

Managing and Mining Uncertain Data

ADVANCES IN DATABASE SYSTEMS

Volume 35

Series Editors

Ahmed K. Elmagarmid

*Purdue University
West Lafayette, IN 47907*

Amit P. Sheth

*Wright State University
Dayton, Ohio 45435*

Other books in the Series:

PRIVACY-PRESERVING DATA MINING: *Models and Algorithms* by Charu C. Aggarwal and Philip S. Yu; ISBN: 978-0-387-70991-8

SEQUENCE DATA MINING by Guozhu Dong, Jian Pei; ISBN: 978-0-387-69936-3

DATA STREAMS: *Models and Algorithms*, edited by Charu C. Aggarwal; ISBN: 978-0-387-28759-1

SIMILARITY SEARCH: *The Metric Space Approach*, P. Zezula, G. Amato, V. Dohnal, M. Batko; ISBN: 0-387-29146-6

STREAM DATA MANAGEMENT, Nauman Chaudhry, Kevin Shaw, Mahdi Abdelguerfi; ISBN: 0-387-24393-3

FUZZY DATABASE MODELING WITH XML, Zongmin Ma; ISBN: 0-387-24248-1

MINING SEQUENTIAL PATTERNS FROM LARGE DATA SETS, Wei Wang and Jiong Yang; ISBN: 0-387-24246-5

ADVANCED SIGNATURE INDEXING FOR MULTIMEDIA AND WEB APPLICATIONS, Yannis Manolopoulos, Alexandros Nanopoulos, Eleni Tousidou; ISBN: 1-4020-7425-5

ADVANCES IN DIGITAL GOVERNMENT: *Technology, Human Factors, and Policy*, edited by William J. McIver, Jr. and Ahmed K. Elmagarmid; ISBN: 1-4020-7067-5

INFORMATION AND DATABASE QUALITY, Mario Piattini, Coral Calero and Marcela Genero; ISBN: 0-7923-7599-8

DATA QUALITY, Richard Y. Wang, Mostapha Ziad, Yang W. Lee; ISBN: 0-7923-7215-8

THE FRACTAL STRUCTURE OF DATA REFERENCE: *Applications to the Memory Hierarchy*, Bruce McNutt; ISBN: 0-7923-7945-4

SEMANTIC MODELS FOR MULTIMEDIA DATABASE SEARCHING AND BROWSING, Shu-Ching Chen, R.L. Kashyap, and Arif Ghafoor; ISBN: 0-7923-7888-1

INFORMATION BROKERING ACROSS HETEROGENEOUS DIGITAL DATA: *A Metadata-based Approach*, Vipul Kashyap, Amit Sheth; ISBN: 0-7923-7883-0

Managing and Mining Uncertain Data

Edited by

Charu C. Aggarwal
IBM T.J. Watson Research Center
USA



Springer

Editor

Charu C. Aggarwal
IBM Thomas J. Watson Research Center
19 Skyline Drive
Hawthorne, NY 10532
charu@us.ibm.com

Series Editors

Ahmed K. Elmagarmid
Purdue University
West Lafayette, IN 47907

Amit P. Sheth
Wright State University
Dayton, Ohio 45435

ISBN 978-0-387-09689-6

e-ISBN 978-0-387-09690-2

DOI 10.1007/978-0-387-09690-2

Library of Congress Control Number: 2008939360

© Springer Science+Business Media, LLC 2009

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.com

Preface

Uncertain data management has seen a revival in interest in recent years because of a number of new fields which utilize this kind of data. For example, in fields such as privacy-preserving data mining, additional errors may be added to data in order to mask the identity of the records. Often the data may be imputed using statistical methods such as forecasting. In such cases, the data is uncertain in nature. Such data sets may often be probabilistic in nature. In other cases, databases may show *existential uncertainty* in which one or more records may be present or absent from the data set. Such data sets lead to a number of unique challenges in processing and managing the underlying data.

The field of uncertain data management has been studied in the traditional database literature, but the field has seen a revival in recent years because of new ways of collecting data. The field of uncertain data management presents a number of challenges in terms of collecting, modeling, representing, querying, indexing and mining the data. We further note that many of these issues are inter-related and cannot easily be addressed independently. While many of these issues have been addressed in recent research, the research in this area is often quite varied in its scope. For example, even the underlying assumptions of uncertainty are different across different papers. It is often difficult for researchers and students to find a single place containing a coherent discussion on the topic.

This book is designed to provide a coherent treatment of the topic of uncertain data management by providing surveys of the key topics in this field. The book is structured as an edited volume containing surveys by prominent researchers in the field. The choice of chapters is carefully designed, so that the overall content of the uncertain data management and mining field is covered reasonably well. Each chapter contains the key research content on a particular topic, along with possible research directions. This includes a broad overview of the topic, the different models and systems for uncertain data, discussions on database issues for managing uncertain data, and mining issues with uncertain data. Two of the most prominent systems for uncertain data have also been described in the book in order to provide an idea how real uncertain data management systems might work. The idea is to structurally organize the topic,

and provide insights which are not easily available otherwise. It is hoped that this structural organization and survey approach will be a great help to students, researchers, and practitioners in the field of uncertain data management and mining.

