Probabilistic Data Integration

Maurice van Keulen
Faculty of EEMCS, University of Twente, Enschede, The Netherlands.
m.vankeulen@utwente.nl

November 3, 2017

Probabilistic data integration is a specific kind of data integration where integration problems such as inconsistency and uncertainty are handled by means of a probabilistic data representation.

The approach is based on the view that data quality problems (as they occur in an integration process) can be modeled as uncertainty [1] and this uncertainty is considered an important result of the integration process [2]. In a sense, data quality problems arising during the data integration process are not solved immediately, but explicitly represented in the resulting integrated data. This data can be stored in a probabilistic database to be queried directly resulting in possible or approximate answers [3]. A *probabilistic database* is a specific kind of DBMS that allows storage, querying and manipulation of uncertain data. It keeps track of alternatives and dependencies among them.

While traditional data integration methods more or less explicitly consider uncertainty as a problem, as something to be avoided, probabilistic data integration treats uncertainty as an additional source of information, which is precious and should be preserved [2]. It effectively allows for postponement of solving data integration problems. When combined with an effective method for data quality measurement, it also has the potential to allow for a pay-as-you-go and good-is-good-enough approach where small iterations reduce overall effort in improving the data quality of the integrated result.

In this presentation, we give an overview of various data integration problems and how a probabilistic approach can improve them, for example, entity resolution [4] and merging of grouping data [5]. We furthermore illustrate how probabilistic data integration as an application asks for more theoretical research on probabilistic database technology, such as more expressive data models and (approximate) querying formalisms. In particular, we present the problem of incorporation of a restricted notion of higher orderedness in datalog without loosing its important properties.

References

[1] M. van Keulen. Managing uncertainty: The road towards better data interoperability. *IT - Information Technology*, 54(3):138–146, 2012.

[2] M. Magnani and D. Montesi. A survey on uncertainty management in data integration. *JDIQ*, 2(1):5:1–5:33, 2010.

[3] N. Dalvi, C. Ré, and D. Suciu. Probabilistic databases: Diamonds in the dirt. *Communications of the ACM*, 52(7):86–94, July 2009.

[4] F. Panse, M. van Keulen, and N. Ritter. Indeterministic handling of uncertain decisions in deduplication. *JDIQ*, 4(2):9:1–9:25, 2013.

[5] B. Wanders, M. van Keulen, and P. van der Vet. Uncertain groupings: Probabilistic com-

bination of grouping data. In *Proc. of DEXA*, LNCS 9261, pages 236–250. Springer, 2015.