International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 3, Issue 4, July-August 2014 **ISSN 2278-6856**

Survey on High Utility Itemset Mining from Large Transaction Databases

Ms. Yogita Khot[1], Asst. Prof. Mrs. Manasi Kulkarni[2]

[1,2] PES's Modern College of Engineering, University of Pune, PUNE, Maharashtra, India

Abstract

Data mining can be defined as an activity that extracts some knowledge contained in large transaction databases. Conventional data mining techniques have focused largely on finding the items that are more frequent in the transaction databases, which is also called frequent itemset mining. These data mining techniques were based on support-confidence model. Itemsets which appear more frequently in the database must be of more meaning to the user from the business point of view. In this paper we present an emerging area called as High Utility Itemset Mining that discovers the itemsets considering not only the frequency of the itemset but also utility associated with the itemset. Every itemset have a value like quantity, profit and other user's interest. This value associated with every item in a database is called the utility of that itemset. Those itemsets having utility values greater than given threshold are called high utility itemsets. This problem can be identified as mining high utility itemsets from transaction database. In many areas of business like retail, inventory etc. decision making is very important. So it can help in mining algorithm, the presence of each item in a transaction database is represented by a binary value, without considering its quantity or an associated weight such as price or profit. However quantity, profit and weight of an itemset are significant for identifying real world decision problems that require increasing the utility in an organization. Mining high utility itemsets from transaction database presents a greater challenge as compared with frequent itemset mining, since anti-monotone property of frequent itemsets is not applicable in high utility itemsets.

In this paper, we present a survey on the current state of research and the various algorithms and techniques for high utility itemset mining.

Keywords: Data Mining, Frequent Itemset Mining, Utility Mining, High Utility Itemset Mining.

1. INTRODUCTION

1.1 Frequent Itemset Mining

High utility itemset mining discovers all high utility itemsets with utility values higher than the minimum utility threshold in a transaction database [14]. The utility of an itemset refers to its associated value such as profit, quantity or some other related measure. Some standard methods for mining association rules [1, 7] that is finding frequent itemsets are based on the support confidence model. They find all frequent itemsets from given database. The problem of frequent itemset mining [1, 2] is finding the complete set of itemsets that appear with high occurrence in transactional databases. However the utility of the itemsets is not considered in ordinary frequent itemset mining algorithms. Frequent itemset mining only considers whether an item has occurred frequently in database, but ignores both the quantity and the utility associated with the item. However, the occurrence of an itemset may not be an adequate indicator of interestingness, because it only shows the number of transactions in the database that contains the itemset. It does not reveal the actual utility of an itemset, which can be measured in terms of cost, quantity, profit, or other expressions of user preference [17]. However, utility of an itemset like profit, quantity and weight are important for addressing real world decision problems that require maximizing the utility in an organization. In many areas of business like retail, inventory, marketing research etc. decision making is very important. So it can help in analysis of sales, marketing strategies, and designing different types of catalog.

EXAMPLE 1

Consider the small example of transaction database, a customer buys multiple items of different quantities in a sale transaction. In general, each item has a certain level of profit. For instance, assume that in an electronic superstore, the profit (in INR) of 'Printer Ink' is 5, and that of 'Laser Printer' is 30. Suppose 'Printer Ink' occurs in 6 transactions, and 'Laser Printer' occurs in 2 transactions in a transactional database. In frequent itemset mining, the occurrence frequency of 'Printer Ink' is 6, and that of 'Laser Printer' is 2. 'Printer Ink' has a higher frequency. Nevertheless, the total profit of 'Laser printer' is 60, and that of 'Printer ink' is 30; therefore, 'Laser Printer' contributes more to the profit than 'Printer Ink'. Frequent itemsets are simply itemsets with high frequencies without considering utility. However, some infrequent itemsets may also contribute more to the total profit in the database than the frequent itemsets. This example shows the fact that frequent itemset mining approach may not always satisfy the retail business objective. In reality a most valuable customers who may buy full priced items or high margin items which may not present from large number of transactions are very important for retail business because they do not buy these items frequently.

1.2 High Utility Itemset Mining

The limitation of frequent itemset mining lead researchers towards utility based mining approach, which allows a user to conveniently express his or her perspectives concerning the usefulness of itemsets as utility and then find itemsets with high utility values higher than given threshold [3]. During mining process we should not identify either frequent or rare itemsets but identify itemsets which are more useful to us. Our aim should be in identifying itemsets which have higher utilities in the database, no matter whether these itemsets are frequent itemsets or not. This leads to a new approach in data mining which is based on the concept of utility called as utility mining. High utility itemset mining refers to the discovery of high utility itemsets. The main objective of high utility itemset mining is to identify the itemsets that have utility values above given utility threshold [14]. The term utility refers to its associated profit or some other related measure [16]. In practice the utility value of an itemset can be profit, quantity, weight, popularity, page-rank, measure of some aesthetic aspect such as beauty or design or some other measures of user's preference [17].

