A Review on Mining Top-K High Utility Itemsets without Generating Candidates

Lekha I. Surana, Professor Vijay B. More

Lekha I. Surana, Dept of Computer Engineering, MET's Institute of Engineering Nashik, Maharashtra, India
Professor V. B. More, Dept of Computer Engineering, MET's Institute of Engineering Nashik, Maharashtra, India

---***---

Abstract - *High utility pattern mining can be defined as discovering sets of patterns that not only co-occurs but they carry high profit. In two-phase pattern mining an apriori algorithm is used for candidate generation. However candidate generation is costly and it is challenging problem that if number of candidate are huge then scalability and efficiency are bottleneck problems. There are lots of efforts has been made to minimize the number of candidates generated in first phase still challenge is remain when raw data contains huge transactions or minimum utility threshold is too small. Therefore, huge number of candidates motives scalability issues not only in the first phase but also second phase also which degrades systems efficiency. Due to inefficient join operations, lack of strong pruning, and scalability issue a well-known HUI mining algorithm is less efficient than two phase algorithm. Extracting combinations of products with maximum profit is more complex than other categories of utility mining problems. In this paper we defined and solve the problem of utility pattern mining without generating candidates.*

Key Words**: Data mining, utility mining, high utility patterns, frequent patterns, pattern mining**

1.INTRODUCTION

Discovering information of products in the market is a crucial task which required complex analysis of information of each product from the user perspective. Identify frequent itemsets from huge database required valuable study. Apriori algorithm is most popular algorithm for pattern mining. It is breadth first search algorithm which scan database as many times as the length of frequent pattern. Author in [2] discussed about the strategy of FP-mining in which frequent pattern tree without generating any candidate gives the frequent itemsets without any candidate key and searches the database only twice. J. Han discussed about FP-pattern tree. It is squeeze form of extended prefix-tree to extract crucial information about frequent patterns. It is depth-first algorithm. FP-growth algorithm used partitioned based method for decomposing the mining task into small set of task [7]. Weighted association ruled framework mines the frequent itemsets by considering its weight. It is outstanding framework but does not considered importance of item quantities in transaction DB [12]. In high beneficial itemset mining considered items frequency, weight and efficiency. It has issue that it generates the large number of candidate which is infancy task. Mining plays an important role to extract hidden information from large dataset. Transaction weighted utilization model is discussed in [2], is used to overcome the problem in pruning search high utility itemsets. It is difficult task as there may chances that superset of low utility can be high utility. Two phase candidate generation algorithm is implemented for utility itemset mining [3] [6] [9] [11] in which first phase extract the candidates of high utility patterns and then again scan the unrefined data for finding more candidates. It is the problematic issue as number of candidates get increased which may caused efficiency and scalability issue. Therefore it degrades the efficiency of system. For mining high utility itemsets HUIMiner algorithm is utilized in [10]. This algorithm is less efficient when there is requirement huge database for mining due its inefficiency of join operation, pruning approach as well as scalability issues with the vertical dataset structure. D^{2}HUP, algorithm is finding to be novel solution for mining utility itemsets in share framework. This algorithm can addresses the scalability and efficiency issues occurred in the existing systems as it directly extracts the high utility patterns from large transactional databases i.e. TWU. Strength of D^2 HUP algorithms is based on the powerful pruning approaches. It tries to find the patterns in recursive enumeration and it utilizes the singleton and closure property to enhance the efficiency of dense data. Linear data structure known as CAUL is used to show the original information of utility in the unrefined data, it helps to discover the root causes of prior algorithm which employs to maintain data structure information of original utility. Constraints based mining is derived approach from frequent pattern mining to mining utility. Its major is to push the constraints into the frequent pattern mining. In [14], constraints are defined same like normalized weighted support. To prune the search space DualMiner algorithm is introduced in [4] with anti-monotone & monotone constraints. In [5], author De Raedt et al. defined the way of applying standard constraint programming techniques on constraint based mining issues. There are some categories of utility mining defined in [17] such as, objective, subjective and semantic measures. Objective measures can be defined as confidence or support for data. In subjective measure unexpectedness is considered which take knowledge domain of users account and the semantic measures are also called as, utilities which consider the both data and user expectations. Among all these three categories are discussed in [8][9] and [13]. In this first objective measure considers the product price in shopping basket then subjective measures are based on count and amount shares. The utility measure is equivalent to the both objective and subjective measures. The concept of weighted itemset mining and association rule mining is proposed by Cai et al and Lin et al. in [4]. A vertical weight is the concept of significance transaction discussed in [12].

