Middle-East Journal of Scientific Research 24 (3): 986-991, 2016
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2016.24.03.23192

A New Hybrid Frequent Pattern-Apriori (FP-AP) Algorithm for High Utility Item Set Mining

R Shyamala Devi and D. Shanthi

Computer Science And Engineering, PSNA College of Engineering and Technology, Dindigul, India

Abstract: Frequent item set mining which deals with the items occurs at sequences of events, High utility item set mining is a confront process in frequent pattern mining. , Frequent Item set Mining (FIM) look at all the items having the equal priority and it is very difficult to find the High utility item (HUI) without any redundancy. If user is in need of finding the rare purchased product or to find the high profitable item that couldn't not be done in frequent item set mining. To overthrown the concern in FIM, high utility items were calculated. The utility denotes the usefulness of the product and also represents the rarely purchased product of the users. Furthermore, in HUI, the huge number of outcome with high utility was composed. So, the user is not able to get the concise result. Meantime for a longer transaction the truncation technique was used which cause to information loss and the privacy is also a main issue. Hence to overcome this controversy, it is proposed to provide a High Utility Item without redundancy and to find the closest item set. Here, the proposed FP-AP algorithm which is the combination of frequent pattern and apriori algorithm.FP is proposed to split the longer transaction rather than truncated and also to find the high profitable item with privacy to that item set without redundancy. AP algorithm helps to provide a concise and lossless representation without affecting the utility.

Key words: Frequent item set mining · Utility Mining · Data Mining · Lossless and concise representation · Closest item set mining

INTRODUCTION

Frequent item set mining is an interesting branch of data mining that have focal point on looking at sequences of actions or events, for example the order in which we purchase the item. Frequent sets play an vital role in many mining tasks is to find interesting patterns from databases, such as association rules, correlations, sequences, episodes, classifiers, clusters and several more of which the mining of association rules. The original impulse for searching frequent set came from the need to study so called supermarket transaction data, that is, to probe customer behavior in terms of the purchased products

Frequent item set is defined as that substance that has minimum support. The subset of a frequent item set must also be an frequent item set i.e., if {AB} is a frequent item set, both {A} and {B} should be a frequent item set then find frequent item sets with cardinally from 1 to k (k-item set).Use the frequent item set to generate association rules. The conventional model of FIM may find a enormous amount of frequent but low revenue item sets and lose the information on beneficial item sets having low selling products. These problems are origin by the facts that the FIM consider all things as having the same importance/unit profit/weight and it assumes that all point in a transaction appears in a binary format i.e., an item can be either present or absent in a transaction, which does not indicate its get quantity in the transaction. For example, if a customer buys a exceedingly expensive wine or just a piece of bread, it is viewed as being equally important[9]. FIM cannot satisfy the requirement of users who desire to discover item set with high utilities such as high profits.

Thus, frequent pattern mining may find many frequent patterns that are not interesting. For example, one may find that {sugar, milk} is a frequent pattern. However, from a business perspective, this pattern may be uninteresting because it does not have much profit. Moreover, frequent pattern mining algorithms may miss the rare patterns that generate a high profit.

Corresponding Author: R Shyamala Devi, Computer Science And Engineering,
PSNA College of Engineering and Technology, Dindigul, India.

The limitations of frequent or rare item set mining motivated researchers to conceive a utility based approach, which allows a user to conveniently express his or her perspectives concerning the need of item sets as utility values and then find item sets with high utility values higher than a threshold. In utility based mining the term utility refers to the quantity of user preference i.e. the utility value of an item set is the evaluation of the importance of that item set in the user's perspective.

High-utility item set mining is to find the item sets (group of items) that generate a high profit in a database when they are sold together. The user have to provide a value for a threshold called "minutil" (the minimum utility threshold) [1]. A high-utility item set mining algorithm outputs all the high-utility item sets, that is those that generates at least "minutil" profit.

The utility and privacy tradeoff can be improved by limiting the length of transactions Thus, the transaction truncating (TT) approach proposed and not suitable to avoid privacy breach, we add noise to the support of item sets.That is, if a transaction has more items than the limit, then divide it into multiple subsets (i.e., sub-transactions) and guarantee that every subset is under the limit. to preserve more frequency information in subsets, we propose a graph-based approach to reveal the items within transactions and utilize such correlation to guide the splitting process.

