International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 02 Issue: 05 | Aug-2015 www.irjet.net p-ISSN: 2395-0072

An Efficient Algorithm for finding high utility itemsets from online sell

Sarode Nutan S, Kothavle Suhas R

[1] *Department of Computer Engineering, ICOER, Maharashtra, India*
[2] *Department of Computer Engineering, ATC Indore,MP , India*

Abstract -

In the databases data mining and knowledge discovery is a new interdisciplinary field. In data mining there is merging ideas from statistics, machine learning, databases .Data Mining is defined as an activity that extracts some new information contained in large databases. In traditional data mining techniques detecting the statistical correlations between the items which are more frequent in the transaction databases termed as frequent itemset mining ,that itemsets which appear more frequently must be of more importance to the user from the business perspective.

The term utility refers to the importance or the usefulness of the itemset in transactions quantified in terms like profit , sales or any other user preferences. By considering user preferences such as profit, quantity and cost from online transaction processing systems, utility mining extracts itemsets with high utilities. Utility Mining which not only considers the frequency of the itemsets but also considers the utility associated with the itemsets. In High Utility Itemset Mining the objective is to identify itemsets that have utility values above a given utility threshold.

Key Words:
Utility mining, High utility itemsets, Frequent itemset mining.

1. Introduction

In recent years data mining and knowledge Discovery from data bases are two of the major areas receiving much attention. In data mining, the extraction of hidden predictive information from large databases, has a great potential to help data owners in focusing on the most important information in their data warehouses. Knowledge Discovery in Databases (KDD) is the process of identifying valid ,previously unknown and potentially useful patterns in data. These patterns becomes useful in order to explain existing data, predict or to put the contents of a large database into a nutshell supporting decision making and graphical representation of data. Plenty of data mining tasks, such as frequent pattern mining, weighted frequent pattern mining, and high utility pattern mining are having important role of useful hidden pattern retrieval from a database. Out of these, frequent pattern mining is a fundamental research topic that has been applied many of databases, such as transactional databases, streaming databases, and time series databases, and in various application domains, such as bioinformatics, Web click-stream analysis, and mobile environments. Taking all this into an account, utility mining is being considered as an important topic in data mining field. Mining high utility itemsets from databases refers to finding the itemsets with high profits.Here, the meaning of itemset utility is attraction, importance or profitability of an item to users. There are two aspects for Utility of transaction database itemsets as:

1. External utility which tells the importance of distinct items
and
2. Internal utility which tells the importance of items in transactions.

Utility Mining

The actual utility of an item is not the occurance of an item. In association rule mining Judging the utility of items by its presence in the transaction set is the older methods. Mining of high utility itemsets efficiently is One of the most challenging data mining tasks. Utility Mining. Is nothing but the identification of the itemsets with high. Cost, quantity, profit or any other user expressions of preference can be used to measure the utility. If item has its utility less than a user-specified minimum utility threshold then this is called a low-utility itemset .

2. Literature Review

A two phase algorithm to find high utility itemsets prunes down the number of candidates and obtains the complete set of high utility itemsets [3]. For two factors speed and memory's cost, this algorithm works efficiently. Association Rule mining model treats all the items in the database by only considering if an item is present in transaction or not.

In Two-phase, focused on traditional databases and is not suited for data streams. In Two-phase we are not finding temporal high utility Itemsets in data streams but this must rescan the whole database when added new transactions from data streams.

Algorithm proposed by R. Agrawal, used to obtain frequent itemsets from the database [2]. Problem with association rule generations to get all association rules

that have confidence and support value is greater than user's specified value. Apriori is a classic algorithm to mine frequent itemset. After identifying the large itemsets, only those itemsets are allowed which have the support greater than the minimum support allowed. Generation of candidate item sets and scaning of database every time takes place in Apriori Algorithm. When a new transaction is added to the database then there is rescan the entire database again. Candidate itemsets are stored in a hash-tree which contains either a list of itemsets or a hash table.

Jiawei Han et al in [7] proposed frequent pattern tree (FP-tree) structure in "Mining frequent patterns without candidate generation," paper for storing crucial information about frequent patterns, compressed and develop an efficient FP-tree based mining method is Frequent pattern tree structure. It constructs a highly compact FP-tree, which is usually substantially smaller than the original database, by which costly database scans are saved in the subsequent mining processes. Efficiency of mining is achived with three techniques 1)large database is compressed into a smaller data structure, for avoiding the costly database scans.2)adopts a pattern frequent growth method to avoid the generation of large no of Itemsets.3)A partitioning based,divide and conquer method for decomposing the mining task into set of smaller tasks for mining confirmed patterns to reduces the search space.in this paper the FP-growth method is efficient and scalable for mining in long and short frequent patterns. Also faster than the Apriori algorithm and recently reported new frequent pattern mining methods.