Contents

Preface	v
List of Figures	xv
List of Tables	xxi
1	
An Introduction to Uncertain Data Algorithms and Applications	1
<i>Charu C. Aggarwal</i>	
1. Introduction	1
2. Algorithms for Uncertain Data	3
3. Conclusions	6
References	7
2	
Models for Incomplete and Probabilistic Information	9
<i>Todd J. Green</i>	
1. Introduction	9
2. Incomplete Information and Representation Systems	13
3. \mathcal{RA} -Completeness and Finite Completeness	14
4. Closure Under Relational Operations	18
5. Algebraic Completion	19
6. Probabilistic Databases and Representation Systems	21
7. Probabilistic $\mathcal{?}$ -Tables and Probabilistic Or-Set Tables	22
8. Probabilistic c -tables	24
9. Queries on Annotated Relations	25
10. K -Relations	27
11. Polynomials for Provenance	30
12. Query Containment	33
13. Related Work	34
14. Conclusion and Further Work	34
References	37
15. Appendix	41
3	
Relational Models and Algebra for Uncertain Data	45
<i>Sumit Sarkar and Debabrata Dey</i>	
1. Introduction	45

2.	Different Probabilistic Data Models	48
2.1	Point-Valued Probability Measures Assigned to Each Tuple	48
2.2	Point-Valued Probability Measures Assigned to Attributes and Attribute Sets	52
2.3	Interval-Valued Probability Measures Assigned to Attribute-Values	54
2.4	Interval-valued Probability Measures Assigned to Tuples	56
3.	Probabilistic Relational Algebra	56
3.1	Basic Definitions	56
3.2	Primary and Foreign Keys	58
3.3	Relational Operations	59
3.4	Relational Algebra	62
3.5	Incomplete Distribution and Null Values	63
4.	Algebraic Implications of the Different Representations and Associated Assumptions	67
4.1	Point Valued Probability Measures in Tuples	68
4.2	Point-Valued Probability Measures in Attributes	69
4.3	Interval-Valued Probability Measures in Attributes	70
4.4	Interval-Valued Probability Measures in Tuples	71
4.5	Observations on Tuple Independence	72
5.	Concluding Remarks	72
	References	75
4		
	Graphical Models for Uncertain Data	77
	<i>Amol Deshpande, Lise Getoor and Prithviraj Sen</i>	
1.	Introduction	78
2.	Graphical Models: Overview	80
2.1	Directed Graphical Models: Bayesian Networks	82
2.2	Undirected Graphical Models: Markov Networks	84
2.3	Inference Queries	85
3.	Representing Uncertainty using Graphical Models	87
3.1	Possible World Semantics	88
3.2	Shared Factors	91
3.3	Representing Probabilistic Relations	91
4.	Query Evaluation over Uncertain Data	92
4.1	Example	94
4.2	Generating Factors during Query Evaluation	96
4.3	Query Evaluation as Inference	99
4.4	Optimizations	100
5.	Related Work and Discussion	101
5.1	Safe Plans	101
5.2	Representing Uncertainty using Lineage	102
5.3	Probabilistic Relational Models	102
5.4	Lifted Inference	104
5.5	Scalable Inference using a Relational Database	104
6.	Conclusions	105
	References	107

Trio: A System for Data, Uncertainty, and Lineage	113
---	-----

Jennifer Widom

1. ULDBs: Uncertainty-Lineage Databases	116
1.1 Alternatives	117
1.2 ‘?’ (Maybe) Annotations	117
1.3 Confidences	118
1.4 Lineage	119
1.5 Relational Queries	121
2. TriQL: The Trio Query Language	122
2.1 Operational Semantics	122
2.2 Querying Confidences	124
2.3 Querying Lineage	124
2.4 Duplicate Elimination	125
2.5 Aggregation	126
2.6 Reorganizing Alternatives	127
2.7 Horizontal Subqueries	128
2.8 Query-Defined Result Confidences	128
2.9 Other TriQL Query Constructs	129
3. Data Modifications in Trio	130
3.1 Inserts	130
3.2 Deletes	131
3.3 Updates	131
3.4 Data Modifications and Versioning	133
4. Confidence Computation	133
5. Additional Trio Features	135
6. The Trio System	136
6.1 Encoding ULDB Data	138
6.2 Basic Query Translation Scheme	140
6.3 Duplicate Elimination	141
6.4 Aggregation	141
6.5 Reorganizing Alternatives	142
6.6 Horizontal Subqueries	142
6.7 Built-In Predicates and Functions	143
6.8 Query-Defined Result Confidences	144
6.9 Remaining Constructs	144

References	147
------------	-----

MayBMS: A System for Managing Large Probabilistic Databases	149
---	-----

Christoph Koch

1. Introduction	149
2. Probabilistic Databases	151
3. Query Language Desiderata	152
4. The Algebra	153
5. Representing Probabilistic Data	159
6. Conceptual Query Evaluation	164
7. The MayBMS Query and Update Language	170
8. The MayBMS System	175
9. Conclusions and Outlook	178

References	181
7	
Uncertainty in Data Integration	185
<i>Anish Das Sarma, Xin Dong and Alon Halevy</i>	
1. Introduction	185
2. Overview of the System	187
2.1 Uncertainty in data integration	187
2.2 System architecture	188
2.3 Source of probabilities	189
2.4 Outline of the chapter	190
3. Uncertainty in Mappings	190
3.1 Motivating probabilistic mappings	190
3.2 Definition and Semantics	192
3.3 Query Answering	197
3.4 Creating P-mappings	201
3.5 Broader Classes of Mappings	204
3.6 Other Types of Approximate Schema Mappings	206
4. Uncertainty in Mediated Schema	207
4.1 P-Med-Schema Motivating Example	207
4.2 Probabilistic Mediated Schema	210
4.3 P-med-schema Creation	212
4.4 Consolidation	214
4.5 Other approaches	216
5. Future Directions	217
References	219
8	
Sketching Aggregates over Probabilistic Streams	223
<i>Erik Vee</i>	
1. Introduction	223
1.1 Aggregates over probabilistic streams	225
1.2 Organization	226
2. The Probabilistic Stream Model	226
2.1 Problem definitions	228
2.2 Frequency Moments and Quantiles	229
3. Overview of techniques and summary of results	231
4. Universal Sampling	235
5. Frequency moments: DISTINCT and REPEAT-RATE	236
5.1 DISTINCT	236
5.2 REPEAT-RATE	238
6. Heavy-Hitters, Quantiles, and MEDIAN	240
7. A Binning Technique for MIN and MAX	241
8. Estimating AVG using generating functions	244
8.1 Generating functions	244
8.2 Estimating AVG	245
8.3 Approximating AVG by SUM/COUNT	248
9. Discussion	252
References	253

