An Improved Approach for Mining High Utility Item Set from Large and Dynamic Data Set

Nidhi Sethi,
Research Scholar Mahatma Gandhi
Chitrakoot Gramodaya Vishwavidhyala
Satna (M.P).,India

Dr. Pradeep Sharma
Head, Dept. Of Computer Science,
Govt. Holkar Science College
Indore (M.P.) India

Dr. Bharat Mishra
Associate Professor, Mahatma Gandhi
Chitrakoot Gramodaya Vishwavidhyala,
Satna (M.P), India

***Abstract:* - Utility mining is an extension of pattern mining. Utility means weight, profit, cost, quantity or any useful entity on which business environment can depend on. Utility mining is an important technique for mining patterns through utility. Utility mining provides sufficient information about the product. Several algorithms have been developed for mining high utility itemsets. Efficiency is a big factor for improvement in the existing algorithms. Efficiency can be measured in term of execution time, memory requirement or arithmetic complexity. In this paper we present a novel approach for mining high utility item set with the help of data compactions techniques. Our proposed algorithm not only reduces the data base size during scanning of data set but also reduces number of candidates and lessens arithmetic calculation that provides a big advantage over the previous algorithms.**

Keywords: - utility mining, weight, profit, cost, quantity.

I. INTRODUCTION

Data mining takes data as input, yields patterns of classification, clustering, association and produces summaries as output. The most significant task in data mining is the process of discovering different types of patterns. Several efficient algorithms have been developed for mining patterns. Utility considerations in data mining tasks are gaining popularity in recent years. Utility-based data mining integrates utility considerations in both predictive and descriptive data mining tasks.

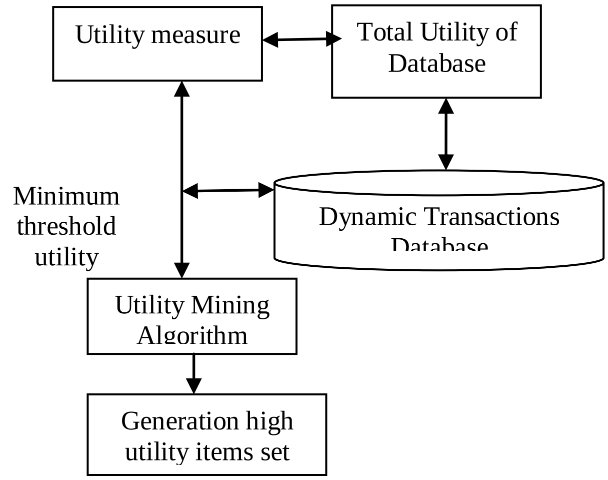
Utility measure
Total Utility of Database
Minimum threshold utility
Dynamic Transactions Database
Utility Mining Algorithm
Generation high utility items set

Figure 1 simple utility mining process

Simple model for mining high utility is shown in figure 1. This shows the steps and process of utility mining

Utility mining process first calculates the total utility of the database and then compares it with given minimum threshold value to generate high utility item set. There are several other terms used in utility mining figure 2 shows some of them.

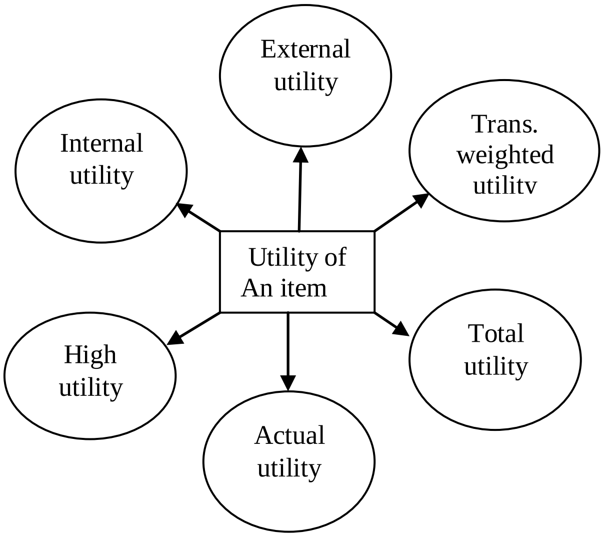

Figure 2 the meaning of utility

Internal utility is the utility of an item in the transaction; External utility is considered as profit and given in the profit table. Transaction weighted utility is calculated by transaction number. Actual utility is the utility of the itemset. Finally overall utility of the database is known as total utility.

