Survey of High Utility Item sets Mining Algorithms

Sharayu H. Fukey, Prof. P. M. Chawan
[1]Student,[2]Associate Professor
[1]Compute Engineering,
[1]VJTI, Mumbai,India

***Abstract*—This days, we come across a lot of things that have profit technically referred as external utility, value greater than the other item sets in the database. Thus, a new concept that needs new work done is utility mining. Rare item sets have been explained that are used for the mining of the items having utility value less than the threshold but are of great use. Utility mining has come as an important topic in data mining and has received extensive research in last some years. In utility mining, each item is associated with a utility that could be profit, quantity, cost or other user preferences. Objective of Utility Mining is to identify the item sets with highest utilities. Basically the utility of an item set represents its importance, which can be measured in terms of weight, values, quantities or any other information depending on the user specification and requirements. Item set is termed to be high utility item set if its utility is greater than min_util i.e. user specified minimum utility threshold. Practically in many applications high utility item sets consists of rare items. Different decision making domains such as business transactions, medical, security, fraudulent transaction, retail etc. make use of rare item sets to get useful information. "High-utility item set mining" utility mining is a popular problem in the field of data mining. Many algorithms have been proposed in this field in the recent years. In this paper, will give an overview of this problem. Organization of this survey is given as follows: In first section we introduced basic terms like Data mining, frequent pattern mining, Association Rule mining, Utility Mining and Rare Item set Mining. In second section we summarizes some important previous research work related to utility mining. A brief literature survey is given so as to show the work done in this field till date.**

IndexTerms—**Utility Mining, Frequent Item set mining, High-utility item sets, rare item sets, Transaction Weighted Utilization. Component**

I. INTRODUCTION

.1. Data Mining

Data mining is basically extracting or mining the knowledge from large amount of data. The term data mining is appropriately named as 'Knowledge mining from data' or "Knowledge mining". Data mining is about the exploration and analysis, by automatic or semiautomatic ways, of large quantities of data in order to discover meaningful patterns and rules. The prospective analyses offered by data mining technology move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve and answer. They search databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations and expertise. There are two general classification of Data Mining: Descriptive Mining and Predictive Mining. Clustering, Association Rule Discovery, Sequential Pattern Discovery are the Descriptive Mining techniques. They are used to find human-interpretable patterns that describe the data. Classification, Regression, Deviation Detection, use some variables to predict unknown or future values of other variables are Predictive Mining techniques.

2. Association rule Mining

One of the important areas of research is Association Rule Mining (ARM) in data mining. It is a prominent part of Knowledge Discovery in Databases (KDD).That's why it requires more concentration to explore. Association rule mining (ARM) is a technique for discovering co-occurrences, correlations, and frequent patterns, associations among items in a set of transactions or a database. We find rules having confidence and support above user defined threshold. The process of Association Rule Mining is divided into two steps: The first is to find all frequent item sets in data base then to generate association rules. Market based analysis widely used ARM. For example, Market basket data are analyzed, frequent item sets are found then association rules can be generated by predicting the purchase of other items by conditional probability. Given a set of transactions where each transaction is a set of items, an association rule is an expression of the form X ->Y, where X and Y are sets of items. The problem of mining association rules was first introduced in [1] and later broadened in [2], for the case of databases consisting of categorical attributes alone.

3. Frequent Item set Mining

Frequent pattern mining is the key area in the data mining concept which reveals the interesting pattern in the large database. Frequent pattern discovers the item set frequently occurs in a dataset and this information can be used in variety of applications such as market dataset analysis, Indexing and retrievals, detection of software bug, web link analysis etc. Frequent pattern mining considers only whether an item is present or not in a transaction. It reveals the pattern which appears more than the user specified support count. The problem of frequent item set mining is popular. But it has some important limitations when it comes to analyzing customer transactions. An important limitation is that purchase quantities are not taken into account. Thus, an item may only appear once or zero time in a transaction. Thus, if a customer has bought five breads, ten breads or twenty breads, it is viewed

