International Refereed Journal of Scientific Research in Engineering (IRJSRE)
Volume 2, Issue 1 (January 2017), PP. 28-32

High Utility Item Sets Mining Algorithms And Application

P. M. Chawan, Sharayu H. Fukey
VJTI, India
VJTI, India

ABSTRACT: 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 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. Such as in supermarket customer purchase washing machine or fridge rarely as compared to butter or sugar. But previous transaction provides more profit for the supermarket. In this paper, will give an overview of high utility mining 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 overview of various algorithms have been presented in this section. In last section we concluded the survey work by discussing some applications of Utility.
Keywords: Utility Mining, High-utility item sets, rare item sets, Frequent Item set mining, Transaction Weighted Utilization. Component

I. INTRODUCTION

Data Mining: Data mining is about the extraction of hidden predictive information from large databases and is a powerful new technology with great potential to help companies focus on the important information in their data warehouses. Data mining tools used to predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. 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.

1.1 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.

1.2 Frequent Itemset Mining

Frequent pattern mining is the key area in the data miningconcept 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 itemset 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 $I = \{a1, a2 \ldots an\}$ be a set of n distinct literals called *items*. An **item set** is a non-empty set of items. An item set $X = (a1, a2, \ldots, ak)$ with k items is referred to as k-item set, A **transaction***T* = <TID, (*a*1, *a*2, …, *ak*)> consists of a transaction identifier (TID) and a setof items (*a*1, *a*2, …, *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.

1.3 Utility Mining

In real life, the merchant are interested in selling the itemset which generate more profits, but the frequent pattern mining generates only frequent itemset without considering quantity of the item sold or the profit of the item. Even though frequent itemset 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 itemset in a database by considering the profit/weight of the itemset.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 itemset is calculated as the product of the profit of the item and the number of item purchased.Utility mining model was proposed in [3] 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 itemset can be generated by considering only K-item set or in other words K+1 itemset will contain only the item set present in the K-itemset. This downward closure property is not satisfied, if k-itemset is low utility itemset 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 [4].

II. UTILITY MINING: EXAMPLE

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 be and item set such that IS كI. 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 Fig.1.1:

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)

Fig.2.1: Transaction Database

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

Fig. 2.2: External Utility Table

TID	Profit
1	20
2	10
3	19
4	12

Fig. 2.3: 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 Fig. 2.1, 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: $u(X,T_d)=\sum_{i_p\in X\wedge X\subseteq T_d} u(i_p,T_d)$ ….. (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:

$HUI \leftarrow \{X \mid \sum_{X\subseteq T_d\wedge T_d\in D} u(X,T_d) \geq minutility\}$ ……. (3)In Fig 2.1 and2.3, 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. ALGORITHMS FOR HIGH UTILITY ITEMSET MINING (HUIM)

In the previous section we have introduced the basic concept of Data Mining, Association Rule mining, Utility Mining and Frequent Item set 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.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: 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. 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 whose 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. An overview of the various Algorithms have been defined in various research publications have been given in this section.

Ying Liu, Wei-keng Liao, and Alok Choudhary in A Two-Phase Algorithm for Fast Discovery of High Utility Item sets [5] states about explained working of Two Phase algorithm as

Phase 1: Discover candidate item sets that is having a TWU ≥ min_util Phase 2: For each candidate, calculate its exact utility by scanning the database. It states that the problem with the algorithm is it requires many scans of database and generates many candidate Item sets.CTU-Mine which uses pattern growth algorithm and also eliminates the expensive second phase of scanning the database was stated in CTU-Mine: An Efficient High Utility Itemset Mining Algorithm Using the Pattern Growth Approach [6] by Alva Erwin, Raj P. Gopalan,

N.R.Achuthan. This approach is suitable for dense dataset with long pattern but was complex for evaluation due to tree structure.UP-Growth: An Efficient Algorithm for High Utility Itemset [7] Mining was proposed by Vincent S. Tseng, Bai-En Shie. It works as (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 through this steps. But UP-Growth still faces the problem of complexity due to tree structure.