9

Probabilistic Join Queries in Uncertain Databases 257

Hans-Peter Kriegel, Thomas Bernecker, Matthias Renz and Andreas Zuefle

1. Introduction 257
2. Traditional Join Approaches 258
 - 2.1 Simple Nested-Loop Join 259
 - 2.2 Nested-Block-Loop Join 260
 - 2.3 Sort-Merge-Loop Join 260
 - 2.4 Other Join Methods 261
 - 2.5 Spatial Join Algorithms 261
 - 2.6 Spatial Join using a spatial index structure for both relations 262
 - 2.7 Spatial Join using a spatial index structure on one relation 262
 - 2.8 Spatial Join using no Spatial-Index Structure 263
3. Uncertainty Models and Join Predicates 263
 - 3.1 The Continuous Uncertainty Model 264
 - 3.2 The Discrete Uncertainty Model 266
 - 3.3 Join Predicates and Score 271
 - 3.4 Probabilistic Join Query Types 272
 - 3.5 Example 273
 - 3.6 Uncertainty Models and Probabilistic Join Queries 274
4. Approaches for Efficient Join Processing on Uncertain Data 277
 - 4.1 Confidence-Based Join Methods 277
 - 4.2 Probabilistic Similarity Joins 280
 - 4.3 Probabilistic Spatial Join 288
5. Summary 289

References 293

10

Indexing Uncertain Data 299

Sunil Prabhakar, Rahul Shah and Sarvjeet Singh

1. Introduction 300
2. Data Models and Query Semantics 302
 - 2.1 Uncertain Attribute types 303
3. Uncertainty Index for Continuous Domains 303
 - 3.1 Probability Threshold Indexing 304
 - 3.2 Special Case: Uniform PDFS 307
 - 3.3 2D mapping of intervals 307
 - 3.4 Join Queries 309
 - 3.5 Multi-dimensional Indexing 311
4. Uncertainty Index for discrete domains 311
 - 4.1 Data Model and Problem Definition 312
 - 4.2 Probabilistic Inverted Index 313
 - 4.3 Probabilistic Distribution R-tree 316
5. Indexing for Nearest Neighbor Queries 318

References 323

11

Querying Uncertain Spatiotemporal Data 327

Yufei Tao

1. Introduction 327

2.	Range Search	331
2.1	Query Definitions	331
2.2	Filter and Refinement	333
2.3	Nonfuzzy Range Search	334
2.4	Fuzzy Range Search	336
2.5	Indexing	339
3.	Nearest Neighbor Retrieval	340
3.1	Query Definition	340
3.2	Query Processing	341
3.3	Variations of Nearest Neighbor Retrieval	343
4.	Summary	347
	References	349
12		
	Probabilistic XML	353
	<i>Edward Hung</i>	
1.	Introduction	353
2.	Sources of Uncertainty in XML Data	354
3.	Modeling Uncertainty using Tags	355
4.	Modeling Uncertainty using Semi-structured Data	358
5.	XML with Independent or Mutually Exclusive Distribution	360
6.	Formal Model with Arbitrary Distributions	363
6.1	Motivating Examples	365
6.2	Probabilistic Semi-structured Data Model	367
6.3	Semantics	372
6.4	PXML Algebra and Comparison with Previous Work	375
6.5	Probabilistic Aggregate Operations	377
6.6	Modeling Interval Uncertainty in Semi-structured Data	379
7.	Summary	382
	References	385
13		
	On Clustering Algorithms for Uncertain Data	389
	<i>Charu C. Aggarwal</i>	
1.	Introduction	389
2.	Density Based Clustering Algorithms	391
3.	The UK-means and CK-means Algorithms	394
4.	UMicro: Streaming Algorithms for Clustering Uncertain Data	395
4.1	The UMicro Algorithm: Overview	397
4.2	Computing Expected Similarity	398
4.3	Computing the Uncertain Boundary	402
4.4	Further Enhancements	402
5.	Approximation Algorithms for Clustering Uncertain Data	403
6.	Conclusions and Summary	403
	References	404
14		
	On Applications of Density Transforms for Uncertain Data Mining	407
	<i>Charu C. Aggarwal</i>	

1.	Introduction	407
2.	Kernel Density Estimation with Errors	409
2.1	Scalability for Large Data Sets	412
3.	Leveraging Density Estimation for Classification	416
4.	Application of Density Based Approach to Outlier Detection	419
4.1	Outlier Detection Approach	420
4.2	Subspace Exploration for Outlier Detection	421
5.	Conclusions	423
References		425
15		
Frequent Pattern Mining Algorithms with Uncertain Data		427
<i>Charu C. Aggarwal, Yan Li, Jianyong Wang and Jing Wang</i>		
1.	Introduction	428
2.	Frequent Pattern Mining of Uncertain Data Sets	429
3.	Apriori-style Algorithms	430
3.1	Pruning Methods for Apriori-Style Algorithms	431
4.	Set-Enumeration Methods	434
5.	Pattern Growth based Mining Algorithms	434
5.1	Extending the H-mine algorithm	435
5.2	Extending the FP-growth Algorithm	436
5.3	Another Variation of the FP-growth Algorithm	446
6.	A Comparative Study on Challenging Cases	446
6.1	Performance Comparison	450
6.2	Scalability Comparison	452
7.	Generalization to the Possible Worlds Model	454
8.	Discussion and Conclusions	455
References		457
16		
Probabilistic Querying and Mining of Biological Images		461
<i>Vebjorn Ljosa and Ambuj K. Singh</i>		
1.	Introduction	461
1.1	An Illustrative Example	462
2.	Related Work	465
3.	Probabilistic Image Analyses	466
3.1	Probabilistic Segmentation	466
3.2	Measuring Neurite Thickness	469
3.3	Ganglion Cell Features	470
4.	Querying Probabilistic Image Data	471
4.1	Range Queries on Uncertain Data	472
4.2	k-NN Queries on Uncertain Data	472
4.3	Adaptive, Piecewise-Linear Approximations	474
4.4	Indexing the APLA	474
4.5	Experimental Results	475
5.	Mining Probabilistic Image Data	476
5.1	Defining Probabilistic Spatial Join (PSJ)	477
5.2	Threshold PSJ Query	478
5.3	Top-k PSJ Query	479