as the same. A second important limitation is that all items are seen as having the same importance, utility of weight. For example, if a customer buys a very expensive bottle of wine or just a piece of bread, it is seen as being equally important. Thus, frequent pattern mining may find many frequent patterns that are not interesting. For example, one may find that {bread, milk} is a frequent pattern. However, from a business perspective, this pattern may be uninteresting because it does not generate much profit. Moreover, frequent pattern mining algorithms may miss the rare patterns that generate a high profit such as perhaps {caviar, wine}. Let I = {a1, a2… an} be a set of n distinct literals called items. An item set is a non-empty set of items. An item set X = (a1, a2, …, ak) with k items is referred to as k-item set, A transaction T = <TID, (a1, a2, …, ak)> consists of a transaction identifier (TID) and a set of items (a1, a2, …, ak), where ij Î I, j = 1, 2, …, k.The frequency of an item set X is the probability of X occurring in a transaction T. A frequent item set is the item set having frequency support greater a minimum user specified threshold.

4. Utility Mining

In real life, the merchant are interested in selling the item set which generate more profits, but the frequent pattern mining generates only frequent item set without considering quantity of the item sold or the profit of the item. Even though frequent item set mining discovers crucial frequent patterns, it leaves the profit and quantity of the item is the disadvantage. To address this issue weighted association rule mining has been developed. It tries to find the association of item set in a database by considering the profit/weight of the item set. To address these, utility mining has been introduced utility mining considers both the profit and the number of items purchased. In this the utility of an item set is calculated as the product of the profit of the item and the number of item purchased. Utility mining model was proposed in [5] to define the utility of item set. The utility is a measure of how useful or profitable an item set X is. The utility of an item set X, i.e., u(X), is the sum of the utilities of item set X in all the transactions containing X. An item set X is called a high utility item set if and only if u(X) >= min_utility, where min_utility is a user-defined minimum utility threshold [11]. Frequent item set mining follows the download closure property, if K- item set is generated, K+1 item set can be generated by considering only K-item set or in other words K+1 item set will contain only the item set present in the K-item set. This downward closure property is not satisfied, if k-item set is low utility item set K+1 can be high utility item set and vice versa. Both the monotonic and antimonotonic property is not supported by high utility item set. This issue is addressed by the over estimation method [6].

5. Rare Itemset Mining

The item sets that occur infrequently in the transaction data set are called rare itemsets. In most business applications, frequent itemsets may not generate much profit while rare itemsets may generate a very high profit. Rare itemsets are very important occasionally and can be further promoted together because they possess high associations and can bring some acceptable profits. Rare itemsets provide very useful information in the real-life applications such as security, business strategies, biology, medicine and super market shelf management. For example in the security field, normal behavior is very frequent, whereas abnormal or suspicious behavior is less frequent

II. UTILITY MINING PROBLEM EXPLANATION

WhereverTimesisspecified,TimesRomanorTimesNewRomanmaybeused.IfneitherisavailableonyourwordprocessorpleaseusethefontclosestinappearancetoTimes.Avoidusingbit-mappedfonts.TrueType1orOpenTypefontsare required.Pleaseembed all fonts, in particular symbolfonts,aswell,formath,etc.

In this section various definitions used in high utility item set mining are introduced. Let I be a set of distinct items I = {x1, x2, x3….xn}. Let IS ك I. Let D be a transaction database which contains set of transaction D= {T1, T2, T3,….Tn}. Each transaction contains a unique identifier Tid. Each item in the transaction xi in the transaction has a quantity or internal utility IU (Tid, xi) associated with it. Each item in the database xi is associated with a profit or external utility EU (xi). For example consider a transaction database TABLE I:

TID	*TRANSACTIONS*
1	(A:2),(B:1),(D:2)
2	(B:2),(C:1)
3	(A:1),(B:2),(C:3)
4	(B:1),(C:1),(D:2)

TABLE 1: Transaction Database

Item	*Profit*
A	5
B	4
C	2
D	3

TABLE 2: External Utility Table

TID	Profit
1	20
2	10
3	19
4	12

TABLE 3: Transaction Utility Table

DEFINITION 1: (Utility of an item in transaction). The utility of an item ij in Td is denoted as u (ij, Td), which is defined as: u (ij, Td) = q (ij, Td) * p (ij) (1) in which q (ij, Td) is the quantity of an item set ij in Td, and p (ij) is the profit of an item set ij. From the running example, in table I, the utility of(A) in TID(=1) is calculated as: u(A,T1)=q(A, T1)*p(A) u(A, T1)=5*2=10.

