Proceedings of the International Conference on Intelligent Computing and Communication for Smart World (I2C2SW-2018)
14-15 December 2018

Contemplate Study of Contemporary Techniques for HUIM

K. Logeswaran
Department of IT
*Kongu Engineering College*Perundurai. India
klogesbtech@gmail.com

P. Suresh
Department of IT
*Kongu Engineering College*Perundurai. India
psuresh.it@kongu.edu

S. Savitha
Department of CSE
*K.S.R. College of Engineering*Tiruchengode. India
infosavi@gmail.com

A. Rajiv Kannan
Department of CSE
*K.S.R. College of Engineering*Tiruchengode. India
rajiv5757@yahoo.co.in

***Abstract*—Inthis internet world, most of the transactions are done in online especially in ecommerce field which creates enormous size of transactional database that grows dynamically at each second. Really it is tough job to analyses the growing transactional database for the betterment of ecommerce business growth. Determining the Interesting pattern form incremental transactional database is evolving research area in the field of Data mining. Frequent ItemsetMining (FIM) approach is followed in earlier days to find the frequent pattern from transactional database. FIM has significant drawback of omitting the interesting factor about each items, such as quantity, price, profit and etc. This drawback is addressed in High Utility ItemsetMining (HUIM) which considers interesting factors of each item. This paper focuses on reviewing the existing state of art algorithms to create a path for the future research in the area of high utility itemset mining.**

Keywords—HUIM, Utility, HUI, FIM

I. INTRODUCTION

Mining task is applied in huge amount of data for finding understandable knowledge in that data. Many methodologies are available to find the meaningful data based on the various business needs. However all the methodologies will have some drawbacks that has to compromised in order to achieve the result. Over the recent years Frequent ItemsetMining(FIM) is used to for business analytics to make profitable business. FIM uses few approaches to find the frequent itemset from the given transactional database. This FIM will identify the itemset which has number of occurrence greater than predefined threshold value in each transaction in the given database, without considering any other relevant information about the each transaction.

However this may create some serious issues in business perspective. The main limitation of this method is that, the quantity of each item is not considered. Next limitation is that all items are assumed to have same weight (cost). Items having high cost may have higher profit even if it is having minimum frequency. Also recency of particular transaction is also not considered. So it is viewed that frequent pattern mining may generate frequent itemset which may not have actual interesting factors.

As a remedial measures to the above said problem alternate approach is proposed by considering the necessary interesting factor about each item in a transaction. This approach of mining interesting itemset from transaction database is called High UtilityItemset (HUI) mining. In HUIM utility are classified into Internal Utility (IU) and External Utility (EU). IU refers to the quantity or count of an item in each transaction. EU refers to the profit generated by each item in transaction which is specified as unit profit.Now the user needs to estimate the minimal utility brink to find the HUIs. HUIM can be viewed as an extension of FIM. Further part of this paper is organized as fundamental concepts of FIM and HUIM, Elemental study of HUI and review of recent techniques used for HUIM.

II. ELEMENTAL STUDY OF HUI

As already discussed, FIM has some critical issues which inhibits the real time application of FIM. In customer buying transaction, the purchase quantity of a product is not considered by FIM. Thus, if a customer has bought four soap, 7 soap or 18 soap, it is consider as items with similar importance. The next obstruction of FIM is that all items are treated as itemset with similar importance that is utility. For example, consider the retail grocery shop transaction where a customer buys a bottle of nutrition supplements to his child which is very costlier or a very low price tagged towel for rough use. Both of the two items in above example are consider as same with similar importance of utility. It is obvious that FIM will the set of frequent patters that are not actually perceived as interesting. Assume that if FIM predicts that a transaction with item {battery, band} is frequent pattern, but from business angle these itemset not worth because it will not earn reasonable profit for the retail shop vendor.Also,FIM may omit the sparse pattern that generates a high profit such as perhaps {health drink, cake}.

As an alternate approach we can use HUIM which prevailsthe drawback of FIM. In HUIM utility of an item is considered while mining the frequent pattern.

A. High Utility ItemsetMining (HUIM)

In recent years many e-commerce business analytics requires confident decision making process based on their past transactional database. In support with this, FIM fails to meet the need, so an alternated strategy called HUIM is adopted in recent years. HUIM includes the utility factor of each transaction such as importance (cost), quantity, recency and ect,. Many algorithms were proposed to perform efficient HUIM. In this HUIM approach, predetermined brink value of the utility of itemset is set. Now if the utility of an itemset is larger than the predetermined brink value, then it is considered as and superior itemset.Also in HUIM, the evaluated itemset contains only the real number value instead of binary value. This is called as utility value which may lead to generation of itemset with greater value of information from the given transaction database when assimilate with FIM.