5.4	Experimental Results	480
6.	Conclusion	481
	References	483
	Index	489

List of Figures

2.1	Boolean c -tables example	25
2.2	Bag semantics example	26
2.3	Minimal witness why-provenance example	26
2.4	Lineage, why-provenance, and provenance polynomials	30
3.1	Probabilistic Database with Employee Information	49
3.2	A Probabilistic Relation Employee	50
3.3	Relations DocTerm and DocAu	52
3.4	Example Probabilistic Relation	52
3.5	Example Probabilistic Relation with Missing Probabilities	53
3.6	A Probabilistic Relation Target with Three Attributes	55
3.7	A Probabilistic Complex Value Relation	56
3.8	EMPLOYEE: A Probabilistic Relation with Null Values	64
3.9	EMPLOYEE Relation after First Moment Operation	67
4.1	(a,b) A simple car advertisement database with two relations, one containing uncertain data; (c) A joint probability function (<i>factor</i>) that represents the correlation between the validity of two of the ads ($prob_e$ for the corresponding tuples in the <i>Advertisements</i> table can be computed from this); (d) A <i>shared</i> factor that captures the correlations between several attributes in <i>Advertisements</i> – this can be used to obtain a probability distribution over missing attribute values for any tuple.	79
4.2	Example of a directed model for a domain with 5 random variables	83
4.3	Example of an undirected model for a domain with 5 random variables	84

- 4.4 (a) A small database with uncertain attributes. For ease of exposition, we show the marginal pdfs over the attribute values in the table; this information can be derived from the factors. (b) Factors corresponding to the database assuming complete independence. (c) Graphical representation of the factors. 88
- 4.5 Possible worlds for example in Figure 4.4(a) and three other different types of correlations. 89
- 4.6 Factors for the probabilistic databases with “implies” correlations (we have omitted the normalization constant \mathcal{Z} because the numbers are such that distribution is already normalized) 90
- 4.7 Representing the factors from Figure 4.6 using a relational database; shared factors can be represented by using an additional level of indirection. 91
- 4.8 Results running the query $\prod_C(S \bowtie_B T)$ on example probabilistic databases (Figures 4.4 and 4.5). The query returns a non-empty (and identical) result in possible worlds D_3 , D_5 , and D_7 , and the final result probability is obtained by adding up the probabilities of those worlds. 93
- 4.9 Evaluating $\prod_C(S \bowtie_B T)$ on database in Figure 4.4(a). 95
- 4.10 An example query evaluation over a 3-relation database with only tuple uncertainty but many correlations (tuples associated with the same factor are correlated with each other). The intermediate tuples are shown alongside the corresponding random variables. Tuples l_2, \dots, l_6 do not participate in the query. 99
- 4.11 PGM constructed for evaluation of $_{count}G(\sigma_{D=\alpha}(L))$ over the probabilistic database from Figure 4.10. By exploiting decomposability of *count*, we can limit the maximum size of the newly introduced factors to 3 (the naive implementation would have constructed a 5-variable factor). 100
- 4.12 A probabilistic relational model defined over an example relational schema. Similar to Bayesian networks, the model parameters consist of conditional probability distributions for each node given its parents. 103

4.13	An instance of the example PRM with two papers: $P1$, $P2$, with the same author $A1$. For $P1$, we use an explicit random variable for representing the mode of $R1.M$ and $R2.M$. No such variable is needed for $P2$ since it only has one review.	104
5.1	TrioExplorer Screenshot.	116
5.2	Relational Queries on ULDBs.	121
5.3	Trio Basic System Architecture.	137
6.1	Tables of Example 6.2.	156
6.2	Two census forms.	160
6.3	A U-relational database.	163
6.4	Complexity results for (probabilistic) world-set algebra. RA denotes relational algebra.	169
6.5	Exact confidence computation.	176
7.1	Architecture of a data-integration system that handles uncertainty.	188
7.2	The running example: (a) a probabilistic schema mapping between S and T ; (b) a source instance D_S ; (c) the answers of Q over D_S with respect to the probabilistic mapping.	191
7.3	Example 7.11: (a) a source instance D_S ; (b) a target instance that is by-table consistent with D_S and m_1 ; (c) a target instance that is by-tuple consistent with D_S and $\langle m_2, m_3 \rangle$; (d) $Q^{table}(D_S)$; (e) $Q^{tuple}(D_S)$.	193
7.4	Example 7.13: (a) $Q_1^{tuple}(D)$ and (b) $Q_2^{tuple}(D)$.	198
7.5	The motivating example: (a) p-mapping for S_1 and M_3 , (b) p-mapping for S_1 and M_4 , and (c) query answers w.r.t. \mathbf{M} and \mathbf{pM} . Here we denote $\{\text{phone}, \text{hP}\}$ by hPP , $\{\text{phone}, \text{oP}\}$ by oPP , $\{\text{address}, \text{hA}\}$ by hAA , and $\{\text{address}, \text{oA}\}$ by oAA .	209
9.1	Order of Accessed Tuple Pairs Using the Simple Nested-Loop Join	259
9.2	Order of Accessed Blocks and Tuples using the Nested-Block-Loop Join	259
9.3	Order of Joins in Sort-Merge-Join	260
9.4	Example of Two Uncertain Object Relations with Given Scores for the ε -Range Join Predicate and the 1-NN Join Predicate	275
9.5	Overview of Uncertainty Models.	276
9.6	Nested-loop-based Join Approaches.	278

