Chapter III

An Overview of Similarity Measures for Clustering XML Documents

Giovanna Guerrini, Università degli Studi di Milano, Italy

Marco Mesiti, Università degli Studi di Genova, Italy

Ismael Sanz, Universitat Jaume I, Spain

Abstract

The large amount and heterogeneity of XML documents on the Web requires the development of clustering techniques to group together similar documents. Documents can be grouped together according to their content, their structure, and the links inside and among the documents. For instance, grouping together documents with similar structure has interesting applications in the context of information extraction, heterogeneous data integration, personalized content delivery, access-control definition, Web site structural analysis, and the comparison of RNA secondary structures. Many approaches have been proposed for evaluating the structural and content similarity between tree-based and vector-based rep-
representations of XML documents. Link-based similarity approaches developed for Web data clustering have been adapted for XML documents. This chapter discusses and compares the most relevant similarity measures and their employment for XML document clustering.

Introduction

XML stands for the eXtensible Markup Language introduced by World Wide Web Consortium (W3C; 1998) that allows one to structure documents by means of nested, tagged elements. The element tag allows the annotation of the semantic description of the element content and can be exploited in order to effectively retrieve only relevant documents. Thus, the document structure can be exploited for document retrieval. Moreover, through the Xlink Language (W3C, 2001), different types of links can be specified among XML documents. In Xlink, a link is a relationship among two or more resources that can be described inside an XML document. These relationships can be exploited as well to improve document retrieval.

The exponential growing of XML-structured data available on the Web has raised the need of developing clustering techniques for XML documents. Web data clustering (Vakali, Pokorný, & Dalamagas, 2004) is the process of grouping Web data into clusters so that similar data belong to the same cluster and dissimilar data to different clusters. The goal of organizing data in such a way is to improve data availability and to make data access faster so that Web information retrieval and content delivery are improved. Moreover, clustering together similar documents allows the development of homogeneous indexing structures and schemas that are more representative of such documents.

XML documents can also be used for annotating Web resources (like articles, images, movies, and also Web services). For example, an image can be coupled with an XML document representing the image author and the date on which it has been shot as well as a textual description of its content or theme. A search engine can be coupled with an XML document containing information on the domain in which it works (e.g., document retrieval, image retrieval, Web-service retrieval) as well as information on the period of time during which the engine is available to answer queries. Web services can be coupled with a description of the services they provide as well as links to analogous providers on the Web. An important activity in this respect is to identify resources on the Web that are similar by considering the similarity of the XML documents containing the annotations in order to provide users with similar resources. Thus, developing approaches for clustering together similar documents that share similar characteristics is an important research direction.

XML document clustering is realized through algorithms that rely on the similarity between two documents computed exploiting a distance metric. The algorithms should guarantee that documents in the same cluster have a high similarity degree (low distance), whereas documents in different clusters have a low similarity degree (high distance). As far as the clustering of XML data is concerned, the document content and the document structure as well as links among documents can be exploited for identifying similarities among documents. Several measures have been proposed for computing the structural and content similarity among XML documents, but few XML-specific approaches exist for computing link similarity (even if the approaches developed for Web data can be easily applied). The purpose of the
UX Challenges and Best Practices in Designing Web and Mobile Solutions
www.igi-global.com/chapter/ux-challenges-and-best-practices-in-designing-web-and-mobile-solutions/203417?camid=4v1a