A Brief Insight on High Utility Rare Itemsets Mining

Sunidhi Shrivastava[1], Punit Kumar Johari[2]
[1, 2] Department of CSE / IT
[1, 2] Madhav Institute of Technology and Science, Gwalior, India

Abstract- *Utility mining is an emerging topic in data mining field. The main objective of Utility Mining is to identify the itemsets whose utility is highest than the user-specified threshold. Association rule mining (ARM) is an approach which works efficiently with the utility mining. It describes frequent itemsets from databases and generates association rules by allowing for each item in equal value. However, items are absolutely different in many aspects in a number of real applications, equally retail marketing, network log, etc. Utility Mining aims to identify itemsets with highest utilities by taking into consideration profit, quantity, cost or other user selections. Rare items are items that occur less frequently in a transaction data set. High Utility Itemsets may either be frequent or rare. Likewise, rare itemset may be of high or low utility. In numerous real-life applications, high-utility item sets which consist infrequent items. In this paper, we provide an overview of the various techniques and algorithms developed for discovering high utility itemsets. A comparative study of all algorithm developed till date for HUI mining also given.*

Keywords- Data mining, utility mining, association rule mining, rare itemset mining.

I. INTRODUCTION

Data mining refers to deriving or mining, useful information or knowledge from large amounts of data. Data mining is the process of discovering the intensive knowledge from a large amount of data which is stored in data warehouses and information repositories. Data mining is the process of knowledge discovery from the huge amount of data stored in various databases. Here the knowledge belongs to the valuable information which can be used further computation. The goal of data mining is to extract higher-level invisible information from an abundance of raw data. Data mining has been used in different data domains. Data mining can be considered as an algorithmic process that takes data as input and yields patterns, similarly classification rules, item sets, association rules, or summaries, as output [1].

1.1 Privacy Preservation in Data Mining:-

Enlarging network complexity, providing greater access, sharing information and a prospering emphasis on the Internet have caused information security and privacy a notable concern for human beings and organizations. Data mining is a well-known technology for automatic and sensibly extracting knowledge from the large amount of data. Such a process can also reveal sensitive information about individuals Compromising the individual's privacy. Privacy preserving data mining (PPDM) is a new age of research in data mining. Its ultimate goal is to develop effective algorithms that allow one to extract appropriate knowledge from large amounts of data, while avoid sensitive information from exposure or interference.

1.2 Utility mining:-

In data mining association rule mining approaches consider the utility of the items by its presence in the transaction set. As we know frequent item set mining is used to indicate the frequent items. But we can't say if any item set which have sold frequently will make a profit. May be those item sets which are less frequent or rare item set can make more profit than frequent item set. One of the most challenging data mining tasks is the mining of high utility item sets conveniently. Identification of the item sets with high utilities is called as Utility Mining. The utility can be measured in terms of cost, profit or other expressions of user selections. For example, a computer system may be more profitable than a telephone in terms of profit. [2]

For example- If in a mobile shop, 100 mobile sets of nokia worth rupees -2000/- are sold frequently, but at the same time in another shop a iPhone sold in 60,000/- rarely so its cleared that if any item which sold frequently but with less prices and at the same time another item which sold rarely can make more profit.

The utility is a degree of how valuable or profitable an itemset X is. The utility of an itemset X, i.e. u(X), which is the sum total of the all utilities of itemset X in all the transactions containing X. An itemset X is called a high utility itemset if and only if u(X) larger than or equal to min_utility, where min_utility is a user defined minimum utility threshold. The main goal of high-utility itemset mining is to find all those itemsets having utility larger or equal to user- defined minimum utility threshold. [3]

1.3 Association rule Mining:-

There are several techniques satisfying these objectives of data mining. Mining Associations is one of the methods involved in the process mentioned above and among the data mining issues it might be the most studied ones. Discovering association rules is the heart of data mining.

Mining for association rules between items in large database of sales transactions has been identified as an important area of database research.

These rules can be effectively used to reveal unknown relationships, producing results that can provide a basis for forecasting and decision making. The original problem defined by association rule mining was to find a correlation among sales of different products from the analysis of a large set of data [4] [5].

