


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

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

© Springer Science+Business Media, LLC 2011

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

Springer is part of Springer Science+Business Media ([www.springer.com](http://www.springer.com))

*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 afterward. 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  
January 2011

*Ming Hua*  
*Jian Pei*

# Contents

<b>1</b>	<b>Introduction</b> .....	1
1.1	Motivation .....	1
1.2	Challenges .....	4
1.3	Focus of the Book .....	5
1.4	Organization of the Book .....	6
<b>2</b>	<b>Probabilistic Ranking Queries on Uncertain Data</b> .....	9
2.1	Basic Uncertain Data Models .....	9
2.1.1	Uncertain Object Model .....	9
2.1.2	Probabilistic Database Model .....	11
2.1.3	Converting Between the Uncertain Object Model and the Probabilistic Database Model .....	12
2.2	Basic Ranking Queries on Uncertain Data .....	12
2.2.1	Ranking Instances in An Uncertain Object .....	13
2.2.2	Ranking Uncertain Instances in Multiple Uncertain Objects .	19
2.2.3	Ranking Uncertain Objects .....	22
2.3	Extended Uncertain Data Models and Ranking Queries .....	22
2.3.1	Uncertain Data Stream Model .....	22
2.3.2	Probabilistic Linkage Model .....	25
2.3.3	Uncertain Road Network .....	27
2.4	Summary .....	31
<b>3</b>	<b>Related Work</b> .....	33
3.1	Uncertain Data Processing .....	33
3.1.1	Uncertain Data Models and Systems .....	33
3.1.2	Probabilistic Queries on Uncertain Data .....	34
3.1.3	Indexing Uncertain Data .....	35
3.2	Ranking (Top- $k$ ) Queries .....	35
3.2.1	Distributed Top- $k$ Query Processing .....	36
3.3	Top- $k$ Typicality Queries .....	36
3.3.1	Typicality in Psychology and Cognitive Science .....	36

3.3.2	The (Discrete) $k$ -Median Problem	37
3.3.3	Clustering Analysis	38
3.3.4	Other Related Models	39
3.4	Probabilistic Ranking Queries	40
3.4.1	Top- $k$ Queries on Uncertain Data	40
3.4.2	Poisson Approximation	42
3.5	Uncertain Streams	43
3.5.1	Continuous Queries on Probabilistic Streams	43
3.5.2	Continuous Ranking and Quantile Queries on Data Streams	44
3.5.3	Continuous Sensor Stream Monitoring	45
3.6	Probabilistic Linkage Queries	46
3.6.1	Record Linkage	46
3.6.2	Probabilistic Graphical Models	47
3.7	Probabilistic Path Queries	48
3.7.1	Path Queries on Probabilistic Graphs	48
3.7.2	Path Queries on Certain Traffic Networks	49
<b>4</b>	<b>Top-<math>k</math> Typicality Queries on Uncertain Data</b>	<b>51</b>
4.1	Answering Simple Typicality Queries	52
4.1.1	Likelihood Computation	52
4.1.2	An Exact Algorithm and Complexity	53
4.1.3	A Randomized Tournament Algorithm	54
4.2	Local Typicality Approximation	56
4.2.1	Locality of Typicality Approximation	56
4.2.2	DLTA: Direct Local Typicality Approximation Using VP-trees	59
4.2.3	LT3: Local Typicality Approximation Using Tournaments	61
4.3	Answering Discriminative Typicality Queries	65
4.3.1	A Randomized Tournament Algorithm	65
4.3.2	Local Typicality Approximation	66
4.4	Answering Representative Typicality Queries	69
4.4.1	An Exact Algorithm and Complexity	69
4.4.2	A Randomized Tournament Method	70
4.4.3	Local Typicality Approximation Methods	70
4.5	Empirical Evaluation	74
4.5.1	Typicality Queries on Real Data Sets	75
4.5.2	Approximation Quality	80
4.5.3	Sensitivity to Parameters and Noise	85
4.5.4	Efficiency and Scalability	86
4.6	Summary	87
<b>5</b>	<b>Probabilistic Ranking Queries on Uncertain Data</b>	<b>89</b>
5.1	Top- $k$ Probability Computation	90
5.1.1	The Dominant Set Property	90
5.1.2	The Basic Case: Independent Tuples	91