High utility mining is the new approach for the problem which mainly concentrates on high utility pattern growth mechanism. In this approach to reduced the drastically pattern growth reverse set enumeration approach and pruning algorithm is proposed. But it is practically impossible to enumerate all patterns and prune search space. Therefore, anti-monoticity property is applied to pruning approach. Pattern growth approach estimates upper bound of utilities of possible patterns represented by nodes of rooted subtree.

2. Related Work

In 1994, R.Agrawal and R. srikant [1], discussed about the problems of extracting association rules between the items in huge databases of sales transactions. Two algorithms namely, Apriori and AprioriTid are proposed in this paper to solve the problem with other algorithms. Both algorithms are integrated together for hybrid algorithm. It is known as, “AprioriHybrid” algorithm. AprioriHybrid algorithm has its own scalability properties. Another is the problem of basket data is also discussed in this paper. It contains the huge applications database. To make discovery of n-number of itemsets there is need of multiple passes over the data. At the very beginning, it determines the individual itemset which has minimum support. The proposed algorithm Apriori and AprioriTid are different from the AIS and SETM algorithms with respect to candidate itemsets. In AprioriTid algorithm one additional property is used to count the support of candidate itemsets after initial pass. For three datasets performance of AprioriHybrid is relevant to the Apriori and AprioriTid algorithm. In all cases the proposed AprioriHybrid outputs the better performance rather than the Apriori. In the last pass switches AprioriHybrid performs the little worst than the Apriori algorithm. Therefore, AprioriTid algorithm is used after each space.

In 1998,R. Hilderman, Colin L. et al. [2], proposed shared confidence framework. It is the framework to discover the knowledge from databases. It also addresses the problem of discovering itemsets from market basket data. In this paper, they concerned on two types of goals such as, first one is to introduce measures of itemset. This measure is useful and practically interactive for commonly used support measure. Secondly, the discovery of profiles of customers buying patterns also to discover profiles of customer which is done by splitting them into individual classes. The proposed mechanism merged the Apriori algorithm to make discovery of association rules between large databases itemsets. In this paper, an experimental results analysis represented that the proposed share confidence framework has ability to give more information feedback than the support confidence framework.

In 2000, M.J. Zaki, C.J. Hsiao [3], represented CHARM. It is an efficient algorithm for mining closest frequent itemsets. The frequent pattern mining includes the discovery of association rules, powerful rules, multidimensional patterns and also other important discovery. To addressed the problem in frequent pattern mining. An apriori algorithm is employs the BFS i.e. Breadth First Search to enumerates the individual frequent itemsets. Downward closure property is used by apriori algorithm to prune the search space. For mining long patterns there two type of solutions are given in this paper, from those solution first is to discover maximum frequent patterns which has the fewer magnitude than all frequent patterns whereas, the other solution mines frequent closed itemsets. The proposed algorithm CHARM, discovered the itemsets and transaction space over novel tree called as, itemsettidset tree (IT). It uses hash-based approach to eliminate non-closed itemsets at the time of subsumption checking. The algorithm is introduced in this paper is CHARM-L to construct a structure of itemsets. It utilizes the intersection-based approach to non-closed itemsets at the time of subsumption checking. For consideration of appeared IT pairings in the prefix class CHARM-EXTEND is responsible. CHARM-EXTEND mainly return the set of closed frequent itemset.

