JCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-6, Special Issue-3, April 2018 E-ISSN: 2347-2693

A Review On High Utility Itemset Mining

D. Divyashree[1*], G.Sunitha[2]

[1*]Department of Computer Science and Engineering, Sree Vidyanikethan Engineering College, JNTUA, Tirupathi, India
[2] Department of Computer Science and Engineering, Sree Vidyanikethan Engineering College, JNTUA, Tirupathi, India

[*]Corresponding Author: divyashreedasi@gmail.com, Tel.: +918897220518

Available online at: www.ijcseonline.org

Abstract- Sequential pattern mining is the imperative data mining problem with expansive application from text analysis to market basket analysis. It is the way towards extricating certain sequential patterns whose support surpasses a predefined limit which is defined by the user according to their interest. With frequent pattern mining, pattern is viewed as fascinating if its event surpasses users determined limit. Notwithstanding, users interest may identify with numerous components that are not really communicated as far as the event recurrence. Since the quantity of sequences can be huge, and users have distinct interest and prerequisites, to get the most fascinating sequential pattern, generally a minimum base support is predefined by clients. Utility mining is a new advancement of data mining innovation. It developed as of late to address the confinement of frequent pattern mining by thinking about the client's desire or objective and in addition the crude information. An efficient algorithm is to be developed for extracting high utility sequential patterns.

***Keywords*—** Data mining, Frequent Pattern Mining, High Utility Itemset mining, sequential pattern mining.

I. INTRODUCTION

Data mining acquire an extensive variety of utilizations; the main task of data mining is extraction of hidden prescient data from extensive databases. Data mining is pre-owned in many industries to convert raw data into useful information. This information is useful for take decision and to understand the data. High utility pattern mining is a new trend emerged in data mining to overcome the problems which are present in frequent pattern mining. Why because frequent pattern mining assumes a vital part in data mining, and a frequent pattern mining is designed to mine the frequent pattern in datasets. Whereas high utility pattern mining is the expansion of the issue of frequent pattern mining. The main task of high utility itemset miningis to mine pattern which produce high utility. The parameters for the utility may be profit, popularity, quantity and cost. Here parameter for utility is considered as profit. Then high utility pattern mining is designed to mine the patterns which having the high utility that is high profit. That profit is more prominent than or parallel to predefined minimum threshold limit. That the high utility pattern is always greater than the users specified minimum threshold.

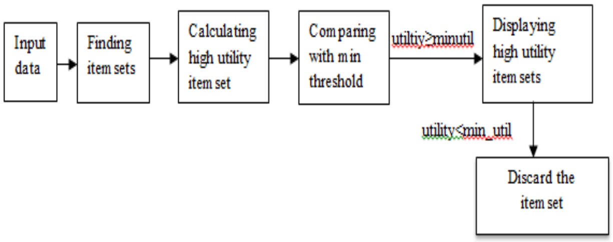
Input data
Finding item sets
Calculating high utility item set
Comparing with min threshold
utiltiy≥minutil
Displaying high utility item sets
utility<min_util
Discard the item set

Figure 1: Process for mining high utility itemsets

Figure 1 explains how high utility itemset is discovered from the datasets [5].

Step 1:Any bench mark datasets (e.g: foodmark, web-view) are going to be considered and fix some value to the min_utility threshold.

Step 2: Analyze the datasets and find the itemsets from the collected dataset and proceed to next step.

Step 3: Calculating utility for each and every individual itemsets which are mined in the step 2.

Step 4: Comparing the utility of each and every itemsets which are calculated in before step with the predefined min_utility threshold.

Step5: (a) If the utility of the itemset is bigger or same to the predefined minimum utility limit by the user than that itemset are going to be displayed as high utility patterns.

(b)If the utility of itemset is less than the user defined minimum threshold limit than simply discard the itemset why because it is not a high utility pattern.

