Knowledge-Based Systems 000 (2017) 1–18

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Damped window based high average utility pattern mining over data streams

Unil Yun[a], Donggyu Kim[a], Eunchul Yoon[b,*], Hamido Fujita[c]

[a] Department of Computer Engineering, Sejong University, Seoul, Republic of Korea
[b] Department of Electronics Engineering, Konkuk University, Seoul, Republic of Korea
[c] Faculty of Software and Information Science, Iwate Prefectural University (IPU), Iwate, Japan

ARTICLE INFO

Article history:
Received 20 April 2017
Revised 18 December 2017
Accepted 27 December 2017
Available online xxx

Keywords:
Data mining
Stream pattern mining
Damped window model
High-average utility
Significant test

ABSTRACT

Data mining methods have been required in both commercial and non-commercial areas. In such circumstances, pattern mining techniques can be used to find meaningful pattern information. Utility pattern mining (UPM) is more suitable for evaluating the usefulness of patterns. The method introduced in this paper employs the high average utility pattern mining (HAUPM) approach, which is one of the UPM approaches and discovers interesting patterns of which the items have more meaningful relations among one another by using a novel utility measure. Meanwhile, past research on pattern mining algorithms mainly focus on mining tasks processing static database such as batch operations. Most continuous, unbounded stream data such as data constantly produced from heart beat sensors should be treated differently with respect to importance because up-to-date data may have higher influence than old data. Therefore, our approach also adopts the concept of the damped window model to gain more useful patterns in stream environments. Various experiments are performed on real datasets in order to demonstrate that the designed method not only provides important, recent pattern information but also requires less computational resources such as execution time, memory usage, scalability and significant test.

1. Introduction

The objective of data mining is to discover valuable information hidden in massive data. Since data mining tasks are hard to be accomplished manually, many kinds of programmatic data mining approaches such as classification, clustering, and association rule mining have been devised in the past few decades. Association rule mining [23,25,39,51,55] is widely used to find meaningful itemsets (such as products in market databases and symptoms in medical databases) from transaction databases.

There are several major approaches of association rule mining: frequent pattern mining, sequential pattern mining, and utility pattern mining. Frequent pattern mining (FPM) [12,22,24,37,38,49,50] is the most common approach in the association rule mining area. It simply finds patterns frequently occurring in databases. Even though this approach has been used in many data mining applications, it is not suitable to analyze databases with non-binary item information because this approach assumes that each item in transactions is expressed as binary information (exist or not, 0 or 1).

On the contrary, in utility pattern mining (UPM), such non-binary information (called utility information) of items in databases can be properly considered to determine the importance of a pattern. Therefore, UPM has been fully utilized by data analysts in many fields, which handle huge data with high complexities. The typical UPM approach is high utility pattern mining (HUPM) [6,20,32,33,36,37,39,40–42,45] evaluating the patterns' utilities through the utility measure, which simply summates the item utilities in patterns. However, this approach can suffer from the generation of a huge number of patterns with long lengths because the summation of item utilities is generally large when a pattern length is long. For mining better utility, high average utility pattern mining (HAUPM) [14,19,36,52] has been researched. This technique reflects pattern lengths into the corresponding patterns' utilities in order to measure their utilities more fairly than the traditional HUPM approach.

Recently, processing time-sensitive stream data has been an important issue because recent data can be more useful for discovering significant pattern information compared to time-elapsed data. For instance, in a given medical database, recent records of the patients are more important to investigate current medical trend

* Corresponding author.
E-mail addresses: yunei@sejong.ac.kr (U. Yun), donggyukim@sju.ac.kr (D. Kim), ecyoon@konkuk.ac.kr (E. Yoon), HFujita-799@acm.org (H. Fujita).

https://doi.org/10.1016/j.knosys.2017.12.029
0950-7051/

compared to old records. In this respect, time-sensitive data need to be dealt via data mining methods with the consideration of time factors. In this regard, the previous studies on HAUPM have fundamental limitations in analyzing such time-sensitive data.

In this paper, motivated by the above problems of the previous studies, we introduce the novel pattern mining algorithm called MPM (*Mining significance utility pattern information from stream data*) adopting the concepts of HAUPM and the damped window model in order to find recent, useful high average utility patterns in stream environments. Based on the exponential damped window model, importance of older data is continually diminished as time passes. Thereby, the importance of data can be differently assigned according to their arrival-times. In addition, in order for our algorithm to have efficient performances, we design novel data structures for storing and processing stream data. In the following points, our work has originality and novelty and its main contributions are as follows:

1) We introduce a novel HAUPM approach that mines recently important HAUPs by considering the time factors of given data in order to find significant, recent pattern information. The approach is useful to analyze stream data, which are continually generated without limitations.
2) In order to facilitate the efficient mining process based on the HAUPM approach and the damped window model, we devise and use new data structures DAT (*Damped average utility tree*) and TUL (*Transaction utility list*) that are different from those proposed in the previous studies, and novel pruning strategies based on decayed average utility values stored in the proposed data structures.
3) We perform various experiments by using actual and synthetic datasets composed of various attributes. Based on the experimental results, we show that the proposed algorithm has better performance compared to other utility-based pattern mining algorithms in terms of runtime, memory, and scalability. Furthermore, we conduct a significance test, which show that our algorithm outperforms state of the art mining algorithms.

