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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

Parallel Approach to High Utility Itemsets Mining from Transactional Databases

Chalak Jyoti V.

PG Student, Dept. of Computer Engineering, VACOE Ahmednagar, Maharashtra, India

ABSTRACT: High utility mining is one of the important tasks of data mining. Existing algorithms for high utility itemset mining have the problem of high execution time and space requirement. These algorithms have to deal with large number of candidate itemsets which tends to degrade the performance of algorithm. V.S.Tseng's UP Growth algorithm is one of the mile stone in this area. This algorithm has effectively pruned the search space by decreasing generation of number of candidate itemsets. Utility pattern tree (UP tree) is used to manage the candidate itemsets. Here, proposed PUP-Growth algorithm tries to enhance the performance of up growth algorithm by applying parallelism. Parallel UP-Growth algorithm improves the performance when database size is large. Here, the concept of threads have been use to provide parallelism.

KEYWORDS: Utility, UP-Growth, Parallel UP-Growth

I. INTRODUCTION

Data mining deals with the discovery of useful data i.e. information from the huge databases. In the world of information technology the data analysis is required to set the goals of business. Therefore information has to be extracted from huge databases for calculations of profit.Various data mining techniques such as frequent pattern mining, weighted association rule mining, utility pattern mining are used for data analysis.Whereas frequent pattern mining is not so effective because in this method only the frequency of items in data bases is focused and the importance of individual items is not taken in account. Weighted association rule mining overcomes the drawback of frequent pattern mining that is importance of each item is considered but only frequency and importance of every item is insufficient for calculation of profit we also have to focus on quantity of each item in database that is the profits are composed of weights and purchased quantity. Therefore something more than frequent pattern mining and weighted association rule mining is required for the calculations of profit which tends to High utility mining. High utility mining is applicable in various sectors like business promotions in cross marketing, chain marketing, e-commerce management, super markets, web click stream analysis and biomedical application. High utility pattern mining discovers the itemsets with high profit. Utility or profitability is given by the product of external utility and internal utility. Where external utility is nothing but the importance of distinct items and internal utility is importance of items in transaction. The itemset is called high utility itemset when its utility is greater than user specified minimum utility threshold else it is called low utility itemset.

There are various algorithms proposed for high utility itemset mining one of them is UP growth algorithm which effectively prunes the search space and gives good results even when very low threshold is set and database is huge. In this algorithm some effective strategies like DGU, DGN, DLU and DLN are suggested these strategies are used to generate utility pattern tree that is UP-tree. UP-tree is a data structure used to find high utility itemset and to maintain useful information from given database.This algorithm gives good result but it is a sequential algorithm and it is well known that, parallel algorithm gives faster results than sequential algorithm. So here, proposed PUP-Growth applies parallelism to UP-Growth algorithm which tends to improve the performance of algorithm in terms of execution time.Here threads are used to provide parallelism. One more important thing is that,here user is going to define threshold.sometimes there is chance that user specified threshold is not in the range of utilities of itemsets under consideration. Proposed algorithm successfully covers this problem.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

II. RELATED WORK

There is lot of research had done in this area of data mining. I have referred following work of some researchers to get details of high utility mining, likeits importance, application area, and different methods to achieve high utility mining.to study high utility mining we have to start with frequent pattern mining [1],[2].in frequent pattern mining frequently occurring items are extracted from databases. Here emphasis is given only on frequency of occurrence of items, but it is not sufficieant to analyse the data in business point of view that is for calculation of profit.
Association rule mining and sequential pattern mining [1],[2] are some of the famous issues of frequent pattern mining. Apriori is well known algorithm for association rule mining as well as FP growth is pattern growth based association rule mining algorithm [3] which tends to have better results just in two database scans. Association rule mining had first proposed the concept of weighted items and weighted association rules that is the importance of items are taken in to consideration but yet quantities of items are not considered which does not improve the mining performance. Hence, the requirement of new algorithm is raised, which is able to consider both the weight (importance) of each item as well as its quantity in the database. This requirement results in 'high utility itemsets mining'. Liu.et al has proposed a two-step algorithm for high utility itemset mining [4]. In first step, the HTWUIs and their TWUs are computed and passed to next 2nd step. Transaction weighted downward closure property is used to identify high utility itemset. The drawback of algorithm was generation of huge number of candidate itemsets and numbers of database scans were required. The problem of generation of huge number of itemset increases the search space and degrades the mine performance in terms of execution time and memory requirement. Various researchers were have taking efforts in direction of pruning the search space. V.S.Tsing, bai.En.shi,Changwei.wu and Philip,S.Yu have done remarkable work in this area. The up growth algorithm they proposed UP growth and UP growth+ algorithm for high utility itemset mining for transactional databases. These algorithms have given better results by pruning search space. Even database contains large number of long transaction and minimum utility threshold is low. These algorithms usetree data structure named UP tree to manage all data. Following are the steps of UP growth algorithm.

