# A Survey Report on High Utility Itemset Mining for Frequent Pattern Mining

**Dhruv Patel[1] Mr. Kaushal Madhu[2]**
[1]P.G. Student [2]Assistant Professor
[1,2]Department of Computer Engineering
[1,2]L. J. Institute of Engineering and Technology, Ahmedabad, Gujarat, India

*Abstract*— Data Mining is the Process to extract knowledge from the data or data stream, Mining high utility itemsets focus on the itemset with high profit only, as data or itemset may be static or in a stream mine this type if data has become a significant research topic. In this research paper we have presented different methods available for mining high utility dataset , to develop algorithm for this requires logic to mine the high utility itemsets from data streams, We have compared following algorithms, One pass Algorithm, Two Phase Algorithm, Mining top k-Utility Frequent itemset, and Sliding window based algorithms like MHUI-BIT(Mining High Utility Itemset Based on BITVector), MHUI-TID(Mining HighUtility Itemset based on TID-List) for lexicographical tree based summary data structure. And the FHM algorithm which we will used for our proposed work "Fast High Utility Miner) Using estimated utility Co-occurrence Pruning.

***Key words:*** Frequent Pattern Mining, High Utility Itemset Mining, Transaction Database

## I. INTRODUCTION

Data Mining is used to extract structure knowledge from large datasets[5]. Frequent Pattern Mining[1] and it can be apply for classification based and cluster based algorithms , it is the method to find the relationship based on the itemset list X-->Y, if we apply the traditional algorithms on transactional dataset each itemset will be mined on based on some values like price ,quantity and weight ,now as per following tradition it will mine only the data set based on the frequent occurrence, but sometime it happens that some product has less frequent purchases but still it gives high profit for organization , for example Customers will not buy 3D printers frequently but the profit for 3D printer is very high so high utility mining is mining the itemsets which has greater utility than min_Utility. For example itemset {3Dink, 3DPrinter} 3D Ink is not frequent but 3D printer has highutility so we cannot prune it.

Different authors have proposed many algorithms, in this paper we have presented the overview of algorithm and proposed the problems and advantages of this work,

## II. LITERATURE REVIEW OF HIGHUTILITY ITEMSET MINING

### *A. [5] CTU-Mine: An Efficient High Utility Itemset Mining Algorithm Using the Pattern Growth Approach.*

Ervin et al[5] proposed a framework for high utility mining using CTU Tree (Compressed Transaction Utility Tree) and Item Table, the approach shows that the algorithm is based on utility mining using the pattern growth approach based on Umine and Umine_H . the advantage of the algorithm is it can mine the dense dataset for too long pattern , it also uses the pattern growth approach instead of the to phase scanning of database for candidate generation so the time required for scanning the database and the removing the spurious high utility dataset. The disadvantage of this method is due to anti-momotoe it requires high memory utilization compare to other algorithm.

Also requires the analysis of sampling based approximations to complement this investigation.

### *B. [4] Fast and Memory Efficient Mining of High – utility Itemset from Data Streams: With and Without Negative Item Profits*

Hua-Fu Li et Al. 2010 has Presented the mining of stream data, author has presented that in real world application data has unique characteristics of data streams , methods for mining such streaming the data and for mining traditional datasets are different because data sets in streams are continuous flows on unique speed so we have to keep track of all the data which comes in stream as well as previously arrived data items , so it requires two phase one to mine the data and another to store the data into the memory , the mining stream algorithms contains only one pass and limited space and real time .

The author proposed efficient algorithms MHUI-BIT(Mining high utility Itemsets based on BIT vector) and MHUI-TIF for mining HU Data, in this they have use item representation methods to improve performance of utility mining the used the generate itemset from sliding window protocol. The pros for this method is to limited memory usage and processing challenges for mining utility itemsets from data streams over time -sensitive sliding windows for negative item profits .

### *C. [3] Mining High Utility Itemsets*

Raymond Chan Et Al. Has proposed mining for OOA data with weaker and anti-Monotic data items this algorithm do not need the min_Utility by user and all frequent patterns produces are closed patterns they have used Apriory top-k algorithm, which is similar to traditional pruning as Apriory and it generates itemset for all sub-itemset .