The research work presented in the paper is also experimented and with the help of exponential analysis we have shown that the number of candidates is reducing at every level. Also removal of these useless candidates makes size of search space smaller and improves the performance of the proposed method.

II. LITERATURE REVIEW

In 2005 Hong Yao, Howard J. Hamilton, and Cory J. Butz proposed”A Foundational Approach to Mining Itemset Utilities from Databases “. They have proposed basic theoretical model. They defined two types of utility of items transaction utility and external utility. The utility of an item can be an integer value, such as the quantity sold, or a real value, such as a profit margin, total revenue, or unit cost. They defined the problem of utility mining by analyzing the utility relationships among itemsets and identified the utility bound property and the support bound property. They further discussed the mathematical model of utility mining based on these properties.

In 2005 to address the drawbacks in MEU Ying Liu Wei-keng Liao AlokChoudhary proposed a novel Two-Phase algorithm that can effectively prune candidate itemsets and simplify the calculation of utility. It substantially reduces the search space, memory cost and requires less computation. They defined a transaction-weighted utilization mining model which holds a “Transaction-weighted Downward Closure Property”. In Phase II, database scan is performed to find out the high transaction-weighted utilization itemsets.

In 2006 Yao H and Hamilton J, proposed “Mining itemset utilities from transaction databases. UMining is one of the well-known algorithms used for mining all high utility itemsets. In this algorithm functions of scan, calculate and store, discover, generate, and prune can be found. This algorithm uses apriori concepts to generate candidate set and then finds high utility itemsets, this process is repeated until no more candidate generation is possible.

In 2007 Alva Erwin, Raj P. Gopalan, N.R. Achuthan “A Bottom-Up Projection Based Algorithm for Mining High Utility Itemsets” proposed a new algorithm called CTU-PRO that mines high utility itemsets by bottom up traversal of a compressed utility pattern (CUP) tree. They developed a compact data representation named Compressed Utility Pattern tree (CUP-tree) for utility mining, and a new algorithm named CTU-PRO for mining the complete set of high utility itemsets. The concept of TWU is used for pruning the search space in CTU-PRO, and it avoids a rescan of the database to determine the actual utility of high TWU itemsets.

In 2008 Hua-Fu Li, Hsin-Yun Huang, Yi-Cheng Chen, and Yu-Jiun Liu, and Suh-Yin Lee proposed one of the most interesting technique for mining of high utility itemsets in many broad apllication. They 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 are developed to improve the efficiency of mining high utility itemsets. The proposed algorithms outperform when compared with the existing algorithms for mining high utility itemsets from data streams.

In 2008 Alva Erwin, Raj P. Gopalan, and N.R. Achuthan proposed “Efficient Mining of High Utility Itemsets from Large Datasets”. High utility itemset mining extends frequent pattern mining to discover itemsets in a transaction database with utility values above a given threshold. They proposed an algorithm that uses TWU with pattern growth based on a compact utility pattern tree data structure. They implement a parallel projection scheme to use disk storage when the main memory is inadequate for dealing with large datasets. They presented the CTU- PROL algorithm to mine the complete set of high utility itemsets from both sparse and relatively dense data sets.

In 2009 S.Shankar, Dr. T. Purusothaman, S.Jayanthi proposed “A Fast Algorithm for Mining High Utility Itemsets”. FUM algorithm is used for mining all high utility itemsets.FUM algorithm generates high utility itemsets by using Combination Generator.It is simple and it executes faster than Umining algorithm. when number of distinct items increases in the input database large number of itemsets are identified as high utility itemsets. The Combination Generator(T) is a method which is used to generate all the combinations of the items. The factorial computation method is defined in this, to generate the factorial of a given number.

In 2010 Vincent S. Tseng, Cheng-Wei Wu, Bai-En Shie, and Philip S. Yu proposed a data structure, named UP-Tree, and then described an algorithm, called UP-Growth, The framework of the UP-Growth: An Efficient Algorithm for High Utility Itemset Mining method. Proposed approach is not based on the traditional framework of transaction-weighted utilization mining model. UP-Tree improves the mining performance and avoids scanning original database repeatedly; proposed algorithm provides a compact tree structure, called UP-Tree to maintain the information of transactions and high utility itemsets.

In 2011 S. Kannimuthu Dr. K. Premalatha S. Shankar proposed iFUM - Improved Fast Utility Mining. The core step of FUM algorithm is Combination Generator (T) which takes significant time to compute. In the existing system FUM combination generation is performed for itemsets and its subset without checking one important condition. Combination Generator (T) - generates all possible combinations of itemset ∈ T. FUM algorithm fails to check this condition so it generates the combinations for the already generated subset of the itemsets too, if it repeats in a later transaction of the input database. Proposed algorithm avoids these extra computations and enhances FUM efficiency.