DEFINITION 2: (Utility of an item set in transaction).The utility of an item set X in transaction is denoted as u(X,Td),which can be defined as: u( ) = (2) From the running example the utility of item set AB is u(AB, T1) = u(A, T1)+ u(B, T1) = q(A, T1)*p(A)+ q(B, T1)*p(B) =(5*2)+(1*4) =10+4 =14.

DEFINITION 3 :(High utility item set, HUI). An item set X is a high-utility item set (HUI) in database D if its utility in D is no less than minimum utility count as: (3)
In table 1 and 3, Let the minimum utility threshold =6 Utility of (C) is calculated as: u(C) =u(C, T2) + u(C, T3) + u(C, T4) =2+6+2 =10>6 Hence C is a high utility item set.

DEFINITION 4: (Transaction-Weighted utility of an item set). The Transaction-Weighted utility of an item set X is the sum of all transaction utility TU (Td) containing item set X in Which is defined as:

$$TWU(X) = \sum_{X \subseteq T_r \wedge T_r \in D} TU(T_r).$$ …….. (4)

TWU (AB) = 20+19 = 39.

III. LITERATURE REVIEW

In the previous section we have introduced the basic concept of Data Mining, Association Rule mining, Utility Mining and Frequent Item set Mining and Rare Itemset Mining. A brief overview of various algorithms of utility mining, previous attempts, concepts and techniques defined in different research papers have been given in this section.
Association Rule Mining (ARM) consists of two-steps. First finding all frequent elements in Database then generating association rules from them. ARM is well studied with methods like Apriori [1][2] Problem of Utility Mining is defined in theoretical model [6] which says finding all elements in a transaction database with utility values greater than minimum utility threshold. [11]Generally depending on semantics of Application utility based measure are classified as item level, transaction level and cell level. High utility frequent item sets helps in objective function or performance boost.
Utility Mining algorithms can be classified as Two Phase and One Phase. In Two Phase algorithm for first phase database is scanned and transaction weighted utility of each transaction is calculated and candidates which are having transaction weighted utilization greater than minimum threshold value are taken in consideration. Now search space of algorithm is limited. In second phase high utility item set are found by scanning database again from high transaction weighted utilization of item set. Basically in first phase it generates candidates with potential high utility item sets in second they calculate exact utility of each every candidate found in first phase and identifies high utility item sets.
Examples of Two Phase Algorithms are Two Phase, IHUP, IID and Up Growth. Unlike two phase algorithm which generate high utility item set using only one phase and produce no candidate.d2HUP and HUI Miner are examples of one phase Algorithm.
Two Phase and HUI Miner are explained in detail below:

A] Two Phase

This algorithm runs in two phases. The algorithm utilizes the downward closure property, of transaction weighted utilization item set. If k+1 item set is high transaction weighted utility item set, only weighted utilization item set. This property also states that the super set of low transaction weighted utility item set is also a low transaction weighted utility item set. By this property this algorithm reduces the number of candidate set. In order to find the high utility item set from the high transaction weighted utility item set. If Transaction weighted high utility item set (TWHUI) be the high transaction weighted utility item set and HUI be the high utility item set then HUI ك TWHUI.
In phase 1 of this algorithm, the database is scanned once; during the scan the transaction weighted utility of each transaction is calculated. The candidate which has transaction weighted utilization value greater than minimum threshold is only taken in account. This limits the search space of this
Algorithm. For larger database this reduces search space shows better performance than MEU. In phase 2 of this algorithm, the database is scanned once again to find the high utility item set. From the high transaction weighted utilization of item sets .When user defined minimum threshold is large, then the high transaction weighted utilization of item set is small. Experiments evaluation shows that, this algorithm shows better performance than previous MEU and it is evident in larger databases. This paper show the important path of utilizing the downward closure property of high transaction weighted utilization of item set.

