Chapter 4

Context-Based Entity Resolution

ABSTRACT

Prior work of entity resolution involves expensive similarity comparison and clustering approaches. Additionally, the quality of entity resolution may be low due to insufficient information. To address these problems, by adopting context information of data objects, the authors present a novel framework of entity resolution, Context-Based Entity Description (CED), to make context information help entity resolution. In this framework, each entity is described by a set of CEDs. During entity resolution, objects are only compared with CEDs to determine its corresponding entity. Additionally, the authors propose efficient algorithms for CED discovery, maintenance, and CED-based entity resolution. The authors experimentally evaluated the CED-based ER algorithm on the real DBLP datasets, and the experimental results show that this algorithm can achieve both high precision and recall as well as outperform existing methods.

INTRODUCTION

Entity resolution (ER) is the process of identifying objects that refer to the same real-world entity. ER is one of the most important problems in data cleaning and arises in many applications. For example, in publications, two identical author names may represent different persons and an author may have different name spellings; same address might have multiple entries due to different spelling; a person can have quite different descriptions on web.

Because of its importance, ER has attracted general attention in the literatures. Most algorithms and frameworks of ER involve two steps. One is similarity comparison between objects (Koudas, Sarawagi & Srivastava 2006), the other is clustering objects (Fan, Wang, Pu, Zhou & Lv 2011, Bansal, Blum & Chawla 2002, Chaudhuri, Ganjam, Ganti & Motwani 2003), in which objects in the same cluster represent the same entity. Even though existing methods can perform ER effectively in many cases, these ER approaches have following five limitations.

1. **Lack of Consideration of Updating:** In many systems, the number of objects increases rapidly. If we search one of the hottest products iPhone 4 on eBay, there are more than 2.5 million results; more than 1.5 million publications have been added in DBLP since 2010. Unfortunately, current ER approaches seldom consider the update problem. Both comparing new objects with old ones and re-clustering are quite expensive.
2. **Expensive Time Complexity:** Computing the similarity between pairs of objects is too expensive when the data size becomes large. Moreover, since the similarity relationship between objects is not transitive, after the pair-wise matching between objects, the clustering step has to be involved next, which is also an expensive operation.

3. **Order Dependent:** Since the similarity relationships are not transitive, the results of ER can be order dependent, which means if different matching orders are selected, the clustering results might be different as well. As the result, inconsistency exists in ER results.

4. **Sensitive to Training Data:** For learning-based strategies of ER (Chaudhuri, Chen, Ganti & Kaushik, R. 2007, Arasu, A., Chaudhuri, S. and Kaushik, R. 2009, Singla, P., Domingos, P. 2005, Verykios, V. S., Moustakides, G. V., Elfeky, M. G. 2003), they usually assume the training data is similar to the test data. In real world, since data can be obtained from various sources continuously, for an entity $e$, more and more objects referring to $e$ with new attributes and quite different values from the old objects may come, thus this assumption cannot always be guaranteed.

5. **Monotonicity Assumption of Similarity Functions:** The similarity functions can be divided into two kinds. One is feature-based similarity, the other is relational similarity. Most approaches of ER are based on feature similarity comparison between object pairs. They are all based on the intuition of the monotonicity of similarity, which means that the matching pairs in real world are more similar in features than the non-matching pairs. We will soon use Example 1 to show that this intuition is incorrect sometimes. Differently with feature similarity, (Fan, X., Wang, J., Pu, X., Zhou, L., Lv, B. 2011, Bansal, N., Blum, A., Chawla, S. 2002, Chaudhuri,S., Ganjam, K., Ganti, V., Motwani, R. 2003) focus on the relationships among objects. They solve the ER problem on a relationship graph, which requires a quite large space and as far as we know, they are not as efficient as feature-based similarity approaches.

We use Example 1 to illustrate the limitations of prior work.

**Example 1:** Consider 8 authors $o_{11}-o_{32}$ with name “wei wang” in Table 1 corresponding to persons $e_1-e_3$ in real world. We suppose each data object $o_i$ refers to the entity $e_i$. The task of entity resolution is to identify the person each author referring to with the information in Table 1. Since all these objects have identical name but different coauthor set in different papers, the similarity between any two objects $o_i$ and $o_j$, denoted by $\text{Sim}(o_i,o_j)$,