A Survey on High Utility Itemset Mining using Transaction Database Technique and Method

Afrin Shaikh[1] Vishal Shah[2]
[1]M.E Student [2]Assistant Professor
[1,2]Department of Computer Engineering
[1,2]Sardar Vallabhbhai Patel Institute of Technology, VASAD, India

Abstract— Data mining is process of analyzing data from different data repository and mine Useful and interesting patterns from them. Data Mining is the process of revealing nontrivial, previously unknown and potentially useful patterns from large database. It is also defined as the use of algorithm to discover hidden patterns and interesting relationship between large itemset. High utility mining is an area research where utility based mining can be done. Mining high utility itemset from a transactional database refers to the discovery of itemset with high utility in terms like weight, unit profit or value in this paper we present related work of currently used algorithms for high utility itemset mining.

Key words: High Utility, HUI_Miner, Transactional Database, Faster High-Utility Itemset Mining (FHM)

I. INTRODUCTION

Data mining is Data mining is the process of revealing nontrivial, previously unknown and potentially useful patterns form large database Data mining can be used to transform the raw data into meaningful and useful information for business analysis processes referred to as Business intelligence. High utility mining refers to finding an itemset with "high profit" in transaction.

High utility item set mining is popular data mining task it consist of enumerating all high utility item set (HUIs) groups of items (itemset) having a high utility in customer transaction databases. HUIM is generalization of the problem of frequent itemset mining (FIM) where item can appear more than once in each transaction and where each items can appear more than once in each transaction and where each item has a weight (eg. Unit profit). High Utility itemset mining is widely viewed as more difficult than frequent item set mining because the utility measures used in HUIM is neither anti-monotonic nor monotonic high utility itemset may have supersets or subsets having lower, equal or higher utilities. HUIM has a wide range of application. However an important issue of traditional HUIM algorithm is that they tend to find itemset containing many iteams, as they are more likely to have a high utility. This is an issue because itemset containing many items are generally less useful than itemsets containing fewer itemsets containing fewer items [2].

Jerry Chun-Wei lin (Jerry chun. 2016) proposed a novel algorithm named FHM+ for mining HUIs, while considering length constraints to discover HUIs efficiently with length constraints FHM+ introduce the concept of Length Upper Bound Reduction (LUR) and two novel upper-bound on the utility of itemsets.an extensive experimental evaluation shows that length constraints are effective at reducing the number of patterns, and the novel upper-bound can greatly decrease the execution time, and memory usage for HUI mining. Moreover result shows that LUR concept greatly improves the algorithm efficiency. Thus prune the search space an extensive experimental evaluation shows that the proposed algorithm can be much faster than the state-of-the-art FHM algorithm. And greatly reduce the number of patterns presented to the user.[2] Previous system is used the Length Upper bound Reduction based algorithms Which results in a large time and memory consumption. Existing algorithm are based on matrix structure. Which use a matrix structure to find all utility itemset. But problem is that it will be generate large number of candidate key and are consume large memory overhead.

A. *Example:*

Example of a transaction database representing the sales data and the profit associated with the sale of each unit of the item

1) Input:

TID	TRANSACTION
T1	(a,1), (b,5),(c,1),(d,3),(e,1),(f,5)
T2	(b,4),(c,3),(d,3),(e,1)
T3	(a,1),(c,1),(d,1)
T4	(a,2),(c,6),(e,2),(g,5)
T5	(b,2),(c,2),(e,1),(g,2)

Table 1: Transaction Database

Item	A	B	C	D	E	F	G
Profit	5	2	1	2	3	1	1

Table 2: Utility Profit Database [2]

minutil: A minimum utility threshold set by the user (a positive integer)

TU (T1) = 5*1+2*5+1*1+2*3+3*1+1*5=30
TU (T2) =2*4+1*3+2*3+3*1=20
TU (T3) =5*1+1*1+2*1=8
TU (T4) = 5*2+1*6+3*2+1*5=27
TU (T5) = 2*2+1*2+3*1+1*2=11

Item	Total Utility
T1	30
T2	20
T3	8
T4	27
T5	11

Table 3: Total Utility [2]

2) Output:

All high-utility itemsets (itemsets having a utility$\geq minutil$)
For example, if minutil = 33$, the high-utility itemsets are:

$$u(\{b,d,e\}) = \underbrace{(5x2)+(3x2)+(3x1)}_{\text{utility in transaction t1}} + \underbrace{(4x2)+(2x3)+(1x3)}_{\text{utility in transaction t2}} = 36\$$$

{b,d,e} 36$ 2 transactions	{b,c,d} 34$ 2 transactions
{b,c,d,e} 40$ 2 transactions	{b,c,e} 37 $ 3 transactions

II. LITERATURE REVIEW

Jerry Chun-Wei lin (Jerry chun. 2016) proposed a novel algorithm named FHM+ for mining HUIs, while considering length constraints to discover HUIs efficiently with length constraints FHM+ introduce the concept of Length Upper Bound Reduction (LUR) and two novel upper-bound on the utility of itemsets.an extensive experimental evaluation shows that length constraints are effective at reducing the number of patterns, and the novel upper-bound can greatly decrease the execution time, and memory usage for HUI mining. Moreover result shows that LUR concept greatly improves the algorithm efficiency. Thus prune the search space an extensive experimental evaluation shows that the proposed algorithm can be much faster than the state-of-the-art FHM algorithm. And greatly reduce the number of patterns presented to the user.[2] Previous system is used the Length Upper bound Reduction based algorithms Which results in a large time and memory consumption. Existing algorithm are based on matrix structure. Which use a matrix structure to find all utility itemset. But problem is that it will be generate large number of candidate key and are consume large memory overhead.