In 2004, J. Pei, J. Han et al [4], discussed about FP-growth algorithm. In this paper, they mainly contribute themselves to show appropriate order of items. In this paper, author represented the effectiveness of the proposed algorithm. The proposed algorithm is systematic way to incorporate two stages of classes' constraints. In this paper, the concept of convertible constraints is introduced. The convertible constraints are divided into three classes such as, convertible anti-monotone, convertible monotone and strongly convertible. Using this number of useful constraints is covered. The convertible constraints cannot be pushed into fundamental apriori framework but they can push into frequent pattern growth mining. Therefore, they were developed fast mining algorithm for various constraints for mining frequent pattern.

In 2005, Ying Liu, W.K. Liao [5], represented the ARM i.e. Association Rule Mining technique. It discovers the frequent itemsets from the large database and considered individual item to generate association rules. ARM only reflects impact of frequency of the presence and absence of an item. An anti-monotone property is used to discover frequent itemsets. Mining using Expected Utility (MEU) is used to prune the search space by anticipating the high utility k-itemsets. In the section of experimental analysis they analyzed the scalability and accuracy of results. Finally it is seems that in this paper, Two-phase algorithm can efficiently extract HUI.

In 2006, L. Geng, H. Hamilton [6], studied the frequent itemsets. They proposed a best well known algorithm for discovering frequent itemsets. Apriori algorithm is used for pruning search space of itemsets. In this paper, different interestingness is measures of domain of data mining have been proposed. There are three objectives discussed in above from them subjective and semantic based measures deals with background knowledge and goals of user's. These measures are suitable for user experience and the interactive data mining. But the problem in the area of frequent mining is that the real human interest remains an open challenging issue. The experimental setup shows that the human needs to measure their interestingness using another method of analysis. User interactions are crucial in the identification of rule interestingness.

In 2008, A. Erwin, R.P. Gopalan et al. [7], proposed TWU algorithm. This algorithm is based on compact utility pattern tree data structures. It implements the parallel projection scheme to utilized disk storage. The algorithm CTU-Mine is proposed for mining HUI from the huge datasets. This algorithm first identifies the TWU items from transaction database. CUP-Tree is the Compressed Utility Pattern Tree for mining complete set of high utility patterns. This algorithm used parallel projection to create subdivision for subsequently mining. TWU has anti-monotone property which is used to discover the pruning space. In this paper the task of HUI mining discovers all the utility which has utility higher than the user specified-utility. CTU-PROL works against the Two Phase algorithm as well as CT Mine. Efficiency of CT-PROL algorithm is improved than the CTU-Mine. In future work to reduce the computation in large database mining they planned to implement a sampling based approximation.

In 2008, Yu-Chiang Li, Jieh-Shan Yeh [8], proposed IIDS i.e. Isolated Items Discarding Strategy. It is implemented to address the problem in previously proposed apriori pruning algorithm which cannot identify high utility itemsets. The proposed IIDS is utility mining algorithm; it reduces the candidates and enhanced the performance. In this paper, IIDS to ShFSM and DCG applies two methods FUM and DCG+. These methods are implemented respectively. IIDS provides an efficient way to designed critical operations by using transaction weighted downward closure. The proposed IIDS can be applied on traditional Apriori algorithms to extend the scope of IIDS to specific classification model. In further implementation they discussed about classification problems in data mining. They were planned to combined classification and the association rule mining i.e. established the connection between mining utility and associative classification.

In 2011, A. Silberschatz, A. Tuzhilin and T.D.Bie [9], classified the measure into actionable, unexpected and examined the relationship between them. They represented the MaxEnt model. It is used to swap randomization and hence it is computationally more efficient. In this paper, a MaxEnt model is proposed for efficient computations. In this paper, they outlined different ways in the MaxEnt model that can be used efficiently for sampling random databases which is helpful to satisfy the prior information. The parallel to this work, in this paper author made the investigation of MaxEnt modeling strategy for different types of data like, relational databases.

