Frequent Queries Computation in Arbitrary Relational Database

Keywords: database, datamining, query, pattern, level-wise algorithm, knowledge discovery.

Abstract

The problem of mining frequent queries in a database has motivated many research e orts during the last two decades. This is so because many interesting patterns, such as association rules, exact or approximative functional dependencies and exact or approximative conditional functional dependencies can be easily retrieved, which is not possible using standard techniques. However, the problem mining frequent queries in a relational database is not easy even when we deal with one table, on the one hand, because 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. In [5, 6, 4], an algorithm for mining Projection-Selection-Join (PSJ) queries has been pro- posed for database defined over a star schema. In this thesis, we focus on Projection-Selection-Join (PSJ) queries, assuming that the database is defined over an arbitrary relational database not only on star schema. In this setting, one goal of this thesis is to de ne a pre-ordering ($q \leq q'$) 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. This two properties will allow us to use algorithm like Apriori [1, 2]. The main contributions of this thesis will be, on the one hand to formally study properties of the pre-ordering and the equivalence relation mentioned above, and on the other hand, to propose a level-wise, Apriori like algorithm for the computation of all frequent queries in a relational database defined over an arbitrary schema. Moreover, the implementation of an algorithm for mining frequent queries in any relational 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 efficient 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 Di- eng. 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. To- wards 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.