Similarity Search and Mining in Uncertain Databases

Matthias Renz
Institute for Informatics,
Ludwig-Maximilians-
Universität
München,
Oettingenstr. 67, D-80538
München, Germany
{renz}@dbs.ifi.lmu.de

Reynold Cheng
Department of Computer
Science,
University of Hong Kong,
Pokfulam Road, Hong Kong
{ckcheng}@cs.hku.hk

Hans-Peter Kriegel
Institute for Informatics,
Ludwig-Maximilians-
Universität
München
Oettingenstr. 67, D-80538
München, Germany
{kriegel}@dbs.ifi.lmu.de

ABSTRACT

Managing, searching and mining uncertain data has achieved much attention in the database community recently due to new sensor technologies and new ways of collecting data. There is a number of challenges in terms of collecting, modelling, representing, querying, indexing and mining uncertain data. In its scope, the diversity of approaches addressing these topics is very high because the underlying assumptions of uncertainty are different across different papers. This tutorial provides a comprehensive and comparative overview of general techniques for the key topics in the fields of querying, indexing and mining uncertain data. In particular, it identifies the most generic types of probabilistic similarity queries and discusses general algorithmic methods to answer such queries efficiently. In addition, the tutorial sketches probabilistic methods for important data mining applications in the context of uncertain data with special emphasis on probabilistic clustering and probabilistic pattern mining. The intended audience of this tutorial ranges from novice researchers to advanced experts as well as practitioners from any application domain dealing with uncertain data retrieval and mining.

1. INTRODUCTION

Searching and mining in uncertain databases has become very popular problem in the recent years. The increasing availability of novel data-collection devices enables to accumulate large amounts of information in unprecedented rates and variability. On the other hand, the collected information is often noisy, incomplete or rendered uncertain due to privacy preserving issues. As a consequence, novel data models representing uncertain data have been developed and integrated into modern database systems [3, 14, 28, 2]. As query evaluation becomes more complex, efficiency issues arise, calling for novel search methodologies that cope with the special nature of uncertain data.

There currently exists a wide range of different approaches caused by different assumptions imposed on the uncertain data. As a consequence, it is a complex task to keep track of the current research results not only because of the large amount of existing approaches, but also because of very different vocabulary used to express similar concepts. This tutorial aims at providing a comprehensive view of the state-of-the-art research in probabilistic similarity search and probabilistic data mining for uncertain data. This tutorial will give a survey of and classify the various approaches for probabilistic similarity queries, indexing uncertain data and data mining on uncertain objects proposed for the different uncertainty models and illustrate relationships among them.

2. MANAGING, QUERYING AND MINING UNCERTAIN DATA

Real-world applications dealing with uncertain data require efficient methods for efficient searching on this data. Example applications are proximity queries in spatial databases with mobile objects, privacy-preserving information retrieval and data mining, searching in scientific databases and sensor databases. There are a bundle of problems emerging from the aforementioned applications that challenge the problem of designing efficient solutions for managing and querying uncertain data.

In this tutorial, we introduce the basic concepts of managing, querying and mining uncertain data. Here we take special emphasize on concepts supporting probabilistic spatial and similarity queries as well as applications on mining uncertain data. Note that as the research area on uncertain databases is very broad such that we cannot address all important issues in this tutorial. However, this tutorial should give an overview of the most significant concepts.

2.1 Modelling and Managing Uncertain Data

The concepts for managing, querying and mining uncertain data mainly rely on the models required to represent uncertain data. One of the most important and prevalent concept to model uncertain data is the *possible worlds model* [1]. This model is designed to reflect all possible instances of a database based on the possible instances of the underlying uncertain data. While this model is primarily designed for discrete uncertain data representation, many approaches using continuous uncertain data representations rely on the same concept when issuing query results.

In the first part of this tutorial, we first give a survey on uncertainty models and then focus on the attribute uncertainty model, specifically the discrete and continuous uncertainty model.

