MHU-ISM: MINING HIGH UTILITY ITEM SETS MINING IN TRANSACTIONAL DATABASES

Mohamed Jamshad K, Assistant Professor of Computer Science, EMEA College of Arts & Science, Kondotty, Malappuram(dist.), Kerala

Abstract

High utility pattern mining is the discovery of groups of patterns that not only co-occur but also carry a high return. For candidate creation in two-phase pattern mining, an Apriori method is employed. However, candidate generation is expensive, and if the number of candidates is large, scalability and efficiency become bottleneck issues. Many attempts have been made to reduce the number of candidates produced in the first phase, but the problem remains when the raw data includes a large number of transactions or the minimal utility criterion is too low. The MHU-ISM technique for High Utility Mining was suggested in this article. This takes into account not just the frequency of the itemsets but also the usefulness connected with the itemsets. The word utility relates to the significance or usefulness of the appearance of the itemset in transactions, as measured in measures such as profit, sales, or any other user preferences. The goal of High Utility Itemset Mining is to find itemsets with utility values greater than a certain threshold. Some high utility itemset mining methods, such as second-Phase and FP-Growth, have been suggested in current systems. This approach is a memory-efficient way for mining high-utility item sets from transactional databases. This method uses less memory space and takes less time to execute than current techniques.

Keywords: HUI, FP-Growth, Utility pattern mining, HTM, MHU-ISM

I. INTRODUCTION

Discovering product information in the market is a critical job that requires a comprehensive examination of each product's information from the user's viewpoint. Identifying common itemsets from a large database required much research. The Apriori algorithm is the most widely used pattern mining algorithm. It is a breadth-first search method that scans the database as many times as the length of the most frequently occurring pattern [1].

a. Mining of Frequent Itemsets

An itemset is a non-empty collection of items. A k-itemset is an itemset that contains k different items. In a grocery transaction, for example, bread, butter, and milk may indicate a three-item set. The itemsets that occur often in transactions are known as frequent itemsets. The objective of frequent itemset mining is to find all itemsets in a transaction dataset. Frequent itemset mining [2] [3] is critical in the theory and implementation of many key data mining tasks, including mining association rules and lengthy patterns. The frequency criteria are stated in terms of the itemsets' support value. An itemset's Support

value is the proportion of transactions that include the itemset. Association rule mining benefits from frequent pattern mining [4].

b. Mining Association Rules

Association Rule Mining is a well-known method for discovering co-occurrences, correlations, common patterns, and connections between sets of objects in a transaction database. The identification of intriguing correlation connections among massive amounts of business transaction data may aid in a variety of corporate decision-making processes, such as catalogue design, consumer shopping behavior analysis, and so on [5]. An association rule is an expression of the form XY, where X and Y are collections of things known as itemsets [6]. It implies that if a client buys X, he or she will also purchase Y. Support and confidence is two metrics that indicate the certainty of found association rules [7].

c. Mining for Utility

Utility mining [11] has the potential to be helpful in a variety of practical applications. To overcome the restriction of association rule mining, a utility-based mining model was developed, which enables a user to express his or her views on the usefulness of itemsets as utility and then identify itemsets with utility values greater than a specified threshold. The word utility in utility-based mining refers to the quantitative representation of user choice, i.e. the utility value of an itemset is a measurement of the significance of that itemset in the user's viewpoint. For example, if a sales analyst doing retail research wants to determine which itemsets in the shops generate the most sales income for the stores, he or she would define the utility of any itemset as the monetary profit earned by the business by selling each unit of that itemset [8] [9]. However, HUI mining is difficult since the downward closure feature of FIM does not apply to utility mining. In other words, since a superset of a low utility itemset may be a high utility itemset, the search area for mining HUIs cannot be immediately decreased as it is in FIM. Many research for mining HUIs have been suggested; however they often provide consumers with a huge number of high utility itemsets [10]. The presence of a significant number of high utility itemsets makes it harder for consumers to understand the findings. It may also lead the algorithms to become inefficient in terms of time and memory requirements, or even cause them to run out of memory. It is generally acknowledged that the higher the usefulness. To create a high utility item set, the utility mining method first calculates the overall utility of the database and then compares it to a specified minimum threshold value.

II. BACKGROUND STUDY

R.Agrawal and R. srikant [1] addressed the difficulties of deriving association rules between products in massive datasets of sales transactions in 1994. In this article, two algorithms, Apriori and AprioriTid, are presented to address the issue using other methods. Both methods are combined to form a hybrid algorithm. It is referred to as the "AprioriHybrid" algorithm. The scalability of the AprioriHybrid algorithm is unique. Another issue addressed in this article is the problem of basket data. It houses the massive application database. Multiple sweeps over the data are required to find an n-number of itemsets. It identifies the individual itemset with the least amount of support right away.