1. Scan the given database twice and obtain global UP tree (using DGU and DGN strategies)
2. Recursively generate PHUIs from global UP tree and local UP tree (using DLU and DLN strategies)
3. Identify actually high utility itemset from PHUIs

III. PROPOSED METHOD

To get high utility itemset from transactional database, parallel UP growth algorithm is proposed.this algorithm is based on UP growth algorithm. I have simply tried to parallelise this UP growth algorithm. Proposed system uses threads of execution of UP growth in parallel. Here this algorithm offers parallel UP tree generation and these trees gives us high utility itemset these high utility itemset are collected together to generate a set of potentially high utility itemset(PHUIs). PHUIs contribute to give highest utility itemset along with its utility as output. Following diagram shows data flow diagram of parallel UP Growth algorithm.

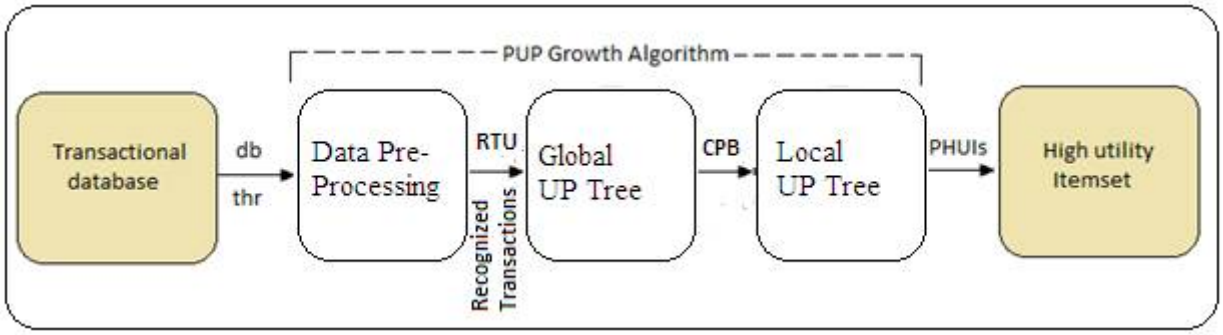

Fig1: Parallel UP-growth Data Flow Diagram

From above figure it is very clear that PUP Growth algorithm works three different phases. In first phase of data pre-processing, the input database is scanned to identify all items with their utilities. User specified threshold is set to

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

appropriate value. Transaction weighted utilization (TWU) of each item is calculated and compared with threshold to discard global unpromising items. Then we will get recognized transactions with recognized transaction utilities (RTU). In second phase of global UP Tree Generation, these recognized transactions are to be inserted into a tree structure. While creating tree, tree nodes are generated and their node utilities are calculated using recognized transactions and their RTUs. Total node utility of each item is compared with threshold and unpromising items are discarded while creating conditional pattern bases CPBs. These CPBs gives reduced paths and their path utilities.
In third phase, local UP Tree is generated by inserting recognized paths in to a tree structure. Recognized paths are got from CPBs. While construction of local UP tree we have to calculate local node utilities. Here also local unpromising items are to be discarded. Here we get set of potentially high utility itemsets (PHUIs) and finally from these potentially high utility itemsets we get high utility itemsets as output of our algorithm.

Parallel UP-growth algorithm

1. Take input transactional database and threshold
2. Scan the database to identify items with their utilities, transaction's transaction utility
3. Convert user specified threshold (%) to appropriate min_util threshold value.
4. Calculate TWU of each item
5. Compare TWU with threshold and discard unpromising items
6. Calculate RTU of each transaction
7. Create global UP tree
8. Create CPBs
9. Create local UP tree
10. Get set of PHUIs
11. Get high utility itemsetsalong with their utilities.

IV. RESULTS AND PERFORMANCE ANALYSIS

The implementation of PUP-Growth algorithm is done and it gives better result than existing serial algorithm. while analysing the performance of parallel UP-Growth algorithm we must have to compare the result with sequential algorithm.Here, a chess database of size 3196 transactions is choosen for experimentation. We have performed high utility mining in sequential as well as parallel way. The corresponding results are shown in the following table and the results are analysed using the graph.For evaluating results we have considereddifferent threshold values against execution time.

