Chapter 6
The Role of Schema and Document Matchings in XML Source Clustering

Pasquale De Meo
University of Messina, Italy

Giacomo Fiumara
University of Messina, Italy

Antonino Nocera
University Mediterranea of Reggio Calabria, Italy

Domenico Ursino
University Mediterranea of Reggio Calabria, Italy

ABSTRACT

In recent years, there has been an increase in the volume and heterogeneity of XML data sources. Moreover, these information sources are often comprised of both schemas and instances of XML data. In this context, the need of grouping similar XML documents together has led to an increasing research on clustering algorithms for XML data. In this chapter, we present an overview of the most popular methods for clustering XML data sources, distinguishing between the intensional data level and the extensional data level, depending whether the sources to cluster are DTDs and XML schemas, or XML documents; in the latter case, we focus on the structural information of the documents. We classify and describe techniques for computing similarities among XML data sources, and discuss methods for clustering DTDs/XML schemas and XML documents.

DOI: 10.4018/978-1-61350-356-0.ch006
INTRODUCTION

The eXtensible Markup Language (XML) has emerged as the de-facto standard for the representation and the diffusion of data both on the Web and within organizations (De Meo, Quattrone, Terracina, & Ursino, 2006; Lee, Yang, Hsu, & Yang, 2002; Thompson, Beech, Maloney, & Mendelsohn, 2004).

Today, various vendors propose software tools to manage native XML (i.e., they use XML documents as the core storage unit) (Tamino, 2010), whereas some software companies, like Oracle, propose tools based on relational database technologies to efficiently store, navigate, and query XML data (Oracle XML DB, 2010). As a result, XML databases are rapidly proliferating and there are significant research efforts devoted to develop scalable techniques for efficiently handling the huge amount of XML data.

An effective solution to organize XML data sources relies on clustering techniques, which are applied based on the XML structural similarities. Clustering is an active research area in data mining (Han & Kamber, 2006). The goal of a clustering algorithm is to group objects into clusters such that objects within a cluster share similar features whereas objects associated with different clusters are dissimilar. Clustering XML data sources is useful to organize data in an unsupervised way and, ultimately, this is important to make the process of retrieving and browsing data easier.

Some of the activities that can benefit from the clustering of XML data sources are data integration and retrieval. The rapid growth of XML data sources explains the need of specific tools for integrating these data sources, i.e., the task of merging data coming from disparate sources with the goal of providing end-users with a unified view of these data (Lee, Yang, Hsu, & Yang, 2002). Data integration hence plays a key role in a wide range of domains, ranging from life sciences to Web mining and e-commerce (De Meo, Quattrone, Terracina, & Ursino, 2007). As soon as the size and complexity of the XML sources to integrate increases, the data integration task must be automatized as much as possible; for this purpose, there is a demand for tools capable of automatically detecting whether portions of two data sources represent the same concepts and of grouping sources describing the same piece of reality into homogeneous clusters. Moreover, the computation of the similarity degree of XML data is useful for designing tools capable of better ranking XML documents on the basis of their similarity with respect to a user query. The final outcome is an increase in terms of both accuracy and completeness of the answers generated for a given user query (Tagarelli & Greco, 2010).

The process of clustering XML data is a non-trivial task as it poses some challenging research issues which are not present when flat data or textual sources are considered. A first research problem regards which data model is to be adopted to represent XML data sources. In the literature, a large variety of data models have been introduced, such as rooted and labeled trees (Dalamagas, Cheng, Winkel, & Sellis, 2006), (weighted) graphs (De Meo, Quattrone, Terracina, & Ursino, 2006), arrays/matrices (Theobald, Schenkel, & Weikum, 2003), or sequential data (Flesca, Manco, Masciari, Pontieri, & Pugliese, 2005).

The second research problem regards the definition of suitable techniques to compute the similarity degree of two XML data sources. This last problem has been considered in the literature at two different levels, specifically:

- At the intensional level (also known as schema-level), DTDs or XML Schemas are taken into account (De Meo, Quattrone, Terracina, & Ursino, 2006; Lee, Yang, Hsu, & Yang, 2002). Moreover, the research line devoted to extract semantic similarities between two DTDs/XML Schemas is part of a more general research activity known as schema matching (Kementsietsidis, 2009; Rahm & Bernstein, 2001).