Chapter 44
Frequent Pattern Discovery and Association Rule Mining of XML Data

Qin Ding
East Carolina University, USA

Gnanasekaran Sundarraj
Pennsylvania State University, USA

ABSTRACT

Finding frequent patterns and association rules in large data has become a very important task in data mining. Various algorithms have been proposed to solve such problems, but most algorithms are only applicable to relational data. With the increasing use and popularity of XML representation, it is of importance yet challenging to find solutions to frequent pattern discovery and association rule mining of XML data. The challenge comes from the complexity of the structure in XML data. In this chapter, we provide an overview of the state-of-the-art research in content-based and structure-based mining of frequent patterns and association rules from XML data. We also discuss the challenges and issues, and provide our insight for solutions and future research directions.

INTRODUCTION

Association rule mining and frequent pattern mining are important problems in data mining (Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1996; Han & Kamber, 2006; Tan, Steinbach, & Kumar, 2006). These two problems are closely related and they aim to discover patterns that occur frequently in large datasets. The problems were originally proposed for market basket transaction data regarding store items purchased on a per-transaction basis. An example is that by analyzing the customer transaction data at Amazon.com, we found that most customers who bought book "A" also bought books "B" and "C". Discovering such customer purchase behaviors can be very useful in decision making and other business-related applications. Many algorithms have been proposed

DOI: 10.4018/978-1-4666-2455-9.ch044

Copyright © 2013, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

to discover association rules and frequent patterns, such as Apriori (Agrawal & Srikant, 1994) and FP-growth (Han, Pei, & Yin, 2000; Han, Pei, Yin, & Mao, 2004). However, most of such algorithms are only applicable to relational data.

In the past decade, XML has become a standard for representing and exchanging information. With the increasing popularity of XML representation and the large amount of XML data available, it becomes important and necessary for researchers to study how to extend association rule mining and frequent pattern mining to XML data so that interesting patterns can be discovered from XML documents. This is a very interesting yet challenging field. The problem was first proposed in 2002 and since then it has gained more attention from an increasing number of researchers.

A simple approach to mining association rules and frequent patterns on XML data is to convert XML data into relational format, and then use the traditional algorithms to perform association rule mining and frequent pattern mining. However, by doing so, the structure information in XML data is mostly lost. As XQuery (W3C XML Query) becomes a standard query language for XML data, researchers have also attempted to use XQuery or extend the features in XQuery to support frequent pattern mining on XML data (Braga, Campi, Ceri, Klemettinen, & Lanzi, 2002a, 2002b; Wan & Dobbie, 2003; Romei & Turini, 2010). The existing or extended features of the XQuery language facilitate the computation needed for mining XML association rules and frequent patterns. However, it also adds an extra layer which in turn brings additional overhead; more importantly, this kind of language-dependent approach lacks flexibility since XQuery is a query language and was not designed for data mining. Therefore it is desired to develop non-XQuery-based approach for frequent pattern discovery on XML data.

Most early work on XML association rule mining and frequent pattern mining focused on mining the content of XML documents (Braga et al., 2002a, 2002b; Wan & Dobbie, 2003; Meo & Psaila, 2002). Besides mining the content in XML documents, it is also interesting in mining the structure (Cong, Yi, Liu, Wang, 2002). For example, frameworks and algorithms have been proposed to discover dynamic structural changes over a collection of historical XML documents (Zhao, Bhowmick, Mohania, & Kambayashi, 2004; Zhao, Bhowmick, & Mohania, 2004; Zhao, Bhowmick, & Gruenwald, 2005; Zhao & Bhowmick, 2005; Zhao, Chen, Bhowmick, & Madria, 2005; Zhao, Bhowmick, & Madria, 2006; Leonardi, Bhowmick, & Madria, 2005; Leonardi, Hoai, Bhowmick, & Madria, 2006, 2007; Leonardi & Bhowmick, 2006a, 2006b, 2007; Leonardi, Budiman, & Bhowmick, 2005; Chen, Bhowmick, & Chia, 2004a, 2004b; Cobena, Abiteboul, & Marian, 2002). Another type of structure-based XML mining is to discover frequent XML query patterns to improve query response time (Yang, Lee, Hsu, & Acharya, 2003; Yang, Lee, & Hsu, 2003; Yang, Lee, Hsu, & Guo, 2004; Yang, Lee, & Hsu, 2004a, 2004b; Yang, Lee, Hsu, Huang, & Wang, 2008; Hua, Zhao, & Chen, 2007; Li, Feng, Wang, Zhang, & Zhou, 2006; Li, Feng, Wang, & Zhou, 2009; Feng, Qian, Wang, & Zhou, 2006; Chen, Yang, & Wang, 2004). Overall, both content-based and structure-based frequent pattern mining on XML data are still at their preliminary stage with many open questions and solutions to be tackled.

The objective of this chapter is to provide an overview of research on content-based and structure-based association rule mining and frequent pattern discovery on XML data. We will discuss various approaches with its advantages, limitations, and issues. We will also provide our insight for future direction in this research field. The remainder of this chapter is organized as follows. In next section, we introduce the background on association rule mining and frequent pattern mining in general, followed by the problem on XML data. Section "Mining frequent patterns and association rules on XML data" details the current state-of-the-art research on association rule mining on XML data, in particular, the content-

19 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/frequent-pattern-discovery-association-rule/73474

Related Content

Exploring “User,” “Video,” and (Pseudo) Multi-Mode Networks on YouTube with NodeXL
Shalin Hai-Jew (2017). *Social Media Data Extraction and Content Analysis (pp. 242-295).*
www.irma-international.org/chapter/exploring-user-video-and-pseudo-multi-mode-networks-on-youtube-with-nodexl/161967

A Clustering Approach to Path Planning for Big Groups
Jakub Szkandera, Ondej Kaasand Ivana Kolingerová (2019). *International Journal of Data Warehousing and Mining (pp. 42-61).*
www.irma-international.org/article/a-clustering-approach-to-path-planning-for-big-groups/225806

Deconstructing Smart Cities
Michael Batty (2016). *Big Data: Concepts, Methodologies, Tools, and Applications (pp. 1957-1969).*
www.irma-international.org/chapter/deconstructing-smart-cities/150251

Network Text Analysis and Sentiment Analysis: An Integration to Analyse Word-of-Mouth in the Digital Marketplace
Veronica Ravaglia, Luca Zanazziand Elvis Mazzoni (2016). *Big Data: Concepts, Methodologies, Tools, and Applications (pp. 1277-1294).*
www.irma-international.org/chapter/network-text-analysis-and-sentiment-analysis/150216

Emotion-Drive Interpretable Fake News Detection
Xiaoyi Ge, Mingshu Zhang, Xu An Wang, Jia Liuand Bin Wei (2022). *International Journal of Data Warehousing and Mining (pp. 1-17).*
www.irma-international.org/article/emotion-drive-interpretable-fake-news-detection/314585