At beginning an algorithm called Two phase [1] technique is suggested. It uses the Transaction Brudern Utility (TBU). Property called anti monotonicity is used for

pruning the search space based on support measure. The decision of extending the itemset to HUIs is regulated by Two phase algorithm.However TWU approach has the defect of scanning the database for multiple times which may incurs high time cost. An HUIM algorithms based on tree structure called IHUP [2], UP-Growth [3], MI-Growth [4], and HUDPI [5] were devised to generate HUIM with more useful information by iron out the defect in TWU.

These approaches generate huge volume of candidate itemset which in turn requires more space in main memory and enormous execution time which will legitimize. In contemporary periods, several algorithms which based on list data structure are proposed. This includes algorithm such as HUI-Miner [6], FHM [7] and HPU-Miner [8] were introduced to mine the HUIs by eradicating the candidate itemsetgeneration, also it eliminates the unnecessary validation of candidate itemset.

B. Average High-Utility ItemsetMining(AHUIM)

Besides providing interesting high utility itmeset, HUIM generates high utility itemset by summing up of utility of all the item in aitemset, which may evaluates to high value utility even though each item in an itemset is having low utility value. This one of notable drawback of HUIM and this is due to not considering the length of each itemset. Average HighUtilityItemsetMining (AHUIM) to quickly fix this issue was introduced. In HAUIM, utility of item is calculated by taking fraction of summation of individual item utility and total length of the itemset.

III. REVIEW ON STATE OF ART RECENT TECHNIQUES FOR HUIM

A. Efficient High Average-Utility Pattern Mining with Tighter Upper Bounds

1) Complication Identified from previous studies:

Nevertheless, HAUI-Miner[13] supports auub model which overrate the utility of itemsets. Hence it performs join operations frequently for mining High average utility itemsets. The auub model performs liberated upper bound on the utilities of itemsets. Auub model failed to trim unpromising itemsets with low relative average utilities from the search space.

Over recent years, numerous algorithms have been deployed to mine high average-utility itemsets (HAUIs). Those algorithms occupied excessive memory and highly time consuming, because they make use of average-utility upper-bound (auub) model to overrate the average utilities of itemsets.

2) Recommended Solution:

HAUIM's performance can be improved by the proposed model called as tighter upper-bound model which can be used instead of auub model to mine HAUIs. The remaining maximum utility itemsets are measured in transactions to decrease upper bound on the utilities of itemsets [14]. The unrelated items in transactions are ignored to tighten the upper bound. Four pruning strategies are applied to decrease the search space to mine High Average Utility Itemsets.

To mine HAUIs we propose a novel algorithm called Improved Proficient algorithm for High Average Utility Pattern Mining (IPHAUPM). A modified middling utility (MMU) catalog is maintained to sketch mining information. Unpromising itemsets are to be trim early to reduce the search space. By utilizing two tighter upper-bounds it can be accomplished. Pruning decreases the complication of the final classifier, and sharpens predictive accuracy.

Search space can be reduced by applying two least upper-bounds for mining high average utility itemsets and Loose Upper bound model has been implemented to judge average utilities of itemsets. Also modified least upper-bound model (MUB) has been implemented to reject irrelevant itemsets in the transactions and also to decrease search space. The modified compact upper bound of an itemset X is denoted as mcub(X), which can be defined as:

$$mcub(X) = \sum_{X \subseteq T_q^{'} \in D} rthu(X, T_q^{'}) \qquad (1)$$

where RHU(Revisited Transaction-highest Utility) of an itemset X in transaction T_q^' is denoted as rhu(X,T_q^'), and defined as:

$$rhu\left(X, T_q^{'}\right) = max\left\{u\left(i_1, T_q\right), u\left(i_2, T_q\right), \ldots, u\left(i_{|T_q|}, T_q\right)\right\},$$

$$\omega\eta\varepsilon\rho\varepsilon\ X \prec i_1 \prec i_2 \prec, \ldots, i_{|T_q|}, T_q^{'} \subseteq T_q, \alpha\nu\delta\ Tq \setminus X = T_q^{'} \qquad (2)$$

A modified middling utility (MMU) catalog is implemented in order to decrease numerous database scans and to maintain needed information to be utilized during the mining high of average utility itemsets. Pruning technique such as alpha-beta pruning along with null move heuristic can be applied to speed up the efficiency of high average utility itemsets mining.

