Working Group: Classification, Representation and Modeling

Anish Das Sarma[1], Ander de Keijzer[2], Amol Deshpande[3], Peter J. Haas[4], Ihab F. Ilyas[5], Christoph Koch[6], Thomas Neumann[7], Dan Olteanu[8], Martin Theobald[7], and Vasilis Vassalos[9]

[1] Stanford University
[2] University of Twente
[3] University of Maryland
[4] IBM Almaden Research Center
[5] University of Waterloo
[6] Cornell University
[7] MPI fr Informatik, Saarbrcken
[8] Oxford University
[9] Athens University of Economics and Business

Abstract. This report briefly summarizes the discussions carried out in the working group on classification, representation and modeling of uncertain data. The discussion was divided into two subgroups: the first subgroup studied how different representation and modeling alternatives currently proposed can fit in a bigger picture of theory and technology interaction, while the second subgroup focused on contrasting current system implementations and the reasons behind such diverse class of available prototypes. We summarize the findings of these two groups and the future steps suggested by group members.

1 Theory and Technologies - The Big Picture

Uncertainty modeling, viewed as an interaction between theory and technologies, is a space of multiple dimensions that cover concepts, theory platforms, technology platforms and application features. More specifically: A *Theory Matrix* that is the product of a vector of uncertainty concepts and a vector of possible theory platforms represents multiple uncertainty formalisms. Example uncertainty concepts include tuple uncertainty (where the existence of a database record is an uncertain event), and value uncertainty (where values of some of the attributes/features of some entity are uncertain or unknown). Example theory platforms include probability theory and fuzzy logic.

The interaction between possible uncertainty formalisms and current technologies (e.g., relational and XML data engines) has produced diverse uncertainty-aware technologies currently proposed by different research groups. For example, introducing uncertainty formalisms based on tuple and value uncertainty and probability theory in relational platform produced probabilistic database engines such as ORION [1] and TRIO [2–4]. Finally, the features implemented

Dagstuhl Seminar Proceedings 08421
Uncertainty Management in Information Systems
http://drops.dagstuhl.de/opus/volltexte/2009/1941

in these uncertainty-aware technologies are highly influenced by application requirements. These interactions can be summarized as follows.

$$Uncertainty\ Concepts \times Theory\ Platforms = Uncertainty\ Formalisms$$

$$Uncertainty\ Formalisms \times Technology\ Platforms = Uncertainty\ Technologies$$

This view, besides helping us understand the various components in current uncertain data solutions, opens multiple research possibilities investigating different theory platforms (e.g., possibility theory in contrast to probability theory) and data models (e.g., RDF data models). It was clear from the discussions that there are research opportunities in these directions that go significantly beyond current approaches, which focus primarily on supporting a limited number of uncertainty concepts in relational databases.

2 Current Uncertainty Models

Various research groups have extensively studied a part of the space in Section 1, focusing primarily on the following aspects:

- Treating tuple membership uncertainty and/or value uncertainty as the basic uncertainty concepts.
- Using relational data engines as the underlying database technology with fewer systems that handle XML data.
- Adopting probability theory to describe and analyze the different uncertain entities modeled as random variables.
- Providing efficient algorithms and techniques to compute answers of a special class of queries, usually inspired by some application domain.

In this working group, the participants tried to list a majority of the current system prototypes. Examples include ORION [1], TRIO [2–4], MYSTIQUE [5–7], MCDB [8], MayBMS [9, 10], BayesStore [11], and PrDB [12]. While these systems share the aforementioned aspects, they adopt different approaches in representing and manipulating uncertain data. Attempting to contrast these systems, the group identified the main features in which these systems show differences. We summarize these features as follows:

- *Supported probability distributions.* Supporting discrete distributions, continuous distributions or both has a direct effect on the class of supported applications and data types. For example, while continuous distributions will be the best to capture uncertainty in sensor readings, discrete probability distributions can capture possible values of uncertain categorical attributes. Models that depend on finite possible world semantics [4, 9] will have technical challenges in modeling continuous random variables, while sampling based techniques as in MCDB [8] can handle continuity in a natural way.

– *Models closure and decomposition.* Uncertainty models that are closed under relational query operators (e.g.,[3] studies the closure of different working uncertainty models under select, project and join operators) offer more flexibility in query optimization and rewrite.
– *Modeling correlated data.* While the ability to model correlated uncertain events widens the scope of supported applications, correlation greatly complicates analysis and query processing. Several proposed systems assume data independence among uncertain data values to simplify analysis, while other systems focus on how to efficiently model these correlations.
– *Space complexity.* Succinct representation of uncertain data models was the main focus of multiple research efforts. Examples include: (a) using compact graphical models to succinctly represent the joint probability distribution of data items [11], (b) using factorization to identify commonalities in uncertainty representation [12], and (c) optimizing lineage and provenance information to succinctly capture data generation and dependency sources [7].
– *Approximate results.* The topic of approximate answers in probabilistic and uncertain data management is controversial. Multiple research efforts define "exact" answers in the context of probabilistic data as producing uncertain answers with exact probability values [5, 4]. For example, Employee "Amy" is part of the query answer with probability 0.72. In contrast, other research efforts (e.g., sampling based systems [8] and approximate lineage based systems [7]) advocate advocate for producing results with approximate probability values since the probability values that are input to an uncertainty management system are often themselves approximate, uncertain, or crudely determined.

References

1. Cheng, R., Kalashnikov, D.V., Prabhakar, S.: Evaluation of Probabilistic Queries over Imprecise Data in Constantly-Evolving Environments. Inf. Syst. **32**(1) (2007) 104–130
2. Widom, J.: Trio: A System for Integrated Management of Data, Accuracy, and Lineage. In: Proceedings of the Second Biennial CIDR. (2005) 262–276
3. Sarma, A.D., Benjelloun, O., Halevy, A., Widom, J.: Working Models for Uncertain Data. In: Proceedings of the 22nd ICDE. (2006) 7
4. Benjelloun, O., Sarma, A.D., Halevy, A., Widom, J.: ULDBs: Databases with Uncertainty and Lineage. In: Proceedings of the 32nd VLDB. (2006) 953–964
5. Dalvi, N., Suciu, D.: Efficient query evaluation on probabilistic databases. The VLDB Journal **16**(4) (2007) 523–544
6. Ré, C., Dalvi, N.N., Suciu, D.: Efficient top-k query evaluation on probabilistic data. In: Proceedings of the 23rd ICDE. (2007) 886–895
7. Ré, C., Suciu, D.: Approximate lineage for probabilistic databases. PVLDB **1**(1) (2008)
8. Jampani, R., Xu, F., Wu, M., Perez, L.L., Jermaine, C., Haas, P.J.: Mcdb: a monte carlo approach to managing uncertain data. In: SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data. (2008) 687–700

9. Antova, L., Koch, C., Olteanu, D.: 10^{10^6} worlds and beyond: Efficient representation and processing of incomplete information. In: ICDE. (2007) 606–615
10. Antova, L., Koch, C., Olteanu, D.: From complete to incomplete information and back. In: SIGMOD Conference. (2007) 713–724
11. Wang, D.Z., Michelakis, E., Garofalakis, M.N., Hellerstein, J.M.: Bayesstore: managing large, uncertain data repositories with probabilistic graphical models. PVLDB **1**(1) (2008) 340–351
12. Sen, P., Deshpande, A., Getoor, L.: Exploiting shared correlations in probabilistic databases. PVLDB **1**(1) (2008) 809–820