A. Frequent Itemset mining:-

Frequent itemsets are the item sets that present frequently in the transactions of any database. The goal of frequent itemset mining is to find out all the itemsets in a transaction dataset. Frequent itemset mining plays an important role in the theory and practice of many important data mining tasks, such as mining association rule, long patterns, emerging pattern, and dependency rules. It has been applied in the field of telecommunications, census analysis and text analysis. The standard of being frequent is expressed in terms of support value of the item sets. The Support value of an itemset is the percentage of transactions that contain the itemset after that the support value will be compared with the predefined threshold value, which was user generated. if support is equal or greater than the minimum threshold value than those values will be further processed for 2k frequent pattern mining, those which not qualify the minimum threshold will be discarded.

In the real life world some applications of frequent item sets are-

- Medical image processing,
- Biological data analysis.

B. Rare itemset mining:-

Itemset that do not occur frequently in the database, or we can say infrequent items in the database. Rare cases deserve special consideration because they represent significant difficulties for data mining algorithms.

The discovery of infrequent itemsets, and in sequence of rare association rules deriving from rare itemsets, may be particularly useful in biology and medicine. Suppose a professional in biology is interested to find out the cause of cardiovascular diseases (CVD) for a given transactional database of medical records. A frequent itemset such as "{elevated cholesterol level, CVD}" may varify the hypothesis that there are two data items are frequently associated, leading to the possible interpretation "citizens having the higher level of cholesterol will be at high risk for CVD". On the other hand, the fact that "{vegetarian, CVD}" is an infrequent itemset may justify that the association of these two itemsets is rather exceptional, leading to the possible interpretation "vegetarian citizens are at a low risk for CVD". Moreover, the itemsets {vegetarian} and {CVD} can be both frequent, while the itemset {vegetarian, CVD} is rare. So in these kind of situation rare item set is become more suitable than frequent item set[6].

The generating of the infrequent item set is valid for the data coming from the different real life application background like:

- Statistical disclosure risk evaluation from census data and
- Fraud detection.

II. LITERATURE SURVEY

The association rules mining for relationship discovery between data items in the big databases are a well studied method in data mining field with representative approaches like Apriori [7], [8]. The ARM procedure can be decomposed into two different levels. The first step includes discovery each frequent item set of databases. The next level contains producing association rules from frequent item sets.
We have introduced the basic of data mining, utility mining, rare item set mining and frequent item set mining. A brief numerous algorithm overview and methods defined in various research papers has been provided in this part.

H. Yao et al proposed in [9], the utility problem based mining is to discover the item sets that are noteworthy according to their utility values. In this paper apriori property and changeable constraint property are not applicable to the utility based item set mining issue. As an outcome, mathematical item set utility value properties were analyzed. Two different new pruning schemes were presented to decrease the finding high utility item sets cost. With these pruning schemes, a k-item set with a utility upper bound less than minutil can be pruned immediately without accessing the database to compute its actual utility value. By exploiting these pruning strategies, the UMining and UMining_H algorithms were developed to provide effective solutions to the utility based item set mining issue.

This approach also has some limitations first, since first depth search approaches such as FP growth have several advantages over level wise approaches. So it can improve in the future scenario. Second, the problem of how to classify high utility rules from the high utility item set could be investigated.

J. Hu et al in [10], proposed Frequent item set mining algorithm which classify high utility item groupings. In contrast to the classical association rule and frequent item mining methods, the aim of the algorithm is to find data segments, defined by few items (rules) groupings, which fulfill various situations present an effective estimate to solve it by particular partition trees, known as high yield partition trees and investigated the various splitting schemes performance.

Pillai, Jyothi, and O. P. Vyas. in [11], proposed HURI which uses the concept of apriori inverse which produces only rare itemsets having support less than maximum support value where as HURI can produce high utility rare itemsets based on support threshold, utility threshold and user's interest. Hence HURI is said to be more beneficial on application of synthetic data set. The future work includes the incorporation of temporal and fuzzy concept in HURI and using it for finding those rare items, which provide maximum profit to a transaction. HURI can also be used as a base for customer utility mining for classifying customers according to some criteria; for example, a retail business may require to identify valuable customers who are major contributors to a Company's overall profit.

