Expert Systems with Applications 41 (2014) 3450–3459

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

On-shelf utility mining with negative item values

Guo-Cheng Lan [a,*], Tzung-Pei Hong [b,c], Jen-Peng Huang [d], Vincent S. Tseng [a]

[a] Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan City 701, Taiwan
[b] Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung City 811, Taiwan
[c] Department of Computer Science and Engineering, National Sun Yat-Sen University, Kaohsiung City 804, Taiwan
[d] Department of Information Management, Southern Taiwan University of Science and Technology, Tainan City 710, Taiwan

ARTICLE INFO

Keywords:
Data mining
Utility mining
On-shelf utility mining
High on-shelf utility itemset
Negative profit

ABSTRACT

On-shelf utility mining has recently received interest in the data mining field due to its practical considerations. On-shelf utility mining considers not only profits and quantities of items in transactions but also their on-shelf time periods in stores. Profit values of items in traditional on-shelf utility mining are considered as being positive. However, in real-world applications, items may be associated with negative profit values. This paper proposes an efficient three-scan mining approach to efficiently find high on-shelf utility itemsets with negative profit values from temporal databases. In particular, an effective itemset generation method is developed to avoid generating a large number of redundant candidates and to effectively reduce the number of data scans in mining. Experimental results for several synthetic and real datasets show that the proposed approach has good performance in pruning effectiveness and execution efficiency.

1. Introduction

Data mining techniques can extract useful information from databases. Among various techniques in data mining, association-rule mining is important due to its consideration of the co-occurrence relationship of items. That is, association-rule mining techniques can be applied to find items with high frequency in a set of transactions (Agrawal, Imielinksi, & Swami, 1993), and thus have many practical applications, such as analyzing purchasing behaviors in retailing stores, traversal behaviors on websites, and so on. Agrawal et al. proposed the most well-known algorithm for mining association rules from a transaction database, called Apriori (Agrawal & Srikant, 1994). However, when the occurrences of items are considered, it is insufficient to evaluate the significance of items in a database. The main reason is that a transaction in a transaction database usually also includes the quantities bought of items and item prices. The same significance in association-rule mining is assumed for all items in a database and thus the actual significance of an itemset cannot be easily recognized. To address the above problem, Chan et al. proposed utility mining (Chan, Yang, & Shen, 2003), which considers both the profits and quantities of products (items) in a set of transactions to evaluate actual utility values of product combinations (itemsets). In their study, itemsets whose actual utility values are larger than or equal to a predefined minimum utility threshold are output as high-utility itemsets. Several studies have modified utility mining for various practical applications, such as improving its performance and the development of incremental utility mining and stream utility mining. The profits of items in these studies were assumed to be positive values.

Temporal data mining has attracted a lot of attention due to its practicality (Ale & Rossi, 2000; Chang, Chen, & Lee, 2002; Lee, Lin, & Chen, 2001; Li, Ning, Wang, & Jajodia, 2003; Ozden, Ramaswamy, & Silberschatz, 1998; Roddick & Spiliopoilou, 2002). For example, consider the product combination {overcoats, stockings}. This combination may not be frequent throughout the entire database, but may have high frequency in winter. Mining time-related knowledge is thus interesting and useful. Of note, since some products in a store may be put on the shelf and taken off it repeatedly, some biases may exist in the discovered association rules. Thus it is necessary to consider the on-shelf time periods of products.

To address this problem, Lan et al. presented a new issue named on-shelf utility mining (Lan, Hong, & Tseng, 2011) to obtain more accurate utility values of itemsets in temporal databases. In Lan et al.'s study (Lan et al., 2011), on-shelf utility mining considered not only quantities and profits of items in transactions but also the on-shelf time periods of the items. Thus, using on-shelf time periods, the actual utility values of itemsets in a temporal database can be accurately evaluated, and also a two-phase algorithm (named *TP-HOU*) was designed to find high-on-shelf-utility itemsets in temporal databases.

* Corresponding author. Tel.: +886 920 231609.
E-mail addresses: rrfoheiay@gmail.com (G.-C. Lan), tphong@nuk.edu.tw (T.-P. Hong), jehuang@mail.stust.edu.tw (J.-P. Huang), tsengsm@mail.ncku.edu.tw (V.S. Tseng).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.10.049

Although the traditional utility mining considered the profits and quantities of items to find useful information with high-profit from transactions databases, the profit values of all items in databases were usually identified as positive values. In real-world applications, the profit values of items in stores may be possibly considered as negative values. For example, when supermarkets want to promote some specific products to attract customers' attentions, the specific products and some free products are usually packaged to sell together. In such scenario, customers can receive not only the specific products but also the free products. With the help of the free products, the supermarkets may earn more profits since customers may buy more relevant products of the free products. However, the profit values of free products in the scenario are usually negative values during supermarket promotion due to the consideration of the costs of the free products. To solve this problem, Chu et al. extended the traditional two-phase utility mining approach to find high-utility itemsets with the consideration of negative item profits from transaction databases (Chu, Tseng, & Liang, 2008). However, since the on-shelf time periods of items were not considered in their study (Chu et al., 2008), accurate utility values of the items were not obtained and thus some itemsets with high utility values may not be found using this approach. To address this, we propose an efficient three-scan mining approach (abbreviated as *TS-HOUN*) for discovering high utility itemsets with negative item values from a transaction database. The major contributions of this work are summarized as follows:

1. This work presents a new issue named on-shelf utility mining with the consideration of negative item values. Different from the existing *TP-HOU* approach in Lan et al. (2011), an efficient three-scan approach for discovering high utility itemsets with negative item values is proposed in this study. To our best knowledge, this is the first work on mining high utility itemsets with considering both of the on-shelf time periods and negative values of items in the field of utility mining.
2. An effective upper-bound model with the consideration of with negative item values is designed in the proposed *TS-HOUN* approach to keep the downward-closure property in mining. In addition, an effective itemset generation method is designed to quickly and directly produce promising itemsets from transactions by using the relationship information of items, and thus unnecessary evaluation in mining can also be avoided. Based on the model and method, the proposed *TS-HOUN* only needs three data scans to finish the mining tasks.
3. In the experimental evaluation, several synthetic and real datasets are used to evaluate the performance of the proposed *TS-HOUN* algorithm with the baseline comparison algorithm. The results show the proposed *TS-HOUN* has better performance in terms of both of pruning effectiveness and execution efficiency with respect to varied parameters.

The rest of this paper is organized as follows. Related works are reviewed in Section 2. The problem to be solved and definitions are stated in Section 3, and the proposed three-scan mining algorithm for finding high-on-shelf-utility itemsets with negative item values is described in Section 4. An example is given to illustrate the execution of the proposed algorithm in Section 5. The experimental results are shown in Section 6, and conclusions and suggestions for future work are given in Section 7.

2. Review of related works

In this section, studies related to temporal association-rule mining, utility mining, and on-shelf utility mining are briefly reviewed.

2.1. Temporal association-rule mining

Traditional association-rule mining (Agrawal & Srikant, 1994; Agrawal et al., 1993) can be used to find the frequency relationships among items in various types of databases. In real-world applications, however, each transaction in a database usually includes a time stamp, such as transactions with occurrence time stamps in retailing business stores. For the analysis of temporal data, temporal data mining was then proposed to find temporal patterns and regularity from a set of data. Forms of temporal patterns include sequential patterns (Pei et al., 2004), periodical association rules (Roddick & Spiliopoilou, 2002), cyclic association rules (Ozden et al., 1998), and calendar association rules (Li et al., 2003).

Ale et al. first proposed a mining approach for finding temporal association rules from a temporal transaction database (Ale & Rossi, 2000). In their study (Agrawal et al., 1993), however, they only considered the transaction periods of products, but not their on-shelf periods. Some products might be frequent if only their transaction periods are considered, but infrequent if their on-shelf periods are considered. Afterward, Chang et al. considered the contiguous time periods of products, which was defined as the contiguous time periods between a product's first and last on-shelf time periods (Chang et al., 2002). That is, the on-shelf periods of a product combination in their study were taken as the common on-shelf periods of all the products in the combination. Based on Chang et al.'s study (Chang et al., 2002), Lee et al. extended the concept of common on-shelf periods to publication databases, such as transaction databases for bookstores or DVD rental stores, to find general temporal association rules (Lee et al., 2001). However, they did not consider that a product might be put on a shelf and taken off it multiple times. Thus, the present study considers the individual on-shelf time periods of products instead of the whole on-shelf period.

2.2. Utility mining

In real-world applications, transactions usually contain the profits and quantities of products. Some high-profit products may rarely occur in a transaction database. Product combinations with high profit but low frequency may not be found using association-rule mining approaches. To address this problem, Chan et al. proposed a new research issue named utility mining (Chan et al., 2003), which considered not only the quantities of items but also their profits in a set of transactions. By using these two kinds of information, the actual utility of an item in a database can be more accurately recognized compared to that obtained using traditional association-rule mining, which only considers the frequencies of items.

However, the downward-closure property in utility mining cannot be maintained due to its utility function. To address this problem, Liu et al. proposed a two-phase mining algorithm for discovering high-utility itemsets from a database by adopting a new downward-closure property (Liu & Qu, 2012) called the transaction-weighted utilization (*TWU*) model. In the *TWU* model, the summation (called transaction utility) of utility values of all items in a transaction is used as the upper-bound value of any sub-itemset in the transaction. The transaction-weighted utility of an itemset is defined as the sum of the transaction utility values of the transactions including the itemset in the database (Liu & Qu, 2012). In addition, Liu et al. proposed a two-phase algorithm to find high-utility itemsets from a database. In the first phase, all high-utility upper-bound itemsets are found from a database by using the *TWU* model (Liu & Qu, 2012). Next in the second phase, an additional data scan is required to find the actual utility values of these high-utility upper-bound itemsets. Itemsets with utility