International Journal of Computer Engineering and Information Technology

VOL. 9, NO. 4, April 2017, 72–75
Available online at: www.ijceit.org
E-ISSN 2412-8856 (Online)

Comparative Analysis of Sequential Pattern Mining and High Utility Pattern Mining

Asst. Prof. Anita A. Bhosale

Computer Sci. and Engg. Dept. TKIET, Warananagar, India

bhosale.anita11@gmail.com

ABSTRACT

Mining high utility itemsets is an important research area in data mining. In this mining area, sequential pattern mining and high utility pattern mining plays an important role. Sequential pattern mining concerned with mining statistically relevant patterns where, data are delivered in a sequence and high utility pattern mining concerned finding itemsets with a high utility like the profit from the database. Different algorithms have been work on these areas, but some of them have a problem of generating a large number of irrelevant patterns. Due to this performance of mining is reduced in the case of execution time and gives a less accurate result. Therefore instead of applying single mining techniques, if both sequential and high utility mining user will get more efficient and useful patterns. In this paper, I have analyzed working of sequential and High utility mining technique.

Keywords: *Data Mining, High Utility Mining, Sequential Pattern Mining.*

1. INTRODUCTION

Data mining is the process of retrieving useful patterns from a dataset. The purpose of this system is to find out High Utility Itemsets[1] with sequential pattern mining. The term utility considers importance, interestingness or profitability of the items. High utility itemsets are those itemsets which have high utility greater or equal to a threshold value.The utility of items is calculated by multiplying internal utility and external utility. Itemset in a single transaction or quantity is called as internal utility and itemset in different transaction database or profit is called as an external utility. This area has many applications like bioinformatics, retail and marketing research, e-commerce management and weblog mining. In retail marketing, each item has different price and there is a possibility that single customer can buy multiple copies of the same item. So considering only frequent patterns cannot fulfill the requirement of finding valuable items which have a contribution in total profit in business. In weblog mining, a sequence of web pages visited by the user can be defined as a transaction. Since a number of visits to web page and time spent on a particular web page is different between different users. The total time spent on a page by a user can be viewed as a utility.

Frequent itemset mining is another important research area of data mining[3]. It gives focus on quantity or frequency of the item. But high utility mining focuses on profit or value of the item. The limitation of frequent itemsets mining motivates researchers for finding high utility itemsets. Sequential pattern mining is an important area related to frequent pattern mining which discovers interesting sequences by considering the frequency of items.

Gaining statistically related patterns of data where the values are delivered in the sequence is called as sequential pattern mining. It is closely related to time series mining and the special case of structural data mining. It has many applications like analysis of customer transactions, DNA sequence detection etc.

Some existing methods often generate a large set of potential high utility itemsets and the mining performance is degraded consequently. In this paper, I have made an analysis of working on both high utility and sequential pattern mining. For this, I have here discussed high utility mining algorithm UP-Growth[1] and Sequential mining algorithm SPAM[13].

2. RELATED WORK

Vicent S. Tseng, Bai-En Shie, Cheng-Wei Wu and Philips Yu discovered two algorithms named as UP-Growth and UP-growth+[1]. It scans database twice to find high utility itemsets. It generates PHUIs efficiently. However, these algorithms have the problem of memory usage and level wise candidate generation.

Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer, Byeong-Soo Jeong, and Young-Koo Lee proposed three novel tree structures for efficiently perform incremental and interactive HUP mining[2]. The first tree structure is Incremental HUP Lexicographic Tree (〖IHUP〗 _L-Tree). This tree arranges the items according to their lexicographic order and It can accept the incremental data without any restructuring operation. The second tree structure is IHUP Transaction Frequency Tree (〖IHUP〗 _TF-Tree), which arranging items in descending order according to item's transaction. The last tree, IHUP-Transaction Weighted Utilization Tree (〖IHUP〗 _TWU-Tree) is originated for minimization of mining time. It is based on the TWU value of items in descending order.

Alva Erwin, Raj P. Gopalan, and N. R. Achuthan proposed CTU-PROL algorithm for efficient mining of high utility itemsets from large datasets[3]. This algorithm finds the large TWU items in the transaction database. If a database is too big which unable to held in main memory, the algorithm creates subdivisions using parallel projections and for each subdivision, a Compressed Utility Pattern Tree (CUP-Tree) is used to mine the complete set of high utility itemsets. If For small dataset, it creates an only one CUP-Tree for discovering high utility itemsets.

Shankar S., Purusothaman T., Jayanthi, S., suggested a novel algorithm for mining high utility itemsets[4]. This fast utility mining (FUM) algorithm deals with finding high utility itemsets within the given utility constraint threshold. This algorithm performs if transactions of the database are too long with regard to the number of distinct items available.

