Mining Frequent Queries for Big Data

Keywords: distributed data mining, mapreduce; hadoop.

Abstract

The problem of mining frequent queries in a relational database defined over a star schema is not easy even when we deal with only one table, because, on the one hand, the size of the search space is huge (because encompassing all possible queries that can be addressed to a given database), and on the other hand, testing whether two queries are equivalent (which entails redundant support computations) is NP-Complete. Therefore, the problem is even more di cult when they are applied to Big Data.

In previous work [5, 6, 4, 8, 7, 3], we defined a pre-ordering ($q \leq q'$) on star schema's queries that satisfies the following basic properties:

1. The support measure is anti-monotonic with respect to $\leq$, and
2. Defining $q \equiv q'$ if and only if $q \leq q'$ and $q' \leq q$

all equivalent queries have the same support. The anti-monotonicity of pre- ordering defined on star schema's queries allow to apply level-wise algorithm like Apriori [1, 2] and we de ne an algorithm named FQF (Frequent Query Finder) to find Frequent Queries in database with small size defined over a Star Schema. However, FQF become problematic when they are applied to huge database (BigData). Fortunately, recent improvements in the field of distributed datamining provide good tools to resolve these issues.

The main goal of this thesis is, on the one hand, to investigate the applicability of FQF on the MapReduce platform, and on the other hand, to propose, implement and test an algorithm based on the MapReduce Platform to mine frequent queries in relational database with huge size defined over a star schema.

References

1. Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant, Hannu Toivonen, and A. Inkeri Verkamo. Fast discovery of association rules. In Advances in Knowledge Discovery and Data Mining, pages 307 328. AAAI/MIT Press, 1996.
2. Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile, pages 487 499, 1994.

3. Cheikh Tidiane Dieng. Etude et implantation de l'extraction de requêtes fréquentes dans les bases de données multidimensionnelles. (Mining pattern in relational databases). PhD thesis, Cergy-Pontoise University, France, 2011.
4. Cheikh Tidiane Dieng, Tao-Yuan Jen, and Dominique Laurent. An effcient computation of frequent queries in a star schema. In Database and Expert Systems Applications, 21th International Conference, DEXA 2010, Bilbao, Spain, August 30 - September 3, 2010, Proceedings, Part II, pages 225 239, 2010.
5. Cheikh Tidiane Dieng, Tao-Yuan Jen, Dominique Laurent, and Nicolas Spyratos. Mining frequent conjunctive queries using functional and inclusion dependencies. VLDB J., 22(2):125 150, 2013.
6. Bart Goethals, Dominique Laurent, Wim Le Page, and Cheikh Tidiane Dieng. Mining frequent conjunctive queries in relational databases through dependency discovery. Knowl. Inf. Syst., 33(3):655 684, 2012.
7. Tao-Yuan Jen, Dominique Laurent, Nicolas Spyratos, and Oumar Sy. Mining frequent queries in star schemes. In Extraction et gestion des connaissances (EGC'2005), Actes des cinquièmes journées Extraction et Gestion des Connaissances, Paris, France, 18-21 janvier 2005, 2 Volumes, pages 331 342, 2005
8. Tao-Yuan Jen, Dominique Laurent, Nicolas Spyratos, and Oumar Sy. Towards mining frequent queries in star schemes. In Knowledge Discovery in Inductive Databases, 4th International Workshop, KDID 2005, Porto, Portugal, October 3, 2005, Revised Selected and Invited Papers, pages 104 123, 2005.