Advanced Engineering Informatics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Advanced Engineering Informatics

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

Efficient algorithms for mining up-to-date high-utility patterns

Jerry Chun-Wei Lin [a,b,*], Wensheng Gan [a], Tzung-Pei Hong [c,d], Vincent S. Tseng [e]

[a] Innovative Information Industry Research Center (IIIRC), School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, HIT Campus Shenzhen University Town Xili, Shenzhen 518055, PR China
[b] Shenzhen Key Laboratory of Internet Information Collaboration, School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, HIT Campus Shenzhen University Town Xili, Shenzhen 518055, PR China
[c] Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung, Taiwan, ROC
[d] Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan, ROC
[e] Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, ROC

ARTICLE INFO

Article history:
Received 10 February 2015
Received in revised form 8 June 2015
Accepted 15 June 2015
Available online xxxx

Keywords:
Data mining
Utility mining
Up-to-date high-utility patterns
Level-wise
UDU-list structures

ABSTRACT

High-utility pattern mining (HUPM) is an emerging topic in recent years instead of association-rule mining to discover more interesting and useful information for decision making. Many algorithms have been developed to find high-utility patterns (HUPs) from quantitative databases without considering timestamp of patterns, especially in recent intervals. A pattern may not be a HUP in an entire database but may be a HUP in recent intervals. In this paper, a new concept namely up-to-date high-utility pattern (UDHUP) is designed. It considers not only utility measure but also timestamp factor to discover the recent HUPs. The UDHUP-apriori is first proposed to mine UDHUPs in a level-wise way. Since UDHUP-apriori uses Apriori-like approach to recursively derive UDHUPs, a second UDHUP-list algorithm is then presented to efficiently discover UDHUPs based on the developed UDU-list structures and a pruning strategy without candidate generation, thus speeding up the mining process. A flexible minimum-length strategy with two specific lifetimes is also designed to find more efficient UDHUPs based on a users' specification. Experiments are conducted to evaluate the performance of the proposed two algorithms in terms of execution time, memory consumption, and number of generated UDHUPs in several real-world and synthetic datasets.

1. Introduction

Knowledge Discovery in Databases (KDD) is a process used to discover meaningful and useful information from a collection of data [1–4]. Depending on different requirements in various domains and applications, the discovered knowledge can be generally classified as association rules (ARs) [2], sequential patterns [5], classification [6], clustering [7] or high-utility itemsets [8–12], among others. Among them, ARs are the most commonly used knowledge in KDD, which can be used to represent the correlation relationships among items or itemsets in transactional databases. Agrawal et al. first presented Apriori algorithm [2] to level-wisely generate and test candidates for mining ARs in two phases. In the first phase, the candidate itemsets are level-wisely discovered to produce frequent itemsets based on minimum support threshold. The remaining frequent itemsets are then used to infer ARs based on minimum confidence threshold. Many algorithms have been proposed to efficiently mine the desired frequent itemsets or ARs in databases [2,4,13].

Temporal data mining [14,15] is another attractive way to find temporal patterns and regularities from temporal databases, which can be used to reveal the ordered correlation of itemsets along with timestamp. For example, the sales of soft drinks in summer and the sales of mittens in winter should be higher than those in the other seasons. The seasonal or periodic behaviors can only be discovered when the window size is properly set. The fixed window size may, however, hide the important information of the purchase itemsets. To solve the limitations of temporal data mining, Hong et al. designed the up-to-date pattern mining to represent not only the frequent itemsets in the entire database but also the up-to-date information from its past timestamp to the current one [16]. Based on the up-to-date concept, an itemset may not be frequent (large) for an entire database but may be large up to date information since the itemset seldom occurs early and may

* Corresponding author at: School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, HIT Campus Shenzhen University Town Xili, Shenzhen 518055, PR China.
E-mail addresses: jerrylin@ieee.org (J.C.-W. Lin), wsgan001@gmail.com (W. Gan), tphong@nuk.edu.tw (T.-P. Hong), tsengsm@mail.ncku.edu.tw (V.S. Tseng).

http://dx.doi.org/10.1016/j.aei.2015.06.002
1474-0346/

often occur lately. The up-to-date patterns include the recent itemsets, which are frequent for a flexible period of time from the current time to its longest past. More useful information of the current usage can thus be provided compared to traditional association rules. For example, a new iPhone may not be considered as a frequent item in the entire database for retailers but may be concerned as a popular sold item during the announcement month or season.

For ARs, only the frequency of an item or itemset in a transaction is considered, which does not reflect any other factors such as price, quantity or profit. The same significance is assumed for all the items in ARs, in which the actual significance of an itemset cannot be easily recognized. High-utility pattern mining (HUPM) was thus proposed to concern both profits (external utility) and sold quantities (internal utility) of the purchase patterns [8–12,17]. A pattern is concerned as a high-utility pattern (HUP) if its utility is no less than the pre-defined minimum utility threshold. Liu et al. first proposed a Two-Phase model [10] to keep *transaction-weighted downward closure* (TWDC) property for efficiently finding HUPs in two phases. It first finds the designed high-transaction-weighted utilization itemsets (HTWUIs) level by level. In the second phase, the remaining HTWUIs are used to reveal the actual HUPs with an additional database scan. Based on Two-Phase model, the unpromising candidates can be early pruned, thus reducing computations for mining HUPs. The meaning of "utility" can be defined as various factor based on the users' specification, such as profit, benefit, weight or risk. The discovered information of HUPs can be generally used in various applications, such as decision support systems [8–12,17], as well as a framework of data mining based analysis [18], or knowledge discovery of material science and engineering [19], to aid mangers or retailers for making the efficient decisions or profitable strategies [20].