A. Features of high utility pattern mining:

- Here in high utility itemset mining purchase quantity and profit are considered to prioritize items.
  For example whenever the customer wants to buy same item with more number of quantities (may be one, two or more in quantities of the same item for example bread) whereas this type of transaction can be considered in frequent pattern mining has the same important. If suppose the same transaction can consider in high utility pattern mining having different priorities and thus each transaction can be consider as different and also it going to find the utility of each pattern, and then it arrange the pattern list in descending order based on profit. So, by this the manager can increase the sales and take the decision for business enhancement.

  One more issue in frequent pattern mining is where the person buy costliest item or cheap product those products can be consider as same priority in frequent pattern mining where as in the high utility pattern mining which is consider having the different levels of priority that is based on the profit. How much profit it going to produce whenever the product is purchased by customer.

  High utility pattern consists of only items that provide high profit. Such patterns are useful further to enhance business decisions, inventory management, marketing strategies.
- Further applications include in genome analysis, condition monitoring, cross marketing.

II. LITERATURE SURVEY

The authors in [1] introduced a novel technique for mining pattern in a database. In this they adopt depth first search tree strategy and they used bucketing concept to mine the large pattern from a database.

Authors in [2] proposed Apriori algorithm which is used to mine the frequent pattern from the database. Because in the previous association rule mining we have the problem that to generate all association rule whose support and certainty is more prominent than the client indicated limit individually. Where proposed algorithm works in two phase approach, in the first pass it checks the event of every item in the database and then next it going to choose the large sequence from the candidate and then it going to scan the database one more time and counts the support of the candidates which are chosen in the previous step. In the second pass it going to generate the association rule between the frequent itemset which are mined in the previous pass, and also in this they used hash table to maintain the itemset list.

Authors in [4] proposed weighted association rule. By using weighted association rule first we mine frequent itemset and weighted association rule for each mined frequent itemset. Here they used two-fold approach in the algorithm that is in the first step it produces frequent itemset, here they overlook the weight related with the item in the transaction. In the second step the weighted association rule finds the support and confidence for each frequent itemset which is generated in the previous step. The weighted association rule doesn't support the downward closure property during mining. The weighted association rule doesn't improve the performance, the performance stays as before.

Authors in [3], these authors proposed algorithm for high utility pattern mining which is two stage calculation algorithms. The utility mining is to identify the itemset which produce the high profit compare to other itemset. Mining high utility pattern is depending upon user specified minimum threshold. This algorithm proficiently reduces the counts of candidate generation and mine total set of high utility itemset. In the first stage it find the transaction weighted utility that is only high weighted itemset are added to candidate set during level wise search. In the second phase it again scans the database to filter overestimated itemset.

Authors in [6] suggested one new algorithm for high utility pattern mining to mine high utility pattern in a single phase whereas the previous all algorithm uses two phase algorithm and that to generate huge number candidates when compared to previous it does not produce the candidates. In this they used CAUL linear data structure this data structure used to store the actual information about their utilities that calculating each item utility and stores this information into the utility list that means in a CAUL data structure. In the previous algorithms they store estimated information means approximate information about the utilities this information is not accurate from this we can face problem during the mining. Problem may be mining uninteresting pattern this problem can be overcome using CAUL linear data structure. This is the one the major advantage using CAUL linear data structure.

Authors in [6]in this paper used a new list named as utility list which is used to maintain utility list, and this list is utilized to store the detail data about an item and furthermore the data about the heuristic for high utility miner search space.

Authors in [9] these authors have undertaken the above mentioned issue and start work on that issue. These authors proposed one new algorithm based on item analysis so that they can easily reduce the join operation. Once the join operation is reduced means we are overcome the above issue. Join operations and cost are directly proportional. If the join operation reduced means automatically cost for the implementation also reduced.

Authors in [10] introduced one property that is ex-antiproperty. The ex-ante property states that if any transaction that doesn't obey the monotone property means that transaction is removed from the database, if accept means that property is going to integrate with apriori based algorithms.