The proposed method contributes to the pattern mining area in that it is the first approach designed to apply both the average utility factor and the time fading concept of data streams. In addition, the newly proposed data structures, and pattern mining and pruning techniques contribute to improving algorithm performance. The extensive experimental results and their analyses also support their effects in empirical aspects. This paper is organized as follows. First, we provide the brief explanation on the previous works related to the topic of this paper in Section 2. Thereafter, the designed data structure and the procedure of the proposed algorithm are minutely described in Section 3. In order to show the effectiveness of our approach, we evaluate the performance of the proposed method through comparisons of ours and previous algorithms in Section 4. Finally, in Section 5, we end the paper with a conclusion offering a summary and plans for future works.

2. Related work

2.1. Utility pattern mining

The major obstacle of mining high utility patterns (HUPs) [47,48,53–56] was that the traditional anti-monotone property [1] in FPM cannot be maintained. To address this problem, the transaction weighted utilization (TWU) downward closure property was employed in the Two-phase algorithm [31]. Thereby, patterns with low TWUs are efficiently excluded from pattern mining processes. However, this method is very inefficient because it requires a large number of database scans caused by the *apriori-like* approach. Therefore, other approaches storing data in their tree-based data structures have been proposed in order to facilitate efficient mining processes. They can minimize execution time and memory usage by reducing the numbers of database scans and generated candidate patterns. HAUI-Tree [36] is used to mine HAUI from transaction databases. The advantage of this method is to reduce candidates efficiently by using HAUI-Tree. IHUP (Incremental high utility pattern mining) [2] inserts all transaction data into its data structure IHUP-Tree without any pruning process in order to satisfy the "*build once, mine many*" property. On the other hand, UP-Growth and UP-Growth++ [44] employ novel pruning strategies (DGU, DGN, DLU, and DLN) to reduce the number of generated candidates by minimizing TWU values stored in their data structures named UP-Tree through the deletions of obsolete item utility values. FHM [7], D^2HUP [27], BAHUI [43], EFIM [57], and HUP-Miner [17] are the other state-of-art HUPM approaches. Based on various techniques such as the estimated utility co-occurrence pruning method [7], the linear data structure-based one phase techniques [27], the bitmap-based efficient techniques [43], the two upper-bound techniques [57], and the other novel pruning methods [17], they mine HUPs in different, efficient ways. In addition, there are the other interesting HUPM approaches such as a method based on the ant colony system [46], a technique using negative unit profits [29], and privacy-preserving utility mining [30]. EHAUPM [36] proposed two novel tighter upper-bound models as alternative to the traditional model for mining High Average Utility Itemset Miing (HAUIs). However, the above approaches still have several limitations because they cannot consider different importance of items with the passage of time and item relations within the mined high utility patterns.

Meanwhile, HAUPM has been studied in order to mine more meaningful utility patterns than HUPM. The TPAU (Two phase average utility) algorithm [10] performs its mining process similar to that of the Two-phase algorithm for finding high average utility patterns (HAUPs) from databases. However, it utilizes the concept of an average utility upper-bound in order to satisfy the downward closure property, while HUPM exploits the TWU downward closure property. Since *apriori-like* algorithms have many drawbacks, as mentioned earlier, other HAUPM algorithms with novel approaches have been designed. As one of them, the PBAU (*Projection-based average utility pattern mining*) algorithm [18] projects sub-databases from a given database and constructs index tables for calculating average utilities of patterns efficiently.

MHAI [54] is the list-based latest HAUPM approach that does not generate any candidate during the mining process. Using its own data structures and mining techniques, it efficiently extracts HAUPs without pattern losses. HAUI-Miner [28] is another list-based recent method similar to MHAI. Through its own data structure, AU-list, it extracts HAUPs in an efficient manner. However, since such methods have been designed to process static databases, they cannot deal with dynamic stream data unlike our approach. SHAU [52] and IMHAUI [53] are dynamic methods designed to deal with stream data. Using their own tree structures, they find HAUPs from given stream data. In contrast to IMHAUI, SHAU focuses on the latest data within a user-given range, called window. By doing so, the algorithm can always provide users with the latest results of HAUPM.

2.2. Damped window model

To process time-sensitive data in stream environments, the ways to discriminate data according to their creation time have been necessitated. As one of the approaches for distinguishing the recent information and old one, the damped window model has been adopted to frequent pattern mining [4,5,15,26], sequence pat-

Please cite this article as: U. Yun et al., Damped window based high average utility pattern mining over data streams, Knowledge-Based Systems (2017), https://doi.org/10.1016/j.knosys.2017.12.029