High Utility Item set is depend on the profit of the product. While considering the utility the redundancy may occur, so need to remove the redundancy in the High Utility Item set. After finding the High Utility Item sets find the closest item set which consist of related item set. In frequent item set , to reduce the computational cost of the mining task present fewer but more important patterns to users, many research focused on developing concise representations.

The set of HUIs is very large, which makes HUI mining algorithms suffer from long execution times and huge memory consumption. To address this issue, comprehend representations of HUIs have been proposed [2]. However, no concise representation of HUIs has been proposed based on the concept of generator despite that it provides several benefits in many applications.

Related Works: Mining of high utility item sets from datasets proposed that the high utility item sets are mined using the pattern growth approach is the new algorithm called CTU Mine[3-7]. For large databases identifying high utility Item sets candidate-generate-and-test approach is not suitable.

High Utility Item sets in Data Streams propose two efficient one pass algorithms MHUI-BIT and MHUI-TID for mining high utility item sets from data streams within a transaction sliding window. For improving the efficiency of mining high utility item sets two effective representations of extended lexicographical tree-based summary data structure and item set information were developed.

Efficient Algorithms used for High Utility point sets from Transactional Databases Mining HUI from transactional database denotes the itemsets with high utility like profits. they incur the problem of produce a large number of candidate itemsets for high utility itemsets [8-10]. Such a huge number of candidate itemsets degrades the mining performance in terms of execution time and space must is used. Improved performance by reducing equally the search space and the number of candidates. Especially UP-Growth, not only reduce the number of candidates effectively except also outperform other algorithms substantially in terms of runtime.

Efficient Mining of Temporal High Utility Itemsets from Data streams proposes a temporal high utility item set mining. The temporal high utility itemsets with less candidate itemsets and higher performance can be given by THUI- mine. To generate a progressive set of itemsets THUI-Mine employ a filtering threshold in each partition.. Large memory requirement and lot of false candidate itemsets be the two problems of THUI- Mine algorithm.

Privbasis Algorithm defines the challenge of high dimensionality by projecting the input dataset onto a small number of selected dimensions that have to be cared. In fact, PrivBasis often uses several sets of dimensions for such projections, to avoid any set containing too many dimensions [6]. Every basis in B corresponds to one such set of dimensions for projection. This techniques enable one to select which sets of dimensions are most helpful for the purpose of finding the k most frequent item sets.

When the column N is larger than the truncated frequency approach is completely ineffective. The reason why the Truncation Frequency(TF) does not scale is that when one needs to select the top k item sets from a large set U of candidates, with large size causes two difficulties. The first is regarding the running time. Even if every single low-frequency item set in U is chosen with only a small probability, the sheer number of such low-frequency item sets means that the k selected item sets likely includes many infrequent ones.

The TF technique tries to notify the running time by pruning the search space, but it does not address the accuracy challenge [5]. This addresses the symptom for a larger candidate set, but not the root cause.

Private Frequent Pattern Mining Algorithm has been well recognized that simple anonymization schemes that remove obvious identifiers carry heavy risks to privacy. Even privacy-preserving graph mining techniques based on k-anonymity are now often considered to offer insufficient privacy under strong attack models [8]. Recently, the model of differential privacy was proposed to restrict the inference of secured information even in the presence of a strong adversary. It needs the output of a differentially private algorithm is identical (in a probabilistic sense), or not a participant contributes her data to the dataset

Background

High Utility Item Set: An item set is called a high utility item set (HUI) if its utility is no less than a user-specified minimum utility threshold; [11] otherwise, it is called a low utility item set.

Transaction Utility and Total Utility: External utility is defined as the importance of distinct items and the Internal utility is the importance of the items in the transaction. Utility is calculated by multiplying the internal and external utility [5]. Total utility is calculated by adding all transaction utility

$U = TQ*PQ \qquad U(X) = \sum U$

Transaction Weight Utilization: The transaction-weighted utilization (TWU) of an item set X is the sum of the transaction utilities of all the transactions containing X, which is denoted as TWU(X) [2].