2.2 Indexing Uncertain Data

Efficient querying uncertain data requires index methods that are appropriate for uncertain data. The probabilistic nature of uncertain data disqualifies traditional indexing methods. Depending on the underlying uncertainty model, there are diverse challenges for

Proceedings of the VLDB Endowment, Vol. 3, No. 2
Copyright 2010 VLDB Endowment 2150-8097/10/09... $ 10.00.

designing suitable index structures. This part of the tutorial gives a brief overview of a number of methods for indexing uncertain data [10, 7, 15, 27, 30].

2.3 Probabilistic (Similarity) Search Paradigms

The problem of processing queries in uncertain data is challenging due to the probabilistic nature of the evaluation of the query predicate. This especially holds for similarity queries, because in general the similarity (distance) between two uncertain objects is uncertain. This presents new challenges that require special types of queries the probabilistic (similarity) query processing. This part shows for various probabilistic similarity search problems current state-of-the-art solutions in particular for probabilistic range queries [10, 16], probabilistic k-nearest neighbor and ranking queries [9, 30, 6, 17, 13, 29, 21], probabilistic reverse k-nearest neighbor queries [8, 23], probabilistic inverse ranking [24] and probabilistic skyline queries [26, 22].

2.4 Data Mining on Uncertain Data

Finally, this tutorial gives an overview of state-of-the-art solutions for various data mining applications such as density-based clustering [18, 5], partition-based clustering [25, 19] and frequent pattern mining [11, 12, 20, 4], emphasizing efficiency issues.

3. ADDITIONAL AUTHORS

Additional authors: Andreas Züfle (Institute for Informatics, Ludwig-Maximilians-Universität München, email: zuefle@dbs.ifi.lmu.de) and Thomas Bernecker (Institute for Informatics, Ludwig-Maximilians-Universität München, email: bernecker@dbs.ifi.lmu.de).

4. REFERENCES

[1] S. Abiteboul, P. Kanellakis, and G. Grahne. On the representation and querying of sets of possible worlds. *SIGMOD Rec.*, 16(3):34–48, 1987.

[2] P. Agrawal, O. Benjelloun, A. D. Sarma, C. Hayworth, S. U. Nabar, T. Sugihara, and J. Widom. Trio: A system for data, uncertainty, and lineage. In *Proceedings of the 32nd International Conference on Very Large Data Bases, Seoul, Korea, September 12-15, 2006*, pages 1151–1154, 2006.

[3] L. Antova, T. Jansen, C. Koch, and D. Olteanu. Fast and simple relational processing of uncertain data. In *ICDE '08: Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, pages 983–992, Washington, DC, USA, 2008. IEEE Computer Society.

[4] T. Bernecker, H.-P. Kriegel, M. Renz, F. Verhein, and A. Züfle. Probabilistic frequent itemset mining in uncertain databases. In *In Proc. 15th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, Paris, France*, 2009.

[5] T. Bernecker, H.-P. Kriegel, M. Renz, and A. Zuefle. Hot item detection in uncertain data. In *Proc. 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, Bangkok, Thailand*, pages 673–680, 2009.

[6] C. Böhm, A. Pryakhin, and M. Schubert. Probabilistic ranking queries on gaussians. In *Proc. SSDBM*, pages 169–178, 2006.

[7] C. Böhm, A. Pryakhin, and M. Schubert. The gauss-tree: Efficient object identification of probabilistic feature vectors. In *Proc. ICDE*, 2006.

[8] M. A. Cheema, X. Lin, W. Wang, W. Zhang, and J. Pei. Probabilistic reverse nearest neighbor queries on uncertain data. *to appear in TKDE*, 22(4):550–564, 2010.

[9] R. Cheng, D. Kalashnikov, and S. Prabhakar. "Evaluating Probabilistic Queries over Imprecise Data". In *Proc. SIGMOD*, pages 551–562, 2003.

[10] R. Cheng, Y. Xia, S. Prabhakar, R. Shah, and J. Vitter. Efficient indexing methods for probabilistic threshold queries over uncertain data. In *Proc. VLDB*, pages 876–887, 2004.

[11] C. K. Chui and B. Kao. A decremental approach for mining frequent itemsets from uncertain data. In *The 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 64–75, 2008.

[12] C. K. Chui, B. Kao, and E. Hung. Mining frequent itemsets from uncertain data. In *11th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, PAKDD 2007, Nanjing, China*, pages 47–58, 2007.

