On entity resolution in probabilistic data

Ayat, S.N.

Citation for published version (APA):
In this chapter, we first revisit the research questions posed in the introduction and discuss how we answered them in our research work presented in Chapters 3 to 6. We then discuss a number of possible extensions and future work for the presented research of this dissertation.

7.1 Answers to Research Questions

The first question posed in the introduction of this thesis was the following.

1. How to match a probabilistic entity against a set of probabilistic entities, while considering both their similarity and probability? In other words, what are the semantics of identity resolution problem over probabilistic data?

We addressed this question in Chapter 3. We adapted the possible worlds semantics of uncertain data to define the novel concepts of most-probable matching pair (MPMP) and most-probable matching entity (MPME), together referred to as most-probable matches.

Chapter 3 also dealt with the second posed question, which was:

2. How can we efficiently deal with the identity resolution problem over probabilistic data?

To propose an efficient solution for computing the most-probable matches, in Chapter 3 we differentiated between two classes of similarity functions: i.e. context-free and context-sensitive. For context-free similarity functions, we proposed the CFA algorithm, which simultaneously computes MPMP and MPME concepts in PTIME. Moreover, we proposed two optimized versions of the CFA algorithm, i.e. CFA-MPMP and CFA-MPME, which respectively compute MPMP
and MPME concepts. For context-sensitive similarity functions, we used the Monte-Carlo approximation algorithm. To speedup the Monte-Carlo algorithm, we proposed a parallel version of it using the MapReduce framework. Moreover, to overcome the high response time of most context-sensitive similarity functions in the literature, which makes them very inefficient for the Monte-Carlo algorithm, we proposed the novel CB similarity function with the following salient features:

- CB is very efficient compared to other context-sensitive similarity functions because it significantly reduces the number of rather costly string comparison operations by working at the attribute level, rather than at the word or q-gram level.

- In contrast to most of the tuple matching methods that work at the attribute level, CB is self-tuning, meaning that it does not need the specification of weights for representing the relative importance of individual attributes.

The third posed research question was the following.

3. How can we efficiently deal with the identity resolution problem over probabilistic data in distributed systems?

Chapter 4 deals with this question. In this chapter, we proposed the FD, a fully distributed algorithm for dealing with the identity resolution problem over distributed probabilistic data, with the objective of minimizing network traffic. FD uses the novel concepts of potential and essential-set to prune data at local nodes. This leads to a significant reduction in network traffic and response time compared to the baseline approaches. FD requires no global information, and does not depend on the existence of certain nodes.

Chapter 5 dealt with the fourth posed question, which was:

4. Does deduplication necessarily improve the quality of a probabilistic database? If not, then how can we improve the quality of a probabilistic database through deduplication?

In Chapter 5, we used entropy as a quality metric for measuring the quality of a probabilistic database, where the higher the entropy of a database, the lower is its quality. We showed that if entropy is not taken into account, deduplication does not necessarily improve the quality of a probabilistic database. Thus, to guarantee the quality improvement, we modeled deduplication problem over probabilistic data as an entropy minimization problem. We then proposed an efficient solution for the deduplication problem in probabilistic data as follows:

- We proposed an efficient technique for computing the entropy of a probabilistic database in the x-relation probabilistic data model [10].
7.1. Answers to Research Questions

- We proposed a merge function for merging x-tuples in the x-relation data model.

- The properties of our proposed merge function, i.e. commutative and associative, as well as entropy-reduction properties, enabled us to propose a PTIME algorithm for dealing with the deduplication problem, and producing a cleaned database with (near) minimum entropy.

The last research question was:

5. How effectively can we deal with the schema matching problem in a fully automated setting?

We answered this question in Chapter 6, where we dealt with the schema matching problem in setting up a fully automated data integration system, denoted by the IFD, from a number of heterogeneous data sources. IFD has two important features which allow it to effectively deal with the schema matching problem. First, IFD is built on a probabilistic data model in order to capture the uncertainty that arises during the schema matching process. Second, IFD takes advantage of the background knowledge which is implied in functional dependencies for finding attribute correlations and using it for matching the source schemas. We showed that it is possible to achieve a fairly high accuracy in dealing with the schema matching problem in a fully automated setting. This lets us to effectively deal with the entity resolution problem when data resides in heterogeneous data sources.

Having answered the sub-questions, we come back to the main research question of this thesis.

How can we, effectively and efficiently, deal with the entity resolution problem for probabilistic data?

As discussed in Chapter 2, dealing with the entity resolution problem, both on deterministic and probabilistic data, greatly depends on the used similarity function. Our proposed methods for the entity resolution problem over probabilistic data is generic, which can thus be applied to any similarity function, suitable for the application in hand. On the other hand, to efficiently deal with the entity resolution problem over probabilistic data, we need to avoid enumerating the possible worlds of uncertain data. Our efficient methods heavily rely on the properties of the x-relation data model, which results in the PTIME time complexity of all of our proposed techniques, except the Monte-Carlo algorithm in Section 3.4. Exploiting the properties of other probabilistic data models for efficient handling of entity resolution problem over them however, remains as a possible direction for future research.
7.2 Future Work

While this thesis has made a number of contributions to the problem of entity resolution over probabilistic data, the general problem still is open. This work, in fact, opens the following directions for future research.

**ERPD over other data models.** First, as discussed in previous section, while the defined semantics of the ERPD problem is not specific to any probabilistic data model, we relied on the properties of the x-relation probabilistic data model for efficient dealing with this problem. Efficient entity resolution over other probabilistic data models however, or even other uncertain data models, is one possible direction for future research.

**Using CB in blocking methods.** Due to its efficiency and effectiveness, we believe that our CB similarity function presented in Chapter 3, can act as a cheap similarity metric in the blocking methods, aiming to further improve the efficiency of ER (see Section 2.2.2). Elaborating on this idea is another possible direction for future research.

**ERPD for special distributed systems.** In Chapter 4, we assume a very general topology for distributed system. However, in some applications, probabilistic data might be fragmented over a distributed system with a particular topology, e.g. distributed hash tables. Attacking the ERPD problem in such distributed systems is a possible future research direction.

**Using functional dependencies in other schema matchers.** Finally, in Chapter 6, we showed that using the background knowledge implied within functional dependencies alone can significantly improve the quality of schema matching. On the other hand, there exist many schema matchers in the literature that use other features, e.g. data types and value ranges, to match the schemas. Integrating other existing high quality schema matchers with our schema matching heuristics is another possible direction for future research.