$Itwu\ (X) = \sum tu(Tq)$

The Transaction-Weighted Downward Closure (TWDC): For any item set X, if X is not a HTWUI, any superset of X is a low utility item set.

Local Promising Item: An item i_{mp} is called a local promising item in $\{a_{ir}\}$-CPB if up(i_{mp}, $\{a_{ir}\}$-CPB) is no smaller than minutia;.

Differential Privacy: For two databases D and D', they are neighboring databases if they differ by at most one record.The amount of injected noise is carefully calibrated to the sensitivity. The sensitivity of count queries is used to measure the maximum possible change in the outputs over any two neighboring databases.

Weighted Splitting Operation: Consider a transaction t whose length exceeds the maximal length constraint Lm. A function f divides t into multiple subsets t1;... ;tk, where ti is assigned a weight wi and the length of ti is under the length constraint Lm.

Proposed System: In the proposed scenario, we introduced the algorithm named as improved splitting algorithm to overcome the privacy issues in the large industrial applications. This proposed algorithm is mainly focused on the achieving the higher privacy for high utility item sets and lower or none privacy for low utility item sets. It also ensures the high instance efficiency by improving the service and privacy tradeoff in the transformed database. And all users do not need high utility item sets information and it is not necessary to show the important utility item sets patterns to all users. Hence apply the privacy concept on the important high utility information and reveal only the particular users.

Architecture

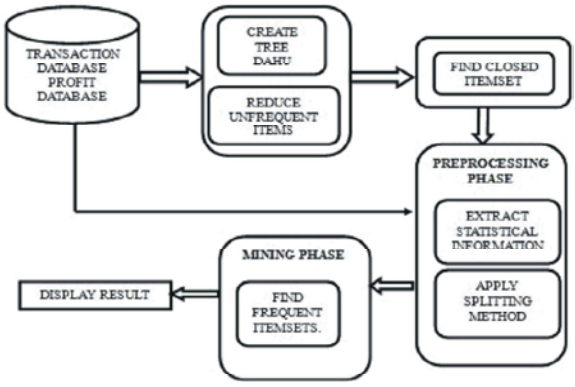


Fig. 4.1: Overall Architecture

To provide a comprehend results to the user without redundancy and also provide a closest item sets FP-AP is used to overcome those issues. And also , for a longer transactions, a new splitting technique is used it maintains the privacy.In first, HUI will be found and then closest item sets will be identified. In second module frequent items will be identified with their utility and in third module transaction will be splitted to maintain the privacy.

Process Flow: Fig.4.2 illustrates , there will be two databases one for storing the transactions and another for profit details. High Utility Items will be found by discarding infrequent and isolated item set. Find the closest item set from the high utility item set.

An item set is defined as closed if none of its immediate supersets has same support as item sets. An item set is maximal frequent if none of its immediate

supersets is frequent. Remove the unpromising items reduce the redundancy

DAHU (Derive All High Utility Item sets)is to derive all the all the high utility from the DAHU tree. For a longer transaction, if truncation is done then sensitive content may loss. Instead, use splitting methods to split the longer transaction. Focusing on those issues the proposed work is to provide a comprehend results to the user without redundancy and also provide a closest item sets.Eventhough, for a longer transactions ,a new splitting technique is used it maintains the privacy

Find the frequent item set with item set which having the high utility. Privacy is the major issue in the real world. So add privacy to the result and display the result.

