Data Warehousing for Association Mining

Yuefeng Li
Queensland University of Technology, Australia

INTRODUCTION

With the phenomenal growth of electronic data and information, there are many demands for developments of efficient and effective systems (tools) to address the issue of performing data mining tasks on data warehouses or multidimensional databases. Association rules describe associations between itemsets (i.e., sets of data items) (or granules). Association mining (or called association rule mining) finds interesting or useful association rules in databases, which is the crucial technique for the development of data mining. Association mining can be used in many application areas, for example, the discovery of associations between customers' locations and shopping behaviours in market basket analysis.

Association mining includes two phases. The first phase is called pattern mining that is the discovery of frequent patterns. The second phase is called rule generation that is the discovery of the interesting and useful association rules in the discovered patterns. The first phase, however, often takes a long time to find all frequent patterns that also include much noise as well (Pei and Han, 2002). The second phase is also a time consuming activity (Han and Kamber, 2000) and can generate many redundant rules (Zaki, 2004) (Xu and Li, 2007). To reduce search spaces, user constraint-based techniques attempt to find knowledge that meet some sorts of constraints. There are two interesting concepts that have been used in user constraint-based techniques: meta-rules (Han and Kamber, 2000) and granule mining (Li et al., 2006).

The aim of this chapter is to present the latest research results about data warehousing techniques that can be used for improving the performance of association mining. The chapter will introduce two important approaches based on user constraint-based techniques. The first approach requests users to inputs their meta-rules that describe their desires for certain data dimensions. It then creates data cubes based these meta-rules and then provides interesting association rules.

The second approach firstly requests users to provide condition and decision attributes that used to describe the antecedent and consequence of rules, respectively. It then finds all possible data granules based condition attributes and decision attributes. It also creates a multi-tier structure to store the associations between granules, and association mappings to provide interesting rules.

BACKGROUND

Data warehouse mainly aims to make data easily accessible, present data consistently and be adaptive and resilient to change (Kimball and Ross, 2002). A data warehouse is an application that contains a collection of data, which is subject-oriented, integrated, non-volatile and time-variant, supporting management's decisions (Inmon, 2005). Data warehousing focuses on constructing and using data warehouses. The construction includes data cleaning, data integration and data consolidation. After these steps, a collection of data in a specific form can be stored in a data warehouse.

Data warehouses can also provide clean, integrated and complete data to improve the process of data mining (Han and Kamber, 2000). Han and Kamber also defined different levels of the integration of data mining and data warehouse. At the loosest level the data warehouse only acts as a normal data source of data mining. While at the tightest level both the data warehouse and data mining are sub-components that cooperate with each other. In a data mining oriented data warehouse, the data warehouse not only cleans and integrates data, but also tailors data to meet user constraints for knowledge discovery in databases. Thus, data mining can return what users want in order to improve the quality of discovered knowledge.

It is painful when we review the two steps in association mining: both take a long time and contain uncertain information for determining useful knowledge. Data mining oriented data warehousing is a promising direction for solving this problem. It refers

to constructing systems, in which both the data mining and data warehouse are a sub-component cooperating with each other. Using these systems, the data warehouse not only cleans and integrates data, but tailors data to fit the requirements of data mining. Thus, data mining becomes more efficient and accurate. In this chapter we discuss how data warehousing techniques are useful for association mining.

MAIN FOCUS

Based on the above introduction, we understand that data warehousing techniques can be helpful for improving the quality of data mining. We will focus on two areas in this chapter: mining patterns and meta-rules through data cubes and mining granules and decision rules through multi-tier structures.

Mining Patterns and Meta Rules through Data Cubes

There are many studies that discuss the research issue of performing data mining tasks on a data warehouse. The main research point is that pattern mining and rule generation can be used together with OLAP (On-Line Analytical Processing) to find interesting knowledge from data cubes (Han and Kamber, 2000) (Imielinski et al., 2002) (Messaoud et al., 2006).

A data cube consists of a set of data dimensions and a set of measures. Each dimension usually has a set of hierarchical levels. For example, a product dimension may have three hierarchal levels: All, Family and Article, where Article could be a set of attributes (e.g., {iTwin, iPower, DV-400, EN-700, aStar, aDream}), Family could be {Desktop, Laptop, MP3} and All is the total aggregation level. The frequency measure in data cube is called COUNT, which is used to evaluate the occurrences of the corresponding data values for data cube cells.

A meta-rule in multidimensional databases usually has the following format:

"$P_1 \wedge P_2 \wedge \ldots \wedge P_m \rightarrow Q_1 \wedge Q_2 \wedge \ldots \wedge Q_n$"

where P_i and Q_j are some predicates which can be sets of data fields in different dimension levels. The meta-rule defines the portion of the data cube to be mined.

The mining process starts to provide a meta-rule and define a minimum support and a minimum confidence. The traditional way to define the support and confidence is the use of the numbers of occurrences of data values. For example, the support can be computed according to the frequency of units of facts based on the COUNT measure. It was also recommended in the OLAP context (Messaoud et al., 2006) that users were often interested in observing facts based on summarized values of measures rather than the numbers of occurrences. Therefore, SUM measure based definitions for the support and confidence were presented in (Messaoud et al., 2006). The second step is to choose an approach (the top-down approach or bottom up approach) to produce frequent itemsets. The last step is to generate interesting association rules to meet the requirements in the meta-rule.

Mining Decision Rules through Multi-Tier Structures

User constraint-based association mining can also be implemented using granule mining, which finds interesting associations between granules in databases, where a granule is a predicate that describes common features of a set of objects (e.g., records, or transactions) for a selected set of attributes (or items).

Formally, a multidimensional database can be described as an information table (T, V^T), where T is a set of objects (records) in which each record is a sequences of items, and $V^T = \{a_1, a_2, \ldots, a_n\}$ is a set of selected items (or called attributes in decision tables) for all objects in T. Each item can be a tuple (e.g., <*name*, *cost*, *price*> is a product item).

Table I illustrates an information table, where $V^T = \{a_1, a_2, \ldots, a_7\}$, $T = \{t_1, t_2, \ldots, t_6\}$.

Given an information table, a user can classify attributes into two categories: condition attributions and decision attributes. For example, the high profit products can be condition attributes and low profit products can be decision attributes. Formally, a decision table is a tuple (T, V^T, C, D), where T is a set of objects (records) in which each record is a set of items, and $V^T = \{a_1, a_2, \ldots, a_n\}$ is a set of attributes, $C \cup D \subseteq V^T$ and $C \cap D = \varnothing$.

Objects in a decision table can also be compressed into a granule if they have the same representation (Pawlak, 2002). Table II shows a decision table of the