In terms of candidate itemsets, the suggested algorithms Apriori and AprioriTid vary from the AIS and SETM algorithms. One extra attribute is utilized in the AprioriTid algorithm to count the support of candidate itemsets after the first pass. The performance of AprioriHybrid is important to the Apriori and AprioriTid algorithms for three datasets. In all instances, the suggested AprioriHybrid performs better than the Apriori. AprioriHybrid performs somewhat worse than the Apriori algorithm in the final pass switching. As a result, the AprioriTid method is applied after each space.

R. Hilderman, Colin L. et al. [2] developed a shared confidence framework in 1998. It is the foundation for extracting information from databases. It also tackles the issue of identifying itemsets in market basket data. They focused on two kinds of objectives in this article, the first of which was to establish measurements of itemset. This is a helpful and practical interactive measure for frequently used support measures. Second, the finding of profiles of consumers' purchasing habits, which is accomplished by categorizing them. The suggested method combined the Apriori algorithm in order to find association rules between big databases itemsets. An experimental results analysis in this article shown that the suggested share confidence framework may provide more information feedback than the support confidence framework.

M.J. Zaki and C.J. Hsiao [4] represented CHARM in 2000. It is a fast method for mining the most frequently occurring itemsets. Pattern mining is often used to uncover association rules, strong rules, multidimensional patterns, and other significant discoveries. Of order to solve the issue in frequent pattern mining. To enumerate the various frequent itemsets, an apriori method uses the BFS, or Breadth First Search. The apriori method uses the downward closure feature to reduce the search space. In this article, two types of methods for mining lengthy patterns are presented. The first approach aims to find maximum frequent patterns with fewer magnitudes than all frequent patterns, while the second technique mines frequent closed itemsets. CHARM, the suggested method, found the itemsets and transaction space over a new tree known as the itemset tree (IT). It employs a hash-based method to remove non-closed itemsets during assumption testing.

J. Pei, J. Han, and others [8] described the FP-growth algorithm in 2004. They mostly contribute to this article by demonstrating proper item ordering. The author demonstrated the efficacy of the suggested method in this article. The suggested technique is a methodical approach to incorporating two phases of class restrictions. The idea of convertible restrictions is presented in this article. Convertible restrictions are classified as convertible anti-monotone, convertible monotone, and highly convertible. This number of appropriate restrictions is discussed. Convertible restrictions cannot be put into the basic Apriori framework, but they may be pushed into frequent pattern growth mining. As a result, they created a quick mining method with different restrictions in order to mine common patterns.

Ying Liu and W.K. Liao [9] represented the Association Rule Mining method in 2005. It finds common itemsets in a big database and considers individual items to create association rules. ARM simply represents the frequency of an item's

presence and absence. To find common itemsets, an anti-monotone characteristic is employed. Mining with Expected Utility (MEU) is a technique for narrowing the search field by expecting high utility k-itemsets. They examined the scalability and correctness of findings in the experimental analysis part. Finally, it seems that the Two-phase method in this study can effectively extract HUI. L. Geng and H. Hamilton [10] investigated the common itemsets in 2006. They developed a well-known method for identifying frequently occurring itemsets. For reducing the search space of itemsets, the Apriori method is employed. Different metrics of interest in the area of data mining have been presented in this article. The three objectives mentioned above are subjective and semantic based measurements that deal with the user's prior knowledge and intentions. These metrics are appropriate for user experience and interactive data mining. However, in the case of frequent mining, the actual human interest remains an open and difficult topic.

a) PROBLEM STATEMENT

Many techniques for mining high utility item sets have been devised. Two major issues are constantly under consideration: how to reduce the number of possibilities and how to eliminate space and time complexity.

III. PROPOSED MODEL

Mining is a major job in every application, according to the suggested framework, and the performance of mining with various Algorithms such as one phase and multiple phases. The suggested method is a very efficient closed HUI mining algorithm. It is a one-phase method developed with the stringent design constraint that all operations for each itemset in the search space must be done in linear time and space. The concept given in this article is to split the database into several divisions, then identify frequent free item sets in each partition, then combine the partitions to create more frequent free item sets and count the support. The method uses minimal memory to store extra support numbers of item sets in each partition while significantly reducing the duration of item set matching, which is the mining process's bottleneck. The high utility item sets have been provided in this suggested system, as have the applications of opinion mining such as positive comments and association rules.

a) The Search Space

The search space of all itemsets may be represented using a set-enumeration tree. Figure 2 depicts the set-enumeration tree of I = a, b, c, d for the lexicographical order. The suggested method investigates this search space by doing a depth-first search from the root (the empty set). During this depth-first search, EFIM-Closed recursively appends one item at a time to according to the order, to create bigger itemsets for each itemset. The order is specified in this implementation as the order of rising TWU since it usually decreases the search space for HUIM. This paper next provides definitions for the depth-first search investigation of itemsets.

b) ASSOCIATION RULES MINING

Let I= $\{i_1, i_2, \ldots, i_n\}$ be a set of items ,i_j ($1 \leq j \leq n$) is an item of the dataset. Given dataset DB= $\{t_1, t_2, \ldots, t_n\}$, DB denotes the set of transactions. Each transaction t_i= {tid, A} has a unique identifier tid and a set of items, A U I.

