Advanced Engineering Informatics 33 (2017) 29–43

Contents lists available at ScienceDirect

Advanced Engineering Informatics

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

Full length article

A two-phase approach to mine short-period high-utility itemsets in transactional databases

Jerry Chun-Wei Lin [a,*], Jiexiong Zhang [a], Philippe Fournier-Viger [b], Tzung-Pei Hong [c,d], Ji Zhang [e]

[a] School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China
[b] School of Natural Sciences and Humanities, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China
[c] Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung, Taiwan
[d] Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan
[e] School of Agricultural, Computational and Environmental Sciences, University of Southern Queensland, Australia

ARTICLE INFO

Article history:
Received 3 May 2016
Received in revised form 9 February 2017
Accepted 29 April 2017

Keywords:
High-utility itemsets
Periodic high-utility itemsets
SPHUIs
Two-phase
Data mining

ABSTRACT

The discovery of high-utility itemsets (HUIs) in transactional databases has attracted much interest from researchers in recent years since it can uncover hidden information that is useful for decision making, and it is widely used in many domains. Nonetheless, traditional methods for high-utility itemset mining (HUIM) utilize the utility measure as sole criterion to determine which item/sets should be presented to the user. These methods ignore the timestamps of transactions and do not consider the period constraint. Hence, these algorithms often finds HUIs that are profitable but that seldom occur in transactions. In this paper, we address this limitation of previous methods by pushing the period constraint in the HUI mining process. A new framework called short-period high-utility itemset mining (SPHUIM) is designed to identify patterns in a transactional database that appear regularly, are profitable, and also yield a high utility under the period constraint. The aim of discovering short-period high-utility itemsets (SPHUI) is hence to identify patterns that are interesting both in terms of period and utility. The paper proposes a baseline two-phase short-period high-utility itemset ($SPHUI_{TP}$) mining algorithm to mine SPHUIs in a level-wise manner. Then, to reduce the search space of the $SPHUI_{TP}$ algorithm and speed up the discovery of SPHUIs, two pruning strategies are developed and integrated in the baseline algorithm. The resulting algorithms are denoted as $SPHUI_{MT}$ and $SPHUI_{TID}$, respectively. Substantial experiments both on real-life and synthetic datasets show that the three proposed algorithms can efficiently and effectively discover the complete set of SPHUIs, and that considering the short-period constraint and the utility measure can greatly reduce the number of patterns found.

1. Introduction

Association rule mining (ARM) [1–3] plays an important role in data mining. The main objective of ARM is to discover interesting associations or patterns in transaction databases. It is performed by first extracting sets of items that appear frequently in databases according to a minimum support threshold, called the frequent itemsets (FIs). Then, the FIs are used to derive association rules (ARs) respecting a minimum confidence threshold [1]. Albeit traditional frequent itemset mining (FIM) and ARM techniques are useful, they also have some important limitations. Some of the main limitations are that items are not allowed to occur more than once in each transaction, and that all items are considered as equally important. But in real-life situations, these assumptions do not often hold [4]. Moreover, FIM and ARM are mainly used to discover patterns that appear frequently in databases, and do not take other criteria into account for discovering patterns such as the importance, unit profits, or weights of items.

Although FIM and ARM have been designed to analyze customer transactions for market basket analysis at first, these tasks are defined quite generally and hence have been applied in many other fields related to science and engineering [1,5,6]. For example, an important application of ARM is clickstream analysis [7], that is the analysis of the behavior of persons visiting a website or using a software. A clickstream is defined as a set of records indicating the webpages or user interface elements that each user has visited or clicked. Analyzing clickstreams allows researchers to discover interesting and useful information about the behavior or

* Corresponding author.
E-mail addresses: jerrylin@ieee.org (J.C.-W. Lin), jiexiongzhang@ikelab.net (J. Zhang), philfv@hitsz.edu.cn (P. Fournier-Viger), tphong@nuk.edu.tw (T.-P. Hong), ji.zhang@usq.edu.au (J. Zhang).

http://dx.doi.org/10.1016/j.aei.2017.04.007
1474-0346/

preferences of persons. This information can be used to support crucial tasks such as improving the design of a website or software, predicting the next webpage that a user would visit to reduce the loading time, predicting the purchase behavior of users, and developing critical marketing strategies. When applying FIM or ARM for clickstream analysis, a user is represented as a transaction, and webpages or user interface elements are represented as items. Besides clickstream and market basket analysis, FIM and ARM can be applied in many other fields [8,9].

Recently, to address the aforementioned limitations of FIM, the task of high-utility itemset mining (HUIM) has been introduced [4]. HUIM generalizes the task of FIM by considering that items may appear more than once in each transaction and that not all items are equally important. In the context of market basket analysis, this means that purchase quantities of items in transactions can be taken into account, as well as the unit profits of items. Differently than FIM and ARM, which aim at finding itemsets that appear frequently in a database, HUIM aims at discovering itemsets that are important to the user, i.e. that have a high utility. In the context of market basket analysis, a high-utility itemset (HUI) is a set of items that yield a high profit. Discovering HUIs is very useful for analyzing customer data to understand customer behavior and develop marketing strategies. Note that although HUIM has been proposed in the context of market basket analysis, it can also applied in many other fields related to science and engineering such as biomedicine and clickstream analysis [4,10].