Two one pass algorithms MHUI-BIT and MHUI-TID are proposed to mine high utility itemsets from data streams within a transaction sensitive sliding window[8]. To improve the efficiency of mining high utility itemsets two effective representations of an extended lexicographical tree-based summary data structure and itemset information were developed. In this data stream concept is used data stream is an infinite sequence of data elements continuously arrived at a rapid rate.so the Frequent pattern mining is the major problem of data mining streams.

Weighted association rule are proposed by Wei Wang et al [1]. For generation of each frequent itemset first discover frequent itemsets and the weighted association rules are used. This rule first proposed the concept of weighted items and weighted association rules. The weighted association rules does not have downward closure property, so the mining performance cannot be improved. By using transaction weight, weighted support can not only reflect the importance of an item set but also maintain the downward closure property during the mining process.

WAR not only improves the confidence of rules as well as provides a mechanism for target marketing by identifying customers based on their potential degree of loyalty or volume of purchases.

J. Hu defines an algorithm in which concept of frequent item set mining is used [4]. In this approach combinations of high utility itemset are find out. In reality algorithm is used to find segment of data, which is defined with the combination of few items i.e. rules, which is different from the frequent item mining techniques and traditional association rule. The problem considered in high utility pattern mining is different from former approaches as it conducts rule discovery with respect to the overall criterion for the mined set as well as with respect to individual attributes.

Another algorithm proposed by Cheng Wei Wu to efficiently discover high utility itemsets from transactional databases [6]. Depending on the construction of a global UP tree the high utility itemsets are generated using UP Growth which is one of the efficient algorithms. In this proposed method , it cannot overcome the screening as well as removal of null transactions i.e the overhead of null transactions are taking place in this algorithm.

S.Shankar presents a algorithm for Fast Utility Mining [5]. for generating of high utility Itemsets different techniques like Low Utility and High Frequency (LUHF) and Low Utility and Low Frequency (LULF), High Utility and High Frequency (HUHF), High Utility and Low Frequency (HULF) are used.

3. EXISTING SYSTEM:

In existing system, performance is degraded because of generation of huge set of PHUIs .Long transaction database is found when threshold value is low, which create worst situation.To overcome this limitation, Utility pattern growth (UP-Growth) . These algorithm mine high utility Itemsets.The information of high utility itemsets is maintained in utility pattern tree (UP-Tree).

Limitations of Existing Systems:

1) Huge no of itemsets which require large memory space.
2) More time is required to check datasets and Null transaction itemsets are also available.
3) Worst performance in case of low memory.

4. PROPOSED WORK :

- ✓ Selecting higher profit itemsets.
- ✓ Eliminate Null transaction itemsets.
- ✓ Provide suggestion to retailers what items can be discarded.

Architecture :

Proposed architecture for high utility itemsets is shown below:

Architecture consists of two steps:-

1) Profit Calculation and filter profit table by applying threshold value

2) Find High utility Itemsets

These steps are done with the help of UP-tree construction. Two scans are performed to get high utility itemsets.

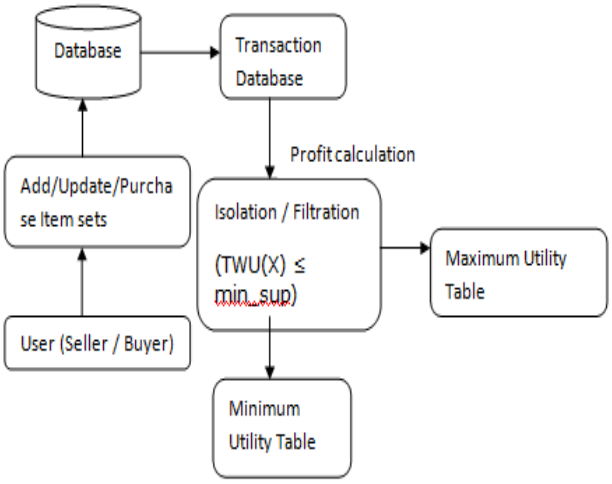


Figure 1:Profit calculation and profit Table filteration

In this Figure we filter out profit table from given transactional database. For this we use profit calculation and then threshold value for filtering out profit table.

