ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

FUP ALGORITHM TO DISCOVER WEIGHTED FREQUENT ITEMSETS FROM TRANSACTIONAL DATABASES

Miss. Shilpa Babhale , Prof. Vipul Bag
Department of Computer Engineering, NK Orchid Collage of Engineering, Solapur, Maharashtra, India

ABSTRACT

Mining high utility itemsets from a transactional database refers to the discovery of itemsets with high utility like profits. Although a number of relevant algorithms have been proposed in recent years, they incur the problem of producing a large number of candidate itemsets for high utility itemsets. Such a large number of candidate itemsets degrades the mining performance in terms of execution time and space requirement. The situation may become worse when the database contains lots of long transactions or long high utility itemsets. In this paper, we propose an algorithm for mining high utility itemsets with a set of effective strategies for pruning candidate itemsets as per periodicity. The information of high utility itemsets is maintained in a tree-based data structure named utility pattern tree such that candidate itemsets can be generated efficiently with only two scans of database. Experimental results show that the proposed algorithm not only reduce the number of candidates effectively but also outperform other algorithms substantially in terms of runtime and frequency based weights, especially when databases contain lots of very long transactions.

KEYWORDS: FUP (Frequent utility pattern), Candidate pruning, frequent itemset, high utility itemset, utility mining, data mining

INTRODUCTION

Data mining is the process of extracting nontrivial, previously unknown and potentially useful information from large databases. Discovering useful patterns hidden in a database plays an essential role in several data mining tasks, such as frequent pattern mining, weighted frequent pattern mining, and high utility pattern mining. Among them, frequent pattern mining is a fundamental research topic that has been applied to different kinds of databases, such as transactional databases [1] streaming databases [1], [2], and time series databases [2], and various application domains, such as bioinformatics [1], [2], Web click-stream analysis [2], [3], and mobile environments [5], [6].Nevertheless, relative importance of each item is not considered in frequent pattern mining. To address this problem, weighted association rule mining was proposed [4], [6]. In this project frequency based weights of items, such as unit profits, periodicity, quantity and seasonal data of items in transaction databases are considered. With this concept, even if some items appear infrequently, they might still be found if they have high weights. An emerging topic in the field of data mining is Utility Mining. The main objective of Utility Mining is to identify the itemsets with highest utilities, by considering profit, quantity, cost or other user preferences. Mining High Utility itemsets from a transaction database is to find itemsets that have utility above a user-specified threshold. Itemset Utility Mining is an extension of Frequent Itemset mining, which discovers itemsets that occur frequently. In many real-life applications, high-utility itemsets consist of rare items. Rare itemsets provide useful information in different decision-making domains such as business transactions, medical, security, fraudulent transactions, and retail communities. For example, in a supermarket, customers purchase microwave ovens or frying pans rarely as compared to bread, washing powder, soap. But the former transactions yield more profit for the supermarket. Similarly, the high-profit rare itemsets are found to be very useful in many application areas. For example, in medical application, the rare combination of symptoms can provide useful insights for doctors [2].

RELATED WORK

The system can not only decrease the overestimated utilities of PHUIs (potential high utility itemsets) [1] but greatly reduce the number of candidates. Different types of both real and synthetic data sets are used in a series of experiments to the performance of algorithm with state-of-the-art utility mining algorithms which show that UP-Growth (utility pattern growth) [1] and UP-Growth+ like other algorithms substantially in term of execution time, especially when databases contain lots of long transactions or minimum utility threshold is set.

R. Agrawal [1] introduced the concept of frequent itemset mining. Frequent itemsets are the itemsets that occur frequently in the transaction data set. The goal of Frequent Itemset Mining is to identify all the frequent itemsets in a transaction dataset.

The mining of association rules for finding the relationship between data items in large databases is a well studied technique in data mining field with representative methods like Apriori [1], [2]. ARM process can be decomposed into two steps. The first step involves finding all frequent itemsets.

H. Yao et al formalized the semantic significance of utility measures in [1]. Based on the semantics of applications, the utility-based measures were classified into three categories, namely, item level, transaction level, and cell level. The unified utility function was defined to represent all existing utility-based measures.
The H.F. Li proposed two efficient one pass algorithms, MHUI-BIT and MHUI-TID, for mining high utility itemsets from data streams within a transaction-sensitive sliding window. Two effective representations of item information and an extended lexicographical tree-based summary data structure were developed to improve the efficiency of mining high utility itemsets [6].