R. Chan, Q. Yang, and Y. Shen suggested mining high utility itemsets[5]. They introduced an innovative concept of top-K objective-directed data mining algorithm, which mines the top-K high utility itemsets that directly fulfill the business objective. To association mining, they add the concept of utility to capture highly desirable statistical patterns and present a level wise itemset mining algorithm. They develop a new pruning strategy based on utilities that allow pruning of low utility itemsets to be done by means of a weaker but anti-monotonic condition.

Ramaraju C., Savarimuthu N., proposed a conditional tree based novel algorithm for high utility itemset mining[6]. A novel conditional high utility tree (CHUT) compress the transactional databases in two stages to reduce search space and a new algorithm called HU-Mine is proposed to mine complete set of high utility itemsets.

Y. Liu, W. Liao, and A. Choudhary proposed a fast high utility itemsets mining algorithm [7]. They are present a Two-Phase algorithm to efficiently prune down the number of candidates and can precisely obtain the complete set of high utility itemsets. The first phase proposes a model that applies the "transaction-weighted downward closure property" on the search space to expedite the identification of candidates. The second phase performs one extra database scan to identify the high utility itemsets.

P. Asha, Dr. T. Jebarajan, G. Saranya, made a survey on an efficient incremental algorithm for mining high utility itemsets in distributed and dynamic database[9]. This proposed system divided into one master node and two slave nodes. The database is partitioned for every slave node for computation. Counts the occurrence of each item is calculated by slave node. These data's are stored in their local table. Then each slave node sends these tables to the master node. This node maintains a global table for storing these data. Depending on the minimum utility threshold value promising and unpromising itemsets are obtained.

Jiaxin Liu[14] have introduced a structure, called frequent sequence tree, and an algorithm Con_FST. The root node of the frequent sequence tree stored the frequent sequence tree support threshold and the path from the root node to any leaf node represents a sequential pattern in the database. Frequent sequence tree stored all the sequential patterns with its support that meet the frequent sequence tree support threshold.

3. SEQUENTIAL PATTERN MINING

3.1 SPAM (Sequential Pattern Mining)

SPAM algorithm uses a vertical bitmap data structure representation of database which is similar to the SPADE[10]. It blends the concept of GSP[11], SPADE[10]and FREESPAN[12] algorithms. It uses a depth-first traversal technique to enhance the performance and it also diminishes the cost of joining but it takes more time and space when compared to other algorithms which can be completely stored in the main memory. It uses two strategies. In the first strategy, it mines sequential patterns by traversing the lexicographical sequence tree in a depth-first fashion. Each node in the tree has sequence-extended children sequences generated in the S-Step of the algorithm, and itemset-extended children sequences generated by the I-Step of the algorithm at each node.

In traversing the tree it checks the support of each sequence-extended or itemset-extended child against min_sup recursively. If the support of a certain child is less than min_sup, there is no need to repeat the depth-first search. Apriori-based pruning is also applied at each S-Step and I-Step of the algorithm, minimizing the number of children nodes and making sure that all nodes corresponding to frequent sequences are visited.

The bitmap has a bit corresponding to each element of the sequences in the database in each bitmap data structure representation of the database. It is for efficient support counting of elements. Each bitmap partition of a sequence

to be extended in the S-Step is first transformed using a lookup table, such that all the bits of the index of the first "1" bit (call it index y) are set to one and all the bits with index less than or equal to y are set to zero.

4. HIGH UTILITY ITEMSETS MINING

1.1 UP- Growth Algorithm

This algorithm finds high utility itemsets from a transactional database. It uses UP-Tree for items in the database. This tree maintains total data of a database as well as high utility itemsets. The following section defines how to construct UP-Tree.

1. UP-Tree elements:

   Each node of UP-Tree has name of the item, utility, support count, parent node & link. These elements of nodes having links to a node in the header table.

2. DGU(Discarding Global Unpromising Items):

   This strategy scans database two times. In first scan Total Utility (TU) and Transaction Weighted Utility (TWU) [1] is calculated. Then unpromising items are removed by using Downward Closure Property [1]. The item which have less utility than threshold value is called unpromising items. Promising item are those items which have utility greater or equal to threshold value.

3. DGN(Decreasing Global Node Utilities):

   In this node utilities are reduced by real utilities of descendant nodes. These nodes are near to root of UP-Tree.
   Using both techniques unpromising items are removed and items in each transaction are arranged in order are called reorganized transactions. UP-Growth algorithm has following approach after constructing Global UP-Tree

4. DLU(Discarding Local Unpromising Items):

   In constructs Conditional Pattern base (CPB) [1] by using paths of original tree and forms a conditional tree. For this, it uses minimum utility table which has only promising items. CPB removes unpromising items which have less minimum utility.

