Chapter 3
XML Similarity Detection and Measures

Sanjay Kumar Madria
Missouri University of Science and Technology, USA

Waraporn Viyanon
Missouri University of Science and Technology, USA

ABSTRACT

XML similarity detection plays an important role in facilitating many applications such as data integration, document classification/clustering, querying, and change management. In this chapter, we present an overview on XML document syntactic and semantic similarity/distance measures along with existing research related to XML similarity detection. The measures are classified into two main categories: structural similarity, and structural and content similarity. We review similarity detection approaches proposed in the literature and discuss some of the challenges and future directions for research on XML similarity detection and related fields.

INTRODUCTION

In recent years, the number of various multi-dimensional data generated and distributed in various information sources as well as the number of users that use these information sources has been increasing. These sources usually use different models for the representation of data, such as the relational model, semistructured models on the web, text files, etc. For efficient data management and exchange, XML has been increasing its relevance as a fundamental standard. As the widespread use of XML for describing and exchanging data on the web is increasing, XML based comparison becomes a central issue in the database and information retrieval. The use of XML similarity in a wide range of applications such as data integration, change management, classification/clustering of XML documents and XML querying is needed (Tekli, Chbeir, & Yetongnon, 2009).

DOI: 10.4018/978-1-61350-356-0.ch003
The objective of data integration is to identify similar XML documents originated from different data source to be integrated so that users can access more complete information. Typically the individuals doing the integration are not experts of the data; they must first understand the nature of the data, what data is available and how good it is. They must determine how the data is represented in the application and decide how to normalize data across the data sources. Data integration task is not easy. A good similarity measurement is required to improve the result of the data integration, otherwise users might discover problems such as unexpected or missing results appear and they have to look through the whole process.

The purpose of change management is to discover changes by finding dissimilarity among documents which helps the user tell what has been inserted, updated, or deleted from the last version. This can be applied in many applications, such as version control or index maintenance. Classification of XML documents can be exploited to improve storage, retrieval and indexing facilities. Only structures of XML documents are measured for the similarity to solve the problem of recognizing different sources providing the same kind of information. For XML query processing, similarity among documents will be calculated to find and rank results according to their similarity in order to retrieve the best results as possible. This is obvious that all of these applications require similarity measurement.

XML similarity detection for XML documents can be classified into two main categories: (i) structural similarity and (ii) structural and content similarity. In addition, semantic similarity becomes important in document similarity particularly in information retrieval and information integration as it supports the detection of conceptually close (but not identical) entities.

The rest of this chapter is organized as follows: Section “Background” introduces background knowledge on views and benefits of using XML, whereas the subsequent section outlines significant challenges in XML similarity detection. Section “Similarity Measures” overviews similarity measures from both syntactic and semantic viewpoints. Section “XML similarity” describes existing approaches and the use of XML similarity. Finally, the chapter ends with a section containing concluding remarks and a discussion on open issues concerning XML similarity detection.

**BACKGROUND**

**Views of XML**

XML documents can be classified as having either a document-centric (text-centric) view or a data-centric view (Bourret, 2005).

Data-centric documents are used to transport data. As such, they are highly structured data marked up with XML tags. Most data-centric XML documents are generated from structured sources such as RDBMS. The data-centric view emphasizes on XML structure since the meaning of a data-centric XML document depends only on the structured data represented inside it, and is usually used to exchange data in a structured form.

Document-centric documents focus on application-relevant objects. They are loosely structured documents marked-up with XML tags, and their meaning depends on the document as a whole. Their structure is more irregular, and their data are heterogeneous. Such documents might not even have a document-type declaration (DTD) or XML schema. For this view, text is a higher priority than structure. Figure 1 shows examples of both document-centric and data-centric documents.

**Benefits of XML**

In (Daly, 2003) the main benefits of XML are outlined, explaining why it is an effective solution for the design of a wide range of applications.
Related Content

A Data Mining Service to Assist Instructors Involved in Virtual Education
www.igi-global.com/chapter/data-mining-service-assist-instructors/73496?camid=4v1a

Knowledge Discovery from Online Communities
www.igi-global.com/chapter/knowledge-discovery-online-communities/73493?camid=4v1a

Improving Classification Accuracy of Decision Trees for Different Abstraction Levels of Data
www.igi-global.com/article/improving-classification-accuracy-decision-trees/1753?camid=4v1a

A Perspective on Data Mining Integration with Business Intelligence
www.igi-global.com/chapter/perspective-data-mining-integration-business/73527?camid=4v1a