Suppose that X,Y_I and $X\ Y=\Phi$ then $X\ Y$ is an association rule. The support of this rule is denoted as Association rules mining means finding all association rules which support the rules and whose confidences are greater than minimum support s_{min} and minimum confidence c_{min}, respectively. The association rules mining that Agrawal had proposed can be divided two steps. The first step is to search the item sets whose supports are large than the threshold value and account the confidence. The second step is to produce the rules taking advantage of the information from the first step. The first step is the main step of association rule mining that is to find all the frequent item sets.

c) FREQUENT PATTERN SETS MINING

To begin, consider this example. S(X) indicates that item set X is supported in database D. If S(AB)=S(A), that is, if the number of times item set AB occurs is the same as the number of times item set B appears, item A and B appear together. S(ABC)=S(AC), S(ABD)=S(AD), and S(ABCD)=S(ABCD) (ACD). To get the support of any superset of item set AB, this paper do not need to count all of the item set AB by scanning the database, but this paper can simply account for it directly by taking use of the free item set feature. This is the idea behind generating the whole frequent item set via mining frequent free item sets.

d) High-utility Transaction Merging (HTM)

To further decrease the cost of database scans, EFIM-Closed introduces High-utility Transaction Merging, an efficient transaction merging method (HTM). The concept of HTM is founded on the fact that transaction databases often include similar transactions (transactions containing exactly the same items, but not necessarily the same internal utility values). The technique consists of replacing a set of identical transactions T r_1, T r_2, ...T r_m in a (projected) database α-D by a single new transaction TM = T r_1 = Tr_2 = ... = Trm where the quantity of each item i TM is defined as q(i, TM) = P k=1...m q(i, Trk).

e) Pruning Non Closed HUIs

This paper will now go through the methods that EFIM-closed employs to prune non-closed HUIs. A naïve method to finding just CHUIs would be to retain all HUIs discovered up to this point in memory. Then, whenever a new HUI is discovered, the algorithm compares it to previously discovered HUIs to determine if (1) the new HUI is included in a previously discovered HUI or (2) some previously discovered HUI(s) are included in the new HUI. The disadvantage of this method is that it may use a significant amount of memory if the number of patterns is high, and it can be extremely time consuming if a large number of HUIs are discovered, since a large number of comparisons must be done. In this work, this paper proposes novel checking methods for determining if a HUI is closed without comparing a new pattern to previously discovered patterns. It is based on a technique similar to that employed in sequential pattern mining.

IV. RESULTS AND DISCUSSION

The proposed method has implemented by using ASP.Net programming language.

