Report on the first VLDB workshop on Management of Uncertain Data (MUD)

Ander de Keijzer• Maurice van Keulen•
•University of Twente
The Netherlands
{a.dekeijzer,m.vankeulen}@utwente.nl

Alex Dekhtyar◇
◇California Polytechnic State University
United States
dekhtyar@csc.calpoly.edu

1 Introduction

On Monday September 24th, we organized the first international VLDB workshop on Management of Uncertain Data [dKvKD07]. The idea of this workshop arose a year earlier at the Twente Data Management Workshop on Uncertainty in Databases [dKvK06]. The TDM is a bi-annual workshop organized by the Database group of the University of Twente, for which each time a different topic is chosen. The participants of TDM 2006 were enthusiastic about the topic "Uncertainty in Databases" and strongly expressed the wish for a follow-up co-located with an international conference. To fulfill this wish, we organized the MUD-workshop at VLDB.

The program committee consisted of 20 members and 1 advisory member, Jennifer Widom from Stanford University. Committee members came from universities and research institutes from Europe and North America. We accepted 6 full papers and invited 2 speakers, Lise Getoor from the University of Maryland at College Park and Sunil Prabhakar from Purdue University.

Both the morning and afternoon session consisted of an invited talk and a research session. In the morning Lise Getoor gave a talk on *Combining Tuple and Attribute Uncertainty in Probabilistic Databases*, which was followed by a research session on *Applications of Uncertain Data*. The afternoon session started with the talk by Sunil Prabhakar on *Supporting Probabilistic Data in Relational Databases*, which was followed by a session on *Querying Uncertain Data*.

Special thanks go to the Centre for Telematics and Information Technology (CTIT) for sponsoring the proceedings.

2 Applications of Uncertain Data

The kick-off of the workshop was given by Lise Getoor from the University of Maryland. She gave an overview of techniques from machine learning and reasoning under uncertainty. These areas have developed quite powerful models for the representation of probability distributions, for example, probabilistic graph models. She showed how these techniques influenced her work on probabilistic relational models especially on how to unify attribute and record level uncertainty.

The first research talk of the workshop was given by Antoon Bronselaer from the University of Ghent. He introduced the application of disaster victim identification for large scale disasters. The problem can be seen as an object matching or entity resolution problem: based on the available data of a victim determine whether or not that data refers to the same real world object as data from a reference list. The focus of the paper was on how to integrate a complex matching technique based on ear biometrics into their object matching framework. It was shown that the framework, which was based on a possibilistic uncertainty model, was capable of effectively capturing and handling the uncertainty resulting from missing data and feature extraction errors.

In the second presentation, Matteo Magnani argued that data integration could be the killer application for uncertain data management systems. One of the main problems in data integration is schema matching. Current approaches combine the judgments of multiple matchers to obtain the most relevant schema mappings. Magnani argues that significant improvement can be obtained by not only finding the correct mappings, but also by managing the incorrect ones properly. They propose to view the mappings as possible mappings with a certain level of uncertainty and treating the accompanying data during querying accordingly, i.e., also with a certain level of uncertainty.

The topic of the third and last presentation of the morning session was fuzzy querying. Ramón Alberto Carrasco presented their language dmFSQL (data mining fuzzy SQL) which allows you to easily verify data mining hypotheses. The paper focused on their latest addition to the language: fuzzy global dependencies. The idea is that the system computes the percentage of tuples which fulfill a given antecedent and consequent together w.r.t. those that only fulfill the consequent. This allows you to validate hypothesized monotonicity of relationships between objects in the data, e.g., which patterns imply higher earnings of a specific share on the stock market. For this particular example, it was presented how the system could obtain the final statement "Greater williams index and roughly equal moving average implies a greater value for the specific enterprise Telefonica with confidence 0.9".

3 Querying Uncertain Data

The afternoon session started with an invited talk by Sunil Prabhakar. The topic of the talk was Supporting Probabilistic Data in Relational Databases and was focused on the ORION DBMS. Sunil Prabhakar provided a nice overview of possible world semantics and the problems that arise with continuous uncertainty. Currently, the ORION system offers the combination of continuous uncertainty and possible world semantics.

The first research talk of the afternoon session on Querying Uncertain Data was given by Patrick Bosc from IRISA/ENSSAT, France. The model of uncertain data he used was a possibilistic model. During the discussion at the end of the talk, many of the questions were addressing the differences between probabilistic and possibilistic theory. One notable difference is that a possibilistic model uses maximum and minimum for combining confidences, while a probabilistic model uses addition and multiplication. From the discussion arose that a possibilistic model does not make assumptions about dependencies between stochastic variables while probabilistic models usually so. The conclusion of the discussion was, that both theories have their advantages and purposes.

The second presentation, given by Jef Wijsen of the University of Mons-Hainaut, was about introducing uncertainty by considering possible repairs for key constraint violations. These violations can be solved in different ways. Each of the minimal solutions can be regarded as a possible world. Jef focused on the notion of relations to 'consistently join', i.e., for all possible repairs the join contains at least one tuple. He used a game theoretic approach to decide on this notion.

The last presentation of the workshop was given by Raghotham Murthy of Stanford University. His presentation on aggregate functions in databases supporting uncertainty used the Trio database system as an example. He presented algorithms for estimating a lower bound, higher bound, and expected value for aggregates on uncertain relations, because these typically produce exponential results. Afterwards, even after the workshop officially ended, several participants continued discussing about the semantics of aggregates. Different views on how aggregates should be interpreted were discussed, and in the end it turned out that people agreed on the main idea of aggregates, although there seemed to be some difference in opinions on the details. All in all, this topic will probably be continued at subsequent workshops.

4 Conclusion and Outlook

Discussions during the workshop showed that the management of uncertain data is a vibrant research area with many promising applications, but also a significant number of open issues. For example, one can distinguish several kinds of underlying data models for uncertain data: fuzzy logic-based models, repair models, possibilistic models and probabilistic models. The relationships, commonalities and differences are not well understood yet. And if theory is not well enough established yet, work on algorithms, scalability and systems is necessarily also still in its infancy. But, the strength of approaches based on properly managing uncertainty in data can already been demonstrated as the application-oriented papers in the MUD workshop clearly show. Moreover, the papers in this workshop also show that the challenges, for example the ones presented in the visionary paper on dataspace systems [HFM06], are being addressed today and significant advances are being made. To continue our efforts to build a rich co-operating community on this topic and support effective exchange of ideas, we plan to organize a second MUD-workshop again co-located with VLDB next year.

References

[dKvK06] A. de Keijzer and M. van Keulen, editors. *Proc. of the 2nd Twente Data Management Workshop (TDM 2006) on Uncertainty in Databases (Enschede, The Netherlands, June 6, 2006)*, number WP06-01 in CTIT Workshop Proceedings Series. Centre for Telematics and Information Technology (CTIT), Univ. of Twente, Enschede, The Netherlands, June 2006. `http://www.cs.utwente.nl/~tdm`.

[dKvKD07] A. de Keijzer, M. van Keulen, and A. Dekhtyar, editors. *Proc. of the 1st Int. VLDB workshop on Management of Uncertain Data (MUD, Vienna, Austria, September 24, 2007)*, number WP-CTIT-07-08 in CTIT Workshop Proceedings Series. Centre for Telematics and Information Technology (CTIT), Univ. of Twente, Enschede, The Netherlands, September 2007. `http://mud.cs.utwente.nl`.

[HFM06] A.Y. Halevy, M.J. Franklin, and D. Maier. Principles of dataspace systems. In *Proceedings of PODS, Chicago, IL, USA*, pages 1–9, 2006.