High Utility Item Set Mining- A Review

Ankit Redwal[1] Anil Patidar[2]
[1,2]Acropolis Institute of Technology & Research, India

Abstract— Data Mining, also called knowledge Discovery in Database, is one of the latest research area, which has emerged in response to the Tsunami data or the flood of data, world is facing nowadays. It has taken up the challenge to develop techniques that can help humans to discover useful patterns in massive data. One such important technique is utility mining. Frequent item set mining works to discover item set which are frequently appear in transaction database, which can be discover on the basis of support and confidence value of different itemset. Using frequent itemset mining concept as a base, many researchers have also proposed different new concept on utility based mining of itemset. This paper presents a comprehensive systematic literature review of present techniques used for mining high utility item sets from huge data set.

Key words: Data Mining, KDD Process, High Utility Mining, Minimum Utility

I. INTRODUCTION

Data mining [1] has become an essential technology for businesses and researchers in many fields, the number and variety of applications has been growing gradually for several years and it is predicted that it will carry on to grow. A number of the business areas with an early embracing of DM into their processes are banking, insurance, retail and telecom. More lately it has been implemented in pharmaceutics, health, government and all sorts of e-businesses.

One describes a scheme to generate a whole set of trading strategies that take into account application constraints, for example timing, current position and pricing [2]. The authors highlight the importance of developing a suitable back testing environment that enables the gathering of sufficient evidence to convince the end users.

These organization sectors include retail, petroleum, telecommunications, utilities, manufacturing, transportation, credit cards, insurance, banking, decision support, financial forecast, marketing policies, even medical diagnosis and many other applications, extracting the valuable data, it necessary to explore the databases completely and efficiently. Due to its versatility (figure 1), we can also define data mining as a woven combination of statistics, artificial intelligence and machine learning.

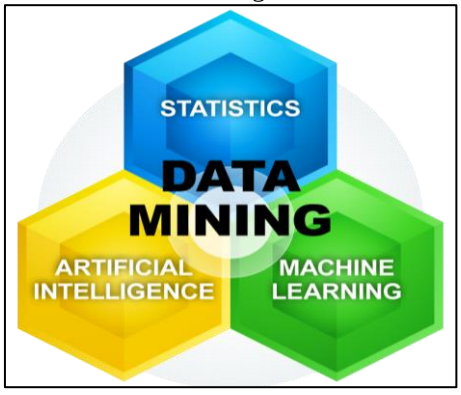

Fig. 1: What is Data Mining [3]

Data mining addresses two basic tasks: verification and discovery. The verification task seeks to verify user's hypotheses. While the discovery task searches for unknown knowledge hidden in the data. In general, discovery task can be further divided into two categories, which are descriptive data mining and predicative data mining. Descriptive data mining presents general summary of data and interesting patterns too. Predictive data mining constructs one or more models to be later used for predicting the behavior of future data sets.

There are a number of algorithmic techniques available for each data mining tasks, with features that must be weighed against data characteristics and additional business requirements. Among all the techniques, in this research, we are focusing on the association rules mining technique, which is descriptive mining technique, with transactional database system. This technique was formulated by [4] and is often referred to as market-basket problem.

In utility mining [5] we concentrate on utility value of itemset while in frequent item set mining we concentrate that how frequently items appears in transactional database. With the help of following example describe in table 1, can easily differentiate utility mining and frequent item set mining:-

Transaction I	Quantity of item sold in Transaction		
T1	0	0	1
T2	2	0	2
T3	1	1	4
T4	0	1	1
T5	5	1	3

Table 1: Transactional Database D1

Unit profit related with each item is described in table 2 as follows:

Item Name	Unit profit
A	6
B	120
C	45

Table 2: Unit Profit Associate with Items

Now with the help of internal utility, external utility and how many times item or itemset appears in transaction, we can calculate support and profits which describe in table 3 as follows:

Itemset	Support (%)	Profit (INR)
A	60	48
B	60	360
C	100	495
AB	40	276
AC	60	768
BC	60	720
ABC	40	456

Table 3: Support & Profits for All Items

If [5] minimum support = 40 % only A, B, C, AC, BC qualify as frequent itemsets. ({ABC}) = (1×6+1×120 +4×45) + (5×6+ 1×120+3×45) = 456.If we specified user threshold value =310 then ABC is a high utility itemset but it is not a frequently accessible itemset.