9.7	Upper Bounding Filter Probability of a Join Predicate	283
9.8	Uncertain Objects in One Page	284
9.9	Representation and Organization of Discrete Uncertain Objects	285
9.10	Example for a Probabilistic Distance Range Join Query.	287
9.11	Refinement Criteria for Uncertain Object Approximations	288
9.12	Example of Score Comparison with Thresholds	290
10.1	Inside an Node N_j , with a 0.2-bound and 0.3-bound. A PTRQ named Q is shown as an interval.	305
10.2	Structure of PTI	307
10.3	Probabilistic Threshold Queries with Uniform pdf	308
10.4	Probabilistic Inverted Index	314
10.5	Highest-Prob-First Search (Example)	315
10.6	Probabilistic Distribution R-tree	317
11.1	An example of irregular object pdf	330
11.2	Range search on uncertain data	332
11.3	Pruning/validating with a 2D probabilistically constrained rectangle	335
11.4	Pruning/validating with PCRs for fuzzy queries (under the L_∞ norm)	337
11.5	Enhanced pruning/validating for fuzzy queries with more "slices" (under the L_∞ norm)	338
11.6	Illustration of calculating an NN probability	340
11.7	Illustration of the filter step	342
11.8	Illustration of calculating an NN probability	342
11.9	NN retrieval by expected distances	346
12.1	A Risk Analysis Application	358
12.2	Three Pattern Trees	363
12.3	Data Trees Matching Example Query Pattern Trees	364
12.4	A Semi-structured Instance for a Bibliographic Domain	365
12.5	A Probabilistic Instance for the Bibliographic Domain	371
12.6	Example of Semi-structured Instances Compatible with a Probabilistic Instance	374
12.7	A Probabilistic Instance for the Surveillance Domain	377
12.8	(a) Graph-structure of a Probabilistic Instance (b) Set of Semi-structured Instances Compatible with a Probabilistic Instance	377
12.9	A Probabilistic Instance for the Surveillance Domain	382
13.1	Density Based Profile with Lower Density Threshold	391
13.2	Density Based Profile with Higher Density Threshold	392

13.3	The UMicro Algorithm	398
14.1	Effect of Errors on Classification	409
14.2	Effect of Errors on Clustering	410
14.3	Density Based Classification of Data	416
14.4	Outlier Detection Issues with Uncertain Data	419
14.5	Outlier Detection Algorithm	422
15.1	H-Struct	435
15.2	An example of a trie tree	445
15.3	Runtime Comparison on Connect4	446
15.4	Runtime Comparison on kosarak	447
15.5	Runtime Comparison on T40I10D100K	447
15.6	Memory Comparison on Connect4	448
15.7	Memory Comparison on kosarak	448
15.8	Memory Comparison on T40I10D100K	449
15.9	Scalability Comparison in terms of runtime	453
15.10	Scalability Comparison in terms of Memory	453
16.1	Retinal Layers.	463
16.2	Confocal Micrograph of Horizontal Cells	463
16.3	Probability Distribution of Neurite Thickness	464
16.4	Probability distribution of Inner Plexiform Layer Thickness	464
16.5	Cell Segmented by the Random-walk-with-restarts Algorithm	467
16.6	The Seeded Watershed Algorithm's Segmentation Result	468
16.7	Cross-section through a Dendrite and the Resulting Projected Signal	469
16.8	Ganglion Cell	470
16.9	An Example pdf and its ED-curve	473
16.10	Maximal Errors for APLA and OptimalSplines	475
16.11	Query Times for Range Queries	475
16.12	Running Time for 10-NN Queries	476
16.13	Running Time for k-NN Queries	476
16.14	Each Point Defines a Triangle	478
16.15	Comparative Performance of Algorithms	480
16.16	Effect of Scheduling on Development of Threshold	480

List of Tables

8.1	Summary of Results	235
9.1	An x-relation containing x-tuples with possible positions of tigers.	268
9.2	Confidence of Different Join Predicates	274
9.3	Query Results of Different Probabilistic Join Queries	275
9.4	List of Publications Elaborated in the Next Section	276
10.1	Example of a relation with x-tuples	302
10.2	Example of a relation with Attribute Uncertainty	302
10.3	Example of Uncertain Relation with an Uncertain Discrete Attribute	312

Chapter 1

AN INTRODUCTION TO UNCERTAIN DATA ALGORITHMS AND APPLICATIONS

Charu C. Aggarwal

IBM T. J. Watson Research Center

Hawthorne, NY 10532

charu@us.ibm.com

Abstract

In recent years, uncertain data has become ubiquitous because of new technologies for collecting data which can only measure and collect the data in an imprecise way. Furthermore, many technologies such as privacy-preserving data mining create data which is inherently uncertain in nature. As a result there is a need for tools and techniques for mining and managing uncertain data. This chapter discusses the broad outline of the book and the methods used for various uncertain data applications.

1. Introduction

In recent years many new techniques for collecting data have resulted in an increase in the availability of uncertain data. While many applications lead to data which contains errors, we refer to *uncertain data sets* as those in which the level of uncertainty can be quantified in some way. Some examples of applications which create uncertain data are as follows:

- Many scientific measurement techniques are inherently imprecise. In such cases, the level of uncertainty may be derived from the errors in the underlying instrumentation.
- Many new hardware technologies such as sensors generate data which is imprecise. In such cases, the error in the sensor network readings can be modeled, and the resulting data can be modeled as imprecise data.

- In many applications such as the tracking of mobile objects, the *future trajectory* of the objects is modeled by forecasting techniques. Small errors in current readings can get magnified over the forecast into the distant future of the trajectory. This is frequently encountered in cosmological applications when one models the probability of encounters with Near-Earth-Objects (NEOs). Errors in forecasting are also encountered in non-spatial applications such as electronic commerce.
- In many applications such as privacy-preserving data mining, the data is modified by adding perturbations to it. In such cases, the format of the output [5] is exactly the same as that of uncertain data.

