ISSN:0975-9646

A Survey on High Utility Itemset Mining Using Transaction Databases

Maya Joshi[1], Mansi Patel[2]

[1]*Computer Engineering Department, Silver oak college of Engineering & Technology, Gujarat Technological University, India*

[2]*Information Technology Department, Silver oak college of Engineering & Technology, Gujarat Technological University, India*

Abstract— **Data Mining can be delineated as an action that analyze the data and draws out some new nontrivial information from the large amount of databases. Traditional data mining methods have focused on finding the statistical correlations between the items that are frequently appearing in the database. High utility itemset mining is an area of research where utility based mining is a descriptive type of data mining, aimed at finding itemsets that devote most to the total utility. Mining high utility itemsets from a database refers to the discovery of itemsets with high utility in terms like weight, unit profit or value, Also entitled as frequent itemset mining with high profit. In High Utility Itemset Mining the goal is to recognize itemsets that have utility values above a given utility threshold. In this paper, we present a literature survey of the present state of research and the various algorithms and its limitations for high utility itemset mining.**

Keywords— **Utility mining, High utility itemsets, Frequent itemset mining.**

I. INTRODUCTION

The objective of frequent itemset mining [1] is to find items that frequently appear in a transaction database [2] and higher than the frequency threshold given by the consumer, without considering profit of the item. However, quantity, weight and value are significant for addressing real world decision problems that require maximizing the utility in an organization.

The restraint of frequent itemset mining [3] is it assumes (1) an item can only appear once in a transaction (2) all items have the same importance/weight (e.g. Profit).

So it may ignore rare itemset having higher profit (e.g. Caviar, wine). To overcome this issue, the problem of FIM [1] has been resolved as High-Utility Itemset Mining (HUIM). The high utility itemset mining problem is to find all itemsets that have utility larger than a user specified value of minimum utility. The value or profit Associated with every item in a database is called the utility of that itemset.

Utility of items in transaction database involves following two aspects:
(1) The importance of distinct items, called external utility(e), and

(2) The importance of items in transactions, called internal utility(i).

Utility of Itemset (U) = external utility (e) * internal utility (i).

In many areas of business like retail, inventory, etc. decision making is very important. In a transaction database each item is represented by a binary value, without considering its profit.

In many applications like cross-marketing in retail stores, online e-commerce management, website click-stream analysis and finding the important pattern in bio-medical applications High utility mining are widely used.

Example1

Consider, example of a transaction database representing the sales data and the profit associated with the sale of each unit of the items.

Table I
Transaction Database

TID	Item sold in a transaction		
	Item A	**Item B**	**Item C**
T1	0	0	18
T2	0	6	0
T3	1	0	1
T4	2	4	8
T5	5	2	4
T6	3	0	2
T7	0	10	0
T8	6	1	25
T9	1	0	0
T10	0	6	2

Table II
Unit Profit Associated with items

Item Name	Unit Profit (in INR)
Item A	5
Item B	10
Item C	3

Let us consider the itemset AB.

Since, there are only 3 transactions T4, T5 and T8 which contains AB itemset out of 10 transactions.

So, support for itemset AB is
Support (AB) = 3 / 10 * 100 = 30 %

In T4 transaction, units gain by item A and B are 2 and 4 respectively, the profit earned from the sale of itemset AB in T4 transaction is given by,

profit (AB, T4) = 2 * profit (A) + 4 * profit (B)
= 2*5 + 4*10
= 50

Since AB appears in transactions T4, T5 and T8,
So, total profit of itemset AB is given by

profit(AB) = profit(AB,T4) + profit(AB,T5) + profit(AB,T8)
= (2*5+4*10) + (5*5+2*10) + (6*5+1*10)
= (10+40)+(25+20)+(30+10)
= 50 + 45 + 40
= 135

Similarly, we can calculate the support values for the different itemsets and also the profit obtained by the sale of those itemsets by all the ten transactions as indicated in table III.

Table III
Support And Profit For All Itemsets

Itemset	Support (%)	Profit
A	**60**	90
B	**60**	**290**
C	**70**	180
AB	30	135
BC	40	**247**
AC	**50**	**205**
ABC	30	**246**

If we consider minimum support 50%, then we can observe that there are only 4 itemsets A, B, C and AC which have the support greater than the threshold value (min_sup). So, they qualify as frequent itemsets. But if we consider it profitwise then we can find out of 4 most profitable itemsets B, BC, AC, ABC only B and AC are frequent itemsets. Itemsets BC and ABC are not frequent but still they fetch the more profit than other itemsets.

As we can see from table III, single unit of item B fetch more profit than single unit of Itemset A and B.

From this Example, we can illustrate frequent Itemset mining may not always satisfy profitwise requirements of sales manager .In this case, the support (%) attribute of the itemsets reflects the the statistical correlation not the semantic significance of items.

II. Data Flow Diagram

The following Diagram shows the chain process of calculating and displaying a high utility Itemsets. From this Fig. , the comparison of frequent Itemset with given threshold value and by considering a profit, gives High utility Itemsets.