Authors in [11] and also the authors in [12] Upper bounds is the another way used for pruning when the itemset share framework and doesn't obey the anti-monotone property. For this the above authors proposed one technique for pruning based on upper bounds. This paper employs the standard technique, when the constraint is neither monotone, anti-monotone, nor convertible.

Authors in [13] proposed FP growth it is well-known depth first algorithm. The main advantage using FP tree is, it going to compress the database into FP tree and stored it into the main memory.

Authors in [14] proposed pseudo projection algorithm this is totally distinct from the existing algorithms. This proposed algorithm utilizes mainly two types of structures one is array and the other one is tree based. These structures are used to represent the pseudo projected transactions subsets. By using this we can make use of CPU efficiently so that we can save the memory. In this depth first search tree is used to arrange the frequent pattern in a tree and to build upper bound they used breadth first search tree. This algorithm is more efficient than other previous existing algorithms when working with sparse and dense data.

Authors in [16] these authors proposed an algorithm for utility mining that is IIDS which stands for isolated items discarding strategy. How this algorithm works mainly it going to discover the itemset with less number of candidates, So that we can improve the performance of pattern mining strategy. The algorithm maintains one array for each candidate during each pass. Here they adopt direct condition generation strategy which is a level wise approach. This algorithm improves the efficiency of pattern mining. The main advantage using this algorithm is it provide less arithmetic complexity, provide accurate result and efficient runtime when compared to other algorithms which are developed for high utility mining before. The disadvantage using this it stills scans the database many when the database is huge.

Authors in [17] in this paper proposed one algorithm that is IHUPM stands for incremental high utility pattern mining, the proposed algorithm is tree based. In this they constructed one tree namely IHUP-tree, the main use of this tree is utilized to maintain the information about the each itemset and utilities. The proposed algorithm maintains a three tree structure to utility mining efficiency. This tree based algorithm comparatively reduces the calculation when the limit is changed or at whatever point the database is refreshed. These three trees are mainly used to arrange the itemset by considering different parameters. The first tree structure us IHUPL-tree abbreviated as incremental as incremental high utility pattern lexicographic tree. Lexicographic is nothing one order based on this order the itemset are going to arranged in a tree. The second tree structure that is IHUPTF-tree which stands for incremental high utility pattern transaction frequency tree [17]. This is second level tree it is very easy and non-complex to maintain. Whereas in this proposed tree, itemset are set based on the transaction frequency. The main advantage using this tree is whenever the database is updated we does not need to reconstruct the tree again. The third tree IHUPTWU-tree stands for incremental high utility pattern transaction weighted utilization tree [17]. The name itself shows that in this tree itemset are arranged based on the transaction weight which is in descending order. This algorithm assumes less memory and execution time when compared to IIDS. The main disadvantage using this is it also works in two phase pattern hence the candidate generation is huge in first phase and the more execution time is required to find the high utility pattern in phase two algorithm.

III. CHALLENGES

- Most of the prior algorithms works based on two phase approach. So the candidate can be huge and number of scans on database is required large.
- Efficiency and scalability issue when dealing with dense data.

IV. CONCLUSION

This paper provides the survey on the existing algorithm which is designed for high utility itemset mining. High utility pattern mining is very imperative topic in data mining field. Various methods and algorithms are discussed above which help in developing efficient and effective high utility itemsets for data mining.

REFERENCES

[1] R. Agarwal, C. Aggarwal, and V. Prasad, "Depth first generation of long patterns," in *SIGKDD*, 2000, pp. 108–118.

[2] R. Agrawal and R. Srikant. "Fast algorithms for mining association rules," in *Proc. of the 20th VLDB Conf.*, pp.487-499, 1994.

[3] Y. Liu, W. Liao and A. Choudhary, "A fast high utility itemsets mining algorithm," in *Proc. of the Utility-Based Data Mining Workshop*, 2005

[4] W. Wang, J. Yang and P. Yu, "Efficient mining of weighted association rules (WAR)," in *Proc. of the ACMSIGKDD Conference on Knowledge Discovery and Data Mining* (KDD 2000), pp. 270-274, 2000.

