ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 5, Issue 4 , April 2018

The Multidimensional Utility Mining for Super Market Applications

Payal Parmar, Anubha Sharma

P.G. Student, Department of Computer Science, Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore, Madhya Pradesh, India

Assistant Professor, Department of Computer Science, Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore, Madhya Pradesh,, India

ABSTRACT: High utility mining is used to find useful and rare items from a set of transactions. Though there are algorithms like association rule mining and frequent pattern mining, these algorithms extracts only the frequent item sets, which only discovers interesting patterns. On other hand, high utility mining finds items based on two parameters that is unit profit and quantity. We are working on multi-dimensional that is three dimensional and concentrates on frequency along with unit profit and quantity to find high utility items. In this paper, we find high utility items with the help of a balanced tree called B+ tree. The data structure B+ tree stores the utility score of each item in the form of key-value pair. B+ tree is used to avoid searching, also save time, easy sorting in ordered form and easy storage.

KEYWORDS: High utility item set mining, Transaction, Tree, B+ Tree

.

I.INTRODUCTION

Data mining [1] has become an essential technology for businesses and researchers in many fields, the number and variety of applications has been growing gradually for several years and it is predicted that it will carry on to grow. A number of the business areas with an early embracing of DM into their processes are banking, insurance, retail and telecom. More lately it has been implemented in pharmaceutics, health, government and all sorts of e-businesses. There are a number of algorithmic techniques available for each data mining tasks, with features that must be weighed against data characteristics and additional business requirements. . Among all the techniques, in this research, we are focusing on the pattern mining technique, which is descriptive mining technique, with transactional database system.

Finding interesting patterns from the datasets is an important research area in the field of data mining. In utility mining we concentrate on utility value of item set. Utility mining has introduced utility mining considers both the profit and the number of items purchased. In these the utility of an item is calculated as the product of the item and the number of item purchased. An item is set to be high utility item if the sum of the utility of the item in the dataset is greater than the user specified minimum threshold. Otherwise it is low utility item. High utility patterns mining is a major research area in recent years and many research are carried in different area of high utility mining which includes sequence high utility mining, lossless representation of HUI, incremental HUI etc.

Figure 1 shows the process of finding high utility items from a transactional dataset. Thus, in these paper, we proposed a method of using B+ tree. And we calculate the utility score of an item considering its unit profit, quantity and frequency/occurrence. That why, it is three dimensional utility mining. B+ tree is a data structure that stores any element in terms of 'key and value' pair. Thus, we store the utility score of an item as key and item as value. B+ tree is used to avoid searching, also save time, easy sorting in ordered form and easy storage.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 5, Issue 4 , April 2018

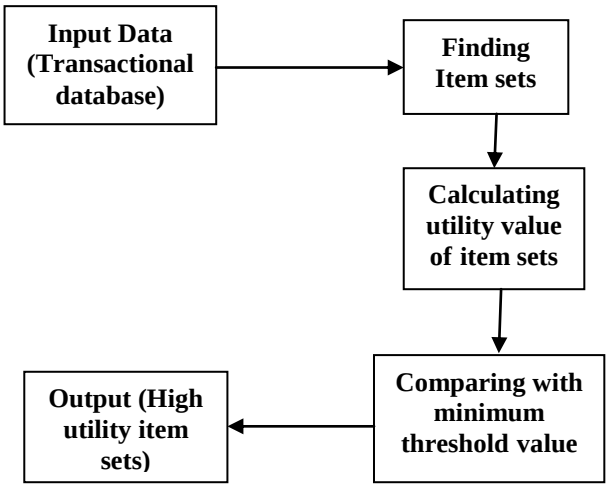

Figure 1: High Utility Mining

II. RELATED WORK

This section of the work comprises of relevant work depending of different approaches used by researches to diagnose and mitigate drawbacks of previous work and also cope in the environment of changing trends in emerging technology.

In 2010 the author ZHOU Jun et al. [2] proposed this algorithm by considering the space as an important factor. Authors used an improved LRU (Least Recently Used) based algorithm. Proposed algorithm omits the infrequent items before taken for the processing. Method increases the stability and the performance. Method is used to find out the frequent items as well as the frequency of those items.

Most of the existing algorithms uses a measure known as TWU (Transaction Weighted Utility). This measure was introduced Liu et al. [3], also they follow the process of two phase candidate generation.
Work done by J. Liu et al. [4] is based on the concept of a tree construction based method. It does not generate candidates. First a tree is constructed and then DFS (Depth First Search) technique is used to visit the nodes of the tree to calculate the utility of items. But construction of tree takes O(n) time. Also searching element in a tree requires O(logn) time. Deletion requires O(logn) time.