5. DLN(Decreasing Local Node Utilities):

   Local UP-Tree is constructed by decreasing node's minimum item utility by the descendant node. Finally, patterns are generated from local UP-tree are high utility itemsets

In the second scan of database high utility itemsets are identified.UP- Growth Algorithm

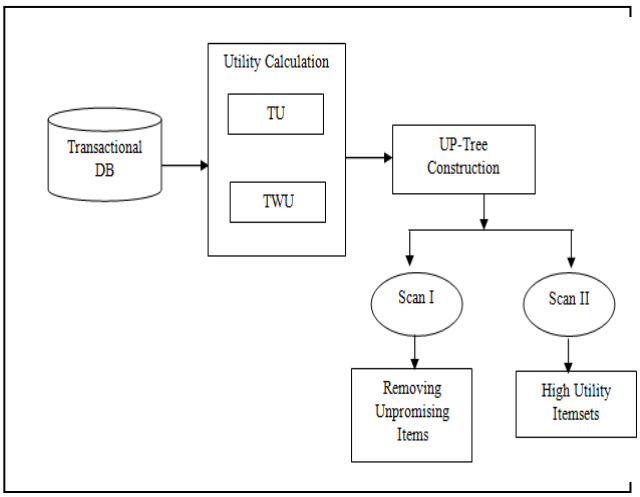

Fig. 1. UP Growth Algorithm

Above figure shows processing of UP-Growth algorithm.

5. CONCLUSION

Sequential pattern mining and high utility pattern mining plays important role in data mining. Sequential pattern mining, it finds which items are brought in a particular order by a single customer; those items come from various transactions. Sequential Pattern Mining is discovering the whole set of a frequent subsequence in the set of sequential transactional database and high utility pattern mining concerned finding itemsets with a high utility like the profit from the database. If both sequential and high utility mining techniques used on the transactional database then a user will get more efficient and valuable patterns.

REFERENCES

[1] Vincent S. Tseng, Bai-En Shie, Cheng-Wei Wu, and Philip S. Yu, Fellow, IEEE "Efficient Algorithms for Mining High Utility Itemsets from Transactional Databases" IEEE Trans. Knowledge and Data Engineering, vol. 25, no. 8, August 2013

[2] Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer, Byeong-Soo Jeong, and Young-Koo Lee, Member, IEEE "Efficient Tree Structures for High

Utility Pattern Mining inIncremental Databases" IEEE Trans. Knowledge and Data Engineering, Vol. 21, Issue 12, pp. 1708-1721, December 2009.

[3] Alva Erwin, Raj P. Gopalan, and N. R. Achuthan, "Efficient Mining of High Utility Itemsets from Large Datasets", In Proc. of PAKDD 2008.

[4] Shankar, S.; Purusothaman, T.; Jayanthi, S. "Novel algorithm for mining high utility itemsets" International Conference on Computing, Communication and Networking, Dec. 2008.

[5] Raymond Chan; Qiang Yang; Yi-Dong Shen, "Mining high utility itemsets" In Proc. of Third IEEE Int'l Conf. on Data Mining ,November 2003.

[6] Ramaraju, C., Savarimuthu N. "A conditional tree based novel algorithm for high utility itemset mining", International Conference on Data mining, June 2011.

[7] Ying Liu, Wei-keng Liao, Alok Choudhary "A Fast High Utility Itemsets Mining Algorithm" In Proc. of the Utility-Based Data Mining Workshop, 2005.

[8] Adinarayanareddy B ,O Srinivasa Rao, MHM Krishna Prasad, "An Improved UP-GrowthHigh Utility Itemset Mining" International Journal of Computer Applications (0975-8887) Volume 58-No.2, November 2012.

[9] P. Asha, Dr. T. Jebarajan, G. Saranya, "A Survey on Efficient Incremental Algorithm for Mining High Utility Itemsets in Distributed and Dynamic Database" IJETAEJournal, Vol.4, Issue 1, January 2014.

[10] M. Zaki, "SPADE: An efficient algorithm for mining frequent sequences", Machine Learning, 2001.

[11] J.Ayres, J. Flannick, J. Gehrke, and T. Yiu, "Sequential pattern mining using a bitmap representation", Proc.Of International Conference on Knowledge Discovery and Data Mining, 2002.

[12] J. Han, G. Dong, B. Mortazavi-Asl, Q. Chen, U. Dayal and M. -C. Hsu,"FreeSpan: Frequent pattern-projected sequential pattern mining", Proc.2000 International Conference of Knowledge Discovery and Data Mining, 2000

[13] Agrawal, R. and Srikant, R. "Mining sequential patterns". In Proceedings of 11th International Conference on Data Engineering (ICDE). Taipei, Taiwan, 1995

[14] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," Proc. 1994 Int'l Conf. Very Large Data Bases, Sept. 1994

[15] Liaxin Liu, "The design of storage structure for sequence in incremental sequential patterns mining," Networked Computing and Advanced Information Management (NCM), 2010.