Liu *et al* proposed Two-Phase algorithm [5] for finding high utility itemsets. In the first phase, a model that applies the "transaction-weighted downward closure property" on the search space to expedite the identification of candidates. In the second phase, one extra database scan is performed to identify the high utility itemsets.

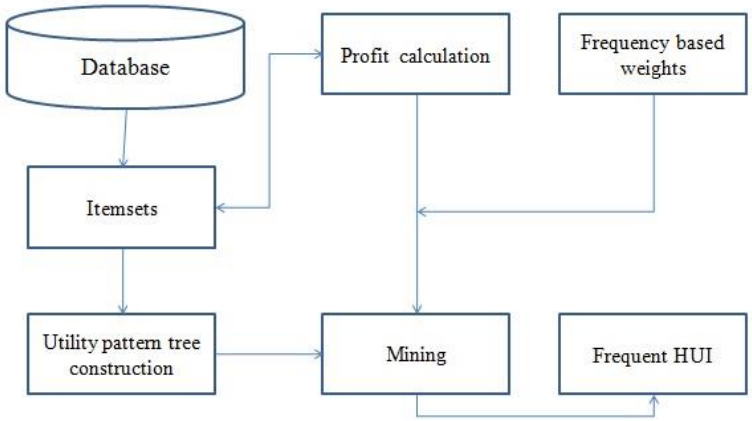


Figure 2.1. System Overview

PORPOSED MODELLING

In this section we present discovery of high utility itemset with not only profit to obtain frequent itemsets but by weighted frequent itemsets periodically. Given novel algorithm of frequent utility pattern growth contains weighted frequent itemsets calculations and time wise observations of transactions.

3.1 The Proposed Data Structure: FUP-Tree-

To discover the mining performance and avoid scanning original database repeatedly, we use a compact tree structure, named FUP-Tree, to maintain the information of transactions and high utility itemsets.

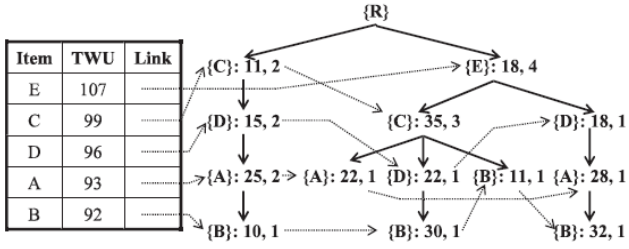


Figure 3.1 Tree generated after using strategies

Two strategies are applied to minimize the overestimated utilities stored in the nodes of global FUP-Tree. In following sections, the elements of FUP-Tree are first defined. Next, the two strategies are introduced. Finally, how to construct an FUP-Tree with the two strategies is illustrated in detail by formulae's.

Item	A	B	C	D	E
Minimum item utility	5	2	1	2	3

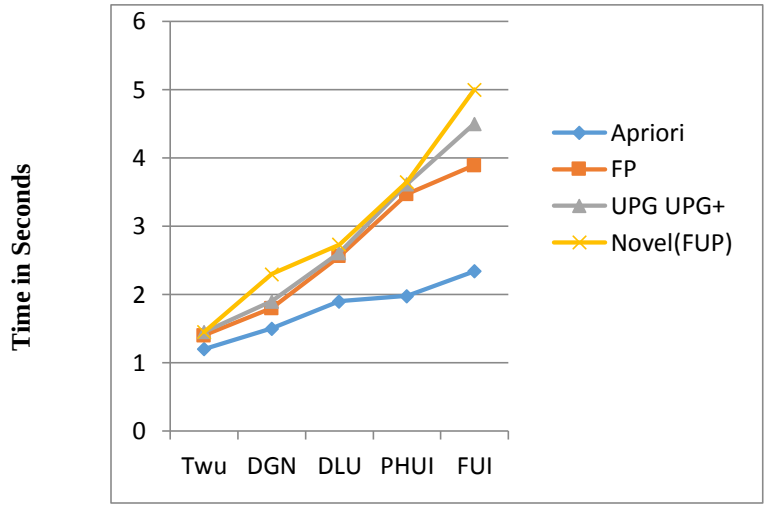
Table 3.1 Minimum utility table

Utility (item) <= OEU(item)

Where OEU is the overestimated utility of itemsets which is less than the minimum utility i.e. Utility(item) < min_utl. That means Utility (i) <OEU(i)<min_util. Only the supersets of promising items are possible to be high utility itemsets.

