香港浸會大學理學院
HKBU Faculty of Science

for favour of posting

DEPARTMENT OF COMPUTER SCIENCE

PhD Degree Oral Presentation

PhD Candidate:	Mr Yun PENG
Supervisor:	Dr Koon Kau CHOI
External Examiner:	Prof Wilfred NG Prof Wei WANG
Time:	23 July 2013 (Tuesday) 10:30 am – 12:30 pm (35 mins presentation and 15 mins Q & A)
Venue:	SCT716, Cha Chi Ming Science Tower, HSH Campus

"Estimation Techniques for Advanced Database Application"

Abstract

Database systems have been ubiquitously used as fundamental facilities to manage a large amount of data efficiently. Nowadays, the data that needs to be managed by database systems is growing explosively. For example, each day Twitter produces more than 300 million tweets; each second Facebook has 7.9 new users and its user number has exceeded one billion. How to efficiently manage and analyze data of such scale is a crucial task of database systems. Among many other solutions, estimation techniques have been proven successful to address these problems. In this thesis, we study the applications of estimation techniques in four important database problems: selectivity estimation on graphs, outsourced subgraph similarity search, optimal graph index prediction and the classical view update problem.

Firstly, we study the selectivity estimation of twig queries on cyclic graphs. Similar to relational database, selectivity estimation plays a crucial role in query optimization of graph database. However, determining the optimal query evaluation plan (QEP) for a graph query is more challenging than its relational counterpart. This is because of the complexity of graph structure analysis. For example, given a twig query against a cyclic graph, computing the cost of a QEP may need to enumerate all the matchings between the subtrees of the twig query and the graph. This is potentially costly as the graph may contain an exponential number of matchings. Many works have been proposed to study the selectivity estimation. However, most of them focus on either the cyclic graphs or the twig queries but not both. In the first part of this thesis, we propose a novel histogram-based method to support selectivity estimation of twig queries on cyclic graphs.

Secondly, we study the estimation in outsourced subgraph similarity search. Subgraph similarity search itself is *estimation-like*, which returns the graphs that have *approximate* substructures with the query graph. This query has been used in a wide range of applications including bioinformatics, chem-informatics, Web topology, etc. Recently, due to the complexity of subgraph similarity search and the explosive growth of graph data, the data owner of graph database is more appealing to outsource their data to third-party service providers, which will process the query on behalf of the data owner. However, the service provider may not be trustable and therefore it is required to return an authentication structure to the user for authenticating the correctness of query results. Since the manipulation of the authentication structure takes the major overhead of outsourced subgraph similarity search, in the second part of this thesis, we propose an estimation-based method to optimize its processing.

Thirdly, estimation is also crucial in the optimal graph index prediction. Recently, an ample body of graph indexes have been proposed to optimize query processing on graphs. However, the performances of such indexes may vary greatly as verified from our experiments with a large number of random and scale-free graphs. In particular, our preliminary experiments show that the runtime of 1,000 random queries on an index, even on the same graph, can often exhibit large variances. Specifically, the mean and standard deviation of `2-hop labeling` are 14.1 seconds and 4.2 and those of `prime labeling` are 11.6 seconds and 59.7, respectively. Moreover, the runtime is often skewed and has a long tail at large values. Therefore, it is desired to predict the optimal index on a graph. However, designing an exact performance model is a daunting task since the structures of graph indexes are often complex and ad-hoc. In the third part of this thesis, we apply statistical distributions to estimate the query performance. Then, the classical data mining techniques are applied to predict the optimal index.

Finally, we study the application of estimation techniques in the classical view update problem, where the updates specified by the user on the view need to be translated to the updates on the source database, such that the new view derived from the updated source database is consistent with the user’s expectation. The state-of-the-art view update analysis methods often involve two interleaving stages: side-effect determination and update translation. It is well-known that the translation problem is NP-complete. Therefore, it is desirable to develop a method that can efficiently estimate the side effects, such that the view updates having side effects can be filtered before they are passed to the costly translation. In this forth part of this thesis, we develop a data-oriented side-effect estimation technique to support such view update analysis.

The works proposed in this thesis verify that the estimation techniques are useful in various major components in database systems.

Keywords: Selectivity estimation, Subgraph similarity search, Graph index prediction,
Side-effect estimation, Graph database, Outsourced database, Relational database

*** ALL INTERESTED ARE WELCOME ***