V.S. Tseng et al. [12] proposed a new technique specifically Temporal High Utility Itemsets (THUI) for temporal high utility item sets mining from data streams effectively and efficiently. The temporal high utility itemsets are the item sets with the support larger than the pre-specified threshold in the present data stream time window. Temporal high utility item sets discovery is a significant mining interesting pattern, procedure for example from data streams association rules. In this paper, we propose a new technique, namely Temporal High Utility Item sets (THUI) -Mine, for the mining of temporal high utility item sets from data streams effectively and efficiently. New THUI-Mine contribution is that it can efficiently identify temporal high utility item sets through high performance. In this way, the discovering procedure each window can be achieved efficiently with limited memory space, less candidate item sets and time of CPU I/O. This meets the critical needs on efficiency of time and space for mining data streams.

G.C.Lan et al. [13] proposed a novel pattern type, known as Rare Utility Item sets in which consider not only individual profits and quantities but also usual current periods and items branches also considered in a multi database atmosphere. A novel approach of mining known as Two-Phase algorithm for mining Rare Utility Item sets in Multiple Databases (TP-RUI-MD) was proposed to effectively discover rare utility item sets. The Two-Phase algorithm for mining Rare Utility Item sets in Multiple Databases (TP-RUI-MD) algorithm is designed to find rare-utility item sets atmosphere. The first one is that we proposed a novel item set type known rare-utility item set in a multi-database atmosphere.

For future scenario, both applied algorithm can be used in the rare-utility item set and proposed mining algorithm into other different practical applications, for example, supermarket promotion, data stream, medical application, etc. to discover the more stimulating and patterns or valuable rules in a multi-database atmosphere.

David j. haglin et al. in [14], proposed Minimal Infrequent Itemsets (MINIT) finding method which was the first algorithm produced specially for identifying Minimal Infrequent Item Set (MIIs). The computational time necessary on the four datasets suggests a correlation between the number of MIIs and the amount of computation required.

Mehdi Addaet described in [15], an ARANIM algorithm for Apriori Rare and Non-Present Item-set Mining. The proposed method is Apriori-like and mining concept behind it is that if the item-set lattice representing the item set space than in traditional Apriori methods is traversed in a bottom-up manner, equivalent properties to the Apriori frequent item-sets exploration is provided to discover rare item-sets. Also author proposed method based on the rare patterns and performances in the web application context.

Younghee Kim et al. in [16], proposed an efficient algorithm named Weighted Support Frequent item sets (WSFI) was proposed which normalized weight mine over the streams of data, along with that a novel tree structure also proposed which is called the WSFP-Tree (weighted support FP-tree), that saves compressed critical knowledge about frequent item sets. The proposed WSFPTree is an extended FP-tree based data structure. It is an extended prefix-tree structure to store compressed, critical knowledge about the frequent patterns. The estimation demonstrates that the WSFI-mine outperforms the DSM-FI and THUI-Mine in mining frequent item sets over the data streams.

Luca Cagliero and Paolo Garza in [17], proposed a paper in which the discovering the rare issue and weighted item sets

were handled. i.e. the IWI (infrequent weighted item set) mining issue. Two new quality measures are proposed to the drive IWI mining procedure. Furthermore, two different algorithms that achieve IWI and Minimal IWI mining efficiencies, driven through proposed measures, were presented. As per the analysis of (Literature Review on Infrequent Item set Mining Algorithms) MIWI is the most efficient algorithm, which computes in less computing time, increases the performance efficiency when large database, computes the weighted transaction among the current algorithms.

High Utility itemsets may contain frequent as well as rare itemsets. Classical utility mining only considers items and their utilities as discrete values. In real world applications, such utilities can be described by fuzzy sets. Thus itemset utility mining with fuzzy modeling allows item utility values to be fuzzy and dynamic over time. In this paper, an algorithm, FHURI (Fuzzy High Utility Rare Itemset Mining) is presented to efficiently and effectively mine very-high (and high) utility rare itemsets from databases, by fuzzification of utility values. FHURI can effectively extract fuzzy high utility rare itemsets by integrating fuzzy logic with high utility rare itemset mining. FHURI algorithm may have practical meaning to real-world marketing strategies. The results are shown using synthetic datasets [18].