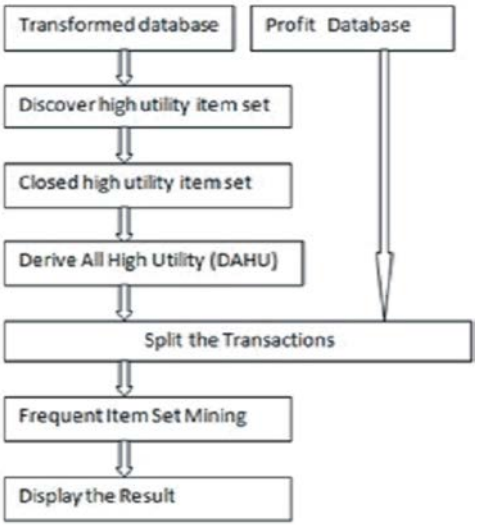


Fig. 4.2. Process flow

Module Description: This proposed algorithm is mainly focused on the achieving the higher privacy for high utility item sets and lower or none privacy for low utility item sets. It also ensures the high time efficiency by improving the utility and privacy transaction in the transformed database. And all users do not need high utility item sets information and it is not necessary to show the important utility item sets patterns to all users. Hence apply the privacy concept on the important high utility information and reveal only the particular users. The proposed method consist of three modules,

- Finding High Utility Item sets
- Preprocessing Phase
- Mining Phase

Finding High Utility Item Sets: From the transaction and profit database find the high utility item set. High utility item set is defined as that, its utility is no less than a user-specified threshold or minutia. The basic meaning of utility is the interestedness/ importance/profitability of items to the users. With the help of derive all high utility (DAHU)tree find HUI and prune the infrequent items.

Algorithm:

Input:P.EIL, the effective information list of item set P, initially empty; EILs, the set of utility-lists of all P's 1-extensions;
minutil, the minimum utility threshold.
Output: all the high utility itemsets with P as prefix.

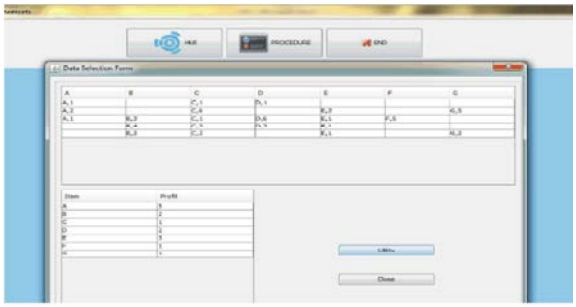
- for each effective information list X in EILs do
- if SUM(X.iutils)=minutil then
- output the extension associated with X;
- end if
- if SUM(X.iutils)+SUM(X.rutils) = minutil then
- exEILs = NULL;
- for each effective information list Y after X in EILs do
- exEILs = exEILs+Build(P.EIL, X, Y);
- end for
- EHUI(X, exEILs, minutil);
- end if
- end for

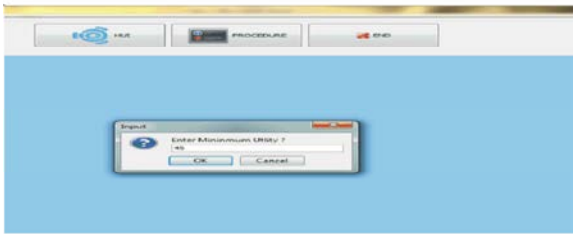
Preprocessing Phase: In preprocessing phase extract some statistical information from the original database.For a longer transaction, if truncation is done then some important content may loss. To improve the utility-privacy tradeoff, we argue that long transactions should be separated rather than cut shorted. That is, we transform the database by dividing long transactions into multiple subsets (i.e., sub-transactions), each of which meets the maximal length constraint. When we divide a long transaction, we assign a weight to each generated subset. The weight of a subset indicates the change to the support of an itemset when adding (removing) this subset into (from) the database. It can be considered as a multiplier. In fact, transaction curtail can be seen as an large case of our weighted splitting operation. Suppose a transaction t is divided into subsets t1;... ; tk. If we assign weight 1 to one subset ti and assign weights 0 to the other subsets, it is equivalent to truncate t items and only keep the items in t.