2. LITERATURE REVIEW

In this section we present a brief review of the different algorithms, techniques, concepts and approaches that have been defined in various research journals and publications. Agrawal, R., Imielinski, T., Swami, A. [1] proposed Frequent itemset mining algorithm that uses the Apriori principle. Standard method is based on Support-Confidence Model. Support measure is used. An anti-monotone property is used to reduce the search space. It generates frequent itemsets and finds association rules between items in the database. It does not identify the utility of an itemset [1]. Yao, H., Hamilton, H.J., Buzz, C.J. [2] proposed a framework for high utility itemset mining. They generalize previous work on itemset share measure [2]. This identifies two types of utilities for items, transaction utility and external utility. They identified and analyzed the problem of utility mining. Along with the utility bound property and the support bound property. They defined the mathematical model of utility mining based on these properties. The utility bound property of any itemset provides an upper bound on the utility value of any itemset. This utility bound property can be used as a heuristic measure for pruning itemsets as early stages that are not expected to qualify as high utility itemsets [2]. Yao, H., Hamilton, H.J., Buzz, C.J. [3] proposed an algorithm named Umining and another heuristic based algorithm UminingH to find high utility itemsets. They apply pruning strategies based on the mathematical properties of utility constraints. Algorithms are more efficient than any previous utility based mining algorithm. Liu, Y., Liao, W.K., Choudhary A. [4] proposed a two phase algorithm to mine high utility itemsets. They used a transaction weighted utility (TWU) measure to prune the search space. The algorithms based on the candidate generation-and-test approach. The proposed algorithm suffers from poor performance when mining dense datasets and long patterns much like the Apriori [1]. It requires minimum database scans, much less memory space and less computational cost. It can easily handle very large databases. Erwin, A., Gopalan, R.P., N.R. Achuthan [5] proposed an efficient CTU-Mine Algorithm based on Pattern Growth approach. They introduce a compact data structure called as Compressed Transaction Utility tree (CTU-tree) for utility mining, and a new algorithm called CTU-Mine for mining high utility itemsets. They show CTU-Mine works more efficiently than TwoPhase for dense datasets and long pattern datasets. If the thresholds are high, then TwoPhase runs relatively fast compared to CTU-Mine, but when the utility threshold becomes lower, CTUMine outperforms TwoPhase. Erwin, A., Gopalan, R.P., N.R. Achuthan [7] proposed an efficient algorithm called CTU-PRO for utility mining using the pattern growth approach. They proposed a new compact data representation named Compressed Utility Pattern tree (CUP-tree) which extends the CFP-tree of [11] for utility mining.TWU measure is used for pruning the search space but it avoids a rescan of the database. They show CTU-PRO works more efficiently than TwoPhase and CTU-Mine on dense data sets. Proposed algorithm is also more efficient on sparse datasets at very low support thresholds.TWU measure is an overestimation of potential high utility itemsets, thus requiring more memory space and more computation as compared to the pattern growth algorithms. Erwin, R.P. Gopalan, and N.R. Achuthan [14] proposed an algorithm called CTU-PROL for mining high utility itemsets from large datasets. They used the pattern growth approach [6]. The algorithm first finds the large TWU items in the transaction database and if the dataset is small, it creates data structure called Compressed Utility Pattern Tree (CUP-Tree) for mining high utility itemsets. If the data sets are too large to be held in main memory, the algorithm creates subdivisions using parallel projections that can be subsequently mined independently. For each subdivision, a CUP-Tree is used to mine the complete set of high utility itemsets. The anti-monotone property of TWU is used for pruning the search space of subdivisions in CTU-PROL, but unlike TwoPhase of Liu et al. [4], CTU-PROL algorithm avoids a rescan of the database to determine the actual utility of high TWU itemsets. The performance of algorithm is compared against the TwoPhase algorithm in [4] and also with CTU-Mine in [5]. The results show that CTU-PROL outperforms previous algorithms on both sparse and dense datasets at most support levels for long and short patterns.

3. CONCLUSIONS

The Frequent itemset mining is based on the principle that the itemsets which appear more frequently in the transaction databases are of more importance to the user.

However in reality the benefit of frequent itemset mining by considering only frequency of itemset is challenged in many research areas such as retail, marketing etc. It has been seen that in many real application domains that the itemsets that contribute the most are not necessarily the frequent itemsets. Utility mining is an era of research which tries to bridge this gap by using item utilities as an analytical measurement of the importance of that item in the user's point of view. In this paper we have presented a brief review of the various algorithms and techniques for mining of high utility itemsets from transaction databases. Most of algorithms are focused on reducing the search space while searching for the high utility itemsets.