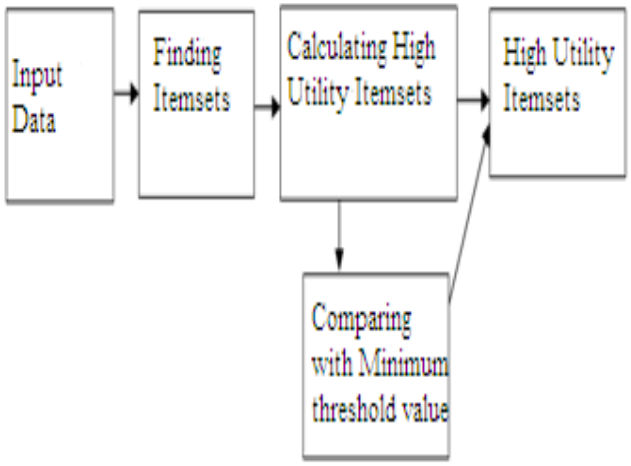

Fig. Data Flow Diagram[9]

III. Literature Review

In the previous section we introduced the overview of Data Mining, Frequent Itemset Mining and High Utility Itemset Mining. A comparison of the various Algorithms, Techniques, approaches and limitations that have been defined in various research publications have been given in this section.

No	Title Of Paper	Year	Author(s)	Datasets	Name of Algorithm	Overview of work/Idea	Limitation	`Idea Of Improvement
1	A Two-Phase Algorithm for Fast Discovery of High Utility Item sets[4]	2005	Ying Liu, Wei-keng Liao, and Alok Choudhary	Transaction dataset	Two-Phase	Phase 1: Discover candidate itemsets, that is having a TWU ≥ minutil, Phase 2: For each candidate, calculate its exact utility by scanning the database	Multiple scans of database and generates many candidate Itemsets	This approach is suitable for sparse database with short Patterns.
2	CTU-Mine: An Efficient High Utility Itemset Mining Algorithm Using the Pattern Growth Approach [5]	2007	Alva Erwin, Raj P. Gopalan, N.R. Achuthan,	Transaction dataset	CTU-Mine	Use pattern growth algorithm and also eliminates the expensive second phase of scanning the database	Complex for evaluation due to the tree structure	This approach is suitable for dense dataset with long pattern
3	UP-Growth: An Efficient Algorithm for High Utility Itemset Mining[6]	2010	Vincent S. Tseng, Bai-En Shie,	Transaction dataset	UP-Growth	(1) construction ofUP-Tree, (2) generation of potential high utility itemsets from the UP-Tree by UP-Growth, and (3) identification of high utility itemsets from the set of potential high utility itemsets	Complex for evaluation due to the tree structure	synthetic and real datasets are used to evaluate the high performance of the algorithm
4	Mining High Utility Itemsets without Candidate Generation[7]	2012	Mengchi Liu, Junfeng Qu	Transaction dataset	Hui-Miner	Single Phase Algorithm. No need to multiple times database scan	Calculating the utility of an itemset joining utility list is very costly.	We should try to avoid performing joins if possible for low-utility itemsets.
5	FHM: Faster High-Utility Itemset Mining using Estimated Utility Co-occurrence Pruning[8]	2014	Philippe Fournier-Viger, Cheng-Wei wu	Transaction dataset	FHM	Estimated-Utility Co-occurrence pruning	Static Database	We should try it using a dynamic database.

Table 1 summarizes the comparison of different existing Approaches

IV. CONCLUSION

Most of research on high utility itemset focuses on static databases (eg. Transaction database). With the emergence of the new application, the data processed may be in the continuous dynamic data streams. Because the data in streams come with high speed and are continuous and unbounded, mining result should be generated as fast as possible and make only one pass over a data.

V. FUTURE WORK

In this Paper we have presented a review on various algorithms, work, idea and limitations of different methods for high utility Itemset mining using a transaction dataset, In the next paper we will present One pass algorithm for High utility Itemset Mining using stream data.

REFERENCES

[1] Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Database. In: ACM SIGMOD International Conference on Management of Data (1993) .

[2] A. Erwin, R.P. Gopalan, and N.R. Achuthan,"Efficient Mining of High Utility Itemsets from Large Datasets", T. Washio et al. (Eds.): PAKDD2008, LNAI 5012, pp. 554–561, 2008. © Springer-Verlag Berlin Heidelberg 2008.

[3] Yao, H., Hamilton, H.J., Buzz, C. J., "A Foundational Approach to Mining Itemset Utilities from Databases", In: 4th SIAM International Conference on Data Mining, Florida USA (2004).

[4] "A Two-Phase Algorithm for Fast Discovery of High Utility Itemsets", Ying Liu, Wei-Keng Liao, and Alok Choudhary, Northwestern University, Evans

[5] "CTU-Mine: An Efficient High Utility Itemset Mining Algorithm Using the Pattern Growth Approach" In: Seventh International Conference on Computer and Information Technology (2007).

[6] "UP-Growth: An Efficient Algorithm or High Utility Itemset Mining ", Vincent S. Tseng, Cheng-Wei Wu, Bai-En Shie, and Philip S. Yu. University of Illinois at Chicago, Chicago, Illinois, USA, 2010.

[7] Mengchi Liu *Junfeng Qu*, "Mining High Utility Itemsets without Candidate Generation", *2012*.

[8] "FHM: Faster High-Utility Itemset Mining using Estimated Utility Co-occurrence Pruning", Philippe Fournier-Viger1, Cheng-Wei Wu 2014.

[9] Smita R. Londhe,, Rupali A. Mahajan,, Bhagyashree J. Bhoyar,"Overview on Methods for Mining High Utility Itemset from Transactional Database", International Journal of Scientific Engineering and Research (IJSER), Volume 1 Issue 4,December2013