II. LITERATURE REVIEW

Some FP tree based and other tree based methods for high utility mining were proposed in [6] [7] [8]. All these methods were simple. They used a special data structure CP pro. Tree construction took extra time specially when the data set is too large.

In 2010 the author ZHOU Jun et al. [9] proposed this algorithm by considering the space as an important factor. Authors used an improved LRU (Least Recently Used) based algorithm. Proposed algorithm omits the infrequent items before taken for the processing. Method increases the stability and the performance. Method is used to find out the frequent items as well as the frequency of those items.

Most of the existing algorithms uses a measure known as TWU (Transaction Weighted Utility). This measure was introduced Liu et al. [10], also they follow the process of two phase candidate generation. The work done in [11] proposed an isolated item discarding strategy. If any size k item set does not contain an item I then item I is termed as an isolated item.

Authors in [12] proposed a projection based method for mining high utility items. This is improvement of two phase algorithm. It speeds up the execution of two phase algorithm. Authors in [13] proposed a hybrid algorithm, a combination of ant monotonicity of TWU and pattern growth approach.

Work done in [14] proposed a FP tree based algorithm, this algorithm uses a tree to maintain the TWU information. It also uses the concept of pruning to eliminate the useless items from the first phase of the algorithm. Number of candidate item sets degrades the mining performance in terms of execution time and space requirement. This algorithm works poor when the data set is having long high utility item sets.

Apriori algorithm for mining high utility items sets was proposed in [15]. It first generates all the probable high utility candidates. Then this algorithm makes use of minimum utility threshold to prune infrequent items. This task is done one size at a time. The problem with this algorithm is too many data base scans. Also it is generate and test kind of method. Such kind of method is suitable only for the small size data sets.

Effective [16] disclosure of item sets with high utility like benefits manages the mining high utility item sets from an exchange database Although various important methodologies have been proposed as of late, these calculation acquire the issue of creating an extensive number of competitor item sets for high utility item sets and most likely debases the mining execution as far as execution time and memory space. In this paper, we propose two calculations, viz., utility example development (UP-Growth) and Improved UP-Growth i.e. Enhanced Utility Pattern Growth, for mining high utility item sets with an arrangement of successful systems for pruning applicant item sets. The data of high utility item sets is kept up in a smaller tree-based information structure utility example tree (UP-Tree), it filter the first database twice to oversee information organized way. Proposed calculations, particularly Improved UP Growth, decrease the quantity of applicants adequately as well as outflank different calculations considerably as far as runtime and memory utilization, particularly when databases contain loads of long exchanges .Mining is only used in "Transactions" Data Set.

Mining [17] exceptionally used thing sets from a value-based dB intends to find the thing sets with high utility as benefits. In spite of the fact that various Algorithms have been created yet they bring about the issue as it produce huge arrangement of applicant Item sets likewise require number of database output. In regular thing set mining the unit benefits and bought amounts of the things are not taken into contemplations and weighted mining benefit is not viewed as just weight is to be considered. Expansive number of Item sets decreases the execution of mining as for execution time and space prerequisite. At the point when database contains countless this circumstance becomes more awful. In proposed framework for make UP-tree and UP-tree mining calculations named as Up-Growth and Improved Up-Growth the data of very used thing sets is recorded in tree based information structure called Utility Pattern Tree which is a minimal tree representation of things in exchange database. With the assistance of Utility Pattern Tree, applicant thing sets produced inside just two sweeps of the database. Proposed calculations not just decrease various competitor thing sets additionally spare memory and time.

It generates huge set of Potential High Utility Item sets.

III. CONCLUSION

High utility frequent pattern mining has a wide range of real world applications. That's why it is one of the most favorite topic of research. Utility mining helps in mining of items which are worthy. This paper presented a systematic literature survey of present high utility frequent pattern mining algorithms. This paper also elaborated the notion of high utility mining in lucrative manner. It is found that although a lot of work is going on in the field of high utility mining but still there is enough scope to improve the performance.

