ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Mining High Utility Patterns in One Phase

Faseela TK

P.G Scholar, Dept. of Computer Science and Engineering, Cochin College of Engineering, Kerala, India

ABSTRACT: Utility mining is a new emerging technology of data mining, but Utility mining does not consider the interestingness measure. High utility pattern growth approach is a look ahead strategy, and a linear data structure. Here linear data structure enables computing a tight bound for powerful pruning search space and to directly identify high utility patterns in an efficient and scalable way. In this it targets the root cause with prior algorithms. Now days, high utility pattern (HUP) mining is one of the most important research issues in data mining due to its ability to consider the no binary frequency values of items in transactions and different profit values for every item. But, incremental and interactive data mining provide the ability to use previous data structures and mining results in order to reduce unnecessary calculations when a database is updated, or when the minimum threshold is changed. In this analytical study, novel tree structures are proposed to efficiently perform incremental and interactive HUP mining.

KEYWORDS: data mining, utility mining, high utility patterns, frequent patterns, pattern mining

I. INTRODUCTION

Frequent pattern mining is provided with the solution, candidate set generation and test paradigm of prior. It has many drawbacks like it requires multiple database scans and generates many candidate item sets. This problem is solved by growth approach by introducing a prefix-tree (FP-tree)-based algorithm without candidate set generation and testing. As frequent pattern mining plays an important role in data mining applications, its two limitations are given as, first, it treats all items with the same importance/weight/price and, second one is, in one transaction, each item appears in a binary (0/1) form, i.e., either it is present or absent. Since, in the real world, each item in the supermarket has a different importance/price and one customer can buy multiple copies of an item. So, items having high and low selling frequencies may have low, and high profit values, respectively. Take a example as, some frequently sold items such as bread, milk, and pen may have lower profit values compared to that of infrequently sold higher profit value items such as gold ring and gold necklace. Therefore, finding only traditional frequent patterns in a database cannot fulfill the requirement of finding the most valuable item sets/customers that contribut1e to the major part of the total profits in a retail business. This gives the motivation to develop a mining model to discover the item sets/customers contributing to the majority of the profit. Now days, a utility mining model was defined to discover more important knowledge from a database. Here the importance of an item set by the concept of utility is measured. The dataset with no binary frequency values of each item in transactions, and also with different profit values of each item is handled. Therefore, utility mining represents real world market data. According to utility mining, several important business area decisions like maximizing the revenue or minimizing the marketing or inventory costs can be considered and knowledge about item sets/customers contributing to the majority of the profit can be discovered. But in real world retail market, takes the biological gene database and web click streams, also there is different importance of each gene or web site and their occurrences are not limited to a 0/1 value. Other application areas, such as stock tickers, network traffic measurements, web server logs, data feeds from sensor networks, and telecom call records can have similar solutions. It is not suitable for large databases.

II. RELATED WORK

High utility pattern mining problem is related to frequent pattern mining and other related topics . Here, we will study how it relates to our work.

Frequent pattern mining

Frequent pattern mining discovers all patterns whose supports are no less than a user defined minimum support threshold. Frequent pattern mining holds the anti-monotonicity property i.e., the support of a superset of a pattern is no

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

more than the support of the pattern. Algorithms for mining frequent patterns as well as algorithms for mining high utility patterns are breadth-first search, depthfirst search, and hybrid search.
This paper uses a depth-first strategy because breadth-first search is typically more memory intensive and more likely to exhaust main memory and thus it is slower. Also, algorithm depth-first searches a reverse set enumeration tree, which can be thought of as exploring a right-to-left in a reverse lexicographic order

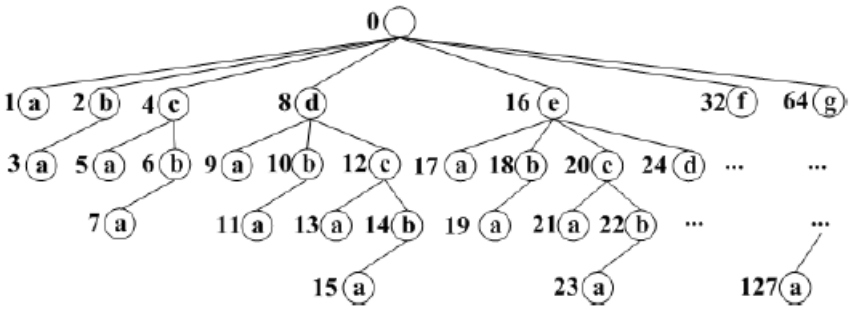

Figure 1: reverse set enumeration tree

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 2000, M.J. Zaki, C.J. Hsiao [2], 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 [3], 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

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

growth mining. Therefore, they were developed fast mining algorithm for various constraints for mining frequent pattern.