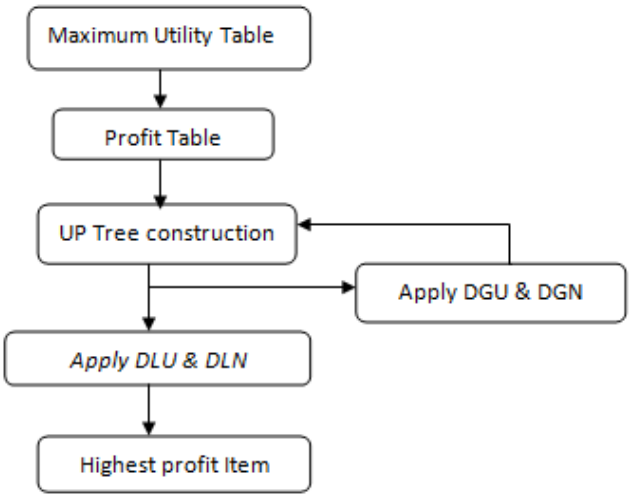


Figure 2: Finding High Profit Items

In the above figure, we filter out the data having frequent transactions and high profit. Profit calculation is applied on Maximum utility table and represent outcome as a UP Tree data structure. After getting, UP Tree DLU and DLN are applied to get highest profit item.

Multiplexor Logic:

In addition to mathematical model the multiplexer logic also discussed here. As the figure indicates, the data from Transactional database is given as input. Processing on the transactional data for profit calculation on Itemsets. Apply threshold value on profit table data. we get low and high profit Itemsets. The final result will be the high profit Itemsets.

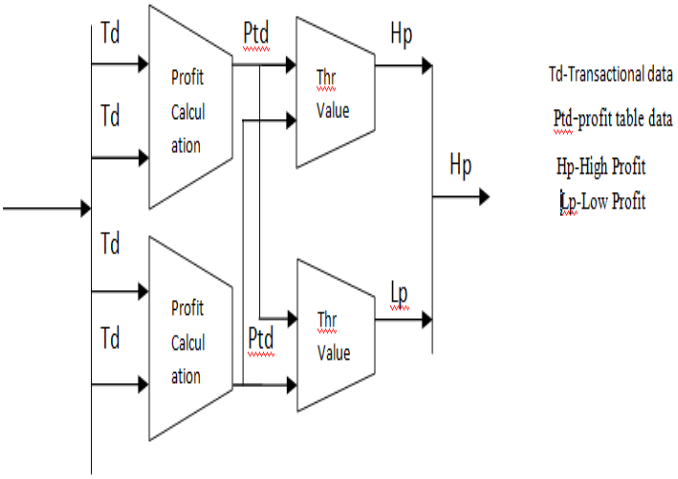


Figure 3: Multiplexor Logic

5. Algorithm

IUPG-Algorithm:

Input: Transaction database D, user specified threshold.

Output: high utility itemsets.

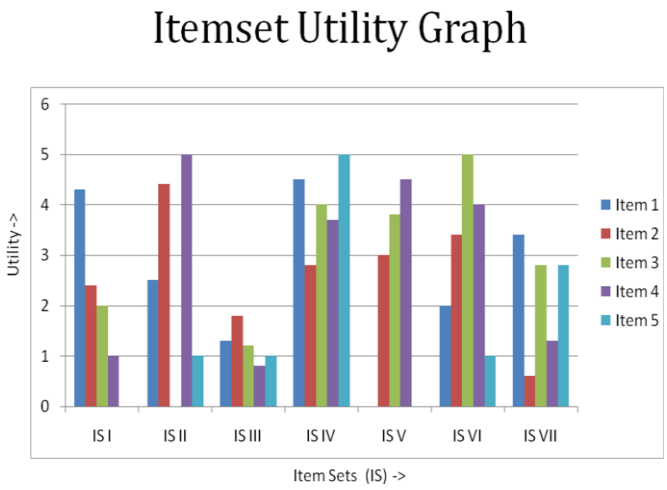
Begin

1. Scan database of transactions $T_d \in D$
2. Determine transaction utility of T_d in D and TWU of itemset (X)
3. Compute min_sup (MTWU * user specified threshold)
4. If (TWU(X) ≤ min_sup) then Remove Items from transaction database
5. Else insert into header table H and to keep the items in the descending order.
6. Repeat step 4 & 5 until end of the D.
7. Insert T_d into global UP-Tree
8. Apply DGU and DGN strategies on global UP-tree
9. Re-construct the UP-Tree
10. For each item a_i in H do
11. Generate a PHUI Y= X U a_i