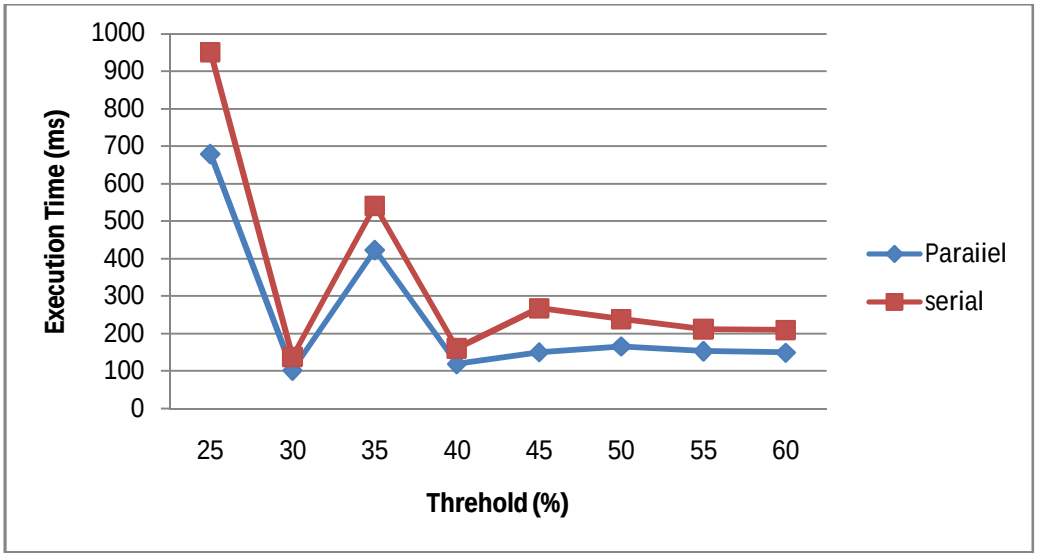

Fig2: Performance comparison of serial and parallel UP-Growth Algorithm.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

From above graph it is very clear that proposed PUP Growth algorithm gives better results than serial UP Growth algorithm.

V. CONCLUSION

In this paper, proposed parallel UP growth algorithm for mining high utility itemsets from transactional databases. Tree data structure named UP-tree is used to maintain all data while execution of algorithm. Parallelism is applied to existing UP growth algorithm using the concept of threads. We got considerable improvement in performance by applying parallelism especially when huge numbers of candidate itemsets are generated.

REFERENCES

1. R. Agrawal and R. Shrikant, "Fast Algorithm for Mining Association Rules," Proc.20th int'l Conf. VLDB 1994
2. R. Agrawal and R. Shrikant, "Mining Sequential Patterns," Proc.11th int'l Conf. Data Eng. Mar.1995
3. J. Han, J. Pei and Y. Yin, "Mining Frequent Patterns Without Candidate Generation," Proc. ACM-SIGMOD Int'l Conf. Management of Data 2000
4. Y. Liu, W. Liao, and A. Choudhary of, "A Fast High Utility Itemsets Mining Algorithm," Proc. Utility-Based Data Mining Workshop, 2005
5. A.Erwin R.P. Gopalan and N.R. Achuthan, "Efficient Mining of High Utility Itemset From Large Dataset," Proc. Conf. 12th PAKDD, 2008
6. 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. Knowledge and Data Eng., Vol.21 No.12, Dec2009
7. R. Chan, Q. Yang and Y. Shen, "Mining High Utility Itemsets,"Proc. IEEE third Int'l Conf. Data Mining,Nov 2003.
8. V.S. Tseng,B.-E.Shie,C.W.Wu,P.S.Yu,"Efficient Algorithms for Mining High Utility Itemset From Transactional Databases" IEEE,Aug2013
9. Mengchi Liu JunfengQu, "Mining High Utility Itemset Without Candidate Generation," CIKM'12 Oct 29-Nov2 2011, Maui, HI,Usa
10. Fournier-Viger, P.Wu, C.W, Zida S. Tseng V.S., "FHM-Faster High Utility Itemset Mining Using Estimated Utility Co-occurrence pruning," Proc.21st ISMIS, 2011 Springer
11. Maya Joshi, Manasi Patel, "A Survey On High Utility Itemset Mining Using Transactional Databases," IJCSIT Vol.5(6),2014.
12. Frequent Itemset Mining Implementation Repository, http://fimi.cs.helsinki.fi/2012.