A Cost-Based Range Estimation for Mapping Top-k Selection Queries over Relational Databases

Anteneh Ayanso, Brock University, Canada
Paulo B. Goes, University of Arizona, USA
Kumar Mehta, George Mason University, USA

ABSTRACT
Finding efficient methods for supporting top-k relational queries has received significant attention in academic research. One of the approaches in the recent literature is query-mapping, in which top-k queries are mapped (translated) into equivalent range queries that relational database systems (RDBMSs) normally support. This approach combines the advantage of simplicity as well as practicality by avoiding the need for modifications to the query engine, or specialized data structures or indexing techniques to handle top-k queries separately. However, existing methods following this approach fall short of adequately modeling the problem environment and providing consistent results. In this article, the authors propose a cost-based range estimation model for the query-mapping approach. They provide a methodology for trading-off relevant query execution cost components and mapping a top-k query into a cost-optimal range query for efficient execution. Their experiments on real world and synthetic data sets show that the proposed strategy not only avoids the need to calibrate workloads on specific database contents, but also performs at least as well as prior methods.

Keywords: Cost Model, Query-Mapping, Query-Processing, Relational Databases, Top-K Query, Tradeoff Analysis, Uncertainty Modeling

INTRODUCTION
Relational databases are increasingly being used to support a wide range of interactive applications that require efficient methods for exploratory search and retrieval (for example, search for airline tickets, hotel rooms, real estates, used cars). In such applications, users specify target values for certain attributes without necessarily requiring exact matches to these values in return. However, relational queries normally establish rigid qualification to deal only with data that exactly match selection conditions (Motro, 1988). Due to the exactness in nature of relational databases and the query...
languages, users face the challenge of routinely specifying value ranges of attributes in search of approximate results, particularly in the event of empty answers or too many answers (Huh & Lee, 2001; Shin, Huh, Park & Lee, 2008). Top-k querying aims to address this problem, and is defined in the recent literature as a method of specifying relational queries by target values of attributes in order to obtain a desired number of best matches based on some ranking (distance) functions (Bruno, Chaudhuri & Gravano, 2002; Chaudhuri & Gravano, 1999). The objective is to retrieve a ranked list of the \( k \) records that are the closest to the query specifications.

Given the enormous installed base of relational database management systems (RDBMSs) in many emerging applications, finding efficient top-k retrieval methods that are easily implementable within these environments is of great importance. Our work in this article is particularly motivated by the mechanism of mapping or translating a top-k query into an equivalent range query that is normally supported by any RDBMS (Bruno et al., 2002; Chaudhuri & Gravano, 1999; Donjerkovic & Ramakrishnan, 1999). The rationale in the query-mapping approach is to exploit existing features of RDBMSs (for example, query language, summary statistics, and available index utilities) and provide a specific method to translate a top-k query into a conventional range query for efficient execution. This approach not only eliminates the need for additional specialized data structures and separate execution mechanisms for top-k queries, but also maintains the transparency in the operations of the RDBMS (Ayanso, Goes & Mehta, 2007; Bruno et al., 2002; Chaudhuri & Gravano, 1999). Nevertheless, existing top-k query-mapping methods are either simple heuristics that lead to significant performance variation by database settings (Chaudhuri & Gravano, 1999) or workload-adaptive that requires learning the data using queries with pre-specified data characteristics (Bruno et al., 2002; Chen & Ling, 2002).

In this article, we present a cost-based, query level model for mapping a top-k query into a cost-optimal range query. To this end, we explicitly formulate the relevant cost components and provide a methodology that allows evaluation of cost tradeoffs for range estimation. The top-k retrieval problem and the query-mapping solution involve many related issues and conflicting efficiency objectives. The use of tradeoff analysis not only captures the complexity of the problem better, but also helps to identify and incorporate the major factors that influence database operations and the associated costs in top-k querying. Our query level model is an improvement over the workload-based approach, which involves calibration of parameters using pre-specified workloads. In the workload-based approach, queries that deviate in characteristics or distribution from pre-specified workloads suffer from performance inefficiency. We adopt an approach that accounts for cost and performance implications of queries individually and show the relative consistency in efficiency under different experimental settings.

In the rest of the article, we review the literature on top-k querying and describe the problem and the main dimensions of the query-mapping approach. Then we present our model, along with detailed descriptions of the model components and assumptions. Following that, we describe the experimental setting used to assess the performance efficiency of our cost-based strategy and discuss the experimental results as well as the theoretical and practical implications. Finally, we provide concluding remarks and discuss limitations and future research issues.

## RELATED LITERATURE

Efficient querying for approximate matches has been extensively studied for document and multi-media retrieval systems (Baeeza-Yates & Ribeiro-Neto, 1999; Bertino, Rabbitti & Gibbs, 1988; Bustos, Keim, Saupe, Schreck & Vranić, 2005; Chan, 2006; Chaudhuri, Gravano & Marian, 2004; Kwok & Zhao, 2006; Salton & McGill, 1983). In contrast, top-k retrieval in