B] HUI Miner

HUI miner algorithm mines the high utility item set in a single phase. It uses a depth first search approach.HUI miner uses a new data structure utility list. The utility list is generated for each item who's TWU>= min_util is generated by scanning the database. The utility list of an item X consists of triples for each transaction the item X participants. The triples contains a transaction id (tid) the utility of item X in the transaction (utility),the utility of items that are present after the item X in
The total order of in that item transaction remains utility. The high utility item set are mined from the utility list constructed on the property, if sum of item utility of an item set X >= min_utilty then X is a high utility item set. Since the item utility is utility of item set the sum of item utility represents the utility of item set in the database. The single high utility items are generated based on the above property. To check whether super set of an item found in a high utility item it uses the property, sum of item utility and min_utility is less than the given minimum utility than all the extension of X in the total order will be low utility item set. This algorithm calculates the sum of item utility and remain utility, if it is greater than min utility, it adds for each item extension y >x in their total order as the extension, by recursion it finds out whether each extension is a HUI.
An overview of the various Algorithms, Techniques, approaches and limitations that have been defined in various research publications have been given in this section.

Two- Phase, This algorithm is suitable for sparse database with short Patterns. It consists of: Phase 1: Discover candidate item sets that are having a TWU $\geqslant$ minutil Phase 2: For each candidate, calculate its exact utility by scanning the database. Limitation is Many scans of database and generates many candidate Item sets were required. It was stated in paper A Two-Phase Algorithm for Fast Discovery of High Utility Item sets [4] by Ying Liu, Wei-keng Liao, and Alok Choudhary in 2005

CTU-Mine, This algorithm is suitable for this approach is suitable for dense dataset with long pattern. It Use pattern growth algorithm and also eliminates the expensive second phase of scanning the database. Limitation was Complex for evaluation due to the tree structure. It was stated in paper CTU-Mine: An Efficient High Utility Item set Mining Algorithm Using the Pattern Growth Approach [5] by Alva Erwin, Raj P. Gopalan, and N.R.Achuthan in 2007

UP-Growth, This algorithm is synthetic and real datasets are used to evaluate the high performance of the algorithm. It includes steps (1) construction of UP-Tree, (2) generation of potential high utility itemsets from the UP-Tree by UP-Growth, and (3) identification of high utility item sets from the set of potential high utility item sets .Limitation is Complex for evaluation due to the tree structure. It was stated in UP-Growth: An Efficient Algorithm for High Utility Item set Mining [6] by Vincent S. Tseng, Bai-En Shie in 2010

Hui-Miner, This algorithm Single Phase Algorithm. No need to multiple times database scan. In this We should try to avoid performing joins if possible for low-utility itemsets. Limitation is calculating the utility of an item set joining utility list is very costly. It was stated in Mining High Utility Itemsets without Candidate Generation [7] by Mengchi Liu, Jun Feng Qu in 2012

FHM, we should try it using a dynamic database. It estimated-Utility Co-occurrence pruning. Limitation is working with Static Database. It was stated in FHM: Faster High-Utility Itemsets Mining using Estimated Utility Co- occurrence Pruning [8] by Philippe Fournier- Viger, Cheng- Wei wu in 2014

This section summarizes the comparison of different existing Approaches

IV. CONCLUSION

The benefit of frequent itemsets mining by considering only frequency of itemsets is challenged in many research areas Utility Mining focuses on profit i.e. utility consideration while item set mining. All aspects of economic utility in data mining are covered in utility mining. It uses item utilities as an analytical measurement of the importance of that item in the user's point of view. Practically in many applications high utility item sets plays an important role. Different decision making domains such as business transactions, medical, security, fraudulent transaction, retail etc. make use of rare item sets to get useful information. Survey on different high utility item set mining algorithms which were proposed are presented in this paper. This survey will be helpful for developing new efficient and optimize technique for high utility item set mining. Reducing the search space while searching for the high utility itemsets is primary concern. This survey will be helpful for developing new efficient and optimize algorithms for high utility item set mining.

REFERENCES

[1] R. Agrawal and R. Srikant, 1994, "Fast Algorithms for Mining Association Rules", in Proceedings of the 20th International Conference Very Large Databases, pp. 487-499.