| OOA 1-itemset | s% |
|---|---|
| tmt = 1 | 0% |
| tmt = 2 | 12.5% |
| tmt = 3 | 12.5% |
| tmt = 4 | 18.75% |
| tmt = 5 | 18.75% |
| med = 1 | 25% |
| med = 2 | 37.5% |

prune →

| OOA 1-itemset ($G_1$) | s% | $u^{(m)}(I)$ |
|---|---|---|
| tmt = 2 | 12.5% | 0.8 |
| tmt = 3 | 12.5% | 0.8 |
| tmt = 4 | 18.75% | 0.9 |
| tmt = 5 | 18.75% | 1.4 |
| med = 1 | 25% | 1.1 |
| med = 2 | 37.5% | 1.4 |

generate closed itemsets from $G_1$ →

| Closure ($C_1$) | s% | c% | u | $u_{(a)}(I)$ |
|---|---|---|---|---|
| tmt = 2, med = 1 | 12.5% | 50% | –0.25 | –1.3 |
| tmt = 3 | 12.5% | 66.67% | –0.067 | –0.5 |
| tmt = 4, med = 2 | 18.75% | 75% | 0.8 | 0.7 |
| tmt = 5 | 18.75% | 75% | 1.2 | 1 |
| med = 1 | 25% | 50% | 0.025 | –1.5 |
| med = 2 | 37.5% | 75% | 0.625 | –0.6 |

Fig. 1: Generation of $G_1$ & $C_1$

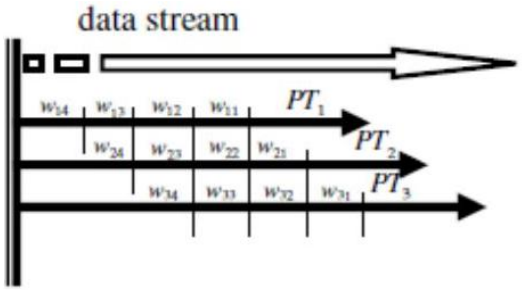
For all other generators that are not qualify for min support are removed in second phase i.e. Infrequent generators, as shown in figure the two data sets which are

used for their purpose they gained 1.9% to 4.9% improvement over 100 or more values for K .

*D. [2] Mining High Utility Itemsets In Data Streams Based On The Weighted Sliding Window Model.*

Puray At El.[4] has represented the weighted sliding window model for efficiently find the high utility itemsets in data streams.




In this model they have provided two features.

1) Time wise window size to avoid the interval length problem.
2) The amount of window considered for mining is decided by user. To distinguish current data from an other to mine weight and influenced.

Advantage of this method is that it can work on multiple pass over the data and it gives efficient result also they proposed single pass algorithm *HUI_W* but it lakes on weighted utility data set on streams.

*E. [1] FHM – Faster High- Utility Itemset Mining Using Estimated Utility Co-occurence Pruning*

Philippe Fournier Et Al. Authors have proposed the mining methodology for Co-occurunce Pruning method for transactional datasets .,High utility itemset mining is very important than the Frequent Itemset Mining, the most algortihms finds itemset in two phases the following definitions they have mentioned, Transactional Database, Utility of Itemset in transaction, and in database, Transaction weighted Utilization, the FHM algorithm finds TWU no less than minutil the procedure takes input as Itemset P and extensions of P having the form pz previously obtained data item, they have derived the itemset by following information.

Execution Time, Pruning Effectiveness, Memory Overhead, and the algorithm reduce the joins operations when using list as data structure. And it found 6 times faster, disadvantage is memory overhead which we will remove in our proposed method. Also we will check the effectiveness for different algorithm.

## III. Conclusion

From the research Paper we have reviewed so far we can conclude that for sequential pattern mining the traditional methods only gives frequent itemsets but to gain more profit for organization mining of high utility datasets is required, all the current methods for high utility dataset mining uses two phase algorithms. We can improve execution time, memory overhead or effectiveness of an algorithm by applying threshold utility for FHM.

## References

[1] Philippe Fournier-Viger, Cheng-Wei Wu, Souleymane Zida, and Vincent S. Tseng, "FHM – Faster High-Utility Itemset Mining Using Estimated Utility Co-occurence Pruning", SSpringer International Publishing Switzerland 2014.
[2] Pauray S.M. Tsai, "Mining High Utility Itemsets In Data Streams Based On The Weighted Sliding Window Model", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.4, No.2, March 2014.
[3] Raymond chan, Qiang Yang, Yi-Dong shen, "Mining High Utility Itemsets", Proceedings of the Third IEEE International Conference on Data Mining,2003 IEEE.
[4] Hua-Fu Li, Hsin-Yun Huang, Suh-Yin Lee, "Fast and Memory Efficient Mining of High–utility Itemset from Data Streams: With and Without Negative Item Profits", Springer-Verlag London Limited 2010.
[5] Alva Erwin, Raj P. Gopalan, N.R. Achuthan , "CTU-Mine: An Efficient High Utility Itemset Mining Algorithm Using the Pattern Growth Approach", Seventh International Conference on Computer and Information Technology,2007 IEEE.