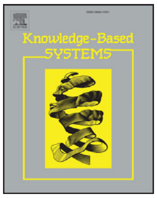
Knowledge-Based Systems 104 (2016) 39–51

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



Discovering highly expected utility itemsets for revenue prediction

Cheng-Hsiung Weng*

Department of Management Information Systems, Central Taiwan University of Science and Technology, Taichung 406, Taiwan, Republic of China.

ARTICLE INFO

Article history:
Received 12 October 2015
Revised 3 April 2016
Accepted 12 April 2016
Available online 13 April 2016

Keywords:
Knowledge discovery
Data mining
Frequent itemset mining
High-utility itemset mining
Revenue prediction

ABSTRACT

Identifying patterns of items that are purchased frequently and generate high profits is crucial for inventory and profit management. However, neither approaches based on frequent itemsets nor those based on high-utility itemsets (HUIs) can meet this requirement alone. Therefore, we propose a new approach, named the FIHUM algorithm, for identifying frequent HUIs. The novel characteristic of the FIHUM algorithm is that it can effectively identify frequent itemsets with high utility (frequent HUIs) without generating many high-utility candidate itemsets. Moreover, experimental results from retail data sets reveal that the FIHUM algorithm integrates the advantages of frequent itemsets and HUIs. Finally, the highly expected utility itemsets (frequent HUIs) generated using the FIHUM algorithm are suitable for predicting patterns of items that are purchased frequently by customers and generate high profits in next-period transactions.

1. Introduction

Association rule mining (ARM) facilitates detecting correlations among items in large transaction data sets, thus facilitating numerous business decision-making processes. ARM is a crucial data mining approach that enables identifying consumer purchasing behaviors in transaction databases [7]. Agrawal *et al.* [1] introduced ARM and defined it as determining all rules from transaction data that satisfy the minimum support and confidence constraints.

The frequency of an itemset in transactions is unsuitable for some applications because it does not reveal the utility of an itemset, which can be measured according to cost, profit, or other expressions of user preferences. Moreover, a retail business may want to identify its most valuable customers who contribute a major fraction of the company profit. However, frequent itemset mining does not consider the utility (such as quantity or profit) of items, which is crucial for addressing real-world decision problems that require maximizing the utility in an enterprise.

For identifying high-utility itemsets (HUIs) that account for a large portion of total utility, Yao *et al.* [19] defined a utility mining model in which utility is considered a measure of how useful (e.g., profitable) an itemset is. The utility $u(X)$ of an itemset X is defined as the sum of the utilities of X in all transactions containing X. Thus, HUI mining identifies the itemsets with a utility greater than a user-specified minimum utility threshold.

Conventional frequency-based ARM algorithms reflect only the frequency at which itemsets exist in transactions; the frequent itemsets that contribute high revenue or profit (high utility) cannot be identified. Conversely, HUIs, such as those for luxury goods, identified by utility-based itemset mining algorithms may contribute a considerable portion of overall revenue or profit; however, the HUIs that appear frequently in the transactions cannot be identified. Therefore, the risk of low revenue or profit generated from frequent itemsets with low utility and risk of inventory costs generated from HUIs with low frequency are major concerns for managers.

In some applications, the stable revenue generated by frequently purchased items with high utility is the primary concern for business managers. Identifying the patterns of items that are purchased frequently and generate high profits is crucial for inventory and profit management. To address this problem, we propose a new approach for identifying frequent itemsets with high utility (frequent HUIs), as shown in Fig. 1. Moreover, we investigated the differences in predictions and utilities among the three types of patterns (frequent HUIs, frequent itemsets, and HUIs) exhibited by customers in next-period transactions.

The remainder of this paper is organized as follows. Section 2 reviews related work. Problem definitions are provided in Section 3. The proposed algorithm and an example are illustrated in Section 4. Section 5 uses survey data for a case

* Corresponding author. Tel.: +886422391647.
E-mail address: chweng@mgt.ncu.edu.tw

http://dx.doi.org/10.1016/j.knosys.2016.04.009
0950-7051/© 2016 Elsevier B.V. All rights reserved.

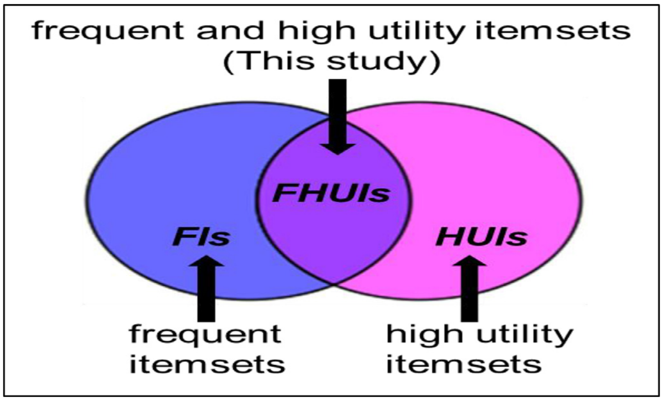


Fig. 1. The aim of this study.

study to demonstrate the usefulness of the proposed algorithm. Conclusions and future work are discussed in Section 6.

