Report on the First International Workshop on Ranking in Databases (DBRank'07)

Ihab F. Ilyas
University of Waterloo
Waterloo, Ontario, Canada
ilyas@cs.uwaterloo.ca

Gautam Das
University of Texas at Arlington
Arlington, Texas, USA
gdas@cse.uta.edu

ABSTRACT

This report summarizes the presentations, keynotes and discussions that took place during the first international workshop on ranking in databases (DBRank'07). The workshop was held on April 16, 2007, in conjunction with ICDE in Istanbul, Turkey.

1. INTRODUCTION

The International Workshop on Ranking in Databases (DBRank) focuses on the semantics, the modeling and the implementation of ranking and ordering in database systems and applications. In recent years, there has been a great deal of interest in developing effective techniques for ad-hoc search and retrieval in a variety of domains such as relational databases, document and multimedia databases, and scientific information systems. In particular, a large number of emerging applications require exploratory querying on such databases; examples include users wishing to search databases and catalogs of products such as homes, cars, cameras, restaurants, and photographs. Traditional database query languages such as SQL follow the Boolean retrieval model, i.e., tuples that exactly satisfy the selection conditions laid out in the query are returned–no more and no less. 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 in these emerging ad-hoc search and retrieval applications, Top-k queries and ranking query results are gaining increasing importance. In fact, in many of these 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. The sheer amount of data makes it almost impossible to process queries in the traditional compute-then-sort approach. Hence, ranking comes as a great tool for soliciting user preferences and data exploration.

Ranking imposes several challenges for almost all data-centric systems. In relational databases, large body of work has been recently proposed to support ranking as a first class construct through rank-aware algebra, ranking operators and new optimization frameworks that integrate ranking in plan enumeration and costing. There has been exciting recent work on automatic learning of appropriate ranking functions for database applications (e.g., based on adaptions of IR ranking functions to leverage dependency information in structured data), on designing expressive languages for user preferences modeling, on adaptation of keyword querying paradigms to relational databases, as well as on exciting new developments in new Top-k algorithms for relational, documents and multimedia databases. Ranking query results in semi-structured and XML databases has been also the focus of many recent contributions.

DBRank'07 solicited full and short papers that describe current research and work-in-progress efforts in enabling ranking in database systems.

2. LOGISTICS

The program committee of DBRank'07 consisted of 19 expert members from academia and industry. We had 23 submissions, each received at least two reviews. 5 full (8 pages) papers and 5 short (4 pages) papers were accepted by the program committee. The submission quality was very high and we wished we could squeeze more papers in the one-day program. However, we decided to accept only 10 papers to leave enough time for each paper to be properly presented and discussed. Each full paper was given a 30 minutes slot, while short papers were assigned 15 minutes slots.

In addition to paper presentations, we had keynotes by two distinguished scholars, Dr. Surajit Chaudhuri (Microsoft Research) and Prof. Gerhard Weikum (Max-Planck Institute for Informatics). For additional information, please refer to the URL of the workshop: `http://www.cs.uwaterloo.ca/dbrank2007`.

We were really happy with the discussions triggered by the keynotes and by the paper presentations. Based on a head-count, we had around 30 attendees throughout the day. In the following sections of this report, we briefly comment on the keynotes and the technical contributions presented in DBRank'07.

3. TECHNICAL CONTRIBUTIONS

Multiple aspects of ranking were discussed in 10 presentations. We roughly categorize the papers as follows:

Preference Specification Kenneth Ross from Columbia University presented two papers on preference specification, formalism and evaluation based on partial orders. The first short paper [5] identified several anomalies in the behavior of conventional notions of composition for preferences defined by strict partial orders. Kenneth showed how these anomalies can be avoided by defining a preorder that extends the given partial order, and by using the pair of orders to define order composition. The presentation included multiple examples to show the unintuitive results of composing strict partial orders using several composition methods, e.g., prioritized and Pareto composition. The second full paper [6] described a study of constraint formalisms for expressing user preferences as base facts in a partial order. The paper proposes a language that allows comparison and a limited form of arithmetic. The paper also shows that the transitive closure computation is required to complete the partial order terminates. Preference query processing was also briefly addressed in this paper, where index structures were presented to allow efficient evaluation over large data sets.