Shiming Guo [5] explores that High utility item set mining (HUIM) can plays important role for incremental data model. They observe that most research works do not consider any kind of updations and modification in analytical model. Furthermore, addition and deletion from database can play lead role in lead generation or knowledge generation. Study explore that main purpose of utility mining is lead generation which can be wrong with outdated transactional data source. To deal with such problems they proposed to have IHUP-Tree instead of conventional HUIM.

Table 1 Comparative Table

AUTHOR	YEAR	TITLE	TECHNIQUE/ALGORITHM USED
ZHOU Jun, CHEN Ming,	2010	A More Accurate Space Saving Algorithm for Finding the Frequent Items	*Least Recently Used (LRU) based algorithm used.*
Y. Liu, W. Liao, and A. Choudhary *[3]*	2005	A fast high utility item sets mining algorithm	*Proposed a TWUIM algorithm.*

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 5, Issue 4 , April 2018

J. Liu, K. Wang, and B. C.M. Fung *[4]*	2016	Mining high utility patterns in one phase without generating candidates	*Proposed a novel reversed set enumeration tree and d2HUP algorithm.*
Shiming Guo, Hong Gao *[5]*	2017	An Efficient Algorithm for Incremental and Interactive High Utility Itemset Mining	*Proposed IHUI Tree data structure and IHUI algorithm.*

III. PROBLEM DOMAIN

A) Problem Statement:

The complete study concludes that:

- Existing solutions do not considered item occurrence frequency and popularity factor for calculating the utility score of an item.
- It only considers two dimensional analyses based on unit profit and quantity of sale. That is,

  Utility of an item = unit profit * quantity
- They used unweighted tree data structure to organize transaction and users which suffer with issue of ordering and searching based on calculated score[weight].

IV. PROPOSED WORK

In our proposed work we will be working on multi dimension utility mining and concentrated on the frequency which is derived on the basis of popularity. Multidimensional filtering in which frequency will be count as one of the dimension is used, that is

Utility of an item = M*N*L

Where, M = Quantity,
N = unit profit and
L = Frequency.

In our proposed work existing tree will be replaced by the B+ tree. B+ tree is a data structure that stores any element in terms of 'key and value' pair. Thus, we store the utility score of an item as key and item as value. B+ tree is used to avoid searching, also save time, easy sorting in ordered form and easy storage. For example, consider a transaction database (Table 2) and net profit table (Table 3). In a transaction database, each row represent a transaction made by particular user. Where A, B and C represent the item name and the number associate with each item represent the quantity of each item purchased by user in each transaction. And in net profit table, there is a net profit value associate with each item.

Tid	Transaction
1	(A,2),(B,1)
2	(B,2),(C,1)
3	(A,1),(B,2),(C,3)
4	(B,1),(C,2)

Table 2

Item	Net Profit
A	5
B	4
C	2

Table 3

Now calculate the utility score of each item. For example the utility score of item A can be calculated as:

U(A) = q(A)*p(A)*f(A)

Where q(A) is the total quantity of A in database. So, item A has value 2 in transaction 1 and value 1 in transaction 2 so, total quantity of A is 2+1 that is 3. p(A) is the net profit of A that is 5 and f(A) is the frequency of A that is item A occur in 2 transactions(tid 1 and tid 2) so frequency of A is 2. Hence, U(A) = 3*5*2 = 30. The utility value of A is 30. In the same way we calculate the utility value of each item, which is shown in table 4. Now construct the B+ tree to insert the utility value of each item as key and the item as value. For the above utility table the B+ tree is as follow:

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 5, Issue 4 , April 2018

Item	Utility value
A	30
B	96
C	36

Table 4: Utility value table

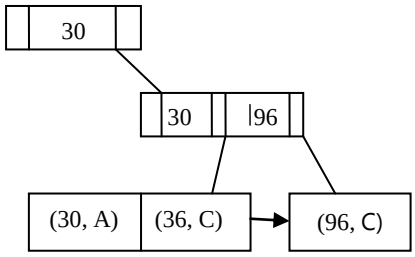

Figure 2 : B+ tree

Since all the values are stored in leaf node of a B+ tree and all the values are sorted in increasing order, it is easy to find the high utility product. And all the leaf nodes of a B+ tree are linked with each other, it is easy for traversing the tree.

V. EXPERIMENTAL RESULTS

In this section, we evaluate the performance of our proposed method and perform experiments on real life transaction dataset. Fig. 3 shows the execution time on the dataset for different minimum threshold varying from 50000 to 250000. It could be easily observed that the execution time of the proposed approach decreased along with the increased of the threshold value. Fig. 4 shows the number of high utility items on the dataset under varied minimum threshold. It could be observed that the number of high utility items decreases as the threshold value increased.

