A Survey on MRPrePost

Sahana. P
Student, M.Tech. 4thsem, CSE
SaiVidya Institute of Technology,
Bangalore, India
E-mail: sahanapandu22@gmail.com

Harisha. D. S
Asst. Prof., Dept. of ISE
SaiVidya Institute of Technology,
Bangalore, India
E-mail: dsharirao@gmail.com

***Abstract*---**Due to the vast amount of processed and unprocessed data that is present in the world and also due to unimaginable amount of data being added continuously there is a need for processing these vast amounts of data. Also the processing capability of any algorithm or tool has to be efficient and fast so as to process this vast data in faster speed consuming less time as possible. MRPrePost is the algorithm that is presented in this survey paper as one of the efficient methods when compared to Apriori with respect to performance and time.

Keywords:-*MRPrePost algorithm, Apriori algorithm, data mining, big data*

*****

I. INTRODUCTION

The processing of large data involved in the development of data mining as the new field that is fast gaining progress and is still in evolution stage. Different methods or tools are present in data mining for efficiently mining data.

Data mining is an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. Data mining can be utilized in various fields one of the best examples is in market analysis or retail industry where data mining helps the retailer to assess the products that are fast moving off the shelves. This can increase or boost the sales of the retailer.

The data mining results in an itemsets called the Frequent Itemsets [4]. Frequent Itemsets is a set that consists of items that appear at regular intervals when scanning the database. The mining of frequent itemsets is referred to as FIM (Frequent Itemsets Mining)[4].

Big data involves certain techniques and technology that need never methods of combinations to unveil huge unseen values from huge datasets that are diverse, complex and massive scale. Characteristics of big data are as follows:

1. Volume – The quantity of data generated is important. It is the size which determines the value and potential is best described under consideration and data generated is big data or not is only related to the size of the data itself.

2. Variety –This describes the variety of the data to which it belongs so that analysts knows it very easily and helps the people analyzing the data to take its advantage and uphold the importance of big data.

3. Velocity - This describes the speed at which data is generated and processed so that it meets the challenges regarding the development.

4. Variability – This tells the inconsistency which can be shown by the data, thus hindering the process of being able to manage the data effective.

5. Complexity –complexity is difficult to manage, especially when large volumes of data that are generated from multiple sources. These data is to be linked, in order to grasp the information that is supposed to.

FIM can be done through algorithms like Apriori- algorithm, FP-growth algorithm, PFP, etc [1][3][10].

The Apriorialgorithm [1] is used to find frequent itemsets using candidate generation. It is a seminal algorithm. The name is based on the prior knowledge of the frequent itemsets and It employs an iterative approach defined as a level-wise search, where k-itemsets are used to explore (k+1)-itemsets. The FP-growth algorithm [3] is an association rule mining and Items in each instance that do not meet minimum converge threshold are discarded.

Comparing with Apriori, FP-growth [3] is an improvedAlgorithm. Its main advantage is that only needs to scan the database twice, and construct a compressed data structure FP-Tree, which reduces the search space, while no candidate set, improved memory utilization. From the algorithm thought, we can see it adapts to depth-first mode policy [10].

PFP [5] is based on the Hadoop [6] parallel algorithms; PFP groups the itemsets, as a condition database divided to each node, each node independentlygenerates the FP-Tree and mines frequent itemsets. PFPreduce the traffic between nodes increases the degree ofpolymerization of node.

However, algorithm is not efficientif the database is discrete. A massively parallel FP-Growthalgorithm. This algorithm is based on a novel data andcomputation distribution scheme, which virtually eliminates communication among computers and makes it possible for us to express the algorithm with the Map Reduce model. Experiments on a massive dataset demonstrated outstanding scalability of this algorithm.

Hadoop[6][8] was created by Doug Cutting and Mike Cafarella in 2005. It uses Java. It is a cross-platform distributed file system being built and used by a global community of contributors and users. It is reliable, scalable, distributed computing. The Hadoop platform is designed to solve the problems where you have a lot of data and it is an open-source software framework that supports big data storage.

II. RELATED WORK

Map Reduce [7] is a framework for processing across huge datasets using a large number of computers (nodes). Processing can occur on data stored either in afile system (unstructured) or in a database (structured) and can take the advantage of locality of data, processing it on or near the storage assets in order to reduce the distance from which it must be transmitted.

- "Map" step: Each worker node applies the "map ()" function to the local data, and output is written to external storage.
- "Shuffle" step: Worker nodes redistribute the data based on the output keys (produced by the "map ()" function).
- "Reduce" step: Worker nodes now process the each group of output data, per key, in parallel method.

Map Reduce framework has two types of algorithm[4]: 1) Dis-Eclat is a Map Reduceimplementation of theEclat algorithm which is optimized for speed in case a specific encoding of the data. Dis-Eclat unlike the previously mentioned algorithm divides database, but the search space will be allocated to each node, which eliminates the communication between nodes. 2) Big-FIM is optimized to deal with truly Big Data by using a hybrid algorithm, combining principles from both Apriori and Eclat's.

The Eclat algorithm is used to perform itemsets mining. Itemsets mining let us find frequent patterns in data like if a consumer buys milk, he also buys bread. This type of pattern is called association rules and is used in many application domains.

