International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 24 Issue 10 – JULY 2017.

PRIVACY PRESERVED CLUSTER BASED MAPREDUCE IN BIG DATA USING SVM CLASSIFIER

SHARMILA [#1], S.VIJAYANAND[*2] and VIJAYAKUMARI[*3]

[#] *PG SCHOLAR THE KAVERY ENGINEERING COLLEGE, TN, India*

[*2] *ASSISTANT PROFESSOR, & HEAD, DEPT., THE KAVERY ENGINEERING COLLEGE, TN, India*

[*3] *ASSISTANT PROFESSOR, THE KAVERY ENGINEERING COLLEGE, TN, India*

***Abstract*— Now-a-days, Big data Mining is an important research topic because it is widely applied in real world to find the frequent item sets. As a volume of database increases day by day, the problems of scalability and efficiency become more severe. As a solution to this problem, Frequent Item set Ultrametric Tree on Map Reduce programming model is proposed. Frequent items ultra metric tree has four major advantages over traditional FP-tree like; it minimizes I/O overhead, reduces the search space by improved way of partition the database, compressed storage because only frequent items are inserted as a nodes and at last reduces computing time without traversing the tree recursively. Parallel mining of frequent item sets using Frequent Item set Ultra metric Tree on Map Reduce framework consists of three Map Reduce job. The first Map Reduce job is responsible for finding frequent one item sets. The second Map Reduce job is responsible of finding all 'k' item sets by pruning infrequent items from each transaction. The third Map Reduce job is responsible for creating all frequent 'k' item sets. The distributed cache technique can be used to store the intermediate result of each Map Reduce job which will significantly improves performance of parallel mining of frequent item sets using FIUT on Map Reduce framework. Frequent itemsetultrametric tree on a cluster is sensitive to data distribution and dimensions. Extensive experiments using synthetic data demonstrate that our proposed solution is efficient and scalable.**

***Index Terms*— Frequent itemsets, Frequent items ultrametric tree (FIU-Tree), MapReduce, DistributedCache.**

I. INTRODUCTION

Data mining is the process of finding patterns among dozens of fields in large database. Discovering a useful patterns in hidden in a large database plays an essential role in several data mining tasks such as frequent pattern mining, high utility mining etc. Frequent pattern mining are not satisfy the user needs. who are interested in discovering the items with the high profits ,the profits are consist of unit profit ie weights and quantity of purchased items.

High utility mining is emerges as important topic in data mining. There are two aspects of utility mining: internal utility and external utility. The internal utility means importance of items in transaction and the external utility means importance of different items. High utility mining itemset mining identifies itemsets whose utility satisfies a given threshold. It allows user to quantify the usefulness or preferences of items using different values. Thus it reflects the impact of different items. High utility itemset mining is useful in decision making process of many application such as retail marketing and web service, since items are actually different in many aspects in real application. The cost of candidate generation of high utility itemsets mining is intolerable in terms of time and memory space.

The pruning of search space in transactional database for high utility mining is more difficult because it has a superset of low utility itemset in database. The efficiently reduce the search space and identify high utility itemsets in large database is a challenging problem in utility mining.

To overcome this issue, propose the effective heuristic rules, for identify high utility patterns from transactional database. The work of this paper is summarized in next section.

II. RELATED WORKS

Frequent pattern mining is the problem of the paper. To avoid this issue, association rule mining is proposed, in this paper apriori algorithm[1] is used. It contains multiple database scan. Discovering many association rules in large database. large number candidate itemsets are generated it will degrade the mining performance.

Frequent pattern growth[2] algorithm was proposed later,it better than the apriori algorithm ,it find frequent items without generate the candidate itemset and it contain two database scan and it consumes more processing time.

To avoid the issue of high utility itemset mining,efficient tree based structure [3]was proposed, the tree structure is used to maintain information items utilities and information about the items. it contains two phase,in phase 1 generate HTWUIs efficiently and to avoid the two many database scans. The algorithm contains Three steps,1.construction of tree structure,and rearranged the transaction in any fixed order.the rearranged transaction are entered in the tree structure.2.Tree structure are generated the HTWUIs by FP growth.phase 1

HTWUIs are found, without generating any candidate itemsets. 3.High utility itemsets are identified by one original database scan.In phase 2,overestimated utilities are produced and required additional database scan to identify high utility itemsets.

To avoid problem of multiple database scan, Isolated Items discarding Strategy[4] is used to decrease the number of candidates. In phase 1, number of candidates itemsets are reduced by level wise search and pruning the isolated items.this algorithm also scan database for many times and generate candidate itemsets for finding high utility itemsets.

To avoid multiple scan, UPgrowth algorithm[5] is proposed, it also contains tree based data structure and is used to maintain about item utilities and item names and four strategies are proposed to enhanced the mining performance and it need two database scans.

III. MINING HIGH UTILITY ITEMSETS:

A. *Tree Structure*

UP tree is used to maintain the information about the transaction and utility items. In a UP tree, two strategies are applied to reduce the overestimated utilities stored in the node of the tree.The elements which consist in aUp tree are N.name,N.nu,N.parent,N.count, N.hlink and child nodes.the header table is used to facilitate the traversal of Up tree. The header table consist of the entry records of an each item name and its link.