Then HUI-Miner [8] was introduced by Mengchi Liu, Junfeng Qu in Mining High Utility Item sets without Candidate Generation. It is single phase algorithm. No need to multiple times database scan.FHM improvisation was stated in FHM: Faster High-Utility Itemset Mining using Estimated Utility Co- occurrence[9] Pruning by Philippe Fournier- Viger, Cheng- Wei wu.It basically stated estimated-Utility co-occurrence pruning. It was limited to Static database.AprioriHC-D/AprioriHC[10] both were about closed high utility item set, lossless and concise representation stated in efficient vertical mining of high utility quantitative item sets by Vincent S. Tseng, Cheng- Wei Wu, Philippe FournierViger, and Philip S. Yu.

Hua-Fu Li, Hsin-Yun Huang, Suh-Yin Lee in Fast and Memory Efficient Mining of High Utility Itemsets[11] in Data Streams stated Mining High Utility Itemsets based on BIT vector to improve the efficiency of mining high utility itemsets two effective representations of an extended lexicographical tree-based summary data structure and itemset information were developed.Recent research done by Vincent S. Tseng, ChengWei Wu, Viger, Philip S. Yu stated in Efficient Algorithms for Mining Top-K High Utility Itemsets[12] two algorithms for Utility Mining. TKO which is one phase and TKU two phase algorithm. Stated framework for top-k high utility itemset mining. Wei Wu, Viger, Philip S. Yu, 2015. Framework for top-k high utility itemset mining they did empirical evaluations on both real and synthetic datasets show that the performance of the proposed algorithms is close to that of the optimal case of state-of-the-art utility mining algorithms. Where k is the desired number of high utility itemsets to be mined.

IV. CONCLUSION

AUtility mining is an evident topic in data mining. It focuses on utility consideration while item set mining. All aspects of economic utility in data mining are covered in utility mining. 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. Some of the Applications of Utility Ming includes: In supermarket customer purchase washing machine or fridge rarely as compared to butter or sugar. But previous transaction provides more profit for the supermarket. Also in medical application the unique or rare combination of symptoms can give a useful information about disease of patient to doctors [13]. Retails business is able to find out their high investing customers with utility mining [14].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.

REFERENCES

[1] R. Agrawal, T. Imielinski and A. Swami, Mining association rules between sets of items in large databases, Proceedings of the ACM SIGMOD International Conference on Management of data, ,1993, 207-216.

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

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

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

[5] Ying Liu, Wei-keng Liao, and Alok Choudhary, A Two-Phase Algorithm for Fast Discovery of High Utility Item sets, in PAKDD'05 Proceedings of the 9th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining,2005, 689-695

[6] A. Erwin, R. P. Gopalan and N. R. Achuthan, CTU-Mine: An Efficient High Utility Itemset Mining Algorithm Using the Pattern Growth Approach, 7th IEEE International Conference on Computer and Information Technology (CIT 2007), Aizu-Wakamatsu, Fukushima, 2007, 71-76.

[7] Vincent S. Tseng, Bai-En Shie, UP-Growth: An Efficient Algorithm for High Utility Itemset Mining, Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 2010,July 25-28

[8] Mengchi Liu,Junfeng Qu, Mining High Utility Itemsets without Candidate Generation, Proceeding CIKM '12 Proceedings of the 21st ACM international conference on Information and knowledge management,2012 55-64

[9] Philippe Fournier- Viger, Cheng- Wei wu, FHM: Faster High-Utility Itemset Mining using Estimated Utility Co- occurrence Pruning, 2014,ISMIS 2014,LNAI 8502 83-92
[10] C. H. Li, C. W. Wu and V. S. Tseng, "Efficient vertical mining of high utility quantitative itemsets," 2014 IEEE International Conference on Granular Computing (GrC), Noboribetsu, 2014, 155-160
[11] H. F. Li, H. Y. Huang, Y. C. Chen, Y. J. Liu and S. Y. Lee, Fast and Memory Efficient Mining of High Utility Itemsets in Data Streams, 2008 Eighth IEEE International Conference on Data Mining, Pisa,881-886.
[12] V. S. Tseng, C. W. Wu, P. Fournier-Viger and P. S. Yu, Efficient Algorithms for Mining Top-K High Utility Itemsets, 2016,IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 1, 54-67
[13] G.C.Lan, T.P.Hong and V.S. Tseng, "A Novel Algorithm for Mining Rare-Utility Itemsets, , November 2016,Knowledge-Based Systems archive Volume 111 Issue C, 283-298
[14] V. S. Tseng, C.J. Chu, T. Liang, "Efficient Mining of Temporal High Utility Itemsets from Data streams",2006 Proceedings of Second International Workshop on Utility-Based Data Mining