By executing empirical analysis, it has been found that the proposed algorithm do better on the real world data as well as the synthetic datasets in terms of speed, reduced size occupation of memory, reduced count of join operations and Elasticity.

B. Efficient Algorithms for Mining High Utility Itemsets from Transactional Databases

1) Complication Identified from previous studies:

Past approaches applied atop estimatedmethodology to facilitate the performanceof utility mining. The candidate itemset with potentially high utility called potentially high utility itemset (PHUI) [15] are identified first and then utility information of those PHUI are obtained during the supplementary scan of the same database. These approaches will generate huge number of PHUIs as consequences of

atrocious situation when transactional database contains more long transactions or low threshold values are set.

2) Recommended Solution:

Dual pattern growth algorithm based on utility such as utility pattern growth (UP-Growth)and UP-Growth+ is proposed [16]. Also,to preserve the crucial information about utility of pattern in a database, a solid tree structure called Utility Patter tree (UP-Tree) is used.UP-Tree enables the mining of high utility itemset efficiently using only double scans of given initial database.

As an initial step, UP-Tree is constructed by applying two strategies called Discarding Global Unpromising Itemset (DGU) strategy during Construction of Global UP-Tree and Decreasing Global Node Utilities (DGN) strategy during Construction of Global UP-Tree.

In succession to the design of UP-Tree, proposed UP-Growth algorithm is used to mine the UP-Tree by applying two strategies such as Strategy LUA: Local Unpromising item Avoiding during Constructing a Localized UP-Tree and Strategy LUND: Localized Utility Nodeduring Constructing a Local UP-Tree. Above said two strategies will reduce the overestimated utilities of itemset which in turn will reduce the number of PHUIs.

Further enhanced mining algorithm called UP-Growth+ is used to effectively reducing over estimated utilities. Minimum item utility table is used in UP-Growth algorithm, alternatively minimal node utilities in each path is used in UP-Growth+ algorithm. Using the minimal node utilities, the estimated pruning values are evaluated to near value of real utility value of pruned items in database.

Performance of proposed algorithm is compared with state-of-the-art utility mining algorithms by using series of experiments done on different type of real and synthetic data sets especially containing lots of long transactions or low minimum utility thresholds. Empirical consequences shows that UP-Growth and UP-Growth+ surpass the other algorithm considerably in terms of execution time.

C. An efficient algorithm for mining top-k on-shelf high utility itemsets

1) Complication Identified from previous studies:

In several approaches, HUIM algorithm used to find the itemset with high utility in an efficient way by considering the interesting factor of each item in a transaction such as profit and quantity. Nonetheless of them have not considered the additional feature of market basket analysis such as profit in terms of negative value which indicates the loss occurred during sale of particular item in a transaction and the on shelf period of an item in a transaction. In real-time, a retail store may have sold the item at loss during offer time in order to increase the familiarity of the store and also some item in a store may not be sold for long time which increases the on shelf period of that item.

Initially, 3-Scan Algorithm for Mining In-shelf High Utility Itemsets with -veprofit(TS-HOUN) [17] algorithm was proposed to mine high utility itemset by considering high on-shelf utility itemset (HOU) with negative/positive unit profit. TS-HOUN generates and maintains huge number of candidate itemset in memory and also performs multiple database scan by using three-phase approach. To overcome the drawbacks of TS-HOUN, algorithm called Fast on-shelf high utility itemsetminer (FOSHU) [18] is used. FOSHU uses single phase to determineitemsetand eliminates merging of patterns found in each time period. Even though FOSHU outperforms TS-HOUN,it consumes more time. However the problem of identifying the exact value for the parameter of minimum threshold is perceived as crucial limitation of conventional method of HOU mining, HUI mining and FIM algorithm.

2) Recommended Solution

To rout these challenges, a novel utility list based algorithm was proposed by (KOSHU) [19] called top-K On-Shelf high Utility itemset miner. This miner algorithm considers either with negative unit profit or without negative unit profits. This methodamalgamate triple offbeat strategies of pruning itemset which are titled as Evaluated Maximal Period Rate Snip (EMPRS), Period Utility Snip (PUS) and Unanimity Extant of pair 2-itemset snipping (UE2S).

