

Fouille Distribuée de Data Streams

C. Raïssi et P. Poncelet

raïssi@lirmm.fr, pascal.poncelet@ema.fr

Le développement des nouvelles technologies permet actuellement de générer de très grands volumes de données issues de différentes sources : trafic TCP/IP, transactions financières, enregistrements médicaux, capteurs. Les données apparaissent alors sous la forme d'un flot (*data streams*) de manière continue, à un rythme rapide et éventuellement de manière infinie. L'extraction de connaissances à partir de tels flots a récemment donné lieu à de nombreux travaux de recherche qui se sont focalisés sur la découverte de motifs fréquents (e.g. [Aggarwal 2007], [Manku et al 2002], [Giannella et al. 2003], [Raïssi et al. 2007]) en utilisant des méthodes telles que le *landmark*, la *fenêtre glissante* ou les *modèles de pondérations temporelles*.

Cependant, de plus en plus d'applications nécessitent une gestion de type distribuée de plusieurs flots (e.g. détection d'intrusion dans des réseaux, capteurs météorologiques, gestion de réseaux télécoms, ...). Dans ce contexte, l'extraction de motifs devient plus complexe car elle suppose de gérer non seulement l'extraction au niveau local (e.g. le capteur) mais également de pouvoir agréger les résultats obtenus dans un modèle global représentatif du système. Récemment, de nouveaux travaux de recherche ont été proposés par [Manjhi et al. 2005] pour extraire des items fréquents en étendant l'approche de [Manku et al. 2002].

L'objectif de ce stage est de proposer une nouvelle approche de recherche de motifs distribués prenant en compte les différentes contraintes telles que le temps et le volume de communications entre les capteurs, l'impossibilité de stocker les données, les différentes approximations nécessaires et la prise en compte de pannes.

Ce travail entre dans le cadre du projet d'ANR MIDAS en collaboration étroite LIRMM/EMALGI2P et en collaboration avec l'ENST Paris et France Télécom (fournisseur de données de flots de communications).

Références bibliographiques

- [Aggarwal 2007] C. Aggarwal. « Data Streams: Models and Algorithms ». Springer Verlag, 2007.
- [Giannella et al. 2003] G. Giannella, J. Han, J. Pei, X. Yan, and P. Yu (2003). Mining frequent patterns in data streams at multiple time granularities. In *Next Generation Data Mining*, MIT Press.
- [Manjhi et al. 2005] A. Manjhi, V. Shkpenyuk, K. Dhamdhere and C. Olston. « Finding (recently) Frequent Items in Distributed Data Streams ». In Proceedings of the International Conference on Data Engineering (ICDE 05), 2005.
- [Manku et al 2002] G. Manku and et R. Motwani (2002). « Approximate frequency counts over data streams ». In *Proceedings of the VLDB 02 Conference*, pp. 346–357.
- [Raïssi et al. 2007] C. Raïssi and P. Poncelet. « Sampling for Sequential Pattern Mining: From Static Databases to Data Streams ». In Proceedings of the IEEE International Conference on Data Mining (ICDM 07), Omaha NB, USA, October 2007.