S.A.R. niha et al [19] proposed UPRI algorithm for generating high utility rare item sets. UPR tree has been used for representing information in the tree based structure. They have proposed algorithm by following four strategies for generating high utility patterns. In this paper, an UPRI has been used to effectively mine high utility rare transactional database. The high utility rare itemsets are the item occur infrequently in the transactional data have a major contribution to the overall business. By analyzing the UPRI algorithm, the itemsets are efficiently generated. The profitability of the company increased by identifying the profitable consisting of high utility rare itemsets an marketing strategies can be developed for knowledge generated from the algorithm crucial business decision-making process catalogue design, cross marketing, final policy.

III. COMPARITVE STUDY

The comparative studies of all the high utility itemset mining technique which are proposed by different authors are given in table1. The comparison table shows the difference between all the techniques which are proposed till the date. We have also found out some drawback of their work with some beneficial ideas for betterment of the work for the future implementation.

These small ideas can make big difference in the future researches; I have also mentioned their techniques and algorithms which were used for their work.

Table 1 Comparison table summarizes the comparison of different existing Approaches

No	Title of Paper	year	Authors	Name of algorithm	Overview of work/idea	Limitation	Idea of improvement
1	A Two-Phase Algorithm for Fast Discovery of High Utility Item sets[20]	2005	Ying Liu, Wei-keng Liao, and Alok Choudhary	TwoPhase	Phase 1: identify all the itemset having TWU>min_utility Phase 2: For each candidate calculate its exact utility by scanning the database	Several number of scans of database and generates many candidate Itemsets	This approach is suitable for wide range database with short Patterns.
2	CTU-Mine: An Efficient High Utility Itemset Mining Algorithm Using the Pattern Growth Approach [21]	2007	Alva Erwin, Raj P. Gopalan, N.R. Achuthan,	CTU-Mine	Use pattern growth approach and reduce the expenses occur in the second phase for scanning the database	Due to the tree stricture complexity will increase	This approach is perfectly useful only for dense data with long patterns
3	UP-Growth: An Efficient Algorithm	2010	Vincent S. Tseng, Bai-	UPGrowth	1) UP-Tree Construction,	1) It requires multiple	synthetic and real datasets

	for High Utility Itemset Mining[22]		En Shie,		2) High utility itemsets are generated from the UP-tree by UP-Growth and 3) identification of high utility itemsets from the set of potential high utility itemsets	database scans & Generate multiple candidate Itemset. 2) It consumes more memory space and performs badly with long pattern data set.	are used to evaluate the high performance of the algorithm
4	Mining High Utility Itemsets without Candidate Generation[23]	2012	Mengchi Liu, Junfeng Qu	Hui-Miner	Single Phase Algorithm. No need to multiple times database scan	Calculating the utility of an itemset joining utility list is very costly.	We should try to avoid performing joins if possible for low-utility itemsets.
5	FHM: Faster High-Utility Itemset Mining using Estimated Utility Cooccurrence Pruning[24]	2014	Philippe FournierViger, ChengWei wu	FHM	Estimated-Utility Co-occurrence pruning	Static Database	We should try it using a dynamic database.
6	High Utility Itemset Mining from Transaction Database Using UP-Growth and UP-Growth+ Algorithm[25]	2015	Komal Surawase1 , Madhav Ingle2	Improved UP-Growth	1) Scan the database twice to construct a global UP Tree with the first two strategies 2) generate PHUIs from global UP -Tree and local UP-Trees by UPGrowth with the third and fourth strategies 3) identify actual high utility item sets from the set of PHUIs	Noisy data	The extension of the temporal utility pattern tree to mine noisy patterns, and developing more efficient techniques to handle genomic data.

IV. CONCLUSION

In this paper, we considered what are high utility itemset mining and their algorithms. The major advantage for identifying infrequent itemset was to advance the profit of rarely recognized datasets in the transactions. All the itemsets which are frequent or infrequent can have high utility. So utility is major concern of this research. Frequent items can make profit we are familiar with that fact but infrequent items can make profit it's a new idea.