3.2 Discarding Global Unpromising Items during tree construction-

The construction of a global UP-Tree can be performed with
two scans of the original database. In the first scan, TU of each transaction is computed. At the same time, TWU of each single item is also accumulated. An item ip is called a promising item if TWU > min. util. Otherwise it is called an unpromising item. Without loss of generality, an item is also called a promising item if its overestimated utility (which is different from TWU in this paper) is no less than min. util. Otherwise it is called an unpromising item. During the second scan of database, transactions are inserted into a UP-Tree. When a transaction is retrieved, the unpromising items should be removed from the transaction. After finding all PHUIs, the third step is to identify high utility itemsets and their utilities from the set of PHUIs by scanning original database once



3.3 The Proposed Mining Method-

UP-Growth achieves better performance than FP-Growth by using DLU and DLN to decrease overestimated utilities of itemsets. However, the overestimated utilities can be closer to their actual utilities by eliminating the estimated utilities that are closer to actual utilities of unpromising items and descendant nodes which helps to generates global FUP-tree.

3.4 Novel algorithm for weighted frequent itemsets mining-

After constructing a global FUP-Tree, a basic method for generating PHUIs is to mine FUP-Tree by FP-Growth [4]. In this section, we propose an improved method (Novel algorithm) for reducing overestimated utilities more effectively by pushing more strategies into the framework of UP-Growth+ algorithm.

1) Generate conditional pattern bases by tracing the paths in the original tree.
2) Construct conditional trees (also called local trees in this paper) by the information in conditional pattern bases.
3) Mine patterns from the conditional trees..

RESULTS AND DISCUSSIONS

Performance of the proposed algorithms is evaluated in this section & results in this section show that the proposed methods outperform the state-of-the-art algorithms almost in all cases.

Figure 1: Comparison of Candidates generation with respect to time for transactions

Apriori	FP Growth	Upg &Upg+	Novel
1.2	1.4	1.45	1.46
1.5	1.8	1.9	2.3
1.9	2.56	2.61	2.73
1.98	3.47	3.62	3.65
2.34	3.89	4.51	5.03

Table 1 Comparison Table

All the values should be observed by research of other articles with this novel algorithm. we have to use recognized subroutine to calculate frequent items from the transactions.

CONCLUSION

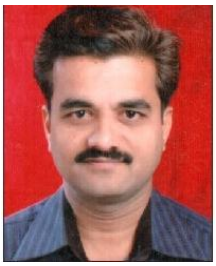
In this paper we propose novel algorithm which is used to mine database to discover high utility frequent itemset by considering frequency based weights, time with profit as well. The algorithm is applied to discover high utility frequent itemsets. Result will give high utility frequent itemsets after mining with profit, time and frequency based weights.

REFERENCES

[1] Vincent S. Tseng, Bai-En Shie, Cheng-Wei Wu, and Philip S. Yu, Fellow,"Efficient Algorithms for Mining High Utility Itemsets from Transactional Databases", Vol.. 25, No. 8, Aug 2013
[2] Y. Liu, W. Liao, and A. Choudhary, "A Fast High Utility Itemsets Mining Algorithm," Proc. Utility-Based Data Mining Workshop, 2005
.[3] V.S. Tseng, C.J. Chu, and T. Liang, "Efficient Mining of Temporal High Utility Itemsets from Data Streams," Proc. ACM KDD Workshop Utility-Based Data Mining Workshop (UBDM '06), Aug. 2006.
[4] Frequent Itemset Mining Implementations Repository, http:// fimi.cs.helsinki.fi/, 2012
.[5] ChowdhuryFarhan Ahmed· Syed KhairuzzamanTanbeer·Byeong-SooJeong· Young-Koo Lee,"HUC-Prune: an efficient candidate pruning technique to mine high utility patterns"2009
[6] Sudip Bhattacharya, DeeptyDubey,"High Utility Itemset Mining", Volume 2, Issue 8, August 2012
[7] JyothiPillai,O.P.Vyas,"Overview of Itemset Utility Mining and its Applications",Volume 5– No.11, August 20

Authors

Ms. Shilpa Nagorao Babhale obtained her bachelor's degree from the Mgm's collage of Engineering, Nanded,Maharashtra,India. She is currently a Master of Engineering (M.E.) student under the supervision of Prof. Vipul Bag. Her research is centered on Data Mining within that discovering frequent itemsets from transactional databases.



Mr. Vipul Bag, , is working as associate professor in Department of Computer Science and Engineering in NK Orchid College of Engineering and Technology , Solapur,Maharashtra,,India.He has 16 years of teaching experience. He has co-authored over 20 International Journal Publications. He is pursuing PhD from SGGSIET, Nanded, Maharashtra, India. The current research interest are recommendation systems, data mining & machine learning.

ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785