International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

A Review on Efficient Algorithms for Mining High Utility Item sets

Nutan Sarode[1], Devendra Gadekar[2]

[1]Research Scholar, Department of Computer Engineering, JSPM's Imperial College of Engineering & Research, Wagholi, Pune

[2] Professor, Department of Computer Engineering, JSPM's Imperial College of Engineering & Research, Wagholi, Pune

Abstract: ***Data mining has been around and all enterprises in the real world need it in order to make well informed decisions. The reason behind this is that analyzing huge data is not possible manually. For mining high utility item sets from databases many techniques came into existence. The discovery of item sets with high utility like profits is referred by mining high utility item sets from a transactional database. A number of data mining algorithms have been proposed, for high utility item sets the problem of producing a large number of candidate item sets is incurred. The mining performance is degraded by such a large number of candidate item sets in terms of execution time and space requirement. There are many Problems Occurs when the database contains lots of long transactions or long high utility item sets. Internet purchasing and transactions is increased in recent years, mining of high utility item sets especially from the big transactional databases is required task to process many day to day operations in quick time. Mining high utility item sets from a transactional database means to retrieve high utility item sets from database. Which item sets have highest profit known as High utility item sets. In existing system number of Algorithm's have been proposed but there is problem like it generate huge set of candidate Item sets for High Utility Item sets. Existing UP-Growth and UP-Growth+ used with aim of improving the performances of high utility itemsts. We will compare the performances of existing algorithms UP-Growth and UP-Growth+ against the improve UP-Growth and UP-Growth+.***

Keywords: Data mining, high utility item sets, candidate pruning, frequent item sets, Utility Mining

1. Introduction

DATA mining is the process of extracting nontrivial, previously unknown and information which is very useful from large databases. Finding useful patterns which are hidden in a database plays an essential role in several data mining tasks, such as frequent pattern mining, weighted frequent pattern mining, and high utility pattern mining. Frequent pattern mining is a fundamental research topic that has been applied to different kinds of databases, such as transactional databases streaming databases and time series databases and various application domains, such as bioinformatics Web click-stream analysis and mobile environments .

Finding of frequent patterns task is very important In large databases The primary goal is to discover hidden patterns, unexpected trends in the data. Data mining is concerned with analysis of large volumes of data to automatically discover interesting regularities or relationships which in turn leads to better understanding of the underlying processes. Data mining activities uses combination of techniques from database artificial intelligence, statistics, technologies machine learning . Identification of the itemsets with high utilities is called as Utility Mining. This the most Challenging data mining tasks in the mining of high utility itemsets efficiently.

To discover the useful patterns from database, frequent pattern mining has been applied to different databases. In recent years, finding of frequent patterns from large databases is very important use full in many applications. The goal of frequent itemset mining is to identify all frequent itemsets and it collects the set of items that occur frequently together. The information of frequent set of items is presented as collection of if-then rules.For finding the association and correlation relationship among the items use the generations of association rules. However the unit profits and the purchased quantities are not considered in the frequent itemsets mining. Therefore, frequent itemset mining cannot not satisfy the needs of customers, who all are wanted the itemsets with high profits. Utility mining concepts is used in data mining for discovering itemsets with high utility like high profits. We can determine utility of the item based on customer behaviors and interestingness.

Many No of methods are used to check the performance of utility mining, potential high utility itemsets (PHUIs) are found first, and then an additional database scan is performed for identifying their utilities. This method generate a huge set of PHUIs and their mining performance is degraded consequently. This is not good when databases contain many long transactions or low thresholds are set. The huge number of PHUIs forms a challenging problem to the mining performance. As the algorithm generates more PHUIs, the higher processing time.

High utility itemsets mining has become one of the most interesting data mining tasks with broad applications and it identifies itemsets whose utility satisfies a given threshold. By using different values it allows users to quantify the usefulness or preferences of items using different values.

A high utility itemset is defined as: A group of items in a transaction database is called itemset. This itemset in a transaction database consists of two aspects: Firstone is itemset in a single transaction is called internal utility and second one is itemset in different transaction database is called external utility. Mining high utility itemsets from

databases is an important task and applications such as website click stream analysis, business promotion in chain hypermarkets, cross-marketing in retail stores, online e-commerce management, mobile commerce environment planning and even finding important patterns in biomedical applications

In the reference papers authors proposed two novel algorithms as well as a compact data structure for efficiently discovering high utility itemsets from transactional databases. This is the solution to mine the large transactional datasets, Experimental results show that UP Growth and UP-Growth+ outperform other algorithms substantially in terms of execution time. But these algorithms further needs to be extend so that system with less memory will also able to handle large datasets efficiently.