In 2012 Mengchi Liu and JunFengQu proposed Mining High Utility Itemsets without Candidate Generation" High utility itemsets refer to the sets of items with high utility like profit in a database, and efficient mining of high utility itemsets plays a crucial role in many real life applications and is an important research issue in data mining area. They proposed an algorithm, called HUI-Miner (High Utility Itemset Miner), for high utility itemset mining.HUI-Miner uses a novel structure, called utility-list, to store both the utility information about itemset and the heuristic information for pruning the search space of HUI-Miner. HUI-Miner can efficiently mine high utility itemsets from the utility lists Constructed from a mined database.

In 2012 Cheng Wei Wu, Bai-En Shie, Philip S. Yu, Vincent S. Tseng "Mining Top-K High Utility Itemsets**"** Mining high utility itemsets from databases is an emerging topic in data mining, which refers to the discovery of itemsets with utilities higher than a user-specified minimum utility threshold minutil. They proposed efficient algorithm named TKU (Top-K Utility itemsets mining).TKU incorporates several novel strategies for pruning the search space to achieve high efficiency.

In 2013 Arumugam P and Jose Proposed "Advance Mining of High Utility Itemsets in Transactional Data". They proposed a novel algorithm for transactional high utility item set mining approach. This method is used to find association, correlation and can generate less number of candidates. So the sales person can use this utility item set transaction for their stocks planning distributor/dealer month wise, product wise, model wise target setting.

In 2014 D. Usha Nandini, Ezil Sam Leni, M. Maria Nimmy proposed Mining of High Utility Itemsets from Transactional Databases. They also proposed a compact tree structure, called Utility pattern tree (UP-Tree) and it maintains the information of high utility itemsets. Performance of UP-Growth and UP-Growth+ become more efficient since database contain long transactions and generate fewer number of candidates than FP-Growth. The experimental results and comparison validate its effectiveness.

In 2014 Rani N. and Anbhule Reshma V proposed "Mining High Utility Item sets From Transaction Database" Mining high utility item sets from a transactional database means to retrieve high utility item sets from database. Proposed system was an efficient Algorithm for Mining High Utility Item sets From Transactional Database when compared with UP-Growth Algorithm. For that algorithm information of high utility item sets is maintained in tree based data structure named Utility Pattern Tree. The proposed tree-based algorithm, called UP-Growth, is used efficiently for mining high utility item sets from transactional databases. This algorithm takes UP-Tree data structure for maintaining the information of high utility itemsets and four effective strategies, DGU, DGN, DLU and DLN, to reduce search space and the number of candidates for utility mining.

In 2014 G. Saranya and A.Deepakkumar proposed "Implementation of Efficient Algorithm for Mining High Utility Itemsets in Distributed and Dynamic Database**".** The traditional method of mining frequent itemset mining embraces the data across and sedentary and imposes extreme communication overheads when the data is distributed, also they waste resources in calculation when the data is dynamic. To overcome this problem, they proposed Utility Pattern Mining Algorithm in which itemsets are maintained in a tree based data structure, called as Utility Pattern Tree. A quick update incremental algorithm is used which scans only the incremental database as well as collects only the support count of newly generated frequent itemsets.

III. PROBLEM STATEMENT

Lots of methods have been developed for mining high utility item set. Two important problems is always in consideration first is how minimize number no of candidates and another is how to remove space and time complexity

IV. PROPOSED METHODS

We used the following step in proposed methods. For each transaction in D, do the following sub steps?

1. Calculate the transaction utility of the transaction.
2. Calculate the transaction-weighted utility of each item as the summation of the transaction utility values of the transactions which include the item.
3 Check whether the transaction-weighted utility of an item is larger than or equal to the minimum utility threshold if it is, put it in the set of high transaction weighted utilization itemsets.
4 Use data compaction techniques to reduce the database size
5 Calculate the transaction utility of each modified transaction.
6 Scan the set of modified transactions to and the transaction-weighted utility
Check whether the transaction-weighted utility of an item is larger than or equal to the minimum utility threshold.