The basic idea for the Eclat algorithm is use tidsets intersections to compute the support of a candidate itemsets avoiding the generation of subsets that does not exist in the prefix tree.

The Eclat algorithm is defined recursively. The initial call uses all the single items with their tidsets. In each recursive call, the function IntersectTidsets verifies each itemset-tidset pair (X,t(X)) with all the others pairs (Y,t(Y)) to generate new candidates N_{XY} . If the new candidate is frequent, it is added to the set P_X. Then, recursively, it finds all the frequent itemsets in the X branch. The algorithm searches in a DFS manner to find all the frequent sets.

Apriori is an algorithm that has been proposed in [4]. The discovery of frequent itemsets is accomplished in several iterations. In each scan, a full scan of training data is required to count new candidate itemsets from frequent itemsets already found in the previous step. Apriori uses the ―Apriori property to improve the efficiency of the search process by reducing the size of the candidate itemsets list for each iteration.

Frequent Itemsets Mining (FIM) [4] has been an essential part of data analysis and data mining. FIM tries to extract information from databases based on frequently occurring events, i.e., an event, or a set of events, is interesting if it occurs frequently in the data, according to a user given minimum frequency threshold.

PrePostalgorithm [2][10] presents a data structure named N-list, which is a modification of the vertical database, storing the association rule mining all the information needed. Pre Post also needs to scan the database twice to construct a PPC-Tree, and make use of PPC-Tree to generate the N-list of FIM. In the mining process, the database does not require rescanning, only need to intersect the merger N-list, and the complexity of the algorithm is O(m+n), m and n are the length of two N-list, Each element of N-list composed by PrePost-Code, which is called after the sequence encoding the preamble, the composition in the form of «pre-order, post-order: count», PrePost-Code is based on the PPC-Tree respectively from the previous order traversal and post order traversal.

The next section will discuss in detail about the MRPrePost which is parallelize to the PrePost. MRPrePost[10] is a

It combines the Dis-Eclat[4] basis PrePost[2] algorithm. The main feature of MRPrePost algorithm is Map Reduce concept [7]. It uses threeMap Reduce to parallelize PrePost.

Parallel algorithm based on Hadoop platform, which improves PrePost by way of adding a prefix pattern, and on this basis into the parallel design ideas, making MRPrePost algorithm can adapt to mining large data's association rules.

Data mining is divided into three stages by MRPrePost algorithm:

l. Statistic I-frequent itemsets, similar to the process of word frequency statistics, raw data sets scattered on each worker nodes, each node independently perform map function, reduce function combined statistical results, and according frequent threshold cropped infrequent items;

2. Using a method similar to building FP-Tree to build PPCTree, traverse and generate N-List frequent one set;

3. The search space division, the N-list is distributed to each worker nodes, in order to ensure the cluster load balance, we make use of Round-Robin[8] to act as partition function, each node using the prefix pattern generates independently frequent itemsets.

Disadvantage of Apriori algorithm is time and memory. First it will generate candidate items then it will scan the items and then generate itemsets so it will take more memory and time. This disadvantage will overcome in MRPrePost algorithm. This will introduce mapper and reducer to items. It will take less time and memory because the above mentioned it will work in three stages at a time.

CONCLUSION

When we compare PrePost, PFP and MRPrePost based on their runtime using two datasets. The sum time is less when the support increases. The performance of parallel algorithm is less compared to PrePost. This is due to the fact that when every node sends data to others, there is a latency of bandwidth which is definable or difficult to point.But, MRPrePost can perform better on all types of datasets. This is due to the making the algorithm run different procedures at the same time.The main with-helding factor in this is that due to the clustering, the processing time of the data is little predictable.

ACKNOWLEDGMENT

I would like to extend my sincere gratitude to Sai Vidya Institute of technology.

REFERENCES

[1] Agrawal R, Srikant R. Fast algorithms for mining association rules[C]lProc. 20th int. conf very large data bases, VLDB. 1994, 1215
[2] Deng Z H, Wang Z H, Jiang J .I. A new algorithm for fast mining frequent itemsets using N-lists[J]. Science China Information Sciences, 2012,55(9) 2008-2030.
[3] Han J, Pei J, Yin Y Mining frequent patterns without candidate generation[C]/ACM SIGMOD Record. ACM, 2000, 29(2): 1-12.
[4] Moens S, Aksehirli E, Goethals B. Frequent Itemset Mining for Big Data[C]/2013 IEEE International Conference on Big Data,. IEEE, 2013: 111-118.
[5] Li H, Wang Y, Zhang D, etalPfp: parallel fp-growth for query recommendation[C]proceedings of the 2008 ACM conference on Recommender systems. ACM,2008 107-114.
[6] Apache Hadoop. http://hadoop.apache.org/, 2013.
[7] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters [J].Communications of the ACM, 2008, 51(1):107-113.
[8] Apache mahout. http://mahout.apache.org/, 2013.
[9] http://adrem.ua.ac.be/bigfim
[10] Li Jinggui, Y Zhao, S Long. MRPrePost-A parallel algorithm adapted for mining big data, 2014 IEEE Workshop on Electronics, Computer and Applications: 564-568.