2. Literature Survey

W. Wang et al in "Efficient mining of weighted association rules (WAR)," [1] proposed weighted association rule. In this rule we first discover frequent itemsets and the weighted association rules for each frequent itemset are generated. Weighted association rule mining first proposed the concept of weighted items and weighted association rules. However, the weighted association rules does not have downward closure property, mining performance cannot be improved. By using transaction weight, weighted support can not only reflect the importance of an itemset but also maintain the downward closure property during the mining process.

In "**Fast algorithms for mining association rules,"R. Agrawal** [2] proposed Apriori algorithm, used to obtain frequent itemsets from the database. in miming the association rules we have the problem to generate all association rules that have support and confidence greater than the user specified minimum support and minimum confidence respectively. Apriori is a classic algorithm for frequent itemset mining and association rule learning over transactional databases. After identifying the large itemsets, only those itemsets are allowed which have the support greater than the minimum support allowed. Apriori Algorithm generates lot of candidate item sets and scans database every time. When a new transaction is added to the database then it should rescan the entire database again. Candidate itemsets are stored in a hash-tree which contains either a list of itemsets or a hash table.

Utility mining is to find all the itemsets whose utility values are beyond a user specified threshold. **"A fast high utility itemsets mining algorithm," by Liu et al in [3]** proposes a Two-phase algorithm for finding high utility itemsets. Two-Phase algorithm, it efficiently prunes down the number of candidates and obtains the complete set of high utility itemsets. It performs very efficiently in terms of speed and memory cost both on synthetic and real databases, even on large databases.

In this there is two phase concept is used. In Two-phase, i focused on traditional databases and is not suited for data streams. In Two-phase we are not finding temporal high utility itemsets in data streams but this must rescan the whole database when added new transactions from data streams.

J. Hu et al in "High-utility pattern mining: A method for discovery of high-utility item sets", [4] defines an algorithm in which concept of frequent item set mining is used which identify high utility item combinations. But actually algorithm is used to find segment of data, which is defined with the combination of few items i.e. rules and different from the frequent item mining techniques and traditional association rule. The problem considered in high utility pattern mining is different from former approaches as it conducts rule discovery with respect to the overall criterion for the mined set as well as with respect to individual attributes.

S.Shankar, A fast algorithm for mining high utility itemsets [5] presents a novel algorithm for Fast Utility Mining. For generating Itemsets techniques like Low Utility and High Frequency (LUHF) and Low Utility and Low Frequency (LULF), High Utility and High Frequency (HUHF), High Utility and Low Frequency (HULF) are used.

Cheng-Wei Wu et al in "UP Growth: An Efficient Algorithm for High Utility Itemsets Mining,"[6] proposed algorithm for efficiently discovering high utility itemsets from transactional databases. Depending on the construction of a global UP tree the high utility itemsets are generated using UP Growth which is one of the efficient algorithms.

J. Han et al in [7] proposed frequent pattern tree (FP-tree) structure in "Mining frequent patterns without candidate generation," paper for storing crucial information about frequent patterns, compressed and develop an efficient FP-tree based mining method is Frequent pattern tree structure. It constructs a highly compact FP-tree, which is usually substantially smaller than the original database, by which costly database scans are saved in the subsequent mining processes. It applies a pattern growth method which avoids costly candidate generation. FP-growth is not able to find high utility itemsets.

H. F. Li et al in "Fast and Memory Efficient Mining of High Utility Itemsets in Data Streams," [8] propose two efficient one pass algorithms MHUI-BIT and MHUI-TID for mining high utility itemsets from data streams within a transaction sensitive sliding window. For improving the efficiency of mining high utility itemsets two effective representations of extended lexicographical tree-based summary data structure and itemset information were developed.

V.S. Tseng et al in "Efficient Mining of Temporal High Utility Itemsets from Data streams," [9] proposes a temporal high utility itemset mining. . The temporal high utility itemsets with less candidate itemsets and higher performance can be discovered by THUI- mine. To generate a progressive set of itemsets THUI-Mine employs a filtering threshold in each partition. . Huge memory requirement and lot of false candidate itemsets are the two problems of THUI-Mine algorithm.

