Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Entity resolution for probabilistic data ☆

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

[a] Informatics Institute, University of Amsterdam, Amsterdam, Netherlands
[b] INRIA and LIRMM, Montpellier, France

ARTICLE INFO

Article history:
Received 5 February 2013
Received in revised form 27 September 2013
Accepted 25 February 2014
Available online xxxx

Keywords:
Entity resolution
Probabilistic database

ABSTRACT

Entity resolution is the problem of identifying the tuples that represent the same real world entity. In this paper, we propose a complete solution to the problem of entity resolution over probabilistic data (ERPD), which arises in many applications that have to deal with probabilistic data. To deal with the ERPD problem, we distinguish between two classes of similarity functions, i.e. context-free and context-sensitive. We propose a PTIME algorithm for context-free similarity functions, and an approximation algorithm for context-sensitive similarity functions. We validated our algorithms through experiments over both synthetic and real datasets. Our extensive performance evaluation shows the effectiveness of our algorithms.

1. Introduction

In recent years, we have been witnessing much interest in uncertain data management in many application areas such as data integration [26,5], sensor networks [15], and information extraction [22]. Much research effort has been devoted to several aspects of uncertain data management, including data modeling [34,6], skyline queries [4], top-k queries [14,35,40,33], nearest neighbor search [41], XML documents [1], etc. Untrusted sources, imprecise measuring instruments, and uncertain methods, are some reasons that cause uncertainty in data. The main difference between a traditional *certain* database and an uncertain database is that an uncertain database represents a set of possible database instances, rather than a single one. Recent uncertain data models tend to represent data uncertainty with probabilistic events [3,34,6].

An important problem that arises in many applications such as information integration is that of Entity Resolution (ER) [19]. ER is the process of identifying tuples that represent the same real-world entity. For instance, consider the two tuples ("Thomas Michaelis", "45, Main street") and ("T. Michaelis", "45, Main st.") on the schema $S(name, address)$ that represent the same person with different conventions. An ER solution is expected to detect that both tuples represent the same 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 ER problem has been well studied in the literature for certain data (refer to [19] for a survey), but it has not been deeply investigated for probabilistic data. Before discussing existing solutions, let us first motivate the need for ER in probabilistic data (which we call ERPD), with two examples from scientific data management and anti-criminal domains.

☆ A preliminary version of this work has been presented to the French conference BDA.

* Corresponding author.

E-mail addresses: s.n.ayat@uva.nl (N. Ayat), Reza.Akbarinia@inria.fr (R. Akbarinia), h.afsarmanesh@uva.nl (H. Afsarmanesh), Patrick.Valduriez@inria.fr (P. Valduriez).

http://dx.doi.org/10.1016/j.ins.2014.02.135
0020-0255/© 2014 Elsevier Inc. All rights reserved.

Example 1 (Finding astronomical objects in astrophysics data). In astrophysics, as well as in other scientific disciplines, the correlation and integration of observational data is the key for gaining new scientific insights. Astronomical observatories produce data about sky surveys, most of which is probabilistic [32]. In its simplified form, an observatory maintains a single probabilistic relation *Objects* that contains data about the observed *astronomical objects* in the sky surveys. Each *object* is represented using a number of alternative tuples each with a membership probability showing its degree of certainty which is computed using some rules. The alternatives are mutually exclusive, meaning that at most one of them can be true. The probabilistic model, which is used in this example, has been widely used in the database literature for representing probabilistic data, e.g. [34,40,7,11], and also for representing astrophysics data [36]. The astrophysics researchers who want to track a particular astronomical object, which has been represented by an uncertain entity *e*, are very interested in queries like the following: *among astronomical objects observed in a sky region, find the object which is most probably the same object as e*.

Example 2 (Suspect detection in anti-criminal police database). The anti-criminal police is faced with many crimes every year. It spends a lot of time and money gathering data about every crime from different sources such as witnesses, interrogations, and police's informants. Some of the gathered data are not certain, for some reasons, e.g. the police cannot completely trust informants and witnesses. To represent this uncertainty, probability values can be attached to the data to show their likelihood of truth according to the confidence on the sources. These probabilistic data are used to find possible suspects, and can greatly speed up the investigation process. In a very simplified form, the police maintains a single relation *Suspects* that contains data about suspects. In this relation, each individual suspect is represented using an entity consisting of a number of alternative tuples each associated with a probability value showing its likelihood of truth. When a crime occurs, detectives gather data about the perpetrator, and the gathered data can be represented in the form of an uncertain entity, say *e*. To get more information about the perpetrator represented by *e*, detectives are interested in the following query: *find in the uncertain database, the person who is most probably the same person as e.* Notice that if there is more than one perpetrator, the police can represent each one using an uncertain entity and repeat the process for each entity.

Existing proposals for the ER problem are not applicable to the above examples since they ignore probability values completely and return the most similar tuples as the solution. To the best of our knowledge, the works presented in [28,31] are the only proposals for dealing with the ERPD problem. The proposal in [28] ignores the probability values and the proposal in [31] uses the probability values only for normalizing the similarity of the uncertain entities. These proposals do not define the ERPD problem based on both similarity and probability of tuples. Consequently, they are not appropriate for answering the queries issued in the above examples.

Inspired by the literature on uncertain data management, in this paper we adopt the well-known possible worlds semantics for defining the semantics for the ERPD problem and propose efficient algorithms for computing it. However, developing an efficient solution for the ERPD problem is challenging, particularly due to the following reasons. First, we must take into account two parameters for matching the entities: the similarity and the probability values. Second, due to the uncertainty in the entity, there may be different similarity values between the entity and the tuples of database. Third, in the case of *context-sensitive* similarity functions (see the definition in Section 2.2), the similarity of two entities may be different in different possible worlds, i.e. in different instances of the database. A naïve solution for ERPD involves enumerating all possible worlds of the uncertain entity and the database. However, this solution is exponential in the number of tuples of the database.

In this paper, we address the ERPD problem and propose a complete solution for it. Our contributions are summarized as follows:

- We adapt the possible worlds semantics of probabilistic data to define the problem of ERPD based on both similarity and probability of tuples.
- We propose a PTIME algorithm for the ERPD problem. This algorithm is applicable to a large class of the similarity functions, i.e. *context-free* functions. For the rest of similarity functions (i.e. *context-sensitive*), we propose a Monte Carlo approximation algorithm.
- We deal with the problem of significant setup time in existing context-sensitive functions, which makes them very inefficient for the Monte Carlo algorithm. We propose a new efficient context-sensitive similarity function that is very appropriate for the Monte Carlo algorithm.
- We propose a parallel version of our Monte Carlo algorithm using the MapReduce framework.

We conducted an extensive experimental study to evaluate our approach for ERPD over both real and synthetic datasets. The results show the effectiveness of our algorithms. To the best of our knowledge, this is the first study of the ERPD problem that adopts the possible world semantics and develops efficient algorithms for it.

The rest of the paper is organized as follows. In Section 2, we present our data model, and define the problem we address. In Section 3, we propose our solution for the ERPD problem for context-free similarity functions. In Section 4, we propose our solution for dealing with the ERPD problem when the similarity function is context-sensitive. In Section 5, we report the performance evaluation of our techniques over synthesis and real data sets. Section 6 discusses related work. Section 7 concludes.

Please cite this article in press as: N. Ayat et al., Entity resolution for probabilistic data, Inform. Sci. (2014), http://dx.doi.org/10.1016/j.ins.2014.02.135