Chapter 3.21
Mining Association Rules from XML Documents

Laura Irina Rusu
La Trobe University, Australia

Wenny Rahayu
La Trobe University, Australia

David Taniar
Monash University, Australia

ABSTRACT

This chapter presents some of the existing mining techniques for extracting association rules out of XML documents in the context of rapid changes in the Web knowledge discovery area. The initiative of this study was driven by the fast emergence of XML (eXtensible Markup Language) as a standard language for representing semistructured data and as a new standard of exchanging information between different applications. The data exchanged as XML documents become richer and richer every day, so the necessity to not only store these large volumes of XML data for later use, but to mine them as well to discover interesting information has became obvious. The hidden knowledge can be used in various ways, for example, to decide on a business issue or to make predictions about future e-customer behaviour in a Web application. One type of knowledge that can be discovered in a collection of XML documents relates to association rules between parts of the document, and this chapter presents some of the top techniques for extracting them.

DOI: 10.4018/978-1-60566-330-2.ch011

INTRODUCTION

The amount of data stored in XML (eXtensible Markup Language) format or changed between fferent types of applications has been growing during the last few years, and more companies are considering XML now as a possible solution for their data-storage and data-exchange needs

(Laurent, Denilson, & Pierangelo, 2003). The first immediate problem for the researchers was how to represent the data contained in the old relational databases using this new format, so various techniques and methodologies have been developed to solve this problem. Next, the users realised that they not only required storing the data in a different way, which made it much easier to exchange data between various applications, but they required getting interesting knowledge out of the entire volume of XML data stored as well. The acquired knowledge might be successfully used in the decisional process to improve business outcomes. As a result, the need for developing new languages, tools, and algorithms to effectively manage and mine collections of XML documents became imperative.

A large volume of work has been developed, and research is still pursued to get solutions that are as effective as possible. The general idea and goal for researchers is to discover more powerful XML mining algorithms that are able to find representative patterns in the data, achieve higher accuracy, and be more scalable on large sets of documents. The privacy issue in knowledge discovery is also a subject of great interest (Ashrafi, Taniar, & Smith, 2004a).

XML mining includes both the mining of structures as well as the mining of content from XML documents (Nayak, 2005; Nayak, Witt, & Tonev, 2002). The mining of structure is seen as essentially mining the XML schema, and it includes intrastructure mining (concerned with mining the structure inside an XML document, where tasks of classification, clustering, or association rule discovering could be applied) and interstructure mining (concerned with mining the structures between XML documents, where the applicable tasks could be clustering schemas and defining hierarchies of schemas on the Web, and classification is applied with name spaces and URIs [uniform resource identifiers]). The mining of content consists of content analysis and structure clarification. While content analysis is concerned with analysing texts within the XML document, structural clarification is concerned with determining similar documents based on their content (Nayak, 2005; Nayak et al., 2002).

Discovering association rules is looking for those interesting relationships between elements appearing together in the XML document, which can be used to predict future behaviour of the document. To our knowledge, this chapter is the first work that aims to put together and study the existing techniques to perform the mining of association rules out from XML documents.

BACKGROUND

The starting point in developing algorithms and methodologies for mining XML documents was, naturally, the existing work done in the relational database mining area (Agrawal, Imielinski, & Swami, 1993; Agrawal & Srikant, 1998; Ashrafi, Taniar, & Smith, 2005; Ashrafi, 2004; Daly & Taniar, 2004; Tjioe & Taniar, 2005). In their attempt to apply various relational mining algorithms to the XML documents, researchers discovered that the approach could be a useful solution for mining small and not very complex XML documents, but not an efficient approach for mining large and complex documents with many levels of nesting.

The XML format comes with the acclaimed extensibility that allows the change of structure, that is, adding, removing, and renaming nodes in the document according to the information necessary to be encoded in. Furthermore, using the XML representation, there are a lot of possibilities to express the same information (see Figure 1 for an example) not only between different XML documents, but inside the same document as well (Rusu, Rahayu, & Taniar, 2005a).

In a relational database, it is not efficient to have multiple tables to represent the same data with different field names, types, and relationships as the constraints and table structures are defined at the design time. In an opposite manner, a new