Advances in Database Systems

Ming Hua
Jian Pei

Ranking Queries on Uncertain Data

Springer

Ranking Queries on Uncertain Data

ADVANCES IN DATABASE SYSTEMS

Volume 42

Series Editors

Ahmed K. Elmagarmid
Purdue University
West Lafayette, IN 47907

Amit P. Sheth
Wright State University
Dayton, OH 45435

For other titles published in this series, please visit www.springer.com/series/5573

Ming Hua • Jian Pei

Ranking Queries on Uncertain Data

Springer

Dr. Ming Hua
Facebook Inc.
S. California Avenue 1601
94304 Palo Alto California
USA
arceehua@fb.com

Dr. Jian Pei
Simon Fraser University
School of Computing Science
University Drive 8888
V5A 1S6 Burnaby British
Columbia
Canada
jpei@cs.sfu.ca

ISSN 1386-2944
ISBN 978-1-4419-9379-3 e-ISBN 978-1-4419-9380-9
DOI 10.1007/978-1-4419-9380-9
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011924205

To my parents {M.H.}
To my wife Jennifer and my daughter Jacqueline for your love and encouragement {J.P.}

Preface

"Maturity of mind is the capacity to endure uncertainty."
— John Finley (1935 – 2006)

"Information is the resolution of uncertainty."
— Claude Elwood Shannon (1916 – 2001)

Uncertain data is inherent in many important applications, such as environmental surveillance, market analysis, and quantitative economics research. Due to the importance of those applications and rapidly increasing amounts of uncertain data collected and accumulated, analyzing large collections of uncertain data has become an important task. Ranking queries (also known as top-k queries) are often natural and useful in analyzing uncertain data.

In this monograph, we study the problem of ranking queries on uncertain data. Specifically, we extend the basic uncertain data model in three directions, including uncertain data streams, probabilistic linkages, and probabilistic graphs, to meet various application needs. Moreover, we develop a series of novel ranking queries on uncertain data at different granularity levels, including selecting the most typical instances within an uncertain object, ranking instances and objects among a set of uncertain objects, and ranking the aggregate sets of uncertain objects.

To tackle the challenges on efficiency and scalability, we develop efficient and scalable query evaluation algorithms for the proposed ranking queries. First, we integrate statistical principles and scalable computational techniques to compute exact query results. Second, we develop efficient randomized algorithms to approximate the answers to ranking queries. Third, we propose efficient approximation methods based on the distribution characteristics of query results. A comprehensive empirical study using real and synthetic data sets verifies the effectiveness of the proposed ranking queries and the efficiency of our query evaluation methods.

This monograph can be a reference for academic researchers, graduate students, scientists, and engineers interested in techniques of uncertain data management and analysis, as well as ranking queries. Although the monograph focuses on ranking queries on uncertain data, it does introduce some general principles and models of

uncertain data management. Thus, the monograph can also serve as introductory reading approaching the general field of uncertain data management.

Uncertain data processing, management, and exploration in general remain an interesting and fast developing topic in the field of database systems and data analytics. Moreover, ranking queries on uncertain data as a specific topic keeps seeing new progress in both research and engineering development. We believe uncertain data management in general and ranking queries on uncertain data in specific are exciting directions, and still have a huge space for further research. This monograph can inspire some exciting opportunities.

This monograph records the major outcomes of Ming Hua's Ph.D. research at School of Computing Science, Simon Fraser University. This research was an interesting and rewarding journey. We started with several interesting problems concerning effective and efficient queries of massive data, where uncertainty, probability, and typicality play critical roles. It turned out that ranking queries provide a simple and nice way to bind those projects and ideas together. Moreover, we considered various application scenarios, including online analytic style exploration, continuously monitoring of streaming data, data integration, and road network analysis. Only after almost three years we realized that the whole bunch of work can be linked together and weave a nice picture under the theme of ranking queries on uncertain data, as presented in this monograph.

A Ph.D. thesis is never easy. Ming Hua's Ph.D. study is exciting and, at the same time, challenging, for both herself and Jian Pei, the senior supervisor (that is, the thesis advisor). We both remember the sleepless nights before the submission deadlines, the frustration when our submissions were rejected, the suffering moments before some new ideas came to our mind, and the excitement when we obtained breakthroughs afterall. This experience has been casted deeply in our memory forever. We are so lucky to be able to work as a team in those four years.

Acknowledgement

This research would not be possible without the great help and support from many people.

Ming Hua thanks Jian Pei, her senior supervisor and mentor, for his continuous guidance and support during her Ph.D. study at Simon Fraser University. Her gratitude also goes to Funda Ergun, Martin Ester, Lise Getoor, Wei Wang, and Shouzhi Zhang for their insightful comments and suggestions on her research along the way.

Jian Pei thanks Ming Hua for taking the challenge to be his first Ph.D. student at Simon Fraser University. He is deeply grateful to his students at Simon Fraser University. He is always proud of working with those talented students. He is also deeply indebted to his colleagues at Simon Fraser University.

Many ideas in this monograph were resulted from collaboration and discussion with Xuemin Lin, Ada Fu, Ho-fung Leung, and many others. We want to thank our

collaborators in the past who have fun together in solving all kinds of data related puzzles. Our gratitude also goes to all anonymous reviewers of our submissions for their invaluable feedback, no matter positive or negative.

We thank Susan Lagerstrom-Fife and Jennifer Maurer at Springer who contributed a lot to the production of this monograph.

This research is supported in part by an NSERC Discovery Grant, an NSERC Discovery Accelerator Supplements Grant, a Simon Fraser University President Research Grant, and a Simon Fraser University Community Trust Endowment Fund. All opinions, findings, conclusions and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

Palo Alto, CA, USA, and Coquitlam, BC, Canada *Ming Hua*
January 2011 *Jian Pei*

Contents

Chapter 1
Introduction

Uncertain data is inherent in several important applications such as sensor network management [1] and data integration [2], due to factors such as data randomness and incompleteness, limitations of measuring equipment and delayed data updates. Because of the importance of those applications and the rapidly increasing amount of uncertain data, analyzing large collections of uncertain data has become an important task.

Ranking queries (also known as top-k queries) [3, 4, 5, 6] are a class of important queries in data analysis. Although ranking queries have been studied extensively in the database research community, uncertainty in data poses unique challenges on the semantics and processing of ranking queries. Traditional queries and evaluation methods on certain data cannot be directly adopted to uncertain data processing. Therefore, practically meaningful ranking queries as well as efficient and scalable query evaluation methods are highly desirable for effective uncertain data analysis.

1.1 Motivation

Recently, there have been an increasing number of studies on uncertain data management and processing [7, 8, 9, 10, 11, 12, 13, 14, 15, 16]. For example, the probabilistic database model [7, 17, 8] and the uncertain object model [18, 19, 20, 21] are developed to describe the uncertainty in data. More details about those models can be found in Chapter 3. In some important application scenarios, various ranking queries can provide intersecting insights into uncertain data.

Example 1.1 (Ranking queries in traffic monitoring applications). Roadside sensors are often used to measure traffic volumes, measure vehicle speeds, or classify vehicles. However, data collected from sensors as such cannot be accurate all the time due to the limitations of equipment and delay or loss in data transfer. Therefore, confidence values are often assigned to such data, based on the specific sensor characteristics, the predicted value, and the physical limitations of the system [22].

M. Hua, J. Pei, *Ranking Queries on Uncertain Data,* ADVANCES IN DATABASE SYSTEMS 42,
DOI 10.1007/978-1-4419-9380-9_1,