Sujet de master recherche « Architectures logicielles distribuées »
2005–2006

Construction et maintenance de synopsis pour l'interrogation approchée de flux de données

Encadrant principal : Noureddine Mouaddib
courriel : Noureddine.Mouaddib@univ-nantes.fr
tél. : 02 40 68 32 02

Co-encadrant(s) : Guillaume Raschia

Objectif du stage

Ces dernières années ont émergés de nouvelles applications pour lesquelles les données sont modélisées, non plus à l'aide de relations persistantes, mais par des flux volatiles. Les applications visées couvrent différents domaines tels que l'analyse financière, la surveillance des réseaux, la sécurité, la gestion de données de télécommunication, les applications web, la manufacture, etc. Dans le modèle de flux de données [1], les objets peuvent être des éléments de relations (n-uplets), cependant leur arrivée continue en flux multiples, rapides, variables, imprévisibles et potentiellement infinis semble soulever un certain nombre de problèmes de recherche en bases de données aussi nouveaux que fondamentaux. Par exemple, il a été établi que l'approximation [2] et l'adaptation [3] sont deux critères essentiels à l'exécution de requêtes et à l'analyse de flux rapides de données, tandis que les SGBD traditionnels se concentrent largement sur des objectifs opposés, à savoir l'obtention de réponses exactes calculées par des plans d'exécution stables. L'émergence de DSMSs (Data Stream Management Systems) tels que STREAM de l'université de Stanford [4] vient renforcer s'il en était besoin, l'intérêt que porte la communauté des bases de données à ces nouveaux enjeux.

Une des solutions retenues pour l'interrogation approchée de flux de données consiste à construire et maintenir en ligne des synopsis [5], i.e. une information synthétique sur les données. Différentes approches ont été proposées, fondées essentiellement sur des modèles statistiques à base d'histogrammes [6] et autres quantiles [7], ou d'ondelettes [8] . Elles permettent de répondre de façon approchée à des requêtes de sélection et d'agrégat, mais sont par essence (statistique) dépourvues de capacité descriptive.

Travail à réaliser

En s'appuyant sur les travaux développés au sein de l'équipe Atlas-GRIM, d'une part sur la construction en ligne de résumés [9], et d'autre part sur l'interrogation approchée [10], nous souhaitons étudier une approche inédite de la construction de synopsis permettant de répondre à de nouvelles familles de requêtes sur des flux de données.

Références

[1] Brian Babcock, Shivnath Babu, Mayur Datar, Rajeev Motwani, and Jennifer Widom. Models and issues in data stream systems. In *Proceedings of 21st ACM Symposium on Principles of Database Systems (PODS 2002)*, pages 1–16, 2002.

[2] Ron Avnur and Joseph M. Hellerstein. Eddies : continuously adaptive query processing. In *Proc. of the ACM Intl. Conf. on Management of Data (SIGMOD 2000)*, pages 261–272, 2000.

[3] Daniel Barbará et al. The new jersey data reduction report. *IEEE Data Eng. Bull.*, 20(4) :3–45, 1997.

[4] The STREAM Group. STREAM : The stanford stream data manager. *IEEE Data Engineering Bulletin*, 26(1), March 2003.

[5] Phillip B. Gibbons and Yossi Matias. Synopsis data structures for massive data sets. *DIMACS : Series in Discrete Mathematics and Theoretical Computer Science : Special Issue on External Memory Algorithms and Visualization*, A, 1999.

[6] Sudipto Guha, Nick Koudas, and Kyuseok Shim. Data-streams and histograms. In *ACM Symposium on Theory of Computing*, pages 471–475, 2001.

[7] A. Gilbert, Y. Kotidis, S. Muthukrishnan, and M. Strauss. How to summarize the universe : Dynamic maintenance of quantiles. In *Int. Conf on Very Large Databases (VLDB 2002)*, pages 454–465, 2002.

[8] Anna C. Gilbert, Yannis Kotidis, S. Muthukrishnan, and Martin Strauss. Surfing wavelets on streams : One-pass summaries for approximate aggregate queries. In *The VLDB Journal*, pages 79–88, 2001.

[9] R. Saint-Paul, G. Raschia, and N. Mouaddib. General purpose database summarization. In *Int. Conf. on Very Large Databases (VLDB 2005)*, pages 733–744, Trondheim, Norway, 2005. Morgan Kaufmann Publishers.

[10] A. Voglozin, G. Raschia, L. Ughetto, and M. Mouaddib. Querying the saintetiq summaries – dealing with null answers. In *IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE 2005)*, Reno, Nevada, USA, 2005.