In HUIM, each item appearing in a transaction is annotated with a purchase quantity, and has its own weight indicating its importance (e.g. unit profit). An item/set is said to be a high-utility item/set if its utility value in a database is no less than a user-specified minimum utility threshold, which is normally defined by the user. It is important to notice that HUIs are not necessarily FIs. For example, the sale of diamonds in a retail store may be quite rare compared to the sale of other items, but it may still be highly profitable, and thus be considered a HUI. Hence, HUIM is a useful mean of discovering patterns that are important in a transaction database (e.g. yield a high profit). To apply HUIM for clickstream analysis, the number of clicks or the time spent on each webpage or user interface element can be viewed as the quantities of items in transactions. Thus, HUIM can be used to discover the sets of webpages or interface elements where users spend most of their time [10]. This information is useful to perform various important activities such as improving website/interface design, and analyzing the most popular content on a website.

Although HUIM is a useful alternative to traditional FIM and ARM, a major limitation is that it does not consider the time dimension, and more specifically the utility of items with respect to time periods. This is a problem because in real-life applications, some items are more popular during specific time periods. For example, in market basket analysis, the sale volume of jackets is higher during the winter, while other products such as ice cream and swimming suits are more popular during the summer. The same observation holds for clickstream analysis, where webpages may be more popular during specific time periods. To address this limitation of HUIM and find items that are important (e.g. profitable) in at least some time period (have a utility that is no less than a threshold in a time period), this paper proposes a method to integrate the concept of periodic constraint in HUIM.

This proposed task is applicable to various real-life scenarios. As an example, in this paragraph, we describe in more details its usefulness for market basket analysis. For retailers, it is important to know the sets of products that have a high sale volume because these products may require to adopt quick changes in terms of inventory and capital expenditure. Having this information can thus help companies to survive in rapidly changing markets. Furthermore, it is common that products that yield a high profit carry a high risk of declining or unstable sales, and thus represent a considerable risk for companies in terms of funding. For instance, some products such as diamonds have very high unit profits. But if a company buys a too large stock of diamonds and cannot sell them, it may lead to major problems for the company, as it may run out of funding for buying other products, and this may even lead to bankruptcy, especially in times of economic crisis. For the sake of obtaining more stable revenue and reducing risks, retailers may prefer to sell products that yield a high profit and have a high sale volume. Besides, another critical aspect for retail stores is the management of supplies. Suppliers must not only consider profits but also the sale cycles of products. In general, products having short sale cycles are more popular and considered more interesting. For example, from the point of view of most businessmen, the sale of electronic products is considered as safer than the sale of diamonds since electronic products are renewed every year as technology changes. The total profit generated by the sale of electronic products such as smartphones is not less than that of diamonds due to the high sale volume of it. Also, products having short sale cycles are interesting for retail stores because they can quickly yield a high profit as the sale cycles of these products are shorter. On the other hand, a considerable risk of products having long sale cycles is that the sale of these products may become profitable only after a long time, and that the risk of having poor sales is greater. Consequently, many companies prefer selling products having short sale cycles such as electronics instead of products having long sale cycles such as diamonds.

To identify patterns while considering the periodic constraint, Tanbeer et al. [11] studied the discovery of regular patterns, that is patterns appearing in all periods. Thereafter, Rashid et al. [12] proposed a pattern-growth approach to discover patterns occurring frequently and regularly in databases. However, a drawback of those approaches is that they discover frequent patterns that occur regularly in data but do not consider the utility of patterns. For example, consider the analysis of customer transaction data. These approaches will tend to find patterns containing items such as bread that have high daily selling frequencies, but are not much profitable. On the other hand, many papers have been published on HUIM to discover high-utility patterns, that is patterns yielding high profits. But an important drawback of these studies is that they do not consider how regularly the patterns appear in a database. Hence, these algorithms may discover many high-utility patterns that have small sale volumes, such as diamonds. To find patterns that are useful, it is important to consider both the utility and regularity of patterns, to find the products that yield a high utility over short periods of time. For example, Apple will launch every year a new model of its iPhone product. The new model is considered has selling well during the holiday season since it is always introduced around September. Discovering this type of patterns in data can help companies or shops to increase their profit by reducing the stocks of older products. To the best of our knowledge, this type of patterns has not been studied yet.

To address these limitations of previous studies, this paper presents a novel framework, named short-period high-utility itemset mining (SPHUIM) to mine short-period high-utility itemsets (SPHUIs). The proposed framework is designed to discover high-utility patterns having high occurrence frequencies in each short period. The proposed framework considers both the short-period and utility constraints to find itemsets that both appear regularly and have a high-utility in databases. The major contributions of this paper are as follows:

1. A novel framework called SPHUIM is presented to mine the set of SPHUIs in which each pattern has a high utility and occurs regularly in each sale period. SPHUIs are different from traditional HUIs since they must satisfy constraints on the utility