Yixin Chen , Guozhu Dong , Jiawei Han , Jian Pei , Benjamin W. Wah , Jianyong Wang. « Online Analytical Processing Stream Data: Is It Feasible? ». ACM DMKD, 2002.
- [Han et al. 2005] J. Han, Y. Chen, G. Dong, J. Pei, B. Wah, J. Wang, Y. Cai. « Stream Cube: An Architecture for Multidimensional analysis of Data Streams », Distributed and Parallel Databases 2005, 18. 173-187, 2005.
- [Muthukrishnan 2005] S. Muthukrishnan, « Data streams: algorithms and applications », In Foundations and Trends in Theoretical Computer Science, Volume 1, Issue 2, August 2005.
- [Plantevit et al. 2005] M. Plantevit, Y.W Choong, A. Laurent, D. Laurent and M. Teisseire. « M²SP: Mining Sequential Patterns among Several Dimensions ». In Proceedings of the International Conference on Principles and Practice of Knowledge Discovery in Databases Conference (PKDD 2005), Porto, Portugal, September 2005.
- [Raïssi et al. 2007] C. Raïssi and P. Poncelet. « Sampling for Sequential Pattern Mining: From Static Databases to Data Streams ». Proceedings of the IEEE International Conference on Data Mining (ICDM 07), Omaha NB, USA, October 2007.