IGI PUBLISHING
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.igi-global.com

ITB15220

This paper appears in the publication, **Data Mining and Knowledge Discovery Technologies**
edited by D. Taniar © 2008, IGI Global

Chapter III

Mining Association Rules from XML Data

Qin Ding, East Carolina University, USA

Gnanasekaran Sundarraj,
The Pennsylvania State University at Harrisburg, USA

Abstract

With the growing usage of XML in the World Wide Web and elsewhere as a standard for the exchange of data and to represent semi-structured data, there is an imminent need for tools and techniques to perform data mining on XML documents and XML repositories. In this chapter, we propose a framework for association rule mining on XML data. We present a java-based implementation of the apriori and the FP-growth algorithms for this task and compare their performances. We also compare the performance of our implementation with an XQuery-based implementation.

Introduction

Advances in data collection and storage technologies have led organizations to store vast amounts of data pertaining to their business activities. Extracting "use-

Copyright © 2008, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

ful" information from such huge data collections is of importance in many business decision-making processes. Such an activity is referred to as *data mining* or *knowledge discovery in databases* (KDD) (Han & Kamber, 2006). The term data mining refers to tasks such as *classification*, *clustering*, *association rule mining*, *sequential pattern mining*, and so forth (Han et al., 2006).

The task of association rule mining is to find correlation relationships among different data attributes in a large set of data items, and this has gained a lot of attention since its introduction (Agrawal, Imieliński, & Swami, 1993). Such relationships observed between data attributes are called *association rules* (Agrawal et al., 1993). A typical example of association rule mining is the *market basket analysis*. Consider a retail store that has a large collection of items to sell. Often, business decisions regarding discount, cross-selling, grouping of items in different aisles, and so on need to be made in order to increase the sales and hence the profit. This inevitably requires knowledge about past transaction data that gives the buying habits of customers. The association rules in this case will be of the form "*customers who bought item A also bought item B*," and association rule mining is to extract such rules from the given historical transaction data.

Explosive use of World Wide Web to buy and sell items over the Internet has led to similar data mining requirements from online transaction data. In an attempt to standardize the format of data exchanged over the Web and to achieve interoperability between the different technologies and tools involved, World Wide Web consortium (W3C) introduced *Extensible Markup Language* (XML) (Goldfarb, 2003). XML is a simple but very flexible text format derived from *Standard Generalized Markup Language* (SGML) (Goldfarb, 2003), and has been playing an increasingly important role in the exchange of wide variety of data over the Web. Even though it is a markup language much like the *HyperText Markup Language* (HTML) (Goldfarb, 2003), XML was designed to describe data and to focus on what the data is, whereas HTML was designed to display data and to focus on how the data looks on the Web browser. A data object described in XML is called an *XML document*.

XML also plays the role of a meta-language, and allows document authors to create customized markup languages for limitless different types of documents, making it a standard data format for online data exchange. This growing usage of XML has naturally resulted in increasing amount of available XML data, which raises the pressing need for more suitable tools and techniques to perform data mining on XML documents and XML repositories.

In this chapter, we study the various approaches that have been proposed for association rule mining from XML data, and present a java-based implementation for the two well-known algorithms for association rule mining: *apriori* (Agrawal & Srikant, 1994) and FP-growth (Han, Pei, Yin, & Mao 2004). The rest of this chapter is organized as follows. In the second section, we describe the basic concepts and definitions for association rule mining. In this section, we also explain the above two

Copyright © 2008, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

11 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/mining-association-rules-xml-data/7513

Related Content

Skeleton Network Extraction and Analysis on Bicycle Sharing Networks
Kanokwan Malang, Shuliang Wang, Yuanyuan Lvand Aniwat Phaphuangwittayakul (2020). *International Journal of Data Warehousing and Mining (pp. 146-167).*
www.irma-international.org/article/skeleton-network-extraction-and-analysis-on-bicycle-sharing-networks/256167

Data Field for Hierarchical Clustering
Shuliang Wang, Wenyan Gan, Deyi Liand Deren Li (2011). *International Journal of Data Warehousing and Mining (pp. 43-63).*
www.irma-international.org/article/data-field-hierarchical-clustering/58637

Discovering Patterns for Architecture Simulation by Using Sequence Mining
Pinar Senkul, Nilufer Onder, Soner Onder, Engin Madenand Hui Meen Nyew (2012). *Pattern Discovery Using Sequence Data Mining: Applications and Studies (pp. 212-236).*
www.irma-international.org/chapter/discovering-patterns-architecture-simulation-using/58682

Working from Claims Data
Patricia Cerrito (2010). *Text Mining Techniques for Healthcare Provider Quality Determination: Methods for Rank Comparisons (pp. 341-356).*
www.irma-international.org/chapter/working-claims-data/36640

Ontology-Based Construction of Grid Data Mining Workflows
Peter Brezany, Ivan Janciakand A Min Tjoa (2008). *Data Mining with Ontologies: Implementations, Findings, and Frameworks (pp. 182-210).*
www.irma-international.org/chapter/ontology-based-construction-grid-data/7578