In 2016, Junqiang Liu, Ke Wang, Benjamin et al. [10], suggested D²HUP, algorithm. It seems to be novel solution for mining utility itemsets in share framework. This algorithm can addresses the scalability and efficiency issues occurred in the existing systems as it directly extracts the high utility patterns from large transactional databases i.e. TWU. Strength of D² HUP algorithms is based on the powerful pruning approaches. It tries to find the patterns in recursive enumeration and it utilizes the singleton and closure property to enhance the efficiency of dense data. Linear data structure known as CAUL is used to show the original information of utility in the unrefined data, it helps to discover the root causes of prior algorithm which employs to maintain data structure information of original utility. Constraints based mining is derived approach from frequent pattern mining to mining utility. Its major is to push the constraints into the frequent pattern mining.

3. System Architecture

Figure 1 represents the system architecture. In this there are three entities presented such as, user, transaction, HUIMiner.

1. User:

-User uploads the transaction dataset.

-Get HUI itemsets

2. Transaction:

-Save user uploaded transaction dataset.

-Generate XUT

-Generate reverse set enumeration tree

- Travel tree using DFS

-Search patterns

-Get d²HUP value.

-Verify threshold

-Add itemset in HUI

1. HUIMiner:

- Calculate item relevance score

-Calculate upper bound

-Apply pseudo random projection

-Send d²HUP value to transaction entity.

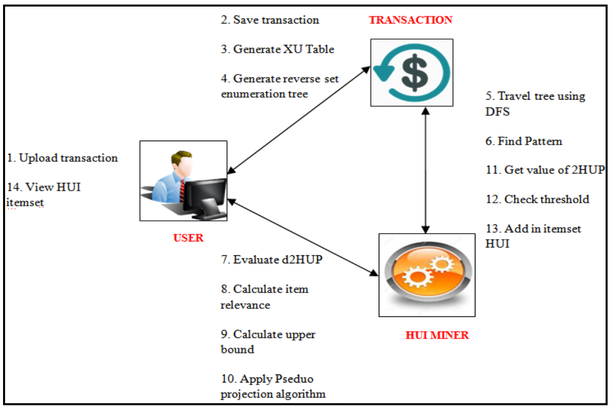

Fig 1: System Architecture

4. CONCLUSIONS

In this survey paper, we discussed about existing techniques used to mine frequent itemsets from the input dataset such techniques are, FP-Growth algorithm, HUIMiner algorithm, MEU, TWU, Apriori pruning algorithm etc. These techniques have some challenging issues such as, large itemset database required more and more scan iteraterations which is time consuming task and degrades the efficiency and system performance. Another scalability is the major issue as large number of itemsets have been generated during processing. From literature survey, we analyze one technique known as, d²HUP algorithm. The scalability issue can be overcome using d²HUP [1] algorithm which is used for utility mining with the itemset share framework which can then enhance system efficiency & performance. It is the technique which has cabability to discover the high utility patterns without candidate generation. Hence from overall review analysis we thought that there is need of such system which can overcome the problems of existing systems and can exhibite better efficiency.

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] 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

[4] 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

[5] 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.

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

[7] 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.

[8] 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.

[9] A.Silberschatz and A. Tuzhilin, “On subjective measures of interestingness in knowledge discovery,” in Proc. ACM 1st Int. Conf. Knowl. Discovery Data Mining, 1995, pp. 275–281.

[10] Junqiang Liu, Ke Wang, Benjamin C.M. Fung,“Mining High Utility Patterns in One Phase without Generating Candidates”, IEEE transaction, vol.28, No.5, 2016.

BIOGRAPHIES

Surana Lekha I. has completed Bachelor’s Degree from SNJB College of Engg.,Chandwad Nashik. Now pursuing ME Degree in Computer Engineering from MET’s Institute of Engineering, BKC, Nashik.

Prof. More Vijay B. is working in MET’s Institute of Engineering, BKC Nashik. He has published/presented research papers in National and International Journals/Conferences. His area of interest is in Data mining.