[5] S.Meenakshi, P.Sharmila, "A review of high utility patterns," Computer and Communication Engineering, vol. 5, Issue 8, August 2017

[6] Junqiang Liu, Ke Wang, and Benjamin C.M. Fung "Mining High Utility Patterns in One Phase without Generating Candidates" IEEE Transactions on knowledge and data engineering, vol. 28, no. 5, May 2016.

[7] Mengchi Liu, Junfeng Qu, “Mining High Utility Itemsets without Candidate Generation”, CIKM‟12, October 29–November 2, 2012, Maui, HI, USA. Copyright 2012 ACM 978-1-503.

[8] Philippe Fournier-Viger, Cheng-Wei Wu, Souleymane Zida, Vincent S., “FHM: Faster High-Utility Itemset Mining using Estimated Utility Co-occurrence Pruning”, 21st International Symposium on Methodologies for Intelligent Systems (ISMIS 2014), Springer, LNAI, pp. 83-92

[9] F. Bonchi, F. Giannotti, A. Mazzanti, and D. Pedreschi, “Exante: A preprocessing method for frequent-pattern mining,” *IEEE Intelligent Systems*, vol. 20, no. 3, pp. 25–31, 2005.

[10] R. Bayardo and R. Agrawal, “Mining the most interesting rules,” in *SIGKDD*. ACM, 1999, pp. 145–154.

[11] S. Morishita and J. Sese, “Traversing itemset lattice with statistical metric pruning,” in *PODS*. ACM, 2000, pp. 226–236.

[12] Han J., Pei J., Yin Y., Mao R., “Mining frequent patterns without candidate generation: a frequent-pattern tree approach,” Data Mining Knowledge Discovery in Data. Vol. 8, No. 1, pp. 53-87, 2004.

[13] Liu J., Pan Y., Wang K., and Han J., “Mining frequent item sets by opportunistic projection,” In Special Interest Group on Knowledge Discovery and Data Mining. Association for Computing Machinery, pp.229–238, 2002.

[14] Fournier-Viger P., Wu C.-W., Zida S., and Tseng V.S., “Fhm: Faster high-utility itemset mining using estimated utility Cooccurrence pruning,” In Proceedings of the 21th International Symposium on Methodologies for Intelligent Systems. Springer, pp.83-92, 2014.

[15] Li Y.-C., Yeh J.-S., and Chang C.-C., “Isolated items discarding Strategy for discovering high utility itemsets,” Data &Knowledge Engineering, Vol. 64, No. 1, pp. 198–217, 2008.

[16] Ahmed C. F., Tanbeer S. K., Jeong B.-S., and Lee Y. -K., “Efficient tree structures for high utility pattern mining in incremental databases,” IEEE Transactions on Knowledge and Data Engineering, Vol. 21, No. 12, pp. 1708– 1721, 2009.

[17] Tseng V. S., Shie B.-E., Wu C.-W., and Yu P. S., “Efficient algorithms for mining high utility itemsets from transactional databases,” IEEE Transactions onData Engineering, Vol. 25, No. 8, pp. 1772–1, 86, 2013.

Authors Profile

Ms D.Divyashree pursued Bachelor of Technology from S.V Engineering College for Women , Tirupathi in 2016 and currently pursuing Master of Technology from Sree Vidyanikethan Engineering College, Tirupathi.

Dr. G. Sunitha has completed her M.Tech in Computer Science from JNT University, Ananthapuramu in 2005 and Ph.D. in Computer Science and Engineering from S. V. University, Tirupati in 2016. She is currently working as Professor in the Department of CSE, Sree Vidyanikethan Engineering College, Tirupati. Her research interests include Data Mining, Big Data Analytics and Health Care Systems. She has published more than 20 papers in various reputed journals and conferences. She is acting as Technical Committee Member for various international journals and conferences