A Hybrid Method for High-Utility Itemsets Mining in Large High-Dimensional Data

Guangzhu Yu, Donghua University, China
Shihuang Shao, Donghua University, China
Bin Luo, Guangdong University of Technology, China
Xianhui Zeng, Donghua University, China

ABSTRACT

Existing algorithms for high-utility itemsets mining are column enumeration based, adopting an Apriori-like candidate set generation-and-test approach, and thus are inadequate in datasets with high dimensions or long patterns. To solve the problem, this paper proposed a hybrid model and a row enumeration-based algorithm, i.e., Inter-transaction, to discover high-utility itemsets from two directions: an existing algorithm can be used to seek short high-utility itemsets from the bottom, while Inter-transaction can be used to seek long high-utility itemsets from the top. Inter-transaction makes full use of the characteristic that there are few common items between or among long transactions. By intersecting relevant transactions, the new algorithm can identify long high-utility itemsets, without extending short itemsets step by step. In addition, we also developed new pruning strategies and an optimization technique to improve the performance of Inter-transaction.

Keywords: high-utility itemset; hybrid model; large high-dimensional data; partition method

INTRODUCTION

Traditional association rule mining (ARM) assumes that the important must be frequent and aims at discovering frequent itemsets. However, in the real world, the frequent is not necessarily important; some infrequent itemsets may have high utility values and thus are important to users. For example, in a transaction database, there are 1000 sale records of milk which occupy 10% of the total transaction number, contributing 1% of the total profit. In the meantime, there are 600 sale records of birthday cake that occupy 6% of the total transaction number, contributing 5% of the total profit. If the support threshold is 8%, according to traditional algorithms for frequent itemset mining, milk will be reported as a frequent itemset and birthday cake will be
ignored. But in fact, market professionals must be more interested in birthday cake because it contributes a larger portion to total profit than milk. The example shows that support is not sufficient to reflect users’ interests and such mining results might not be satisfactory.

According to Expectancy Theory (Vroom, 1964), we have the well-known equation “motivation = probability * utility”, which says that motivation is determined by the utility of making a decision and the probability of success. In retailing field, users are not only interested in the frequency of occurrence of an itemset (support), but also its utility. So a decision-oriented ARM algorithm should output both the support and the utility of interesting patterns. For this reason, utility-based ARM (or utility mining for short) has been proposed to discover all the itemsets in a database with utility values higher than a user-specified threshold.

Utility of an item is a subjective term, depending on users and applications; it could be measured in terms of profit, cost, risk, aesthetic value or other expressions of user preference. For easy understanding, in this paper, “utility” is viewed as economic utility such as sales profit, and all databases are regarded as transaction databases, so that we can define the utility of an item as the product of quantity sold and the unit profit of the item. Table 1 is an example of a simplified transaction database where the total utility value is 162. The number in each transaction in Table 1 is the sales profit of each item. If \( s(X) \) and \( u(X) \) represent the support and utility of itemset \( X \) respectively (for details, refer to definition 4 in Section 2), then \( u(A,B)=43 \), \( s(A,B)=5 \), \( u(A,B,C)=54 \), \( s(A,B,C)=3 \), \( u(A,B,C,D)=45 \), \( s(A,B,C,D)=2 \), \( u(A,B,C,D,E)=57 \), and \( s(A,B,C,D,E)=2 \).

If the support threshold is 3 and the utility threshold is 50, \( \{A, B\} \) is a frequent but not a high-utility itemset. On the other hand, \( \{A,B,C\} \) is both a frequent and high-utility itemset, \( \{A,B,C,D\} \) is neither a frequent nor a high-utility itemset and \( \{A,B,C,D,E\} \) is a high-utility but non-frequent itemset.

From the above example, we can draw a conclusion: downward closure property, which states that if an itemset is frequent by support, all its nonempty subsets must also be frequent by support, does not apply to utility mining. Relevant studies have shown that utility constraint is neither anti-monotone, monotone, succinct, nor convertible (Shen, Zhang, & Yang, 2000; Yao, Hamilton, & Geng, 2006). Because of this property, most algorithms for frequent pattern mining such as FP-Tree (Han, Pei, Yin, & Mao, 2004), CARPENTER (Pang, Cong, Tung, Yang, & Zaki, 2003), Tree-projection (Agarwal, Agrawal, & Prasad, 2000) and so on cannot be used to find high-utility itemsets.

Lots of researches have been conducted to improve the usefulness of traditional ARM. Ngan, Lam, Wong, and Fu (2005) proposed an algorithm called COFI+BOMO to mine N-most interesting itemsets, but the interestingness measure still depends on the support. Value added association rule (Wang, Zhou, & Han, 2002; Lin, Yao, & Louie, 2002) extends traditional association rule by taking the semantics of data into consideration. The difference between Wang et al. (2002) and Lin et al. (2002) is that price and quantity of supermarket sales are considered in the former, while the latter tries to attach a value to every item in the database and use the added values to rank the association rule. Weighted association rule gives up treating all the items and all the transactions uniformly

<table>
<thead>
<tr>
<th>Table 1. A transaction database</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>T1</td>
</tr>
<tr>
<td>T2</td>
</tr>
<tr>
<td>T3</td>
</tr>
<tr>
<td>T4</td>
</tr>
<tr>
<td>T5</td>
</tr>
<tr>
<td>T6</td>
</tr>
<tr>
<td>T7</td>
</tr>
<tr>
<td>T8</td>
</tr>
<tr>
<td>T9</td>
</tr>
<tr>
<td>T10</td>
</tr>
</tbody>
</table>

Copyright © 2009, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
Related Content

Finding Associations in Composite Data Sets: The CFARM Algorithm
[www.irma-international.org/article/finding-associations-composite-data-sets/55077/](http://www.irma-international.org/article/finding-associations-composite-data-sets/55077/)

Data Mining in Aviation Safety Data Analysis
[www.irma-international.org/chapter/data-mining-aviation-safety-data/29071/](http://www.irma-international.org/chapter/data-mining-aviation-safety-data/29071/)

Discovering Surprising Instances of Simpson's Paradox in Hierarchical Multidimensional Data
[www.irma-international.org/article/discovering-surprising-instances-simpson-paradox/1762/](http://www.irma-international.org/article/discovering-surprising-instances-simpson-paradox/1762/)

Time Series Analysis and Structural Change Detection
[www.irma-international.org/chapter/time-series-analysis-structural-change/39649/](http://www.irma-international.org/chapter/time-series-analysis-structural-change/39649/)

Incremental Algorithm for Discovering Frequent Subsequences in Multiple Data Streams
[www.irma-international.org/article/incremental-algorithm-discovering-frequent-subsequences/58635/](http://www.irma-international.org/article/incremental-algorithm-discovering-frequent-subsequences/58635/)