REFERENCES

[1] Tan P.-N., Steinbach M., and Kumar V. ―Introduction to data mining, Addison Wesley Publishers‖. 2006
[2] Fayyad U. M., Piatetsky-Shapiro G. and Smyth, P. ―Data mining to knowledge discovery in databases, AI Magazine‖. Vol. 17, No. 3, pp. 37-54, 1996.
[3] https://www.sas.com/en_us/insights/analytics/data-mining.html
[4] C. F. Ahmed, S. K. Tanbeer, B.-S. Jeong, and Y.-K. Lee. Efficient tree structures for high utility pattern mining in incremental databases. In IEEE Transactions on Knowledge and Data Engineering, Vol. 21, Issue 12, pp. 1708-1721, 2009.
[5] A. Erwin, R. P. Gopalan, and N. R. Achuthan. Efficient mining of high utility itemsets from large datasets. In Proc. of PAKDD 2008, LNAI 5012, pp. 554-561.
[6] Y. G. Sucahyo and R. P. Gopalan. "CT-ITL: Efficient Frequent Item Set Mining Using a Compressed Prefix Tree with Pattern Growth". Proceedings of the 14th Australasian Database Conference, Adelaide, Australia, 2003.

[7] Y. G. Sucahyo and R. P. Gopalan. "CT-PRO: A Bottom Up Non Recursive Frequent Itemset Mining Algorithm Using Compressed FP-Tre Data Structure‖. In proc Paper presented at the IEEE ICDM Workshop on Frequent Itemset Mining Implementation (FIMI), Brighton UK, 2004.
[8] A.M.Said, P.P.Dominic, A.B. Abdullah. ―A Comparative Study of FP-Growth Variations‖. In Proc. International Journal of Computer Science and Network Security, VOL.9 No.5 may 2009.
[9] ZHOU Jun, CHEN Ming, XIONG Huan A More Accurate Space Saving Algorithm for Finding the Frequent Items.IEEE-2010.
[10] Y. Liu, W. Liao, and A. Choudhary, "A fast high utility itemsets mining algorithm," in Proc. Utility-Based Data Mining Workshop SIGKDD, 2005, pp. 253–262.
[11] 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, 2008.
[12] G.-C. Lan, T.-P. Hong, and V. S. Tseng, "An efficient projectionbased indexing approach for mining high utility itemsets," Knowl. Inf. Syst., vol. 38, no. 1, pp. 85–107, 2014.
[13] A. Erwin, R. P. Gopalan, and N. R. Achuthan, "Efficient mining of high utility itemsets from large datasets," in Proc. 12th Pacific-Asia Conf. Adv. Knowl. Discovery Data Mining, 2008, pp. 554–561.
[14] V. S. Tseng, B.-E. Shie, C.-W. Wu, and P. S. Yu, "Efficient algorithms for mining high utility itemsets from transactional databases," IEEE Trans. Knowl. Data Eng., vol. 25, no. 8, pp. 1772–1786, Aug. 2013
[15] Cheng-Wei Wu, Philippe Fournier-Viger, Philip S. Yu, Fellow, IEEE, Vincent S. Tseng, "Efficient algorithms for mining the concise and lossless representation of high utility item sets," IEEE Trans. Knowl. Data Eng., vol. 27, no. 3, pp. 726–739, Mar. 2014.
[16] Miss. A. A. Bhosale , S. V. Patil, Miss. P. M. Tare, Miss. P. S. Kadam"High Utility Item sets Mining on Incremental Transactions using UP-Growth and UP-Growth+ Algorithm":
[17] Switi Chandrakant Chaudhari, Vijay Kumar Verma, Mining High Utility Item Set From Large Database: A Recent Survey, International journal of Emerging Technology and Advanced Engineering, Website: www.ijetae.com (ISSN 2250-2459,ISO 9001:2008 Certified Journal, Volume 3, Issue 5, May 2013)