A detailed survey of uncertain data mining and management algorithms may be found in [2]. In this book, we discuss techniques for mining and managing uncertain data. The broad areas covered in the book are as follows:

- **Modeling and System Design for Uncertain Data:** The nature of complexity captured by the uncertain data representation relies on the model used in order to capture it. The most general model for uncertain data is the *possible worlds model*[1], which tries to capture all the possible states of a database which are consistent with a given schema. The generality of the underlying scheme provides the power of the model. On the other hand, it is often difficult to leverage a very general representation for application purposes. In practice, a variety of simplifying assumptions (independence of tuples or independence of attributes) are used in order to model the behavior of the uncertain data. On the other hand, more sophisticated techniques such as probabilistic graphical models can be used in order to model complex dependencies. This is a natural tradeoff between representation power and utility. Furthermore, the design of the system used for representing, querying and manipulating uncertain data critically depends upon the model used for representation.
- **Management of Uncertain Data:** The process of managing uncertain data is much more complicated than that for traditional databases. This is because the uncertainty information needs to be represented in a form which is easy to process and query. Different models for uncertain data provide different tradeoffs between usability and expressiveness. Clearly, the best model to use depends upon the application at hand. Furthermore, effective query languages need to be designed for uncertain data and index structures need to be constructed. Most data management operations such as indexing, join processing or query processing need to be fundamentally re-designed.
- **Mining Uncertain Data:** The uncertainty information in the data is useful information which can be leveraged in order to improve the quality

of the underlying results. For example, in a classification application, a feature with greater uncertainty may not be as important as one which has a lower amount of uncertainty. Many traditional applications such as classification, clustering, and frequent pattern mining may need to re-designed in order to take the uncertainty into account.

This chapter is organized as follows. In the next section, we will discuss the broad areas of work in the topic of uncertain data. Each of these areas is represented by a chapter in the book. The next section will discuss a summary of the material discussed in the chapter and its relationship to other chapters in the book. Section 3 contains the conclusions.

2. Algorithms for Uncertain Data

This section will provide a chapter-by-chapter overview of the different topics which are discussed in this book. The aim is to cover the modeling, management and mining topics fairly comprehensively. The key algorithms in the field are described fairly comprehensively in the different chapters and the relevant pointers are provided. The key topics discussed in the book are as follows:

Models for Uncertain Data. A clear challenge for uncertain data management is underlying data representation and modeling [13, 16, 20]. This is because the underlying representation in the database defines the power of the different approaches which can be used. Chapter 2 provides a clear discussion of the several models which are used for uncertain data management. A related issue is the representation in relational databases, and its relationship with the query language which is finally used. Chapter 3 also discusses the issue of relational modeling of uncertain data, though with a greater emphasis on relational modeling and query languages. While chapter 2 discusses the formal definitions of different kinds of models, chapter 3 discusses some of the more common and simplified models which are used in the literature. The chapter also discusses the implications of using different kinds of models from the relational algebra perspective.

Probabilistic Graphical Models. Probabilistic Graphical Models are a popular and versatile class of models which have significantly greater expressive power because of their graphical structure. They allow us to intuitively capture and reason about complex interactions between the uncertainties of different data items. Chapter 4 discusses a number of common graphical models such as Bayesian Networks and Markov Networks. The chapter discusses the application of these models to the representation of uncertainty. The chapter also discusses how queries can be effectively evaluated on uncertain data with the use of graphical models.

Systems for Uncertain Data. We present two well known systems for uncertain data. These are the *Trio* and *MayBMS* systems. These chapters will provide a better idea of how uncertain data management systems work in terms of database manipulation and querying. The *Trio* system is described in chapter 5, whereas the *MayBMS* system is discussed in chapter 6. Both these chapters provide a fairly comprehensive study of the different kinds of systems and techniques used in conjunction with these systems.

Data Integration. Uncertain data is often collected from disparate data sources. This leads to issues involving database integration. Chapter 7 discusses issues involved in database integration of uncertain data. The most important issue with uncertain data is to use schema mappings in order to match the uncertain data from disparate sources.

Query Estimation and Summarization of Uncertain Data Streams.

The problem of querying is one of the most fundamental database operations. Query estimation is a closely related problem which is often required for a number of database operations. A closely related problem is that of resolving *aggregate queries* with the use of probabilistic techniques such as sketches. Important statistical measures of streams such as the quantiles, minimum, maximum, sum, count, repeat-rate, average, and the number of distinct items are useful in a variety of database scenarios. Chapter 8 discusses the issue of sketching probabilistic data streams, and how the synopsis may be used for estimating the above measures.

Join Processing of Uncertain Data. The problem of join processing is challenging in the context of uncertain data, because the join-attribute is probabilistic in nature. Therefore, the join operation needs to be redefined in the context of probabilistic data. Chapter 9 discusses the problem of join processing of uncertain data. An important aspect of join processing algorithms is that the uncertainty model significantly affects the nature of join processing. The chapter discusses different kinds of join methods such as the use of *confidence-based join methods*, *similarity joins* and *spatial joins*.

Indexing Uncertain Data. The problem of indexing uncertain data is especially challenging because the diffuse probabilistic nature of the data can reduce the effectiveness of index structures. Furthermore, the challenges for indexing can be quite different, depending upon whether the data is discrete, continuous, spatio-temporal, or how the probabilistic function is defined [8, 9, 12, 22, 23]. Chapter 10 provides a comprehensive overview of the problem of indexing uncertain data. This chapter discusses the problem of indexing both continuous and discrete data. Chapter 11 further discusses the problem of

indexing uncertain data in the context of spatiotemporal data. Chapters 10 and 11 provide a fairly comprehensive survey of the different kinds of techniques which are often used for indexing and retrieval of uncertain data.

Probabilistic XML Data. XML data poses a number of special challenges in the context of uncertainty because of the structural nature of the underlying data. Chapter 12 discusses uncertain models for probabilistic XML data. The chapter also describes algebraic techniques for manipulating XML data. This includes probabilistic aggregate operations and the query language for XML data (known as PXML). The chapter discusses both special cases for probability distributions as well as arbitrary probability distributions for representing probabilistic XML data.