1) *DGU*

The global Up Tree is constructed by only two scans of the original database. In the first scan,TU of each item is found and at the same time,of each single items are also found. By TWDC property,the unpromising itemsets are found.the unpromising itemset means which TWU is less than the minimum utility threshold.during the second scan of the database,the transactions are entered into a tree. After retrieved the transaction, the unpromising items should be removed from the transaction and its utilities are also removed from the transaction.NewTU,after pruning unpromising item and sorting the remaining items in any order is known as RTU

2) *DGN*

By using this strategy DGN, the utilities of the nodes that are closer to the root of a global up tree are reduced.DGN is suitable for the database contains the long transactions. They use the divide and conquer technique in mining processes. The search space aredivided into smaller subspaces.

For example,

{b}'s conditional tree

{a} does not contain {b}tree

{d}does not contain {b}and{a}

{c}does not contain {b},{a}and{d}

{e} does not contain {b},{a},{d}and{c}

The searching is starts from bottom of the tree. The nodes does not appear the descendant nodes.the proposed strategies is used for decreasing overestimated utilities is remove the descendant nodes in a tree.

B. *DLU and DLN*

They are pushing the two more strategies into the FP Growth.By pushing these two strategies overestimated utilities are decreased and the number of PHUIs can be reduced.

1) *DLU*

The algorithm contains tree steps1. Generate the conditional pattern bases for tracing the trees original path,2.conditional tree are to be constructed is calle local tree. 3.mine the patterns from conditional trees. By using DLU, minimum item utilities are utilized to reduced utilities of local unpromising items in conditional pattern bases. the local unpromising items are subtracted from the path utility of an extracted path.

2) *DLN*

In DLN, the path are reorganized by pruning unpromising items and resorted in any fixed order. These paths are known as reorganized path.DLU and DLN are can be local version of the DGU. By using , these two strategies, overestimated utilities for itemsets can be locally reduced without losing an actual high utility itemsets.

C. *Heuristic rules*

Heuristic rules are used for better decision making process. Potential high Utility itemsets are found by four strategies. An association rules for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness. A rule states as follows A-> B (The logs A are being visited followed by B) , The support and count is being measured.

An item L with their minimal node utility, path utility and counts, The second step is to find all valid rules with the itemsets . The rules we need to find out in this step can be classified into two groups by verifying whether the item ir is the least frequent item in the whole rule or not: 1. the rules with i_r on the antecedence and ir is the least frequent item in the rule. 2. the rules with i_r on the antecedence, but there exists at least one item with its occurrence count less than i_r on the consequence. We can find out the first kind of rule directly by using the information found in the first step. Our method is to compare any two itemsets IS_A and IS_B found in the first step. If one is the subset of the other, say IS_A is the subset of IS_B. Then let ISC be the difference of IS_B and IS_A. To verify the validity of the rule IS_A !IS_C , we need to verify the following rule

IS_A-> IS_C =P(IS_A ->IS_c)/P(IS_A)

On solving this equality the subsequent strength of the rule and its validity is being found. Thus on adopting all these the high utility items are evaluated , with strongly saying in terms of rules. Thus Proposed methodology will adopt for many number of transactions/logs.

For Example,

If totally 5 logs are being depicted as high utility logs say (A,B,C,I,K), the possible sets are extracted,which contains 2 combination,3 combination, 4 combination, etc

(A,C) -> 2combination

(C,J)->2 combination

(C,J,K)-> 3 combination

Finally evaluating high utility patterns which means(A,C,K,J).Many user has used A,C,K,J items

Repeatedly, predicted by highest confidence.

D. Identify High Utility Pattern

After finding the PHUIs,now identify actual high utility itemsets and sets of high utilities are produced by scanning the original database one time.

IV. CONCLUSION

In this paper, proposed rule based method named heuristic rule for extracting high utility patterns. Potential High utility Itemsets are efficiently generated by two scans of the original database. Here use four strategies to reduced the overestimated utilities and increase the performance of the high utility mining.

REFERENCES

[1] Agrawal.R and Srikant.R,"Fast Algorithms For Mining Association Rules",Proc.20th int'l conf.very Large Data,Bases(VLDB),pp.487-499.1994.

[2] Pei.J,Han.J,Mortazavi-Asl,Pinto.H,Chen.Q,Moal.U and Hsu.M.C,"Mining sequential Patterns By Pattern Growth: The prefix Approach,"IEEE Trans.Knowledge", Eng,Vol.16,no.10,pp:1424-1440,oct 2004.

[3] C.F. Ahmed, S.K. Tanbeer, B.-S., and Y.-K. Lee, "EfficientTree Structures for High Utility Pattern Mining in Incremental Databases," IEEE Trans. Knowledge and Data Eng., vol. 21, no. 12, pp. 1708-1721, Dec. 2009.

[4] Li Y.C, Yeh .J.S and Chang.C.C, "Isolated items discarding strategy for discovering high utility itemsets," Data & Knowledge Engineering, Vol. 64, Issue 1, pp. 198-217, Jan., 2008.

[5] V.S. Tseng, C.-W.Wu, B.-E. Shie, and P.S. Yu, "UP-Growth: An Efficient Algorithm for High Utility Itemsets Mining," Proc. 16thACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD'10), pp. 253-262, 2010.

[6] R. Chan, Q. Yang, and Y. Shen, "Mining High Utility Itemsets,"Proc. IEEE Third Int'l Conf. Data Mining, pp. 19-26, Nov. 2003.

[7] A. Erwin, R.P. Gopalan, and N.R. Achuthan, "Efficient Mining ofHigh Utility Itemsets from Large Data Sets," Proc. 12th Pacific-AsiaConf. Advances in Knowledge Discovery and Data Mining (PAKDD),pp. 554-561, 2008.