References

[1] Agrawal, R., Imielinski, T., Swami, A., "Mining Association Rules between Sets of Items in Large Database", In: ACM SIGMOD International Conference on Management of Data (1993).

[2] Yao, H., Hamilton, H.J., Buzz, C. J., "A Foundational Approach to Mining Itemset Utilities from Databases", In: 4th SIAM International Conference on Data Mining, Florida USA (2004).

[3] Yao, H., Hamilton, H.J., "Mining itemset utilities from transaction databases", Data & Knowledge Engineering 59(3), 603–626 (2006).

[4] Liu, Y., Liao, W.K., Choudhary, A., "A Fast High Utility Itemsets Mining Algorithm", In: 1st Workshop on Utility-Based Data Mining, Chicago Illinois (2005).

[5] Erwin, A., Gopalan, R.P., N.R. Achuthan, "CTU-Mine: An Efficient High Utility Itemset Mining Algorithm Using the Pattern Growth Approach", In: IEEE CIT 2007.Aizu Wakamatsu, Japan (2007).

[6] Han, J., Wang, J., Yin, Y., "Mining frequent patterns without candidate generation", In: ACM SIGMOD International Conference on Management of Data (2000).

[7] Erwin, A., Gopalan, R.P., Achuthan, N.R, "A Bottom-Up Projection Based Algorithm for Mining High Utility Itemsets", In: International Workshop on Integrating AI and Data Mining. Gold Coast, Australia (2007).

[8] CUCIS. Center for Ultra-scale Computing and Information Security, Northwestern University, http://cucis.ece.northwestern.edu/projects/DMS/MineBenchDownloa d.html.

[9] Yao, H., Hamilton, H.J., Geng, L., "A Unified Framework for Utility Based Measures for Mining Itemsets", In: ACM SIGKDD 2nd Workshop on Utility-Based Data Mining (2006).

[10] Pei, J., "Pattern Growth Methods for Frequent Pattern Mining", Simon Fraser University (2002).S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," IEEE Electron Device Lett., vol. 20, pp. 569–571, Nov. 1999.

[11] Sucahyo, Y.G., Gopalan, R.P., CT-PRO: "A Bottom-Up Non Recursive Frequent Itemset Mining Algorithm Using Compressed FP-Tree Data Structure", In: IEEE ICDM Workshop on Frequent Itemset Mining Implementation (FIMI), Brighton UK (2004).

[12] G. Salton, Automatic Text Processing, Addison-Wesley Publishing, 1989.

[13] J. Pei, J. Han, L.V.S. Lakshmanan, "Pushing convertible constraints in frequent itemset mining", Data Mining and Knowledge Discovery 8 (3) (2004) 227–252.

[14] A. Erwin, R.P. Gopalan, and N.R. Achuthan, "Efficient Mining of High Utility Itemsets from Large Datasets", T. Washio et al. (Eds.): PAKDD 2008, LNAI 5012, pp. 554–561, 2008. © Springer-Verlag Berlin Heidelberg 2008.

[15] Bin Chen, Peter Hass, Peter Scheuermann, "A New Two-Phase Sampling Based Algorithm for Discovering Association Rules", SIGKDD '02 Edmonton, Albetra, Canada © 2002 ACM 1 58113 567 X/02/2007.

[16] Ming-Yen lin, Tzer-Fu Tu, Sue-Chen Hsueh, "High utility pattern mining using the maximal itemset property and, lexicographic tree structures", Information Science 215(2012) 1-14.

[17] Sudip Bhattacharya, Deepty Dubey, "High utility itemset mining, International Journal of Emerging Technology and advanced Engineering", ISSN 2250-2459, Volume 2, issue 8, August 2012.

AUTHOR

Ms. Yogita Khot received the B.E. degree in Computer Science and Engineering from Government College of Engineering Aurangabad, Maharashtra, India in 1999 and pursuing M.E. in Computer Engineering from University of Pune. She is working as a Lecturer in Computer Engineering Department in PES's Modern College of Engineering, Pune, Maharashtra, India.

Prof. Mrs. Manasi Kulkarni currently working as Assistant Professor in PES's Modern College of Engineering, Pune, Maharashtra, India. She received B.E. degree in Computer Science and Engineering from PDA College of Engineering, Gulbarga and M.Tech. in Computer Science and Engineering from Visvesvaraya Technological University, Belgaum, Karnataka. She worked as a Lecturer in National Institute of Engineering College, Mysore, Karnataka. She has total teaching experience of 15 years. Her area of interest is Computer Graphics, Databases.