Clustering Uncertain Data. Data mining problems are significantly influenced by the uncertainty in the underlying data, since we can leverage the uncertainty in order to improve the quality of the underlying results. Clustering is one of the most comprehensively studied problems in the uncertain data mining literature. Recently, techniques have been designed for clustering uncertain data. These include the *UMicro* algorithm, the UK-means algorithms, the FDBSCAN, and FOPTICS algorithms [6, 18, 19, 21]. Recently, some approximation algorithms [7] have also been developed for clustering uncertain data. Chapter 13 discusses a comprehensive overview of the different algorithms for clustering uncertain data.

General Transformations for Uncertain Data Mining. A natural approach to uncertain data management techniques is to use general transformations [3] which can create *intermediate representations* which adjust for the uncertainty. These intermediate representations can then be leveraged in order to improve the quality of the underlying results. Chapter 14 discusses such an approach with the use of density based transforms. The idea is to create a probability density representation of the data which takes the uncertainty into account during the transformation process. The chapter discusses two applications of this approach to the problems of classification and outlier detection. We note that the approach can be used for any data mining problem, as long as a method can be found to use intermediate density transformations for data mining purposes.

Frequent Pattern Mining. Chapter 15 surveys a number of different approaches for frequent pattern mining of uncertain data. In the case of transactional data, items are assumed to have *existential probabilities* [4, 10, 11], which characterize the likelihood of presence in a given transaction. This includes Apriori-style algorithms, candidate generate-and-test algorithms, pat-

tern growth algorithms and hyper-structure based algorithms. The chapter examines the uniqueness of the tradeoffs involved for pattern mining algorithms in the uncertain case. The chapter compares many of these algorithms for the challenging case of high existential probabilities, and shows that the behavior is quite different from deterministic algorithms. Most of the literature [10, 11] studies the case of low existential probabilities. The chapter suggests that the behavior is quite different for the case of high-existential probabilities. This is because many of the pruning techniques designed for the case of low existential probabilities do not work well for the case when these probabilities are high.

Applications to Biomedical Domain. We provide one application chapter in order to provide a flavor of the application of uncertain DBMS techniques to a real application. The particular application picked in this case is that of biomedical images. Chapter 16 is a discussion of the application of uncertain data management techniques to the biomedical domain. The chapter is particularly interesting in that it discusses the application of many techniques discussed in this book (such as indexing and join processing) to an application domain. While the chapter discusses the biological image domain, the primary goal is to present an example of the application of many of the discussed techniques to a particular application.

3. Conclusions

In this chapter, we introduced the problem of uncertain data mining, and discussed an overview of the different facets of this area covered by this book. Uncertain data management promises to be a new and exciting field for practitioners, students and researchers. It is hoped that this book is able to provide a broad overview of this topic, and how it relates to a variety of data mining and management applications. This book discusses both data management and data mining issues. In addition, the book discusses an application domain for the field of uncertain data. Aside from the topics discussed in the book, some of the open areas for research in the topic of uncertain data are as follows:

- **Managing and Mining Techniques under General Models:** Most of the uncertain data mining and management algorithms use a variety of simplifying assumptions in order to allow effective design of the underlying algorithms. Examples of such simplifying assumptions could imply tuple or attribute independence. In more general scenarios, one may want to use more complicated schemas to represent uncertain databases. Some models such as probabilistic graphical models [15] provide greater expressivity in capturing such cases. However, database management and mining techniques become more complicated under such models.

Most of the current techniques in the literature do not use such general models. Therefore, the use of such models for developing DBMS techniques may be a fruitful future area of research.

- **Synergy between Uncertain Data Acquisition and Usage:** The utility of the field can increase further only if a concerted effort is made to standardize the uncertainty in the data to the models used for the general management and mining techniques. For example, the output of both the privacy-preserving publishing and the sensor data collection fields are typically uncertain data. In recent years, some advances have been made [5, 14] in order to design models for data acquisition and creation, which naturally pipeline onto useful uncertain representations. A lot more work remains to be done in a variety of scientific fields in order to facilitate model based acquisition and creation of uncertain data.

Acknowledgements

Research was sponsored in part by the US Army Research laboratory and the UK ministry of Defense under Agreement Number W911NF-06-3-0001. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies of the US Government, the US Army Research Laboratory, the UK Ministry of Defense, or the UK Government. The US and UK governments are authorized to reproduce and distribute reprints for Government purposes.

References

- [1] S. Abiteboul, P. C. Kanellakis, G. Grahne. "On the Representation and Querying of Sets of Possible Worlds." in *Theoretical Computer Science*, 78(1): 158-187 (1991)
- [2] C.C. Aggarwal, P. S. Yu. "A Survey of Uncertain Data Algorithms and Applications," in *IEEE Transactions on Knowledge and Data Engineering*, to appear, 2009.
- [3] C. C. Aggarwal, "On Density Based Transforms for Uncertain Data Mining," in *ICDE Conference Proceedings*, 2007.
- [4] C. C. Aggarwal, Y. Li, J. Wang, J. Wang. "Frequent Pattern Mining with Uncertain Data." *IBM Research Report*, 2008.
- [5] C. C. Aggarwal, "On Unifying Privacy and Uncertain Data Models," in *ICDE Conference Proceedings*, 2008.
- [6] C. C. Aggarwal and P. S. Yu, "A Framework for Clustering Uncertain Data Streams," in *ICDE Conference*, 2008.

