Chapter 1
A Study of XML Models for Data Mining:
Representations, Methods, and Issues

Sangeetha Kutty
Queensland University of Technology, Australia

Richi Nayak
Queensland University of Technology, Australia

Tien Tran
Queensland University of Technology, Australia

ABSTRACT

With the increasing number of XML documents in varied domains, it has become essential to identify ways of finding interesting information from these documents. Data mining techniques can be used to derive this interesting information. However, mining of XML documents is impacted by the data model used in data representation due to the semi-structured nature of these documents. In this chapter, we present an overview of the various models of XML documents representations, how these models are used for mining, and some of the issues and challenges inherent in these models. In addition, this chapter also provides some insights into the future data models of XML documents for effectively capturing its two important features, structure and content, for mining.

INTRODUCTION

Due to the increased popularity of XML in varied application domains, a large number of XML documents are found in both organizational intranets and Internet. Some of the popular datasets such as English Wikipedia contains 3.1 million web documents in XML format with 1.74 billion words, and the ClueWeb dataset used in Text Retrieval Conference (TREC) tracks contains 503.9 million XML documents collected from the web in January and February 2009. In order to discover useful knowledge from these large amount of XML documents, researchers have used data mining techniques (Nayak, 2005). XML data mining techniques have gained great deal of

DOI: 10.4018/978-1-4666-2455-9.ch001

interest among researchers due to their potential to discover useful knowledge in diverse fields such as bioinformatics, telecommunication network analysis, community detection, information retrieval, social network analysis (Nayak, 2008).

Unlike structured data where the structure is fixed because the data is stored in structured format as in relational tables, XML data has flexibility in its structure as the users are allowed to use custom-defined tags to represent the data. An XML document contains tags and the data is enclosed within those tags. A tag usually describes a meaningful name to the content it represents. Moreover, tags present in the document are organised in hierarchical order showing the relationships between elements of the document. Usually, the hierarchical ordering of tags in an XML document is called as the *document structure* and the data enclosed within these tags is called as the *document content*.

XML data can be modelled in various forms namely vectors (or transactional data models), paths, trees and graphs based on its structure and/or content. The focus of this chapter is to present an overview of the various models that can be used to represent XML documents for the process of mining. This chapter also addresses some of the issues and challenges associated with each of these models.

Organisation of this chapter is as follows. In the next section, it explains various XML data models in detail. The third section discusses about the roles of models in diverse mining tasks such as frequent pattern mining, association rules mining, clustering and classification. The chapter then details about the issues and the challenges in using these models for mining. It concludes with the needs and opportunities of new models for mining on XML documents.

DATA MODELS FOR XML DOCUMENT MINING

To suit the objectives and the needs of XML mining algorithms, XML data has been represented in various forms. Figure 1 gives taxonomy of XML data showing various data models that facilitate XML mining with different features that exist in the XML data.

There are two types of XML data: *XML document* and *XML schema definition*. An XML schema definition contains the structure and data definitions of XML documents (Abiteboul, Buneman, & Suciu, 2000). An XML document, on the other hand, is an instance of the XML schema that contains the data content represented in a structured format.

Figure 1. Data models facilitating mining of XML data