[2] Attila Gyenesei, "Mining Weighted Association Rules for Fuzzy Quantitative Items", Lecture notes in Computer Science, Springer, Vol.1910/2000, pages 187-219, TUCS Technical Report No.346, ISBN 952-12-659-4,ISSN 1239-1891, May 2000.

[3] R. Chan, Q. Yang, Y. D. Shen, "Mining High utility Itemsets", In Proc. of the 3rd IEEE Intel.Conf. on Data Mining (ICDM), 2003.

[4] H. Yun, D. Ha, B. Hwang, and K. Ryu. "Mining association rules on significant rare data using relative support". Journal of Systems and Software, 67(3):181–191, 2003.

[5] H.Yao, H. J. Hamilton, and C. J. Butz, "A Foundational Approach to Mining Itemset Utilities from Databases",Proceedings of the Third SIAM International Conference on Data Mining, Orlando, Florida, pp. 482-486, 2004.

[6] G. Weiss. "Mining with rarity: a unifying framework", .SIGKDD Explor. Newsl., 6(1):7–19, 2004.

[7] Liu, Y., Liao, W., and A. Choudhary, A., "A Fast High Utility Itemsets Mining Algorithm", In Proceedings of the Utility- Based Data Mining Workshop, August 2005.

[8] Lu, S., Hu, H. and Li, F. 2005. "Mining weighted association rules. Intelligent Data Analysis", 5(3):211–225.

[9] V. S. Tseng, C.J. Chu, T. Liang, "Efficient Mining of Temporal High Utility Itemsets from Data streams", Proceedings of Second International Workshop on Utility-Based Data Mining, August 20, 2006

[10] H. Yao, H. Hamilton and L. Geng, "A Unified Framework for Utility-Based Measures for Mining Itemsets", In Proc. Of the ACM Intel. Conf. on Utility-Based Data Mining Workshop (UBDM), pp. 28-37, 2006.

[11] A. Erwin, R.P.Gopalan and N. R. Achuthan, 2007, "A Bottom-up Projection based Algorithm for mining high utility itemsets", in Proceedings of 2nd Workshop on integrating AI and Data Mining(AIDM 2007)", Australia, Conferences in ResearchandPracticeinInformation Technology(CRPIT), Vol. 84.

[12] J. Hu, A. Mojsilovic, "High-utility pattern mining: A method for discovery of high-utility item sets", Pattern Recognition 40 (2007) 3317 –3324.

[13] L. Szathmary, A. Napoli, P. Valtchev, "Towards Rare Itemset Mining" Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence, 2007, Volume 1, Pages: 305-312, ISBN ~ ISSN:1082-3409 , 0-7695-3015-X

[14] Kriegel, H-P et al. 2007. "Future Trends in Data Mining, Data Mining and Knowledge Discovery", 15:87–97.

[15] M. Adda, L. Wu, Y. Feng, "Rare Itemset Mining", Sixth International conference on Machine Learning and Applications, 2007, pp 73-80.

[16] H.F. Li, H.Y. Huang, Y.Cheng Chen, and Y. Liu and S. Lee, "Fast and Memory Efficient Mining of High Utility Itemsets in Data Streams", 2008 Eighth IEEE International Conference on Data Mining.

[17] M. Sulaiman Khan, M. Muyeba, Frans Coenen, 2008. "Fuzzy Weighted Association Rule Mining with Weighted Support and Confidence Framework", to appear in ALSIP (PAKDD),pp. 52-64.

[18] S. Shankar, T.P.Purusothoman, S.Jayanthi and N.Babu, "A Fast Algorithm for Mining High Utility Itemsets", Proceedings of IEEE International Advance Computing Conference (IACC 2009), Patiala, India, pages : 1459 - 1464

[19] Hu, J., Mojsilovic, A. "High-utility Pattern Mining: A Method for Discovery of High-utility Item" Sets, Pattern Recognition, Vol. 40,3317-3324.

[20] G.C.Lan, T.P.Hong and V.S. Tseng, "A Novel Algorithm for Mining Rare-Utility Itemsets in a Multi-Database Environment"

[21] J. Pillai, O.P. Vyas, S. SoniM. Muyeba "A Conceptual Approach to Temporal Weighted Itemset Utility Mining", 2010 International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 28