Erwin et al in "Efficient mining of high utility itemsets from large datasets," [10] proposed that the high utility itemsets are mined using the pattern growth approach is the novel algorithm called CTU Mine. For large databases identifying high utility Itemsets candidate-generate-and-test approach is not suitable.

3. Conclusion

In this paper various High Utility Itemsets mining algorithms are discussed. In data mining there are many algoritms for mining high utility Itemsets. Mainly two efficient algorithms named UP-Growth and UP-Growth+ and data structure named UP-Tree was proposed for maintaining the information of high utility itemsets.

References

[1] W. Wang, J. Yang and P. Yu, "Efficient mining of weighted association rules (WAR)," in Proc. of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2000), pp. 270-274, 2000.
[2] R. Agrawal, Imielinski. T and A. Swami, "Mining association rules between sets of items in large databases", in proceedings of the ACM SIGMOD International Conference on Management of data, pp. 207-216, 1993.
[3] Y. Liu, W. Liao and A. Choudhary, "A fast high utility itemsets mining algorithm," in Proc. of the Utility-Based Data Mining Workshop, 2005.
[4] Liu. Y, Liao. W, A. Choudhary,"A Fast High Utility Itemsets Mining Algorithm," In: 1st Workshop on Utility-Based Data Mining.Chicago Illinois, 2005.
[5] S.Shankar, T.P.Purusothoman, S. Jayanthi,N.Babu, A fast agorithm for mining high utility itemsets, in: Proceedings of IEEE International Advance Computing Conference (IACC 2009), Patiala, India, pp.1459-1464
[6] J. Han, J. Pei, Y. Yin, "Mining frequent patterns without candidate generation," in Proc. of the ACM-SIGMOD Int'l Conf. on Management of Data, pp. 1-12, 2000.
[7] 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 Proc. of the 8th IEEE Int'l Conf. on Data Mining, pp. 881-886, 2008.
[8] V. S. Tseng, C. J. Chu and T. Liang, "Efficient Mining of Temporal High Utility Itemsets from Data streams," in Proc. of ACM KDD Workshop on Utility-Based Data Mining Workshop (UBDM'06), USA, Aug., 2006.
[9] Tseng V.S, C.W. Wu, B.E. Shie, and P.S. Yu, "UP-Growth: An Efficient Algorithm for High Utility Itemsets Mining," Proc. 16th ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD"10), pp. 253-262, 2010.
[10] Han. J, J. Pei, Yin. Y, "Mining frequent patterns without candidate generation," In: ACM SIGMOD International Conference on Management of Data, 2000. International Journal of Engineering and Advanced Technology (IJEAT)ISSN: 2249 – 8958, Volume-3, Issue-4, April 2014 Published By:Blue Eyes Intelligence Engineering& Sciences Publication Pvt. Ltd.
[11] S.J. Yen and Y.S. Lee, "Mining High Utility Quantitative Association Rules." Proc. Ninth Int"l Conf. Data Warehousing and Knowledge Discovery (DWK), pp. 283-292, Sept. 2007.
[12] U. Yun, "An Efficient Mining of Weighted Frequent Patterns with Length Decreasing Support Constraints," Knowledge-Based Systems, vol. 21, no. 8, pp. 741-752, Dec 2008.
[13] C. H. Lin, D. Y. Chiu, Y. H. Wu and A. L. P. Chen, "Mining frequent itemsets from data streams with a time sensitive sliding window," in Proc. of the SIAM Int'l Conference on Data Mining (SDM 2005), 2005.
[14] B.-E. Shie, V. S. Tseng and P. S. Yu, "Online mining of temporal maximal utility itemsets from data streams," in Proc. of the 25th Annual ACM Symposium on Applied Computing, Switzerland, Mar., 2010.
[15] K. Sun and F. Bai, "Mining Weighted Association Rules without Preassigned Weights," IEEE Trans. on Knowledge and Data Engineering, Vol. 20, No. 4, 2008.
[16] S. K. Tanbeer, C. F. Ahmed, B.-S. Jeong and Y.-K. Lee, "Efficient frequent pattern mining over data streams," in Proc. of the ACM 17th Conference on Information and Knowledge Management, 2008.