V. EXPERIMENTAL ENVIRONMENT

We have used VB dot net 2010 as front end and SQL server as back end for evaluating and validating the algorithm. All the experiments were performed on a i3 4M Cache, 2.50 GHz Intel PC machine with 2 gigabyte main memory, running Microsoft Windows 7.To evaluate the performance real life dataset is used. We have compared three algorithms first one Two Phase (TP), iFUM (Improved Fast Utility Mining) and proposed method by taking 50 different items and records of 1000 customers from store and analyze the results that can be used in practice.

VI GRAPHS ANALYSIS

1. Comparison between level and number of candidate

On the x axis we have taken level number and on y axis numbers of candidates are displayed.

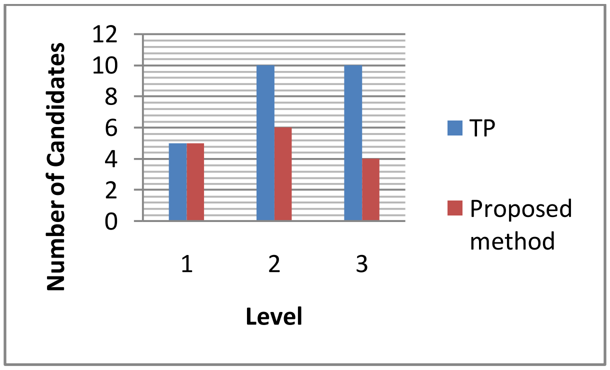

Figure 3 Comparison on the basis of level and number of candidate

VII .CONCLUSION AND FUTURE WORKS

Several algorithms have been developed sofar for mining high utility items set. But improving efficiency and reducing complexity is still a matter of consideration for researchers. In proposed work we have reduced the number of candidates at different level that further shortens the execution time. We have also used concept of merging the transactions that contains similar items this has reduced the number of transactions as well as lessen the calculation, as high utility item set require multiplication operation and addition operations. Our algorithms works for large dataset and the dataset can be scalable too if required. In future it can be implemented on data set based on time factors or simply temporal data sets.

REFERENCES

[1] Hong Yao, Howard J. Hamilton, and Cory J. Butz A Foundational Approach to Mining Itemset Utilities from Databases Department of Computer Science University of Regina
Regina, SK, Canada 2005.

[2] Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer, Byeong-Soo Jeong, and Young-Koo LeeAn Efficient Candidate Pruning Technique forHigh Utility Pattern Mining T. Theeramunkong et al. (Eds.): PAKDD 2009, LNAI 5476, pp. 749–756, 2009.Springer-Verlag Berlin Heidelberg 2009

[3] JyothiPillaiO.P.Vyas Overview of Itemset Utility Mining and itsApplications International Journal of Computer Applications (0975 – 8887)Volume 5– No.11, August 2010

[4] Guangzhu Yu, Shihuang Shao And Xianhui Zeng Mining Long High Utility Itemsets in Transaction Databases Wseas Transactions onInformation Science & Applications Issue 2, Volume 5, Feb. 2008.

[5] Alva Erwin, Raj P. Gopalan, and N.R. Achuthan Efficient Mining of High Utility Itemsets from Large Datasets Department of Computing, Department of Mathematics and Statistics Curtin University of Technology, Kent St. Bentley Western Australia T. Washio et al. (Eds.): PAKDD 2008, LNAI 5012,

[6] Cheng Wei Wu, Bai-En Shie, Philip S. Yu, Vincent S. Tseng1 Mining Top-K High Utility Itemsets KDD'12, August 12–16, 2012, Beijing, China 1Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan

[7] Shankar S Dr.T.Purusothaman A Novel Utility and Frequency Based Itemset Mining Approach for Improving CRM in Retail Business 2010 International Journal Of Computer Applications (0975 - 8887)
Volume 1 No. 16.

[8] D. UshaNandini, Ezil Sam Leni, M. Maria Nimmy Mining of High Utility Itemsets from Transactional Databases Mining of High Utility Itemsets from Transactional Databases. International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-3, Issue-4, April 2014

[9] Mengchi Liu, Junfeng Qu Mining High Utility Itemsets without Candidate Generation State Key Lab of Software Engineering, School of Computer, Wuhan University, Wuhan 430072, China CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

[10] More Rani N. Anbhule Reshma V Mining High Utility Item sets From Transaction Database International Journal of Latest Trends in Engineering and Technology (IJLTET)Vol. 3 Issue 3 January 2014.

[11] Bai-En Shie1, Philip S. Yu2 and Vincent S. Tseng Efficient Algorithms for Mining Maximal High Utility Itemsets from Data Streams with Different Models University of Illinois at Chicago, Chicago, Illinois, USA