Scoring and Ranking Functions Scoring is a fundamental challenge in supporting ranking of database objects. While several rank aggregation and preference handling algorithms have been proposed, all these techniques depend on some sort of scores or a scoring function to be provided by the application or by the data generating process. Three papers focused on providing such scoring mechanism in different contexts.

Aparna Varde of Virginia State University presented a full paper [9] on learning the relative importance of similarity features in the context of image retrieval. The paper proposes a method called `FeaturesRank` for learning a distance function between the query image and images stored in the database. A training sample with pairs of images is used and the extent of similarity is identified for each pair. `FeaturesRank` clusters the given images in levels. It then adjusts the distance function based on the error between the clusters and training samples using multiple heuristics. `FeaturesRank` was evaluated with real image data from nanotechnology and bioinformatics.

Gultekin Ozsoyoglu of Case Western Reserve University presented a full paper [4] on comparing the quality of scoring functions in the context of searching literature in digital libraries. The paper discusses three different functions that assign scores to papers based on their context. The extensive experimental study compares the quality of these functions based on accuracy (precision) and separability (uniformity of output scores), and shows that the text-based and the pattern-based scores yield better accuracy and separability than the citation-based scores.

Arthur Van Bunningen of the University of Twente presented a full paper [1] that proposes a novel explanatory approach of looking at context-aware relevance by defining the context-aware relevance of features as a probabilistic function of past choices. Context-aware preference is introduced as a way to capture the changes in user needs and preferences with respect to the search context. The paper shows that this approach goes well together with traditional probabilistic information retrieval and uncertainty of context information.

Ranking in XML Ranking is a natural way to explore and query large volumes of XML documents. In contrast to keyword search in text documents or ranking query results in relational database systems, the characteristics of XML data impose unique combination of ranking based on value and structural similarity. Two prototypes for enabling ranking in XML database were discussed in DBRank'07: the `ArHeX` and the `TReX` systems.

Ismael Sanz of the Universitat Jaume I, Spain presented a short paper [7] describing the features of the `ArHeX` similarity-oriented XML processing toolkit. `ArHeX` is designed to assist in the engineering of XML similarity-oriented applications, and to support the design and evaluation of suitable similarity measures and their associated indexes for each specific application.

Mariano Consens from the University of Toronto presented a full paper [2] that addresses retrieval queries that combine structural constraints with keyword search in XML database. The paper describes the `TReX` system an XML retrieval system that can exploit multiple structural summaries (including newly proposed ones) . `TReX` can also self-manage small, redundant indexes to speed up the evaluation of workloads of top-k queries. The redundant indexes are maintained to enable `TReX` to select among different evaluation strategies.

Skyline Queries Skyline queries are natural ways to express ranking requirements in the absence of a concrete scoring function that aggregates the scores of multiple ranking criteria. Marcel Karnstedt of the Technische Universitt Ilmenau, Germany presented a short paper [3] that proposes three variants of a skyline operator and two extensions, especially suitable for efficient determination of skylines in structured overlays peer-to-peer environments.

Ranking in Other Domains Two short papers discussed the application of ranking in domains other than traditional relational and XML database systems. Jiawei Han of the University of Illinois at Urbana-Champaign presented a short paper [10] that addresses efficient evaluation of ranking queries in OLAP environments, introducing the *ranking cube*: a semi off-line materialization and semi-online computation model for answering top-k queries. On the other hand, Sharma Chakravarthy of the University of Texas at Arlington presented a short paper [8] that highlights multiple ranking issues in the context of information integration.

