Chapter XIV

Frequent Mining on XML Documents

Sangeetha Kutty
Queensland University of Technology, Australia

Richi Nayak
Queensland University of Technology, Australia

ABSTRACT

With the emergence of XML standardization, XML documents have been widely used and accepted in almost all the major industries. As a result of the widespread usage, it has been considered essential to not only store these XML documents but also to mine them to discover useful information from them. One of the very popular techniques to mine XML documents is frequent pattern mining, which has huge potential in varied domains such as bio-informatics, network analysis. This chapter presents some of the existing techniques to discover frequent patterns from XML documents. It also covers the applications and addresses the major issues in mining XML documents.

INTRODUCTION

With the emergence of XML (eXtensible Markup Language) standardization, XML documents have been widely used and accepted in almost all the major industries. This has resulted in a large amount of XML documents. In order to discover useful knowledge from the huge collections, various data mining techniques have been used (Nayak, 2005). One of the popular data mining techniques is frequent pattern mining on XML documents (Tan, Dillon, Hadzic, Chang, & Feng, 2005), which has created a great deal of interest among researchers due to its potential of application in diversified fields. Some of them include data warehouse integration (Termier, Rouset, & Sebag, 2002), bio-informatics (Hadzic, Dillon, Sidhu, Chang, & Tan, 2006b; Zaki, 2005), network analysis and information retrieval. Further-
more, XML frequent pattern mining serves as an important function for other data mining tasks such as classification (Zaki, 2005), clustering (Kutty, Tran, Nayak, & Li, 2007; Nayak & Iryadi, 2007) and association rules mining (Rusu, Rahayu, & Taniar, 2007).

In spite of its potential, there are major challenges that are yet to be fully addressed while mining XML documents in comparison to the established field of frequent pattern mining from relational databases. Firstly, unlike the relational databases which has structured or the unstructured data such as text, sound or image files, XML documents model the hierarchical relationships between data items. XML data represent semi-structured format consisting of values and their relationships. An XML document contains tags and the data enclosed within those tags. The tag, that describes the element name, includes the semantics in the form of text data. The tags define the structure of an XML document showing the relationships between elements of the document. Hence, while mining the XML documents, the structure of the data items has to be preserved and taken into consideration. This contributes to the second challenge that is, dealing with huge volume of data compared to structured data due to the storage of the structural information (Nayak, 2005).

This chapter presents a discussion of the current trends in frequent pattern mining on XML documents, addresses the related issues and challenges and finally presents the applications and opportunities for this type of mining. The background of frequent mining is provided in Frequent Pattern Mining section. The XML Frequent Pattern Mining section covers the analysis of various types of frequent pattern mining. Also, we will present some of the potential opportunities of XML frequent pattern mining from a technological as well as an application-based perspective in Applications and Issues of XML Frequent Mining and highlight some of the issues and challenges.

**FREQUENT PATTERN MINING**

**Overview: On Relational Databases**

Frequent pattern mining was first introduced along with association rules mining to analyze customer-buying behaviour from retail transaction databases. Frequent pattern mining in a transaction database involves identifying patterns, which occurs quite often, and hence these patterns are called as frequent. The frequent patterns are types of itemsets, sequences, subtrees and subgraphs when extracted from relational databases, sequential databases, trees and graphs respectively.

In relational databases, frequent pattern mining refers to identifying a frequent item (eg. Milk) or an itemset (Milk and Butter). An item or itemset is said to be frequent if they occur more than a user-specified threshold limit denoted by \( \min\_supp \). In order to check whether an item or itemset is frequent, first the frequency or support count of an itemset is calculated. Frequency or support count is the percentage of transactions that contain the itemset. If the support count is greater than \( \min\_supp \), then the itemset is said to be frequent.

Consider the transaction database given in Table 1(a), with TID representing the transaction id and Items representing the items in the customer’s basket. The task is to identify itemsets whose support is greater than \( \min\_supp \), that is set as 40%. This implies that the item or itemset should occur twice or more in the given database to become frequent. The 1-frequent itemsets are \{Milk\}, \{Butter\} and \{Fish\} as they have supports greater than 50% (as shown in Table 1 (b)). The 1-Length frequent itemsets are combined to form 2-Length candidate itemsets and their supports are calculated to determine whether