UNIVERSITY OF AMSTERDAM

On Entity Resolution for Probabilistic Data
S.N. Ayat

Abstract

Entity resolution (ER) is the problem of identifying duplicate tuples, which are the tuples that represent the same real-world entity. There are many real-life applications in which the ER problem arises. These applications range from *news* aggregation websites, identifying the news that cover the same story, in order to avoid presenting one story several times to the user, to the integration of two companies' customer databases in the case of a merger, where identifying the tuples that refer to the same *customer* is crucial.

Due to its diverse applications, the ER problem has been formulated in different ways in the literature. The two main ER's related problem formulations include: 1) *identity resolution*, and 2) *deduplication*. In identity resolution, the aim is to find duplicate(s) of a given tuple in a given database, while in deduplication, the aim is to find groups of duplicate tuples in a given database, and merge them in order to increase the quality of the database itself.

The ER problem is however not limited to deterministic (ordinary) databases, rather it also arises in applications that deal with probabilistic databases, i.e. databases in which each tuple or attribute value is associated with a probability value to, for instance, indicate its confidence level. In this thesis, we study the ER problem in probabilistic databases. More specifically, we address five challenges described in the following paragraphs.

The first challenge is that in contrast to deterministic data, in probabilistic data, the semantics of identity resolution problem is not clear. In identity resolution over *deterministic* data, the aim is to match the *most similar* tuple in the database to a given tuple. However the aim is not so clear when matching probabilistic entities, since we have to deal with the two concepts of the *most similar* and the *most probable*, at the same time.

Efficient dealing with the identity resolution problem in probabilistic data is the second challenge that we address in this thesis. In order to define the semantics of the identity resolution problem over probabilistic data, we use the *possible worlds* semantics of uncertain data, treating a probabilistic database as

the probability distribution over a set of *deterministic* database instances, each of which is called a *possible world*. Each possible world thus, is a deterministic database which occurs with certain probability. The number of possible worlds of a probabilistic database might easily be exponential, which thus makes the naïve computation of the defined semantics impractical.

In many applications that the identity resolution problem arises, probabilistic data is distributed among a number of nodes. Dealing with the identity resolution problem in distributed probabilistic data, while reducing the amount of exchanged data among nodes, is quite challenging. Efficient dealing with the identity resolution problem over probabilistic data in *distributed systems* is the third challenge that we address in this thesis.

The fourth challenge is raised by considering the *deduplication* problem in probabilistic data. Similar to *deterministic* data, the aim of deduplication in probabilistic data is to improve the quality of the database. However, we observe that deduplication does not necessarily improve the quality of the probabilistic database. Therefore, guaranteeing the quality improvement of probabilistic databases by a deduplication approach is another challenge that we address in this thesis.

In many applications where the ER problem arises, data resides in a number of heterogenous data sources. In such applications, matching the heterogeneous schemas of data sources, to which we refer as *schema matching*, is an inevitable step in dealing with the ER problem. On the other hand, dealing with the schema matching problem requires human knowledge about the context, which is in contrast to the full automated resolution setting, which we propose for applications in which the ER problem arises. Thus, effective dealing with the schema matching problem in a fully automated setting is the fifth challenge addressed in this thesis.

The thesis is structured as follows.

In Chapter 1, we elaborate on the motivation of this work, and present the research questions and contributions of this research.

Chapter 2 provides preliminary definitions and concepts that are used throughout the thesis. We first present the uncertain data models, and then review the related work on entity resolution area.

In Chapter 3, we deal with the identity resolution problem over probabilistic data. We adapt the possible worlds semantics of uncertain data to define the semantics of identity resolution problem in probabilistic data. Our approach for computing the defined semantics depends of the similarity function, which is used for computing the similarity between tuples. We differentiate between two classes of similarity functions, i.e. *context-free* and *context-sensitive*. We propose a PTIME algorithm for context-free similarity functions, and a Monte Carlo approximation algorithm for the context-sensitive similarity functions. We deal with the problem of high response time of existing context-sensitive similarity functions, which makes them very inefficient for the Monte Carlo algorithm, by

proposing a new efficient context-sensitive similarity function that is very appropriate for the Monte Carlo algorithm. We further speed up our proposed Monte Carlo algorithm by parallelizing it using the MapReduce framework.

Chapter 4 deals with the identity resolution problem over distributed probabilistic data. We propose a fully distributed algorithm for computing the semantics of the identity resolution problem, as defined in Chapter 3, in a distributed system. Our algorithm prunes data at local nodes, which thus results in significant reduction in bandwidth usage and the response time compared to the baseline approaches. Moreover, it requires no global information, and does not depend on the existence of certain nodes.

In Chapter 5, we deal with deduplication problem over probabilistic data with the aim of improving the quality of probabilistic data. We use the amount of uncertainty, i.e. entropy, of the probabilistic database as a quality metric and propose an efficient technique for computing it. We then propose a merge function for merging probabilistic duplicate tuples. Further, we propose an efficient algorithm that uses our proposed merge function and produces a cleaned database with (near) minimum entropy. This leads to a significant improvement in the results of the queries which are posed over the database.

Chapter 6 aims at building a data integration system in a fully automated setting. The main problem that is dealt with in this chapter is the schema matching problem, which arises in many applications that need to deal with the entity resolution problem. We propose an algorithm that takes advantage of the background knowledge implied in *functional dependencies* for finding attribute correlations and using it for matching the source schemas. and generating the mediated schema. Our algorithm is built on a probabilistic data model in order to model the uncertainty in data integration systems.

Finally, Chapter 7 concludes and gives some research directions for future work.