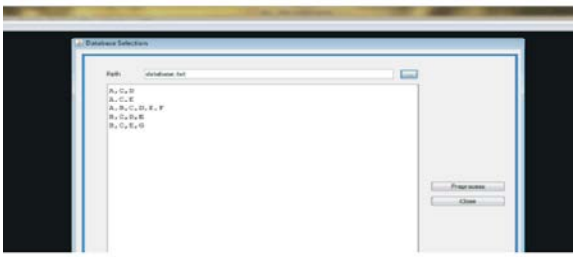
Mining Phase: In mining phase find the frequent item sets which has the high utility and add privacy to the result and display it. In the mining process, if a frequent itemset is mislabeled as infrequent itemset, due to the downward closure property, all its supersets are regarded infrequent. It will negatively affect the quality of the results. To solve this problem, given an i-itemset X, we estimate its

"maximal" support in the original database to determine whether we need to compute the noisy support of X's supersets. To estimate the "maximal" support, we utilize the r-lower bound [12-17]. By treating v0 as the r-lower bound, we can estimate the "maximal" support of X in the original database

Sample Results:







CONCLUSION

Frequent item set mining is used to denote the items purchased together but the limitation like it doesn't refer the item's purchase quantity. In FIM it treats all the items or purchased product in the same level. So this makes the motivation to find the HUI which have the high importance. So finding high utility refers the items purchased by the user interestedness. In this paper FP-AP is used to calculate the HUI along with the closest item sets and also prunes the infrequent item set.So far, truncation technique is used ,here splitting technique is used.

REFERENCES

1. Sen su Shengzhi zu, 2015. Differentially Private Frequent Item Set Mining Via Transaction Splitting , IEEE Trans. Knowl. Data Eng, 27(7): 10
2. Tseng, V.S., 2015. Efficient Algorithm For Mining Concise And Lossless Representation For Mining High Utility Item sets, IEEE Trans. Knowl. Data Eng, 27(3): 1.
3. G.C., T.P., V. and S. Tseng, 2014. An efficient projection-based indexing approach for mining high utility item sets, Knowl. Inf. Syst, 38(1): 85-107.
4. Ahmed, C.F., S.K. Tanbeer, B.S. Jeong and Y.K. Lee, 2009. Efficient tree structures for high utility pattern mining in incremental databases, IEEE Trans. Knowl. Data Eng., 21(12): 1708 -1721.
5. Chan, R., Q. Yang and Y. Shen, Mining high utility item sets, Proc. IEEE Int. Conf. Data Min., 19-26.
6. Li, N., W. Qardaji, D. Su and J. Cao, 2012. Privbasis: Frequent item set mining with differential privacy,Proc. VLDB Endowment, 5(11): 1340-1351.
7. Bonomi, L. and L. Xiong, 2013. A two-phase algorithm for mining sequential patterns with differential privacy, Proc. 22nd ACM Conf. Inf. Knowl. Manage., pp: 269-278.
8. Shen, E. and T. Yu, 2013. Mining frequent graph patterns with differential privacy, Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp: 545-553.
9. C.W., T.P. and W.H., 2011. An effective tree structure for mining high utility item sets, Expert Syst. Appl., 38(6): 7419-7424.
10. Agrawal, R. and R. Srikant, Fast algorithms for mining association rules, Proc. 20th Int. Conf. Very Large Data Bases, pp: 487-499.
11. Liu, Y., W. Liao and A. Choudhary, A fast high utility item sets mining algorithm, Proc. Utility-Based Data Mining Workshop, 90-99.

12. Lucchese, C., S. Orlando and R. Perego, 2006. Fast and memory efficient mining of frequent closed item sets, IEEE Trans. Knowl. Data Eng., 18(1): 21-36.
13. Tseng, V.S., C.W. Wu, B.E. Shie and P.S. Yu, UP-Growth: An efficient algorithm for high utility item set mining, Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, pp: 253-262.
14. Wu, C.W., B.E. Shie, V.S. Tseng and P.S. Yu, Mining top-k high utility item sets, Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 78-86.
15. Wu, C.W, P. Fournier-Viger, P.S. Yu and V.S. Tseng, Efficient mining of a concise and lossless representation of high utility item sets, Proc. IEEE Int. Conf. Data Mining, 824-833.
16. Antonio Gomariz, Manuel Campos and Roque Marin, clasp: an "Efficient algorithm for mining frequent closed sequences, Springer-Verlag Berlin Heidelbergadvances in knowl discovery and data mining, 7818: 50-61.
17. Zaki Mohammed, J., 2001. spade:an efficient algorithm for mining frequent sequences, springer article, 42(1): 31-60.