In 2005, Ying Liu, W.K. Liao [4], 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 [5], 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. [6], 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 [7], 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 [8], 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, suggested d^2HUP, 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 d^2HUP 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

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

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.

From the problems of utility mining, utility mining with the item set share framework is a hard one as it does not consider interestingness measure. Prior works for this problem with a two-phase candidate generation approach. But it is having one exception that is inefficient and not scalable with large databases. The two-phase approach suffers from scalability issue due to the huge number of candidates. To implement an effective share Framework of using new algorithm, d^2HUP, for utility mining with the item set share framework, which finds high utility patterns in big data without candidate generation.

Algorithm 1: d^{2}HUP$(D, XUT, minU)$

```
1   build TS({}) and Ω from D and XUT
2   N ← root of reverse set enumeration tree
3   DFS(N, TS(pat(N)), minU, Ω)
```

Subroutine: DFS$(N, TS(pat(N)), minU, \Omega)$

```
4   if u(pat(N)) ≥ minU then output pat(N)
5   W ← {i | i ≺ pat(N) ∧ uB_item(i, pat(N)) ≥ minU}
6   if Closure(pat(N), W, minU) is satisfied
7   then output nonempty subsets of W ∪ pat(N)
8   else if Singleton(pat(N), W, minU) is satisfied
9   then output W ∪ pat(N) as an HUP
10  else foreach item i ∈ W in Ω do
11        if uB_fpe({i} ∪ pat(N)) ≥ minU
12        then C ← the child node of N for i
13             TS(pat(C)) ← Project(TS(pat(N)), i)
14             DFS(C, TS(pat(C)), minU, Ω)
15        end foreach
```

d^2HUP, i.e. Direct Discovery of High Utility Patterns, which is an integration of the depth-first search of the reverse set enumeration tree, which prune the techniques that drastically reduces the number of patterns to be enumerated, and a novel data structure that enables efficient computation of utilities and upper bounds.

We can overcome the disadvantage of the existing method. In the existing system a single dataset is used for utility mining. But this is not suitable for large databases. So here we will overcome this problem by taking big data as a input. So for this first of all we need to do the parallel mining. So we need to partition the whole mining tasks into smaller independent subtasks and mining them independently and finally combining the results. So for the high utility mining we will provide a big data as a input for d^2HUP algorithm.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

III. CONCLUSION

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 iterate rations which is time consuming task and degrades the efficiency and system performance. Scalability is the major issue as large number of itemsets have been generated during processing. From literature survey, we analyse one technique known as, d^2HUP algorithm. The scalability issue can be overcome using d^2HUP [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 capability 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 exhibit 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 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
3 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
4 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.
5 L. Geng and H. J. Hamilton, "Interestingness measures for data mining: A survey," *ACM Comput. Surveys*, vol. 38, no. 3, p. 9, 2006.
6 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.
7 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.
8 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.
9 R. Agarwal, C. Aggarwal, and V. Prasad, "Depth first generation of long patterns," in *SIGKDD*, 2000, pp. 108–118.
10 R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in *SIGMOD*. ACM, 1993, pp. 207–216.
11 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 TKDE*, vol. 21, no. 12, pp. 1708–1721, 2009.
12 R. Bayardo and R. Agrawal, "Mining the most interesting rules," in *SIGKDD*. ACM, 1999, pp. 145–154.
13 F. Bonchi, F. Giannotti, A. Mazzanti, and D. Pedreschi, "Exante: A preprocessing method for frequent-pattern mining," *IEEE Intelligent Systems*, vol. 20, no. 3, pp. 25–31, 2005.
14 F. Bonchi and B. Goethals, "Fp-bonsai: The art of growing and pruning small fp-trees," in *PAKDD*, 2004, pp. 155–160
15 F. Bonchi and C. Lucchese, "Extending the state-of-the-art of constraint-based pattern discovery," *Data and Knowledge Engineering*, vol. 60, no. 2, pp. 377–399, 2007