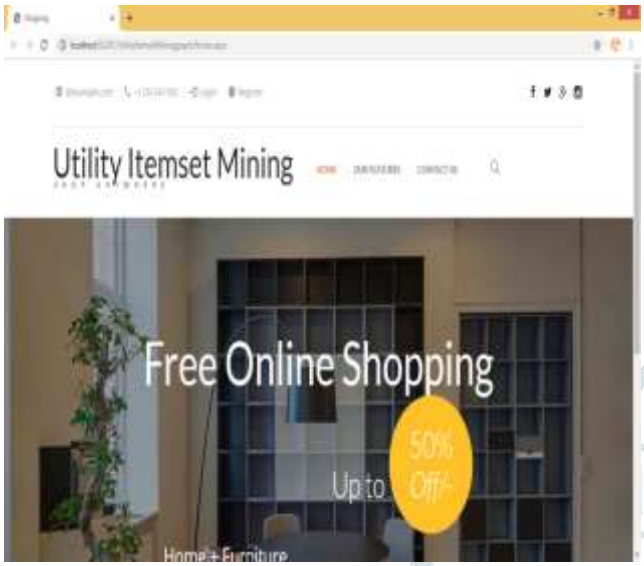

Figure 1: User home page

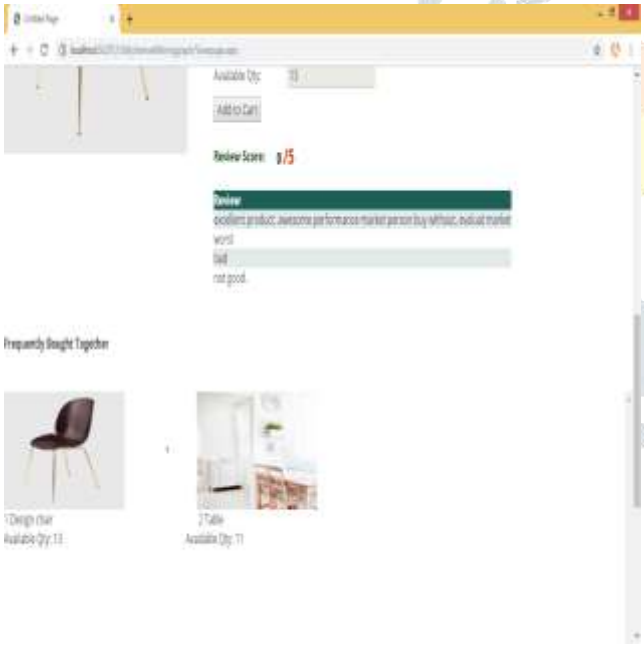

Figure 2: User searching the product with reviews

Figure 3: Applying the Association Rule

Figure 4: Support condition applied Result

Figure 5: Support and confidence applied result

V. CONCLUSION

This paper developed the MHU-ISM technique for mining high utility patterns from transactional databases using the FP-Growth and Apriori algorithms. Mining of association rules is a significant issue in data mining; given a huge collection of data, extracting frequent item groups from this set is a difficult task in data mining. The main issue in extracting frequent item sets is matching item sets. And item set matching is the mining process's bottleneck. It has also been shown that extracting frequent free item sets is an effective approach. In the literature, several efficient methods have been suggested. The concept given in this article is to split the database into several divisions, then identify frequent free item sets in each partition, then combine the partitions to create more frequent free item sets and count the support. The extraordinary development of Websites has resulted in an incredible rise of social media and e-commerce. The availability of a large number of goods online has defined the growth of e-commerce. The websites that offer these goods enable their consumers to openly voice their opinions about a bought product, and these reviews may number in the hundreds. As a result, it becomes very difficult for consumers to read each and every review and keep track of all the pros and drawbacks of a certain product in terms of all the characteristics that it has. This is the focus of the current research. As a result, as a remedy, a feature-based opinion summary of the product evaluations is suggested. The suggested method, which is based on frequent item-set mining, generates a list of potential characteristics for a product. To eliminate duplicate and extremely basic characteristics, several feature filtering techniques such as subset filtering, superset filtering, and distance based filtering are used.

More study is required to discover how the HUI-With positive utility thresholds may be adjusted from the user provided utility to minimize this overestimation. Because the data for mining is often extremely big, this paper wants to investigate sampling-based approximations to minimize computation.

VI. REFERENCES

[1] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules,” in Proc. 20th Int. Conf. Very Large Databases, 1994, pp. 487–499.

[2] R. J. Hilderman, C. L. Carter, H. J. Hamilton, and N. Cercone, “Mining market basket data using share measures and characterized itemsets,” in Proc. PAKDD, 1998, pp. 72–86.

[3] C. F. Ahmed, S. K. Tanbeer, B.-S. Jeong, and Y.-K. Lee, "Efficient tree structures for high utility pattern mining in incremental databases," IEEE Trans. on Knowl. and Data Eng., vol. 21, no. 12, pp. 1708–1721, Dec. 2009. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2009.46

[4] M. J. Zaki and C. Hsiao, "Efficient algorithms for mining closed itemsets and their lattice structure," IEEE Trans. Knowl. Data Eng., vol. 17, no. 4, pp. 462–478, Apr. 2000

[5] 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," in 2008 Eighth IEEE International Conference on Data Mining, Dec 2008, pp.881–886.

[6] M. Zihayat and A. An, "Mining top-k high utility patterns over data streams," Information Sciences, vol. 285, pp. 138 – 161, 2014, processing and Mining Complex Data Streams. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020025514000814

[7] D. Lee, S.-H. Park, and S. Moon, "Utility-based association rule mining: A marketing solution for cross-selling," Expert Systems with Applications, vol. 40, no. 7, pp. 2715 – 2725, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417412012316

[8] J. Pei, J. Han, and V. Lakshmanan, "Pushing convertible constraints in frequent itemset mining," Data Mining Knowl. Discovery, vol. 8, no. 3, pp. 227–252, 2004

[9] 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.

[10] L. Geng and H. J. Hamilton, "Interestingness measures for data mining: A survey," ACM Comput. Surveys, vol. 38, no. 3, p. 9, 2006.

[11] C. W. Wu, P. Fournier-Viger, P. S. Yu, and V. S. Tseng, "Efficient mining of a concise and lossless representation of high utility itemsets," in 2011 IEEE 11th International Conference on Data Mining, Dec 2011, pp.824–833.