Chapter 7
XML Object Identification

ABSTRACT

For the ability to represent data from a wide variety of sources, XML is rapidly emerging as the new standard for data representation and exchange on Web and e-government. To effectively use XML data in practice, entity resolution, which has been proven extremely useful in data fusion, inconsistency detection, and data repairing, must be in place to improve the quality of the XML data. In this chapter, the authors deal specifically with object identification on XML data, the application of which includes XML document management in highly dynamic applications like the Web and peer-to-peer systems, detection of duplicate elements in nested XML data, and finding similar identities among objects from multiple Web sources. The authors survey techniques of pairwise and groupwise entity resolution for XML data, which adopt structured information to describe the similarity or distance of XML data, like XML document and XML elements in document, and find the matching pairs which describe same object or classify them into separate groups, each group corresponding to the same object in real world. There are a lot of ways to describe the XML structure and content, such as a tree, Bayesian network, and set. The authors introduce some well-known algorithm base on these structures to solve matching XML data problems. Finally, the authors discuss directions for future research.

INTRODUCTION

The object identification problem is a central problem arising in data cleaning and data integration, where different objects must be compared to determine if they refer to the same real-world entity, even in the presence of errors such as misspellings. As the spread of the XML format as a data model increases, the need to develop effective strategies for XML object identification grows.

For the ability to represent data from a wide variety of sources, XML is rapidly emerging as the new standard for data representation and exchange on web and e-government etc. There are quantities of data represented in XML forms in information system nowadays, and XML document has becomes the standard of describing information of integrating different kind of data from sources in which XML data has different schema or there exists different forms of the same object. Query on data which has not been processed will lead to duplicate or inconsistent results. Statistics analysis on such data will count one object many times, which causes wrong decision making. If integrating XML data using object identification
method, making each class describing the same object, it will improve the quality of query and accuracy of statistics.

Although there exist a lot of object identification techniques, they mainly focus on duplicate record in literature or in a database. The research on object identification of complex data especially those in XML form is starting recently, the application of which mainly includes XML document management in highly dynamic applications like the Web and peer-to-peer systems, detection of duplicate elements in nested XML data and finding similar identities among objects from multiple web sources.

Our focus in this chapter is on object identification on XML data, the rest of the paper is organized as follows. We begin with a brief overview of the definition of object identification on complex data and present our framework for pairwise and groupwise entity resolution on XML. Next we introduce several algorithms to solve corresponding problems separately. We conclude with a discussion of the difficulties and challenges when solving XML object identification problem.

Since XML has been widely used in information integration due to its flexibility, entity resolution techniques could be applied to information integration based on XML (Chapter 14).

**BACKGROUND**

According to the different identification results, object identification on XML data can be classified into two categories, namely pairwise entity and groupwise entity resolution. The former is to determine whether two object \( o_1 \) and \( o_2 \) describe the same object in the real world; the latter is to classify the data set \( S \) into subset \( S_1, S_2, ..., S_k \), which satisfy:

\[
S_1 \cap S_2 \cap ... \cap S_k = \emptyset \quad (1)
\]

and

\[
S_1 \cup S_2 \cup ... \cup S_k = S \quad (2)
\]

making that \( \forall i \in [1, k], \forall o_1, o_2 \in S_i \), they describe the same object in real world.

**MAIN FOCUS OF THE CHAPTER**

In pairwise entity resolution, which is also named as XML document matching or element matching, the main work concentrates on the similarity or distance of XML data. Compared with structured or unstructured data, the outstanding feature of XML data is its abundant structure information, with this respect, the most used matching approach is to describe the similarity or distance of XML document with structure information. There are many ways to describe the similarity of XML document structure, such as similarity of tree when using tree to simulate XML document structure, XMLDup system using Bayesian network similarity (Leit et al. 2007), similarity of sets when extract XML documents into set.

In groupwise entity resolution, people often adopt object identification method based on similarity function, which means by definition of distance function or similarity function sim, given two objects \( o_1, o_2 \) and threshold \( \varepsilon \), when \( \text{sim}(o_1, o_2) \geq \varepsilon \) or distance\((o_1, o_2) \leq \varepsilon \), \( o_1 \) and \( o_2 \) will be regarded as the same object. According to objects, XML group-wise entity resolution can be classified into two categories, document level identification and element level identification. The former is to classify XML documents, making each category describes same object, which can be done with either neighbour ordering method (Puhlmann S et al. 2006) or identification based on tuple matching; the latter is to classify elements in same XML document, making each category describing same entity.