Chapter 2.6
A Framework for Efficient Association Rule Mining in XML Data

Ji Zhang
University of Toronto, Canada

Han Liu
Carnegie Mellon University, USA

Tok Wang Ling
National University of Singapore, Singapore

Robert M. Bruckner
Microsoft, USA

A Min Tjoa
Vienna University of Technology, Austria

ABSTRACT

In this article, we propose a framework, called XAR-Miner, for mining ARs from XML documents efficiently. In XAR-Miner, raw data in the XML document first are preprocessed to transform either to an Indexed XML Tree (IX-tree) or to Multirelational Databases (Multi-DB), depending on the size of the XML document and the memory constraint of the system, for efficient data selection and AR mining. Concepts that are relevant to the AR mining task are generalized to produce generalized metapatterns. A suitable metric is devised for measuring the degree of concept generalization in order to prevent undergeneralization or overgeneralization. Resulting generalized metapatterns are used to generate large ARs that meet the support and confidence levels. A greedy algorithm is also presented in order to integrate data selection and large itemset generation to

enhance the efficiency of the AR mining process. The experiments conducted show that XAR-Miner is more efficient in performing a large number of AR mining tasks from XML documents than the state-of-the-art method of repetitively scanning through XML documents in order to perform each of the mining tasks.

INTRODUCTION

The eXtensive Markup Language (XML) has become a standard for representing and exchanging information on the Internet. The fast-growing amount of XML-based information on the Web has made it desirable to develop new techniques to discover patterns and knowledge from XML data. Association Rule (AR) mining frequently is used in data mining to reveal interesting trends, patterns, and rules in large datasets. Mining ARs from XML documents thus has become a new and interesting research topic.

Association rules first were introduced in the context of retail transaction databases (Agrawal & Srikant, 1994). Formally, an AR is an implication in the form of $X \rightarrow Y$, where X and Y are sets of items in database D that satisfy $X \cap Y = \phi$ and $X \cup Y \subseteq I$. D is a set of data cases, and I is the complete set of distinct items that appear in D. X and Y are called the antecedent and subsequent of the rule. The rule $X \rightarrow Y$ has support of s in D if $s\%$ of the data cases in D contain both X and Y, and the rule has a confidence of c in D if $c\%$ of the data cases contain Y and if they also contain X. Association rule mining aims to discover all large rules whose support and confidence exceed these user-specified minimum support and confidence thresholds: minsup and minconf.

A recent trend of mining ARs is the development of query language in order to facilitate association rule mining. The MINE RULE operator is introduced by Meo, Psaila, and Ceri (1996) and extended by Meo, Psaila, and Ceri (1998). In addition, Imielinski and Virmani (1999) introduced an SQL extension called MSQL, which is equipped with various operators over association rules such as generation, selection, and pruning. DMSQL is another query language used for association rule specification (Han & Kamber, 2000). There already has been research in mining ARs from semi-structured documents (Amir, Feldman, & Kashi, 1997; Feldman & Hirsh, 1996; Singh, Scheuermann, & Chen, 1997; Singh, Chen, Haight, & Scheuermann, 1999), among which the works in Amir et al. (1997) and Feldman et al. (1996) use unsupervised procedures, and Singh et al. (1997) and Singh et al. (1999) use single and multi-constraints in mining ARs. The constraint-based methods, compared to the non-constraint ones, are able to achieve better efficiency in the mining process and to generate a manageable number of rules.

Though we have witnessed intensive research work in AR mining in the past years, there has been very little work in the domain of AR mining from XML documents. For instance, the work by Braga, Campi, Klemettinen, and Lanzi (2002) uses the MINE RULE operator introduced by Meo et al. (1996) for AR mining purposes in native XML documents. The Predicative Model Markup Language (PMML) is proposed to present various patterns such as association rules and decision trees extracted from XML documents (PMML, 2000). Feng, Dillon, Weigand, and Chang (2003) present an XML-enabled AR mining framework but do not give any details on how to implement this framework efficiently. Wan and Dobbie (2003) claim that XML AR can be accomplished simply by using XQuery language.

However, the major problems with the state-of-the-art methods are twofold: (1) These approaches select data from native XML data; thus, the efficiency of these approaches is low because of the normally huge volume of XML data that need to be scanned in the process of AR mining. Each AR mining task involves a scan of XML data sources, which is not practical when performing a large number of AR mining tasks. Data orga-