Chapter 15
Entity Resolution in Bibliography Information Management

ABSTRACT
Entity resolution, that is to build corresponding relationships between objects and entities in dirty data, plays an important role in data cleaning. In bibliography information management system, the confusion between authors and their names often results in dirty data. That is, different authors may share the identical name, and different names may correspond to the identical author. Therefore, the major task of entity resolution is to distinguish entities sharing the same name and recognize different names referring to the same entity. However, current research focuses on only one aspect and cannot solve the problem completely. To address this problem, in this chapter, EIF, a framework of entity resolution with the consideration of both kinds of confusions, is proposed. With effective clustering techniques, approximate string matching algorithms, and a flexible mechanism of knowledge integration, EIF can be widely used to solve many different kinds of entity resolution problems. In this chapter, as an application of EIF, the authors solve the author resolution problem. The effectiveness of this framework is verified by extensive experiments.

INTRODUCTION
In bibliography information management, entities are often queried by their names. For example, researchers are often queried by their names on dblp. Unfortunately, dirty data often lead to incomplete or duplicated results for such queries. From different aspects, there are two major problems. On the one hand, a name may have different spellings and one entity can be represented by multiple names. For example, the name of a researcher "Wei Wang" can be both written as “Wei Wang” and “W. Wei”. Another example of this confusion is movie names. Such as the movie called “Hong lou meng” can also be represented as “A Dream in Red Mansions”. On the other hand, one name can represent multiple entities. For example, when querying an author named “Wei Wang” in dblp, the database system will output seven different authors all named “Wei Wang”. In this chapter, the former problem is called name variant for brief and the latter is called name sharing.

DOI: 10.4018/978-1-4666-5198-2.ch015
Entity resolution techniques are to deal with these problems. This is a basic operation in data cleaning and query processing with quality assurance. Given a set of objects with name and other properties, the goal of this operation is to split the set into clusters, such that each cluster corresponds to one real-world entity.

Some techniques for entity resolution have been proposed. Some of them have been introduced in Chapter 3 and Chapter 9. However, each of these techniques focuses on one of the two problems. The techniques for the first problem are often called “duplicate detection” (Newcombe, Kennedy & Axford 1959). These techniques usually find duplicate records by measuring the similarity of individual fields (e.g. objects with similar names). Different approaches are used to compare the similarity. These techniques are based on the assumption that duplicate records should have equal or similar values. With the second problem existing simultaneity, records with the same name referring to different entity cannot be distinguished. As far as we know, the only technique for the second problem is presented in (Yin, Han & Yu 2007). It identifies entities using linkage information and clustering method. From the experiment in (Yin, Han & Yu 2007), it takes a long time for object distinction; therefore it is not suitable for entity resolution on large datasets. Besides, this method distinguishes objects by assuming that the objects have identical names. If this method is used to solve entity resolution problem, in which the assumption is unsatisfied, the results might be inaccurate. In summary, when these two aspects of problems both exist, current techniques cannot distinguish objects effectively.

For entity resolution in general cases, new techniques with the ability of dealing with both these problems are in demand. For effective entity resolution, this chapter proposes EIF, an entity resolution framework. With effective clustering techniques, approximate string matching algorithms and a flexible mechanism of knowledge integration, EIF can deal with both the two aspects of the problems. Given a set of objects, EIF split them into clusters, such that each cluster corresponds to one entity. In this chapter, as an application of EIF, we process an author resolution algorithm for identifying authors from the database with dirty data. For the simplicity of discussion, in this chapter, we only focus on relational data. The techniques in this chapter can also be applied to semi-structured data or data in OO-DBMS by representing each object as a tuple of attributes. The content of this chapter is from Li, Wang, Gao & Li (2010).

The contributions of this chapter can be summarized as following:

1. EIF, a general entity resolution framework by using name and other attributes of objects, is presented in this chapter. Both approximate string matching and clustering techniques can be effectively embedded into EIF, domain knowledge integration mechanism as well. This framework can deal with both name variant and name sharing problems. As we know, it is the first strategy with the consideration of both problems.
2. As an application of EIF, an author resolution algorithm is proposed by using the information of author name and co-authors to solve author resolution problem. It shows that by adding proper domain information, EIF is suitable to process problems in practice.
3. The effectiveness of this framework is verified by extensive experiments. The experimental results show that the author resolution algorithm based on EIF outperforms the existing author resolution approaches both in precision and recall.

The rest of this chapter is organized as follows. The entity resolution framework EIF is introduced in Section 2. In Section 3, we demonstrate how to apply EIF on author resolution. Related work