In KOSHU, instead of mining HUIs based on relative minimumthreshold which is set by user, threshold raising strategy is followed to increase the minutil threshold variable from its initial value zero. The above proposed approach utilize the different scenario of assigning beginning value and aggressively modifying the internal minimal proportionate threshold utility of k-top utility mining work. This is done by RRIU and RRI2U. Also, it introduces the process of constructing the list of utility during the process of mining which has very less complexity. The index value corresponding to the final search position is represented using the rapid binary search method. This approach facilitates the upcoming search to instantiate from this position. Through this mechanism a significant performance gain is achieved.

Different real-time and synthetic datasets have be considered to assess the above proposed methodology. The performance measure including execution time and memory consumptions has significantly improved and the same is observed form the empirical results of comprehensive evaluation.

D. Evolutionary Algorithms inspired by Bio science for Mining High Utility Itemsets: A Optimal Diverse Value Framework

1) Complication Identified from previous studies:

The problem of HUIM in incremental database deteriorate the performance of various contemporary approaches based on pattern growth method which includes the generating candidate itemset in level-wise manner [20], top k HUIs identification and average high utility itemset.Also due to increase in distinctness of item in itemset performance bottleneck may be traced in the empirical results.

Inorder to demolish the enactment hindrance available in those methodologies Evolutionary algorithm is used. It is a

metaheuristicapproach admired by the natural science process. It is also called genetically inspired algorithm is suggested to find the HUIM. The available approaches of evolutionary algorithm such as Genetic Algorithm (GA), bio- admired algorithm and swarm optimization(PSO) are used to isolate the high utility itmeset. All the above said approaches for computing the HUIM follows the ancestral procedure of using the optimal value of on population as target value for next population.

Those ancestral approaches are not suitable for HUIM. In HUIM all itemset with utilities not less than threshold must be discovered which is considered as relatively best value. Due to uneven distribution of HUIs, the target value for next population is calculated from the optimal value of current instance population. This may lead to missing of results.

2) Recommended Solution:

To riddle this dilemmaaoffbeat bio-admired-algorithm-based HUIM framework (Bio-HUIF) is used to devise the HUIs.Poker wheel selection approach is used in Bio-HUIF to devise the HUIs of current population which determine next population initial target. The modified version of GA, PSO, and Bat Algorithm (BA) is called as Bio-HUIF-GA, Bio-HUIF-PSO, and Bio-HUIF-BArespectively which employees Bio-HUIF framework.

Based on the proportional to the utility of the itemset and total utility of all identified HUIs, the introductory target of next population is selected for every algorithm.The suggested framework uses a bitmap representation of given original database which is an effective representation method for HUI mining.

Assume that a transaction database contains list of transaction such as $D = \{T_1, T_2, \dots, T_N\}$. Each transaction consist of finite set of itemset such as $I = \{i_1, i_2, \dots, i_M\}$. D is represented using bitmap approach which contains M X Nmatrix B(D). B(D)is a Boolean matrix which has domain value {0,1}. The notation (j, r) represent the each entry of B(D) conform to the individual transaction T_i $(1 \leq j \leq N)$ and item $i_j (1 \leq r \leq M)$. Also that particular item is location is identified from i^{th}column and j^{th} column of BD. The notation (i, j)gives the value which can be defined as

$$B_{j,k} = f(x) = \begin{cases} 1, & if\ i_k \in T_k \\ 0, & Otherwise \end{cases} \qquad (3)$$

From the equation (3) it is observed that the notation (j, r) of BD will be 1 iff item i_j is added in the transaction $T_k$which is assigned to $0.\mathrm{Bit}(Y) = \mathrm{bitwisemap} - \mathrm{AND}_{i \in X}(\mathrm{Bit}(j))$represent the itemset Y of an bitmap cover. $\mathrm{Bit}(j_k)$represent the k^{th}vector column which will behave as cover bitmap of an item j_k.

IV. CONCLUSION

In this revolutionary world it is important to understand the customer requirement and their interest inorder to improve the ecommerce business. From past many methodologies are adapted in ecommerce business out of which HUIM plays major rule to improve the business profit. In this paper we studied four approaches to improve the performance of HUIM. Each approaches mentioned above has proposed solution to address the bottleneck performance issues in HUIM. However as the importance of the HUIM increases as time goes, it is necessary to fine-grain on demand impediment that may occur during real time analogy. It is perceived form above study that an alternate methodology to traditional approach is required to assess the importance of HUIM. In future the natural inspired algorithm can be used along with machine learning to address evolving need of ecommerce business needs.