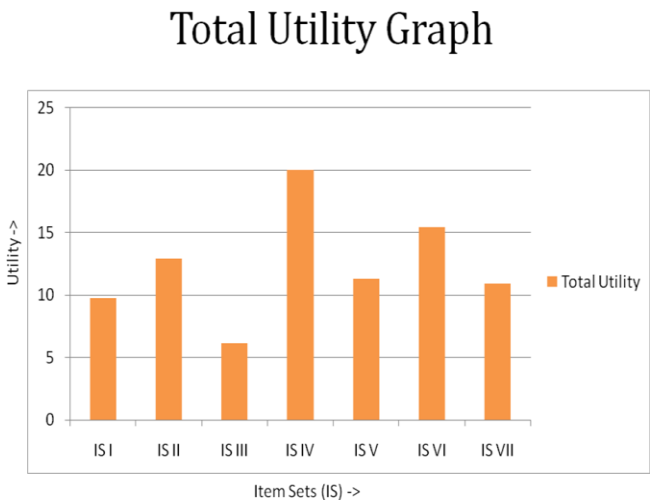
12. Estimate utility of Y is set as a_i's utility value inH
13. Put local promising items in Y-CPB into H
14. Apply strategy DLU to reduce path utilities of the paths
15. Apply strategy DLN and insert paths into T_d

6. RESULTS

The graph (1) there is no of itemsets with respect to there utility. In the itemsets there are no of items which shows their different utilities in different itemsets.



Graph 1: Itemset Utility Graph



Graph -2: Total Utility Graph

The graph states the total utility of the itemsets.From This graph we can find out which itemset have more utility and then discard the low utility itemsets.

6. CONCLUSION

In this Paper we focused on the big size data of Transactional database. This approach is useful to group profitable items and non-profitable items. Also we can find the high utility items.These results can be shown to retailer and he can taken decision to put/discard those items in his shop.

ACKNOWLEDGEMENT

The success of any project is never restricted to any individual .This project is the results of thoughts contributed by many people and would like to acknowledge them here. I express sincere thanks to all those who have provided me with valuable guidance towards the completion of this report .I deeply thank my guide Prof. D.P.Gadekar for his useful guidance. The support that he gave truly helped me, without whom this project would have been a far realism. I also thank to HOD prof S.R.Todmal for giving me good quality support, suitable remarks and conversation in all phase of the project .I would also like to widen sincere thanks to all teachers and staff for their valuable suggestions and feedback.

REFERENCES

[1]Y. Liu, W. Liao and A. Choudhary, "A fast high utility itemsets mining algorithm," in Proc. of the Utility-Based Data Mining Workshop, 2005.

[2] R. Agrawal, Imielinski. T and A. Swami, "Mining association rules between sets of items in large databases", in proceedings of the ACM SIGMOD International Conference on Management of data, pp. 207-216, 1993.

[3] H. F. Li, H. Y. Huang, Y. C. Chen, Y. J. Liu and S. Y. Lee, "Fast and Memory Efficient Mining of High Utility Itemsets in Data Streams," in Proc. of the 8th IEEE Int'l Conf. on Data Mining, pp. 881-886, 2008.

[4] V. S. Tseng, C. J. Chu and T. Liang, "Efficient Mining of Temporal High Utility Itemsets from Data streams," in Proc. of ACM KDD Workshop on Utility-Based Data Mining Workshop (UBDM'06), USA, Aug., 2006.

[5]W. Wang, J. Yang and P. Yu, "Efficient mining of weighted association rules (WAR)," in Proc. of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2000), pp. 270-274, 2000.

[6] Liu. Y, Liao. W, A. Choudhary, "A Fast High Utility Itemsets Mining Algorithm," In: 1st Workshop on Utility-Based Data Mining.Chicago Illinois, 2005.

[7] H. F. Li, H. Y. Huang, Y. C. Chen, Y. J. Liu and S. Y. Lee," Fast and Memory Efficient Mining of High Utility Itemsets in Data Streams",in Proc. of the 8th IEEE Int'l Conf. on Data Mining, pp. 881-886, 2008.

[8] S.Shankar, T.P.Purusothoman, S. Jayanthi,N.Babu, A fast agorithm for mining high utility itemsets, in: Proceedings of IEEE International Advance Computing Conference (IACC 2009), Patiala, India, pp.1459-1464

[9] J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation", Proc. ACM -SIGMOD Int‟l Conf. Management of Data, pp. 1-12, 2000.

[10] J. Hu, A. Mojsilovic, "High-utility pattern mining: A method for discovery of high-utility item sets", Pattern Recognition 40 (2007) 3317 –3324.