

## Guest editorial: special issue on ranking in databases

Ihab Ilyas

Published online: 8 July 2009  
© Springer Science+Business Media, LLC 2009

In recent years, there has been a great deal of interest in developing effective techniques for ad-hoc search and retrieval in databases systems. In particular, a large number of emerging applications require exploratory querying on general-purpose or domain-specific databases; examples include users wishing to search bibliographic databases or catalogs of products such as homes, cars, cameras, restaurants, and photographs.

Current database query languages such as SQL follow the Boolean retrieval model, i.e., tuples or elements that exactly satisfy the selection conditions laid out in the query are returned. While extremely useful for the expert user, this retrieval model is inadequate for ad-hoc retrieval by exploratory users who cannot articulate the perfect query for their needs; either their queries are very specific, resulting in no (or too few) answers, or are very broad, resulting in too many answers.

To address the limitations of the Boolean retrieval model, top-k queries and ranking query results are gaining increasing importance. In fact, in many applications, ranking is an integral part of the semantics, e.g., keyword search, similarity search in multimedia as well as document databases. The increasing importance of ranking is directly derived from the explosion in the volume of data handled by current applications, which makes it almost impossible to process queries in the traditional compute-then-sort approach.

This special issue of the Distributed and Parallel Databases solicited contributions that address novel and important challenges in supporting ranking in database systems. We received 17 submissions, out of which 4 original papers were accepted to the special issue.

---

I. Ilyas (✉)

School of Computer Science, University of Waterloo, 200 University Ave, West Waterloo, Ontario N2L3G1, Canada

e-mail: [ilyas@uwaterloo.ca](mailto:ilyas@uwaterloo.ca)

*Rank aggregation in distributed environments* The paper titled “Distributed Top-k Aggregation Queries at Large” focuses on optimizing top-k query processing algorithms in distributed environments. The proposed optimization focuses on forming top-k operators’ trees by hierarchical grouping of input sources, and on estimating the scan depths of inputs. The solution is based on statistical cost model that uses sampling to create local synopsis, convolutions and order statistics.

*Continuous monitoring in data streams* The paper titled “Continuously Monitoring Top-K Uncertain Data Streams: A Probabilistic Threshold Method” formulates a novel and challenging problem of continuously monitoring top-k uncertain data streams. The paper proposes a probabilistic threshold approach and introduces multiple efficient query evaluation algorithms to solve the problem.

*Semantics of ranking in probabilistic databases* The paper titled “Semantics and Evaluation of Top-k Queries in Probabilistic Databases” addresses fundamental issues involved in top-k query evaluation in probabilistic databases. The authors formulate three intuitive postulates that the semantics of top-k queries in probabilistic databases should satisfy, and introduce a new semantics, Global-Topk, that satisfies those postulates to a large degree. The paper also shows how to evaluate queries under the Global-Topk semantics for both simple probabilistic databases in which probabilities are associated with individual tuples, and general probabilistic databases in which, additionally, exclusivity relationships between tuples can be represented.

*Skyline computation in Web applications* The paper titled “Ranking Strategies and Threats: A Cost-based Pareto Optimization Approach” addresses an interesting problem in skyline query processing; the goal of the paper is to identify the cost-minimal Pareto-optimal improvement strategy for non-skyline objects to be part of the skyline results. The paper abstracts this problem as a mixed integer-programming problem and develops a novel algorithm for efficiently identifying the optimal solution.