Entity Resolution for Distributed Uncertain Data

Naser Ayat[#], Reza Akbarinia[*], Hamideh Afsarmanesh[#], Patrick Valduriez[*]

[#]Informatics Institute, University of Amsterdam, Amsterdam, Netherlands
[*]INRIA and LIRMM, Montpellier, France

1 Abstract

The main difference between a traditional certain database and an uncertain (probabilistic) database is that an uncertain database represents a set of possible database instances, called *possible worlds*, rather than a single one.

An important problem that arises in many applications such as information integration is that of Entity Resolution (ER).ER is the process of identifying tuples that represent the same real-world entity. The problem of ER is challenging since the same entity can be encoded in different ways due to a variety of reasons such as different formatting conventions, abbreviations, and typographic errors.

The problem of *entity resolution over uncertain data* (which we call ERUD) arises in many application domains that have to deal with uncertain data, ranging from sensor databases to scientific data management. In this paper, we are interested in the following formulation of the ERUD problem[1] [1]. Let e be an uncertain entity represented by multiple possible alternatives, i.e. tuples, each with a membership probability. Let D be an uncertain database composed of a set of uncertain entities. Then, given e, D, and a similarity function F, the problem is to find the entity-tuple pair (t, t_i) (where $t \in e, t_i \in D$) such that (t, t_i) has the highest cumulative probability to be the most similar in all possible worlds.

There have been recent proposals dealing with the ERUD problem [2, 3, 1]. They all deal with the uncertain data which is stored in a central database. However, many real-life applications, in which the ERUD problem arises, produce uncertain data distributed among a number of databases. Dealing with the ERUD problem for distributed data is quite important for such applications.

A straightforward approach to answer the above queries is to ask all distributed nodes to send their databases to a central node who deals with the problem of ER by using one of the existing centralized solutions, e.g. [1]. However, this approach is very expensive and does not scale well neither in the size of the local databases, nor in the number of nodes. Therefore, using a distributed algorithm for dealing with the ERUD problem over distributed data is inevitable.

In this paper, we propose FD, a fully distributed algorithm for dealing with the ERUD problem over distributed data, with the goal of minimizing bandwidth usage and reducing processing time. To the best of our knowledge, FD is the first proposal that deals with the ERUD problem over distributed data. It has the following salient features. First, it uses the novel concepts of *Potential* and *essential-set* to prune data at local nodes. This leads to a significant reduction of bandwidth usage compared to the baseline approaches. Second, its execution is completely distributed and does not depend on the existence of certain nodes. We validated FD through simulation and the results show very good performance, in terms of bandwidth usage and response time.

References

1. N. Ayat, R. Akbarinia, H. Afsarmanesh, and P. Valduriez. Entity resolution for uncertain data. In *BDA*, 2012.
2. D. Menestrina, O. Benjelloun, and H. Garcia-Molina. Generic entity resolution with data confidences. In *Proc. of CleanDB*, 2006.
3. F. Panse, M. van Keulen, A. de Keijzer, and N. Ritter. Duplicate detection in probabilistic data. In *Proc. of ICDE Workshops*, 2010.

[1] Some texts refer to this formulation of the ER problem as *Identity Resolution*.