2. Related work

2.1. Frequent item set mining

Frequent pattern mining reveals the intrinsic and essential properties of data sets and is the foundation for ARM. Association rules are valuable and unexpected relationships among attributes that satisfy the minimum support and confidence constraints in a database [7]. Agrawal *et al.* [1] introduced ARM, defining it as determining all rules from transaction data that satisfy the minimum support and confidence constraints.

The Apriori algorithm has been widely and successfully used to identify all frequent itemsets in a transaction database. Because of the high effectiveness and widespread use of ARM, many variants of ARM algorithms have been proposed on the basis of the Apriori algorithm. These algorithms assume that each item's weight is equal; that is, that the quantity is 0 or 1.

To consider the difference in each item's weight, a different concept that allows multiple minimum supports for various itemsets has been proposed [14]. However, it is difficult for a user to assign each item's minimum support. In market basket analysis, determining the minimum supports for all items can be challenging. Moreover, conventional ARM methods may generate less important rules with high support and light weight. To solve this problem, Tao *et al.* [18] proposed weighted ARM for considering the weight of each item.

However, the practical usefulness of frequent itemsets is limited by the significance of the discovered itemsets. A frequent itemset reflects only the statistical correlation between items and does not reflect the semantic significance of the items, such as profit.

2.2. High-utility item set mining

To overcome the limitations of conventional ARM, Chan *et al.* [5] proposed HUI mining (HUIM) by defining itemset utility according to price and quantity. HUIM entails predefining an objective function that a user seeks to achieve and then determining useful rules that satisfy that function. Yao *et al.* [19] proposed an HUIM method for market basket analysis. Yao and Hamilton [20] proposed a utility-based itemset mining approach for quantifying their preferences concerning the usefulness of itemsets by using utility values. Hu and Mojsilovic [10] presented an algorithm for identifying itemsets through combinations of a few high-utility items (rules) that satisfy certain conditions as a group and maximize a predefined objective function.

Existing utility mining methods produce several patterns, making it difficult for users to identify useful patterns in a large set of patterns. Ahmed *et al.* [3] proposed a new tree structure, namely a utility tree based on frequency affinity, and a novel algorithm, namely high-utility interesting pattern mining, for the single-pass mining of HUIPs from a database. Lin *et al.* [12] designed a high-utility pattern (HUP) tree and proposed the HUP-growth mining algorithm, which derives HUPs effectively and efficiently. Lin *et al.* [13] used a maximal itemset property and proposed an algorithm called UMMI (high utility mining by using the maximal itemset property) to considerably reduce the number of potential itemsets in the first step. Shie *et al.* [17] proposed GUIDE (generation of maximal high-utility itemsets from data streams) to identify maximal HUIs from data streams with different models.

Mining HUPs over data streams has become challenging because of the level-wise candidate generation-and-test problem in data mining. Chu *et al.* [6] proposed a novel method, namely temporal high-utility itemset (THUI) mining, for mining THUIs from data streams efficiently and effectively. Ahmed *et al.* [4] proposed a HUS tree (high-utility stream tree) for executing HUPMS (high-utility pattern mining over stream data) for incremental and interactive HUP mining over data streams with a sliding window. Most recent studies have focused on improving the efficiency of HUI mining, as summarized in Table 1.

Table 1 shows that most utility itemset mining algorithms apply heuristics to improve performance rather than identify frequent HUIs. The algorithm proposed here was developed according to the concept of identifying frequent HUIs.

2.3. Comparison of the proposed algorithm with those in the literature

Table 2 lists the differences between our algorithm and those proposed in prior studies.

3. Problem definitions

In this section, we define the problem of identifying frequent HUIs.

Definition 1. Let $I=\{i_1, i_2, \ldots, i_m\}$ be a set of items and $D=\{T_1, T_2, \ldots, T_m\}$ be a transaction data set, where each transaction T is a set of items such that $T \subseteq I$. Let X be a set of items. A transaction T is said to contain X if and only if $X \subseteq T$. The itemsets XY are present in the data set D with *support sup*, where *sup* is the percentage of the transaction in D that contains $X \cup Y$. The formal expression for $sup(X \cup Y)$ is as follows:

$$sup(X \cup Y) = \frac{|X \cup Y|}{|D|},$$

where $|D|$ denotes the number of transactions in D and $|X \cup Y|$ denotes the number of transactions containing $X \cup Y$ in D.

Example 1. Assume that we have a data set (D) containing 10 transactions, as shown in Table 3. The $sup(a)$ and $sup(ae)$ in the transaction set are as follows: $sup(a)=5/10=0.5$, $sup(e)=8/10=0.8$, and $sup(ae)=3/10=0.3$.

Definition 2. Given a user-specified support threshold σ_{sup}, an itemset X is frequent if $sup(X) \geq \sigma_{sup}$.

Example 2. Assume that the support threshold $\sigma_{sup}=0.3$. The itemsets *a*, *e*, and *ae* are all frequent itemsets because $sup(a)=0.5$, $sup(e)=0.8$, and $sup(ae)=0.3$ are not smaller than σ_{sup}.

Definition 3. The utility of item i_p in transaction T_q can be defined as follows:

$$u(i_p, T_q) = o(i_p, T_q) \times s(i_p, T_q),$$