Mengchi liu (2012) proposed an algorithm called HUI-Miner (High Utility Itemset Miner) for high utility itemset mining. To identify high utility itemsets, most existing algorithms first generate candidate itemset by overestimating their utilities state of the art algorithm on various databases, and result shows that HUI-Miner its time and memory consumption .[5]

Tseng et al. (Viscent 2014) proposed a novel strategy based on the analysis of item co-occurrences to reduce the number of join operations that need to be performed. An extensive experimental study with four real-life datasets shows that the resulting algorithm named FHM(Fast High Utility Miner) reduce the number of join operations by up to 95% and is up to six times faster than the state-of-art algorithm HUI-Miner. Frequent Itemset Mining (FIM) is popular data mining task that is essential to a wide range of application given a transaction database FIM consist of discovering frequent itemset as group of item appering frequently in transaction [1]

Frequent Itemset Mining (FIM) is popular data mining task that is essential to a wide range of application. Given a transaction database FIM consist of discovering frequent itemset as group of item appering frequently in transaction [1]. However an important limitation of FIM is that it assumes that each item cannot appear more than once in each transaction and that all item have the same importance (weight, unit profit or value).

These assumptions often do not hold in real applications the problem of HUIM is widely recognized as more difficult than the problem of FIM in FIM the downward-closure property states that the support of an itemset is anti-monotonic, that is the supersets of an infrequent itemset are infrequent and subset of a frequent itemset are frequent. The property is very powerful to prune the search space.

Souleymane Zidal (Zidal 2015) proposed a novel algorithm named EFIM (EFficient high-utility Itemset Mining), which introduces several new ideas to more efficiently discovers high-utility itemsets both in terms of execution time and memory. EFIM relies on two upper-bounds named sub-tree utility and local utility to more effectively prune the search space. It also introduces a novel array-based utility counting technique named Fast Utility Counting to calculate these upper-bounds in linear time and space. Moreover, to reduce the cost of database scans, EFIM proposes efficient database projection and transaction merging techniques. An extensive experimental study on various datasets shows that EFIM is in general two to three orders of magnitude faster and consumes up to eight times less memory than the state-of-art algorithms d2HUP, HUI-Miner, HUP-Miner, FHM and UP-Growth.[3]

Sunidhi Shrivastava (Sunidhi 2016) In this paper gives a brief idea about the work completed till the date of infrequent item sets and utility mining also proposes a latest advance for their preservation. In this paper, analysis on the high utility infrequent itemsets using Utility Pattern Rare Itemset (UPRI) algorithm has been done. message. Infrequent itemsets finds the hidden an association among the data items. The rare consolidation of the itemsets can be interesting and more profitable. [4]

Hong Gao (Gao 2017) proposed a an algorithm a novel tree structure IHUI-Tree and an efficient algorithm IHUI-Miner for incremental and interactive HUIM. Different from the algorithm based on IHUP-Tree, IHUI-Miner does not generate any candidate. Extensive performance analysis show our proposed tree structure is efficient, and our algorithm is at least one order of magnitude faster than the state-of-the-art algorithm in interactive. [6]

Supachai Laoviboon (Laoviboon. 2017) proposed to discover itemsets giving high utility (such as high profit, low cost/risk and other factors) this can help to extract hidden-knowledge from buying behavior of customer. However, HUIM may not sufficiently give hidden-Knowledge and observe occurrence behavior of itemsets in some applications, since it only considers utilities of items/itemsets. Thus we propose to mine high utility itemsets with irregular occurrence (also called High Utility Irregular Itemsets, HUIIs). [14]

III. COMPARISON BETWEEN ALGORITHMS

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	Name of Algorithm	Limitation
1	Mining High Utility Itemsets Without Candidate Generation	2012	Mengchi Liu, Jufeng Qu	HUI_Miner	Costly Join Operations on each
2	Efficient Algorithm for Mining High Utility Itemset From Transactional Database	2013	Vincent Tseng, Bai- Shie, Cheng wu,	UP_Growth UP_Growth+	Complex for evaluating due to tree structure
3	FHM: Faster High Utility Itemset	2014	Philipe Viger,	FHM	Large Memory