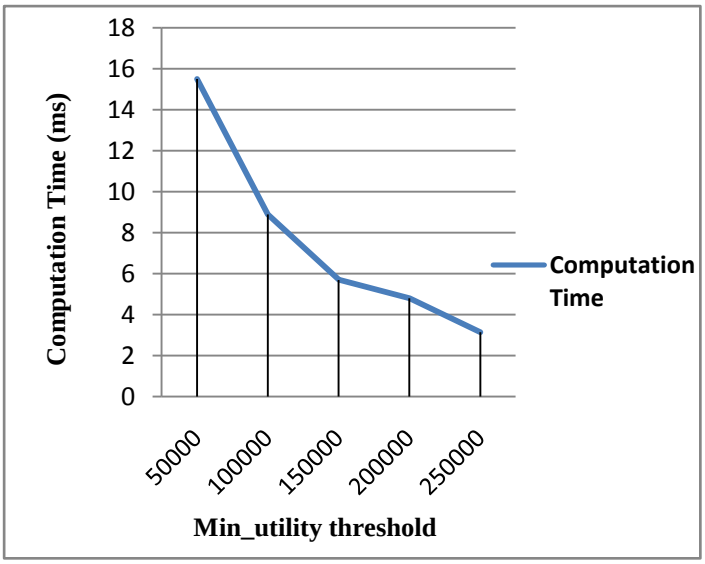

Figure 3: Execution time under different threshold

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 5, Issue 4 , April 2018

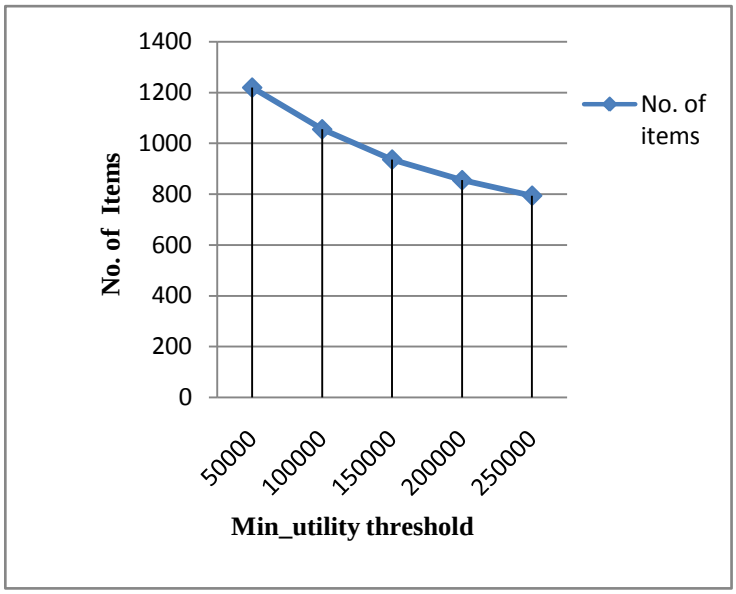

Figure 3: Execution time under different threshold

VI. CONCLUSION AND FUTURE WORK

High utility frequent pattern mining has a wide range of real world applications. That's why it is one of the most favourite topic of research. Utility mining helps in mining of items which are worthy. In our proposed work we will be working on multi dimension utility mining and concentrated on the frequency which is derived on the basis of popularity. A Multidimensional evaluation system will be used to calculate impact. It is found that although a lot of work is going on in the field of high utility mining but still there is enough scope to improve the performance.

REFERENCES

[1] Tan P.-N., Steinbach M., and Kumar V., ''Introduction to data mining'', Addison Wesley Publishers, 2006
[2] ZHOU Jun, CHEN Ming, XIONG Huan, ''A More Accurate Space Saving Algorithm for Finding the Frequent Items'', IEEE-2010.
[3] Y. Liu, W. Liao, and A. Choudhary, "A fast high utility itemsets mining algorithm," in Proc. Utility-Based Data Mining Workshop SIGKDD, 2005, pp. 253–262.
[4] J. Liu, K. Wang, and B. C.M. Fung, ''Mining high utility patterns in one phase without generating candidates," IEEE Transaction Knowledge And Data Eng., vol. 28, no. 5, pp. 1245–1257, May 2016.
[5] Shiming Guo, Hong Gao, "An Efficient Algorithm for Incremental and Interactive High Utility Itemset Mining," published in International Conference on Image, Vision and Computing, 2017 pp 996-1001
[6] C.F. Ahmed, S.K. Tanbeer, and B.S. Jeong et aI., "Efficient tree structures for high utility pattern mining in incremental databases," In TKDE 2009 Vol. 21(12) , pp. 5 5-64.
[7] M. Zihayat and A. An, "Mining top-k high utility patterns over data streams," In Inf. Sci. 2014 Vol. 285, pp. 138-161.