In retail markets, the utility is the most crucial part of marketing. The key to attain this objective is to understand the consumer's behavior. For achieving the preset mission and vision of the new business strategies have to be developed frequently. To formulate the right business strategy, the key part is to understand the dynamic nature of the environment in which the business operates. In future identification of high utility rare itemsets will be a great idea by applying some optimization techniques for privacy preservation.

REFERENCES

[1] Abhijit Raorane, R.V.Kulkarni, "Data Mining Techniques: A Source For Consumer Behavior Analysis", International Journal of Database Management Systems, Vol.3,No.3,Aug. 2011, pp.45-56.

[2] Sudip Bhattacharya1 , Deepty Dubey, "High Utility Itemset Mining", International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, Volume 2, Issue 8, August 2012.

[3] H. Yao and H. J. Hamilton, "Mining itemset utilities from transactio databases," Data and Knowledge Engineering, vol. 59, pp. 603-626 2006.

[4] Farah Hanna AL-Zawaidah and Yosef Hasan Jbara, "An Improved Algorithm for Mining Association Rules in Large Databases", World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 1, No. 7, 311-316, 2011.

[5] Endu Duneja and A.K. Sachan, "A Survey on Frequent Itemset Mining with Association Rules", International Journal of Computer Applications (0975 – 8887) Volume 46– No.23, May 2012.

[6] Sujatha Kamepalli, Raja Sekhara Rao Kurra and Sundara Krishna.Y.K, "Apriori Based: Mining Infrequent and Non-Present Item Sets from Transactional Data Bases," International Journal of Electrical & Computer Science IJECS-IJENS Vol:14 No:03.

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

[8] 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.

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

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

[11] Pillai, Jyothi, and O. P. Vyas. "High Utility Rare Item Set Mining (HURI): An Approach for Extracting High Utility Rare Item Sets." Journal on Future Engineering and Technology 7, no. 1 (2011).

[12] 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

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

[14] Haglin, David J., and Anna M. Manning. "On Minimal Infrequent Itemset Mining." In DMIN, pp. 141-147. 2007.

[15] Mehdi Adda, Lei Wu, Sharon White(2012), Yi Feng ‖ Pattern detection with rare item-set mining‖ International Journal of Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol.1, No.1, August 2012.

[16] Younghee Kim, Wonyoung Kim and Ungmo Kim- "Mining Frequent Itemsets with Normalized Weight in Continuous Data Streams", Journal of Information Processing Systems, Vol.6, No.1, March 2010

[17] Luca Cagliero and Paolo Garza, " Infrequent Weighted Item set Mining Using Frequent Pattern Growth" IEEE

[18] Jyothi Pillai1 , O.P. Vyas2 and Maybin K. Muyeba, Int. J. on Recent Trends in Engineering and Technology, Vol. 10, No. 1,Jan,2014.

[19] Niha, S. A. R., and Uma N. Dulhare. "Extraction of high utility rare itemsets from transactional databases." Computer and Communications Technologies (ICCCT), 2014 International Conference on. IEEE, 2014.

[20] Itemsets", Ying Liu, Wei-Keng Liao, and Alok Choudhary, Northwestern University, Evans

[21] "CTU-Mine: An Efficient High Utility Itemset Mining Algorithm Using the Pattern Growth Approach" In: Seventh International Conference on Computer and Information Technology (2007).

[22] "UP-Growth: An Efficient Algorithm or High Utility Itemset Mining ", Vincent S. Tseng, Cheng-Wei Wu, Bai-En Shie, and Philip S. Yu. University of Illinois at Chicago, Chicago, Illinois, USA, 2010.

[23] Mengchi Liu Junfeng Qu, "Mining High Utility Itemsets without Candidate Generation", 2012.

[24] “FHM: Faster High-Utility Itemset Mining using Estimated Utility Co-occurrence Pruning”, Philippe Fournier-Viger1, Cheng-Wei Wu 2014.

[25] Dewarde, D. H., S. A. Kahate, and P. R. Chandre. "High Utility Itemsets Mining from Transactional Databases using UP-Growth and UP-Growth+ Algorithm."