4. KEYNOTES

We were fortunate to have two outstanding keynote talks. Prof. Gerhard Weikum of Max-Planck Institute for Informatics reviewed the advantages and disadvantages of *TA* (the Threshold Algorithm) and its many extensions, putting them in perspective against algorithmic alternatives and pointing out unsolved technical issues and research opportunities. The family of Threshold Algorithms has become a very popular method for top-k query processing and ranked retrieval of unstructured, semi-structured, and structured data. *TA* has many elegant properties, such as instance optimality, and is extremely versatile. However, Gerhard's talk demonstrated that it also has specific limitations and is competing with alternative methods for top-k queries. The talk gave a nice overview and a timeline for the evolution of ranking algorithms in the last decade highlighting their strengths and weaknesses, and described recent work on optimizations and variations of *TA* in tackling several technical challenges. Examples of discussed challenges are choosing the best order for rank aggregation in nested top-k queries and relaxing ranking criteria.

The second keynote address was given by Dr. Surajit Chaudhuri of Microsoft Research who discussed various aspects of enabling general search over relational databases. The talk reviewed semantics and efficiency issues in supporting keyword search and ranking over databases, and critically examined past and current research. The talk highlighted important issues that require more attention in future research, e.g., (a) role of applications/business objects (b) architectural considerations - separation of functionality in database server vs. middleware. This interesting talk by Surajit intrigued the audience to engage in an active discussion debating the applicability of recent research contributions to real-world database engines. This unique industry perspective of Surajit raised many flags which will help shaping the emerging area of ranking in database systems.

5. FUTURE OF DBRANK

After receiving many unsolicited positive comments from the workshop attendees, we were encouraged to try making DBRank a yearly event. DBRank'08 will be held in conjunction with ICDE 2008 in Cancun, Mexico on the 11^{th} and 12^{th} of April. DBRank'08 is co-chaired by Vagelis Hristidis (Florida International University) and Ihab F. Ilyas (University of Waterloo). We are looking forward to having another interesting and successful round of DBRank.

6. ACKNOWLEDGMENTS

We would like to express our sincere gratitude to our distinguished program committee members who provided timely, high quality and constructive reviews to all the papers. We were really impressed by their enthusiasm and dedication to make DBRank'07 a great success. This workshop would not have been successful without the huge support by our two keynote speakers who gave exciting talks full of research challenges and future research directions to this emerging community. Last but not least, we would like to thank all the authors who submitted their self-selected high quality papers to DBRank'07.

7. PAPERS PRESENTED IN DBRANK'07

[1] Arthur Van Bunningen, Maarten Fokkinga, Peter Apers, and Ling Feng. Ranking query results using context-aware preferences.

[2] Mariano Consens, Xin Gu, Yaron Kanza, and Flavio Rizzolo. Self managing top-k (summary, keyword) indexes in xml retrieval.

[3] Marcel Karnstedt, Jessica Muller, and Kai-Uwe Sattler. Cost-aware skyline queries in structured overlays.

[4] Nattakarn Ratprasartporn, Sulieman Bani-Ahmad, Ali Cakmak, Jonathan Po, and Gultekin Ozsoyoglu. Evaluating different ranking functions for context-based literature search.

[5] Kenneth Ross. On the adequacy of partial orders for preference composition.

[6] Kenneth Ross, Peter Stuckey, and Amelie Marian. Practical preference relations for large data sets.

[7] Ismael Sanz, Rafael Berlanga, Marco Mesiti, and Giovanna Guerrini. Flexible composition of indexes and similarity measures in xml.

[8] Aditya Telang, Roochi Mishra, and Sharma Chakravarthy. Ranking issues for information integration.

[9] Aparna Varde, Elke Rundensteiner, Giti Javidi, Ehsan Sheybani, and Jianyu Liang. Learning the relative importance of features in image data.

[10] Dong Xin and Jiawei Han. Integrating olap and ranking: The ranking-cube methodology.