A CRITICAL REVIEW OF VARIOUS METHODOLOGIES FOR MINING HIGH UTILITY ITEM SETS FROM A UTILITY DATA SET

Arundhati Roy, Amitav Ghosh
The God of Small Things Studied at Delhi School of Architecture; Political activist; Booker Prize winner
The Shadow Lines Former professor .

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 item set. Using frequent item set mining concept as a base, many researchers have also proposed different new concept on utility based mining of item set. This paper presents an analysis of various methodologies used for mining high utility item sets from a utility data set.

Keywords: *Data Mining, High Utility Mining, Minimum Utility, 2 phase algorithm.*

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.

Figure 1: Data Mining [3]

In utility mining [4] we concentrate on utility value of itemset while in frequent item set mining we concentrate that how frequently items appears in transactional database.

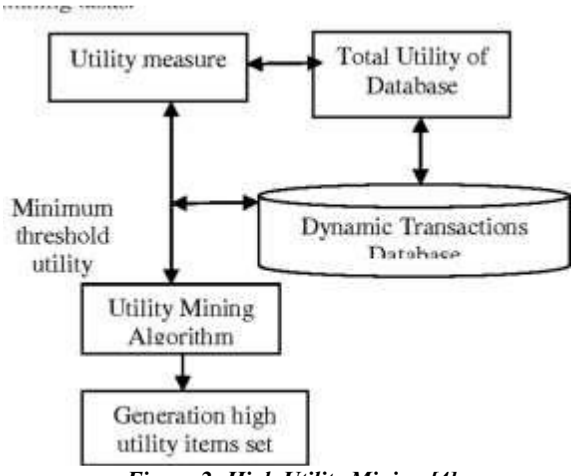

Figure 2: High Utility Mining [4]

II. LITERATURE SURVEY

The main objective of Utility Mining is to identify the item sets with highest utilities, by considering profit, quantity, cost or other user preferences. Mining High Utility item sets from a transaction database is to find item sets that have utility above a user-specified threshold. Item set Utility Mining is an extension of Frequent Item set mining, which discovers item sets that occur frequently. In many real-life applications, high-utility item sets consist of rare items. Rare item sets provide useful information in different decision-making domains such as business transactions, medical, security, fraudulent transactions, retail communities. For example, in a supermarket, customers purchase microwave ovens or frying pans rarely as compared to bread, washing powder, soap. But the former transactions yield more profit for the supermarket. Similarly, the high-profit rare item sets are found to be very useful in many application areas.

CT-PRO is also the variation of classic FP-tree algorithm [6]. It is based upon the compact tree structure [6, 7]. This algorithm uses bottom up approach for performing tree traversal. This is not a recursive method. Compress tree structure is also the prefix tree in which all the items are stored in the descending order of the frequency with the field index, frequency, pointer, item-id [8]. 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 Weigheted 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 antimonotonicity 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.

III. CONCLUSION

Mining High Utility item sets from a transaction database is to find item sets that have utility above a user-specified threshold. Itemset Utility Mining is an extension of Frequent Itemset mining, which discovers item sets that occur frequently. In many real-life applications, high-utility item sets consist of rare items. Rare item sets provide useful

information in different decision-making domains such as business transactions, medical, security, fraudulent transactions, retail communities. This paper presented a review of high utility item set mining in a lucrative manner.

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 item sets 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 Compress ed 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 item sets 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 item sets," Data Knowl. Eng., vol. 64, no.1, pp. 198–217, 2008.*
12. *G.-C. Lan, T.-P. Hong, and V. S. Tseng, "An efficient projection based indexing approach for mining high utility item sets," 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 item sets 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 item sets from transactional databases," IEEE Trans. Knowl. Data Eng., vol. 25, no. 8, pp. 1772–1786, Aug. 2013.*