Available online at www.sciencedirect.com
ScienceDirect

Decision Support Systems 42 (2006) 859–878

Decision Support Systems

www.elsevier.com/locate/dsw

Comprehensive data warehouse exploration with qualified association-rule mining

Nenad Jukić [a,*], Svetlozar Nestorov [b,1]

[a]School of Business Administration, Loyola University Chicago, 820 N. Michigan Avenue, Chicago, IL 60640, USA
[b]Department of Computer Science, The University of Chicago, Chicago, IL, USA

Received 10 May 2004; received in revised form 9 July 2005; accepted 11 July 2005
Available online 11 August 2005

Abstract

Data warehouses store data that explicitly and implicitly reflect customer patterns and trends, financial and business practices, strategies, know-how, and other valuable managerial information. In this paper, we suggest a novel way of acquiring more knowledge from corporate data warehouses. Association-rule mining, which captures co-occurrence patterns within data, has attracted considerable efforts from data warehousing researchers and practitioners alike. In this paper, we present a new data-mining method called qualified association rules. Qualified association rules capture correlations across the entire data warehouse, not just over an extracted and transformed portion of the data that is required when a standard data-mining tool is used.

Keywords: Data warehouse; Data mining; Association rules; Dimensional model; Database systems; Knowledge discovery

1. Introduction

Data mining is defined as a process whose objective is to identify valid, novel, potentially useful and understandable correlations and patterns in existing data, using a broad spectrum of formalisms and techniques [9,23]. Mining transactional (operational) databases, containing data related to current day-to-day organizational activities could be of limited use in certain situations. However, the most appropriate and fertile source of data for meaningful and effective data mining is the corporate data warehouse, which contains all the information from the operational data sources that has analytical value. This information is integrated from multiple operational (and external) sources, it usually reflects substantially longer history than the data in operational sources, and it is structured specifically for analytical purposes.

The data stored in the data warehouse captures many different aspects of the business process across

[*] Corresponding author. Tel.: +1 312 915 6662.
E-mail addresses: njukic@luc.edu (N. Jukić), evtimov@cs.uchicago.edu (S. Nestorov).
[1] Tel.: +1 773 702 3497.

0167-9236/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2005.07.009

various functional areas such as manufacturing, distribution, sales, and marketing. This data explicitly and implicitly reflects customer patterns and trends, business practices, organizational strategies, financial conditions, know-how, and other knowledge of potentially great value to the organization.

Unfortunately, many organizations often underutilize their already constructed data warehouses [12,13]. While some information and facts can be gleaned from the data warehouse directly, through the utilization of standard on-line analytical processing (OLAP), much more remains hidden as implicit patterns and trends. The standard OLAP tools have been performing well their primary reporting function where the criteria for aggregating and presenting data are specified explicitly and ahead of time. However, it is the discovery of information based on implicit and previously unknown patterns that often yields important insights into the business and its customers, and may lead to unlocking hidden potential of already collected information. Such discoveries require utilization of data mining methods.

One of the most important and successful data mining methods for finding new patterns and correlations is association-rule mining. Typically, if an organization wants to employ association-rule mining on their data warehouse data, it has to use a separate data-mining tool. Before the analysis is to be performed, the data must be retrieved from the database repository that stores the data warehouse, transformed to fit the requirements of the data-mining tool, and then stored into a separate repository. This is often a cumbersome and time-consuming process. In this paper we describe a direct approach to association-rule data mining within data warehouses that utilizes the query processing power of the data warehouse itself without using a separate data mining tool.

In addition, our new approach is designed to answer a variety of questions based on the entire set of data stored in the data warehouse, in contrast to the regular association-rule methods which are more suited for mining selected portions of the data warehouse. As we will show, the answers facilitated by our approach have a potential to greatly improve the insight and the actionability of the discovered knowledge.

This paper is organized as follows: in Section 2 we describe the concept of association-rule data mining and give an overview of the current limitations of association-rule data mining practices for data warehouses. Section 3 is the focal point of the paper. In it we first introduce and define the concept of qualified association rules. In 3.1 we discuss how qualified association rules broaden the scope and actionability of the discovered knowledge. In 3.2 we describe why existing methods cannot be feasibly used to find qualified association rules. In 3.3 3.4 and 3.5 we offer details of our own method for finding qualified association rules. In Section 4 we describe an illustrative experimental performance study of mining real world data that uses the new method we introduced. And finally, in Section 5 we offer conclusions.

2. Association-rule data mining in data warehouses

The standard association-rule mining [1,2] discovers correlations among items within transactions. The prototypical example of utilizing association-rule mining is determining what products are found together in a basket at a checkout line at the supermarket; hence the often-used term: market basket analysis [4]. The correlations are expressed in the following form:

Transactions that contain X are likely to contain Y as well noted as $X \rightarrow Y$, where X and Y represent sets of transaction items. There are two important quantities measured for every association rule: *support* and *confidence*. The support is the fraction of transactions that contain both X and Y items. The confidence is the fraction of transactions containing items X, which also contain items Y. Intuitively, the support measures the significance of the rule, so we are interested in rules with relatively high support. The confidence measures the strength of the correlation, so rules with low confidence are not meaningful, even if their support is high. A rule in this context is the relationship among transaction items with enough support and confidence.

The standard association rule mining process employs the basic a-priori algorithm [1,3] for finding sets of items with high support (often called *frequent itemsets*). The crux of the a-priori approach is the observation that a set of items I can be frequent only if all proper subsets *sub(I)* are also frequent in recorded transactions. Based on this observation, a-priori applies step-wise pruning to the sets of items