[13] G. Cormode, F. Li, and K. Yi. Semantics of ranking queries for probabilistic data and expected results. In *Proceedings of the 25th International Conference on Data Engineering, ICDE 2009, March 29-April 2, 2009, Shanghai, China*, pages 305–316, 2009.

[14] N. Dalvi and D. Suciu. Efficient query evaluation on probabilistic databases. *The VLDB Journal*, 16(4):523–544, 2007.

[15] B. Kanagal and A. Deshpande. Indexing correlated probabilistic databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2009, Providence, Rhode Island, USA, June 29 - July 2, 2009*, pages 455–468, 2009.

[16] H.-P. Kriegel, P. Kunath, M. Pfeifle, and M. Renz. "Probabilistic Similarity Join on Uncertain Data". In *Proc. 11th Int. Conf. on Database Systems for Advanced Applications, Singapore, pp. 295-309*, 2006.

[17] H.-P. Kriegel, P. Kunath, and M. Renz. "Probabilistic Nearest-Neighbor Query on Uncertain Objects". In *Proc. 12th Int. Conf. on Database Systems for Advanced Applications, Bangkok, Thailand*, 2007.

[18] H.-P. Kriegel and M. Pfeifle. Density-based clustering of uncertain data. In *KDD '05: Proc. 11th ACM SIGKDD international conference on Knowledge discovery in data mining, Chicago, Illinois, USA*, pages 672–677, 2005.

[19] S. D. Lee, B. Kao, and R. Cheng. Reducing uk-means to k-means. In *Proc. 7th IEEE International Conference on Data Mining Workshops*, pages 483–488, 2007.

[20] C. K.-S. Leung, C. L. Carmichael, and B. Hao. Efficient mining of frequent patterns from uncertain data. In *ICDMW '07: Proceedings of the Seventh IEEE International Conference on Data Mining Workshops*, pages 489–494, 2007.

[21] J. Li, B. Saha, and A. Deshpande. A unified approach to ranking in probabilistic databases. *PVLDB*, 2(1):502–513, 2009.

[22] X. Lian and L. Chen. Monochromatic and bichromatic reverse skyline search over uncertain databases. In *SIGMOD '08: Proc. 28th international conference on Management of data, Vancouver, Canada*, pages 213–226, 2008.

[23] X. Lian and L. Chen. Efficient processing of probabilistic reverse nearest neighbor queries over uncertain data. *VLDB J.*, 18(3):787–808, 2009.

[24] X. Lian and L. Chen. Probabilistic inverse ranking queries over uncertain data. In *Proc. 14th International Conference on Database Systems for Advanced Applications, Brisbane, Australia.*, pages 35–50, 2009.

[25] W. K. Ngai, B. Kao, C. K. Chui, R. Cheng, M. Chau, and K. Y. Yip. Efficient clustering of uncertain data. *Data Mining, IEEE International Conference on*, 0:436–445, 2006.

[26] J. Pei, B. Jiang, X. Lin, and Y. Yuan. Probabilistic skylines on uncertain data. In *VLDB '07: Proceedings of the 33rd international conference on Very large data bases, Vienna, Austria*, 2007.

[27] Y. Qi, S. Singh, R. Shah, and S. Prabhakar. Indexing probabilistic nearest-neighbor threshold queries. In *Proceedings of the International Workshop on Quality in Databases and Management of Uncertain Data, Auckland, New Zealand, August 2008*, pages 87–102, 2008.

[28] P. Sen and A. Deshpande. Representing and querying correlated tuples in probabilistic databases. 2007.

[29] M. Soliman and I. Ilyas. Ranking with uncertain scores. In *Proceedings of the 25th International Conference on Data Engineering, ICDE 2009, March 29-April 2, 2009, Shanghai, China*, pages 317–328, 2009.

[30] Y. Tao, R. Cheng, X. Xiao, W. Ngai, B. Kao, and S. Prabhakar. "Indexing Multi-Dimensional Uncertain Data with Arbitrary Probability Density Functions". In *Proc. VLDB*, pages 922–933, 2005.