- 5.1.3 Handling Generation Rules . . . . . 92
- 5.2 Exact Query Answering Methods . . . . . 94
  - 5.2.1 Query Answering Framework . . . . . 94
  - 5.2.2 Scan Reduction by Prefix Sharing . . . . . 95
  - 5.2.3 Pruning Techniques . . . . . 100
- 5.3 A Sampling Method . . . . . 102
- 5.4 A Poisson Approximation Based Method . . . . . 104
  - 5.4.1 Distribution of Top- $k$  Probabilities . . . . . 104
  - 5.4.2 A General Stopping Condition . . . . . 105
  - 5.4.3 A Poisson Approximation Based Method . . . . . 106
- 5.5 Online Query Answering . . . . . 107
  - 5.5.1 The *PRist* Index . . . . . 107
  - 5.5.2 Query Evaluation based on *PRist* . . . . . 110
  - 5.5.3 *PRist+* and a Fast Construction Algorithm . . . . . 115
- 5.6 Experimental Results . . . . . 117
  - 5.6.1 Results on IIP Iceberg Database . . . . . 117
  - 5.6.2 Results on Synthetic Data Sets . . . . . 120
- 5.7 Summary . . . . . 127
- 6 Continuous Ranking Queries on Uncertain Streams . . . . . 129**
  - 6.1 Exact Algorithms . . . . . 129
    - 6.1.1 Top- $k$  Probabilities in a Sliding Window . . . . . 130
    - 6.1.2 Sharing between Sliding Windows . . . . . 133
  - 6.2 A Sampling Method . . . . . 138
  - 6.3 Space Efficient Methods . . . . . 138
    - 6.3.1 Top- $k$  Probabilities and Quantiles . . . . . 139
    - 6.3.2 Approximate Quantile Summaries . . . . . 142
    - 6.3.3 Space Efficient Algorithms using Quantiles . . . . . 144
  - 6.4 Experimental Results . . . . . 145
    - 6.4.1 Results on Real Data Sets . . . . . 145
    - 6.4.2 Synthetic Data Sets . . . . . 146
    - 6.4.3 Efficiency and Approximation Quality . . . . . 147
    - 6.4.4 Scalability . . . . . 150
  - 6.5 Summary . . . . . 150
- 7 Ranking Queries on Probabilistic Linkages . . . . . 151**
  - 7.1 Review: the Probabilistic Linkage Model . . . . . 151
  - 7.2 Linkage Compatibility . . . . . 152
    - 7.2.1 Dependencies among Linkages . . . . . 152
    - 7.2.2 Probabilistic Mutual Exclusion Graphs . . . . . 153
    - 7.2.3 Compatibility of Linkages . . . . . 155
    - 7.2.4 Resolving Incompatibility . . . . . 157
    - 7.2.5 Deriving All Possible Worlds . . . . . 158
  - 7.3 Ranking Queries on Probabilistic Linkages . . . . . 160
    - 7.3.1 Predicate Processing . . . . . 161

7.3.2	Dominant Subgraphs	162
7.3.3	Vertex Compression	163
7.3.4	Subgraph Probabilities	164
7.4	Tree Recurrence: Subgraph Probability Calculation	165
7.4.1	A Chain of Cliques	165
7.4.2	A Tree of Cliques	167
7.4.3	Integrating Multiple Components	168
7.5	Exact Query Answering Algorithms	169
7.5.1	An Exact Algorithm	169
7.5.2	Reusing Intermediate Results	169
7.5.3	Pruning Techniques	171
7.6	Extensions to Aggregate Queries	172
7.6.1	Aggregate Queries on Probabilistic Linkages	172
7.6.2	Count, Sum and Average Queries	173
7.6.3	Min and Max Queries	177
7.7	Empirical Evaluation	178
7.7.1	Results on Real Data Sets	178
7.7.2	Results on Synthetic Data Sets	182
7.8	Summary	184
<b>8</b>	<b>Probabilistic Path Queries on Road Networks</b>	<b>185</b>
8.1	Probability Calculation	186
8.1.1	Exact $l$ -Weight Probability Calculation	186
8.1.2	Approximating $l$ -Weight Probabilities	188
8.1.3	Estimating $l$ -Weight Probabilities	191
8.1.4	A Depth First Search Method	191
8.2	P*: A Best First Search Method	192
8.2.1	The P* Algorithm	193
8.2.2	Heuristic Estimates	194
8.3	A Hierarchical Index for P*	199
8.3.1	HP-Tree Index	199
8.3.2	Approximating Min-Value Estimates	200
8.3.3	Approximating Stochastic Estimates	200
8.4	Experimental Results	201
8.4.1	Simulation Setup	202
8.4.2	Efficiency and Memory Usage	204
8.4.3	Approximation Quality and Scalability	205
8.5	Summary	206
<b>9</b>	<b>Conclusions</b>	<b>207</b>
9.1	Summary of the Book	207
9.2	Future Directions: Possible Direct Extensions	209
9.2.1	Top- $k$ typicality queries for uncertain object	210
9.2.2	Top- $k$ queries on probabilistic databases	210
9.2.3	Probabilistic ranking queries	210

- 9.2.4 Probabilistic path queries on road networks ..... 212
- 9.3 Future Directions: Enriching Data Types and Queries ..... 213
  - 9.3.1 Handling Complex Uncertain Data and Data Correlations .. 213
  - 9.3.2 Answering More Types of Ranking and Preference  
Queries on Uncertain Data ..... 213
- References ..... 215

# 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].*