- [7] G. Cormode, and A. McGregor, "Approximation algorithms for clustering uncertain data," in *PODS Conference*, pp. 191-200, 2008.
- [8] R. Cheng, Y. Xia, S. Prabhakar, R. Shah, and J. Vitter, "Efficient Indexing Methods for Probabilistic Threshold Queries over Uncertain Data," in *VLDB Conference Proceedings*, 2004.
- [9] R. Cheng, D. Kalashnikov, S. Prabhakar: "Evaluating Probabilistic Queries over Imprecise Data" in *SIGMOD Conference*, 2003.
- [10] C.-K. Chui, B. Kao, E. Hung. "Mining Frequent Itemsets from Uncertain Data." *PAKDD Conference*, 2007.
- [11] C.-K. Chui, B. Kao. "Decremental Approach for Mining Frequent Itemsets from Uncertain Data." *PAKDD Conference*, 2008.
- [12] D. Pfozer, C. Jensen. Capturing the uncertainty of moving object representations. in *SSDM Conference*, 1999.
- [13] A. Das Sarma, O. Benjelloun, A. Halevy, and J. Widom, "Working Models for Uncertain Data," in *ICDE Conference Proceedings*, 2006.
- [14] A. Deshpande, C. Guestrin, S. Madden, J. M. Hellerstein, W. Hong. "Model-Driven Data Acquisition in Sensor Networks." in *VLDB Conference*, 2004.
- [15] A. Deshpande, S. Sarawagi. "Probabilistic Graphical Models and their Role in Databases." in *VLDB Conference*, 2007.
- [16] H. Garcia-Molina, and D. Porter, "The Management of Probabilistic Data," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 4, pp. 487-501, 1992.
- [17] B. Kanagal, A. Deshpande, "Online Filtering, Smoothing and Probabilistic Modeling of Streaming data," in *ICDE Conference*, 2008.
- [18] H.-P. Kriegel, and M. Pfeifle, "Density-Based Clustering of Uncertain Data," in *ACM KDD Conference Proceedings*, 2005.
- [19] H.-P. Kriegel, and M. Pfeifle, "Hierarchical Density Based Clustering of Uncertain Data," in *ICDM Conference*, 2005.
- [20] L. V. S. Lakshmanan, N. Leone, R. Ross, and V. S. Subrahmanian, "ProbView: A Flexible Probabilistic Database System," in *ACM Transactions on Database Systems*, vol. 22, no. 3, pp. 419-469, 1997.
- [21] W. Ngai, B. Kao, C. Chui, R. Cheng, M. Chau, and K. Y. Yip, "Efficient Clustering of Uncertain Data," in *ICDM Conference Proceedings*, 2006.
- [22] S. Singh, C. Mayfield, S. Prabhakar, R. Shah, S. Hambrusch. "Indexing Uncertain Categorical Data", in *ICDE Conference*, 2007.
- [23] Y. Tao, R. Cheng, X. Xiao, W. Ngai, B. Kao, S. Prabhakar. "Indexing Multi-dimensional Uncertain Data with Arbitrary Probability Density Functions", in *VLDB Conference*, 2005.

Chapter 2

MODELS FOR INCOMPLETE AND PROBABILISTIC INFORMATION

Todd J. Green

Department of Computer and Information Science

University of Pennsylvania

tjgreen@cis.upenn.edu

Abstract We discuss, compare and relate some old and some new models for incomplete and probabilistic databases. We characterize the expressive power of c -tables over infinite domains and we introduce a new kind of result, algebraic completion, for studying less expressive models. By viewing probabilistic models as incompleteness models with additional probability information, we define completeness and closure under query languages of general probabilistic database models and we introduce a new such model, probabilistic c -tables, that is shown to be complete and closed under the relational algebra. We also identify fundamental connections between query answering with incomplete and probabilistic databases and data provenance. We show that the calculations for incomplete databases, probabilistic databases, bag semantics, lineage, and why-provenance are particular cases of the same general algorithms involving semi-rings. This further suggests a comprehensive provenance representation that uses semi-rings of polynomials. Finally, we show that for positive Boolean c -tables, containment of positive relational queries is the same as for standard set semantics.

Keywords: Incomplete databases, probabilistic databases, provenance, lineage, semi-rings

1. Introduction

This chapter provides a survey of models for incomplete and probabilistic information from the perspective of two recent papers that the author has written with Val Tannen [28] and Grigoris Karvounarakis and Val Tannen [27]. All the concepts and technical developments that are not attributed specifically to another publication originate in these two papers.

The representation of incomplete information in databases has been an important research topic for a long time, see the references in [25], in Ch.19 of [2], in [43], in [48, 36], as well as the recent [45, 42, 41, 4]. Moreover, this work is closely related to recently active research topics such as inconsistent databases and repairs [5], answering queries using views [1], data exchange [20], and data provenance [9, 8]. The classic reference on incomplete databases remains [30] with the fundamental concept of *c*-table and its restrictions to simpler tables with variables. The most important result of [30] is the query answering algorithm that defines an algebra on *c*-tables that corresponds exactly to the usual relational algebra (\mathcal{RA}). A recent paper [41] has defined a hierarchy of incomplete database models based on finite sets of choices and optional inclusion. We shall give below **comparisons** between the models [41] and the tables with variables from [30].

Two criteria have been provided for comparisons among all these models: [30, 41] discuss *closure* under relational algebra operations, while [41] also emphasizes *completeness*, specifically the ability to represent all finite incomplete databases. We point out that the latter is not appropriate for tables with variables over an infinite domain, and we describe another criterion, **\mathcal{RA} -completeness**, that fully characterizes the expressive power of *c*-tables.

We outline a method for the study of models that are not complete. Namely, we consider combining existing models with queries in various fragments of relational algebra. We then ask how big these fragments need to be to obtain a combined model that is complete. We give a number of such **algebraic completion** results.

Early on, probabilistic models of databases were studied less intensively than incompleteness models, with some notable exceptions [10, 6, 39, 34, 17]. Essential progress was made independently in three papers [22, 33, 47] that were published at about the same time. [22, 47] assume a model in which tuples are taken independently in a relation with given probabilities. [33] assumes a model with a separate distribution for each attribute in each tuple. All three papers attacked the problem of calculating the probability of tuples occurring in query answers. They solved the problem by developing more general models in which rows are **annotated** with additional information (“event expressions,” “paths,” “traces”), and they noted the similarity with the conditions in *c*-tables.

We go beyond the problem of individual tuples in query answers by defining **closure** under a query language for probabilistic models. Then we describe **probabilistic *c*-tables** which add to the *c*-tables themselves probability distributions for the values taken by their variables. Here is an example of such a representation that captures the set of instances in which Alice is taking a course that is Math with probability 0.3; Physics (0.3); or Chemistry (0.4), while Bob takes the same course as Alice, provided that course is Physics or