REFERENCES

[1] Liu Y, Liao W,Choudhary, "A A two-phase algorithm for fast discovery of high utility itemsets. In: Advances in knowledge discovery and data mining," pp. 689–695, 2005.

[2] Ahmed CF, Tanbeer SK, Jeong B, Lee Y, "Efficient tree structures for high utility pattern mining in incremental databases,"IEEE Trans Knowl Data Eng 21(12):1708–1721, 2009.

[3] Tseng V, ShieBE,Wu CW, Yu PS, "Efficient algorithms for mining high utility itemsets from transactional databases," IEEE Trans Knowl Data Eng 25(8):1772–1786, 2013.

[4] Yun U, Ryang H, "Incremental high utility pattern miningwith static and dynamic databases," ApplIntell 42(2):323–352, 2015.

[5] Yun U, Ryang H, Ryu K,"High utility itemset mining with techniques for reducing overestimated utilities and pruningcandidates,"Expert SystAppl 41(8):3861–3878, 2014.

[6] Liu M, Qu J, "Mining high utility itemsets without candidate generation," In: Proceedings of the 21st ACM internationalconference on Information and knowledge management, pp 55–64, 2012.

[7] Fournier-Viger P, Wu C, Zida S, Tseng V, "FHM: faster high-utility itemset mining using estimated utility co-occurrencepruning. In: ISMIS, pp 83–92, 2014.

[8] Krishnamoorthy S, "Pruning strategies for mining highutility itemsets," Expert SystAppl 42(5):2371–2381, 2015.

[9] Hong T, Lee C, Wang S, "Effective utility mining with themeasure of average utility," Expert SystAppl 38(7):8259–8265, 2011.

[10] Lan G, Hong T, Tseng V,"A projection-based approach for discovering high average-utility itemsets," J InfSciEng 28:193– 209, 2012.

[11] Lan G, Hong T, Tseng V,"Efficiently mining high average utility itemsets with an improved upper-bound strategy," Int J InfTechnolDecis Making 11(5):1009–1030, 2012.

[12] Lu T, Vo B, Nguyen HT, Hong T, "A new method for mining high average utility itemsets,"In: Computer Information Systemsand Industrial Management, pp 33–42, 2014.

[13] J. C.-W. Lin, T. Li, P. Fournier-Viger, T. P. Hong, J. Zhan, and M. Voznak, "An efficient algorithm to mine high average-utility itemsets," Adv. Eng.Inform., vol. 30, no. 2, pp. 233-243, 2016.

[14] Jerry Chun-Wei Lin, ShifengRen, Philippe Fournier-Viger, and Tzung-Pei Hong, "EHAUPM: Efficient High Average-Utility PatternMining With Tighter Upper Bounds," IEEE Access, vol.5, no.8 VOLUME 5, pp.12927 – 12940, 2017.

[15] V.S. Tseng, C.-W. Wu, B.-E. Shie, and P.S. Yu, "UP-Growth: An Efficient Algorithm for High Utility Itemsets Mining," Proc. 16th ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD'10), pp. 253-262, 2010.

[16] Vincent S. Tseng, Bai-En Shie, Cheng-Wei Wu, and Philip S. Yu, "Efficient Algorithms for Mining High UtilityItemsets from Transactional Databases," IEEE Transactions on Knowledge and Data Engineering, vol.25,no.8, pp.1772 – 1786, 2013.

[17] Lan GC, Hong TP, Huang JP, Tseng VS, "On-shelf utility mining with negative item values," ExpertSystAppl 41(7):3450–3459, 2014.

[18] Fournier-Viger P, Zida S, "FOSHU: faster on-shelf high utility itemset mining—with or without negative unit profit. In: Proceedings of the

30th annual ACM symposium on applied computing," SAC'15. ACM, New York, pp 857–864, 2015.

[19] Thu-Lan Dam1, Kenli Li1, Philippe Fournier-Viger, Quang-Huy Duong, "An efficient algorithm for mining top-k on-shelf highutility itemsets," Springer - KnowlInfSyst, vol.52, no.3, pp. 621–655, 2017.

[20] Y. C. Li, J. S. Yeh, and C. C. Chang, "Isolated items discarding strategy for discovering high utility itemsets," Data Knowl. Eng., vol.64, no. 1, pp. 198-217, Jan. 2008.