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,* *VipulKashyap, AmitSheth*; ISBN:0-7923-7883-0

For a complete listing of books in this series, go to http://www.springer.com

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

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

List of Figures

List of Tables

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, pattern 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 particular 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. Prabhaker: “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 Probabality 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