Although HUPM can reveal more useful information in entire databases than traditional association-rule mining, the discovered HUPs may be irrelevant to decision making if they only occurred in long past. A pattern is not a HUP in the entire databases but is a HUP in the recent intervals by considering timestamp factor. For example, the combination of {jacket, stocking} may not be concerned as a HUP in an entire database but only in winter season. It is thus important to find the seasonal or periodic HUPs than the entire ones. In the past, several studies have been addressed the problem of temporal (or temporal maximal) HUPM in data stream [21–23]. The sliding window model of data stream was discussed to discover HUPs. In the sliding window model, the past information is ignored but the focus is on the data within the window size for mining HUPs. The size of the sliding window in a data stream is, however, very difficult to manually set as a correct period of time such that the associations of particular interest can be found from the transactional databases. If the window size is set smaller, the discovered patterns are frequently changed (quickly inserted then deleted); otherwise, it is hard to present the different exhibition periods of each item due to the size of the window being set larger. In addition, since different items may have different exhibition periods in a log database, it is hard to carry out a fair measurement by considering a fixed window size of each item.

It is an important task to make the real-time decisions or strategies for managers or retailers. Recent information of trends is concerned as a critical factor to make the efficient and up-to-date decisions instead of the elder or out-of-date information. In this study, the concept of up-to-date high-utility pattern (UDHUP) is designed to reveal more useful and meaningful HUPs in recent trend. It discovers not only HUPs in the entire databases but also the recently up-to-date HUPs within its lifetime from the past timestamp to the current one. The UDHUP-apriori and UDHUP-list algorithms are respectively developed in this paper to efficiently mine UDHUPs based on a level-wise approach and the list structures with an enumeration tree. Since the UDHUPs indicate the recent HUPs within their lifetime, more HUPs are produced if their last appeared transaction is close to the current time. A flexible minimum length (*minLen*) strategy with two specific lifetimes is also designed to find UDHUPs based on users' specification. The major contributions in this paper are described below.

1. A new knowledge representation of up-to-date high-utility pattern (UDHUP) is designed to reveal more useful and meaningful HUPs within their lifetime from the past timestamp to the current one.
2. Two algorithms, UDHUP-apriori and UDHUP-list, are respectively proposed to mine UDHUPs. The UDHUP-apriori algorithm is used as a baseline approach to level-wisely mine UDHUPs. An improved UDHUP-list algorithm is also proposed to efficiently derive UDHUPs without candidate generation based on the developed up-to-date utility-list (UDU-list) structure and a designed pruning strategy.
3. A flexible minimum length (*minLen*) with two lifetime strategies is also designed to efficiently and effectively set the intervals for discovering UDHUPs based on users' specification.
4. The proposed UDHUP framework and the developed two algorithms can be served as an efficient tool, especially in decision support systems (DSS), for managers or retailers to discover more useful, meaningful and recent information in many real-life applications. Moreover, based on the various definitions of "utility" (profit, benefit, weight, risk), the proposed technologies can be applied to other various domains, such as multi-dimensional data analysis [18], benefits evaluation, possibilities and risks of engineering informatics [19].

The rest of this paper is organized as follows. Related works are respectively reviewed in Section 2. The developed up-to-date high-utility pattern (UDHUP) mining framework is described in Section 3. The designed UDHUP-apriori and UDHUP-list algorithms are respectively mentioned in Sections 4 and 5. Experiments are conducted in Section 6. Conclusions are given in Section 7.

2. Related work

In this section, some preliminaries and related works of up-to-date pattern mining (UDPM) and high-utility pattern mining (HUPM) are briefly reviewed.

2.1. Up-to-date pattern mining

2.1.1. Problem definition

Let $I = \{i_1, i_2, \ldots, i_m\}$ be a finite set of distinct items in a log database $D = \{T_1, T_2, \ldots, T_n\}$, where each transaction $T_q \in D$ is a subset of I and has a unique identifier, called TID. An itemset X is a set of k distinct items $\{i_1, i_2, \ldots, i_k\}$, where k is the length of an itemset called k-itemset. An itemset X is said to be contained in a transaction T_q if $X \subseteq T_q$. A minimum support threshold is set as ε. A log database used as an example is shown in Table 1.

Table 1
Log database (TID: transaction ID).

TID	Transaction time	Item
1	2011/4/21 10:00	b, c, d, g
2	2011/4/21 11:10	a, b, c, d, e
3	2011/4/22 08:00	a, c, d
4	2011/4/22 15:00	c, e, f
5	2011/4/28 08:30	a, b, d, e
6	2011/5/01 10:00	a, b, c, f
7	2011/5/02 13:00	d, g

Please cite this article in press as: J.-C.W. Lin et al., Efficient algorithms for mining up-to-date high-utility patterns, Adv. Eng. Informat. (2015), http://dx.doi.org/10.1016/j.aei.2015.06.002