	Mining using Estimated Utility Co-occurrence Pruning		Cheng Wu		Overhead
4	EFIM: A Highly Efficient Algorithm For High-Utility Itemset Mining	2015	Souleymane Zida, PhilippeFournier Viger	Efficient high-utility Itemset Mining(EFIM)	Occupies large memory overhead
5	FHM+: Faster High-Utility Itemset Mining using Length Upper-Bound Reduction	2016	Philippe Fournier-Viger, Jerry Chun-Wei Lin,	Length Upper Bound Reduction (LUR)	It required two database scan
6	Discovering High Utility Itemset Using Map Reduce	2016	Wei Song, Jaipei Xu	HUIMR	Huge Set of map reduce Low Threshold for long Transaction
7	Improving method for graphical analysis and representation of high utility itemsets using UP++Growth	2016	Payal swamy, Amit pimpalkar	Closed high utility itemset discovery algorithm (CHUD)	Dealing with long complexity
8	An efficient algorithm for incremental and interactive high utility itemset mining	2017	Shiming guo, Hong gao	IHUI-Tree	Its database are no less than user-specified threshold
9	Mining high-utility itemset with irregular occurrence	2017	Supachai laoviboon	HUIM	Large Memory Overhead

IV. CONCLUSION

In Data Mining, Association Rule Mining is one of the most important tasks. A large number of efficient algorithms are available for association rule mining, which considers mining of frequent itemsets. But an emerging topic in Data Mining is Utility Mining, which incorporates utility considerations during itemset mining. In this paper we detailed study about the different High utility mining algorithm, their work flow and their limitations. This paper provides an overview a comparative study of various algorithms that are used to improvise the efficiency of mining high utility itemsets. In the future scope, we will be proposing algorithms for mining high utility itemset which reduces memory overhead and execution time parameters.

REFERENCES

[1] Philippe Viger, Cheng Wu, Souleymane Zida, Vincent S. Tseng- "FHM: Faster High Utiltiy Mining Itemset mining Using Estimated Utility Co-occurrence purning." , Springer International, Switzerland 2014,ISMIS 2014.
[2] Philippe Fournier-Viger, Jerry Chun-Wei Lin,Quang-Huy Duong, Thu-Lan Dam, "FHM+: Faster High-Utility Itemset Mining using Length Upper-Bound Reduction.", Springer International Publishing Switzerland 2016.
[3] Souleymane Zida, Philippe Fournier-Viger(B), Jerry Chun-Wei Lin, Cheng-Wei Wu, and Vincent S. Tseng- "EFIM: A Highly Efficient Algorithm for High-Utility Itemset Mining.", Springer International Publishing Switzerland 2015.
[4] Sunidhi Shrivastava, Punit Kumar Johari "Analysis on High Utility Infrequent ItemSets Mining Over Transactional Database.", IEEE International Conference On Recent Trends In Electronics Information Communication Technology, May, 2016.
[5] Menghchi Liu, Junfeng Qu-"Mining High Utility Itemsets without Candidate Generation", CIKM, Maui. USA, Nov 2012.
[6] Shiming Guo, Hong Gao- "An Efficient Algorithm For Incremental and Interactive High Utility Itemset Mining" ,International Conference on Image, 2017.
[7] Jerry Chun-Wei Lin, Wensheng Gan, Philippe Fournier-Viger, Tzung-pei Hong "Mining High-Utility Itemsets with Multiple Minimum Utility Thresholds", C3S2E 2015.
[8] Thang Mai, Bay Vo, Loan T.T. Nguyen "A Lattice-based approach for mining high utility association rules", Information science Elsevier, 2017.
[9] Vincent S. Tseng, Bai-En Shie, Cheng Wu, philip S. Yu- "Efficient Algorithms For Mining High Utility Itemset from Transactional Databases.", IEEE transactions on knowledge and data engineering (Vol 25, no. 8), Aug 2013, ISSN No: 1041-4347, DOI: 10.1109/TKED.2012.
[10] Jerry Chun-Lin, Member, IEEE, Shifeng Ren, Philippe Fournier-Viger Tzung-Pei Hong,"EHAUPM: Efficient High Average-Utility Pattern Mining with Tighter Upper-Bounds"IEEE 2016
[11] Junqiang Liu, Member, IEEE, Ke Wang, Senior Member, IEEE, and Benjamin C.M. Fung, Senior Member, IEEE" Mining High Utility Patterns in one phase without generating candidates" IEEE 2015.
[12] Philippe Fournier-Viger, Jerry Chun-Wei Lin, Ted Gueniche, Prashant Barhate, "Efficient Incremental High Utility Itemset Mining" ACM 2015.
[13] Guo-Cheng Lan , Tzung-Pei Hong , Vincent S. Tseng "An efficient projection-based indexing approach for Mining High Utility Itemsets" Springer 2013.
[14] Supachai Laoviboon, Komate Amphawan, "Mining High-Utility Itemsets with Irregular Occurrence" IEEE 2017.
[15] Wei Song, jaipei Xu "Discovering High Utility Itemset Using Map Reduce" IEEE 2016.

[16] P.Payal Swamy, Amit Pimpalkar "Improving method for graphical Analysis and representation of high utility itemsets using UP++ Growth. IEEE 2016.

