A Survey on Association Rule Mining Algorithm with Hadoop MapReduce and Spark: A Big Data Perspective

A.Senthil kumar* & Dr.D.Hariprasad**
*Research Scholar, Sri Ramakrishna College of Arts and Science, Coimbatore-641 006
**Professor and Head, Department of Computer Application, Sri Ramakrishna College of Arts and Science, Coimbatore-641 006

Received: February 14, 2019 Accepted: March 24, 2019

ABSTRACT: *Association Rule Mining(ARM) is used to find frequent patterns, correlations, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories. Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It is used to find the frequently generated item sets and its corresponding association. These associations are used to make various business decisions. Traditional Association rule mining algorithm lacks performance and support when it comes to deal with large volume of data and also it is impossible to generate rules at a faster pace. These algorithms do not have mechanisms like Load balancing, data distribution etc. Big data is used to analyze the huge volume of data and get the association between the items and also checks if there are any trend in the data. This paper presents a survey on Association Rule Mining algorithms with MapReduce framework and Spark. An analysis is also made on the role of cloud computing for big data processing.*

Key Words:

1. Introduction

Big data is the term for collection of large and complex data sets that it becomes difficult to process using on-hand database system tools or traditional data processing application. It is concerned with the storage, retrieval, transmission and processing of extremely high volume, high velocity and high variety of data[1][2]. In the present world, huge volume of data is being generated almost in the industries, right from Internet, social media to retail stores, finance banking etc.. These datas are being generated from various source and they are in various format like files images, twitter etc. These massive data are very difficult to handle with the traditional formats. In these cases Big Data Analytics can be used to discover the insights from the available large volume and huge variety of data. And business can take decisions based on the findings. A business can understand its customers better with the help of big data analysis. It can predict what its customers want in advance and provide them a better service in addition to better products.

1.1 Hadoop

Hadoop[3] is a framework that distributes the processing of large data sets across different clusters of computers using simple models. It is used in solving the problems that involves massive amount of data and computation. Hadoop is the best framework for processing data in batches. The two main components of Hadoop are MapReduce and HDFS. MapReduce is a model that are used in the applications which involves processing of huge amount of data in parallel on large clusters. It is highly reliable and fault tolerant. In Map reduce the data is being stored in HDFS(Hadoop Distributed File System). As the name implies MapReduce has two different phase, map and reduce phase. The main components of MapReduce model are

Map Phase: Block of data is read and processed to produce key value pairs as an intermediate output.

Reduce Phase: Intermediate results which are key value pairs are taken as input to reduce phase. It aggregates all intermediate results and final output is produced.

Combiner : Mini reducer in the map phase. They perform a local reduce on the mapper results before they are distributed further. Once the combiner functionality is executed. it is then passed on to the reducer for further work.

Partitioner : used when more than one reducer are used. Decides which reducer is responsible for the particular key. It takes the Mapper or combiner result and sent it to the responsible reducer based on the key. The default partitioning function is the hash partitioning function where hashing is done on the key.

Shuffle and Sort : The process of transferring data from the mappers to reducers is shuffling. It is also the

process by which the system performs the sort.

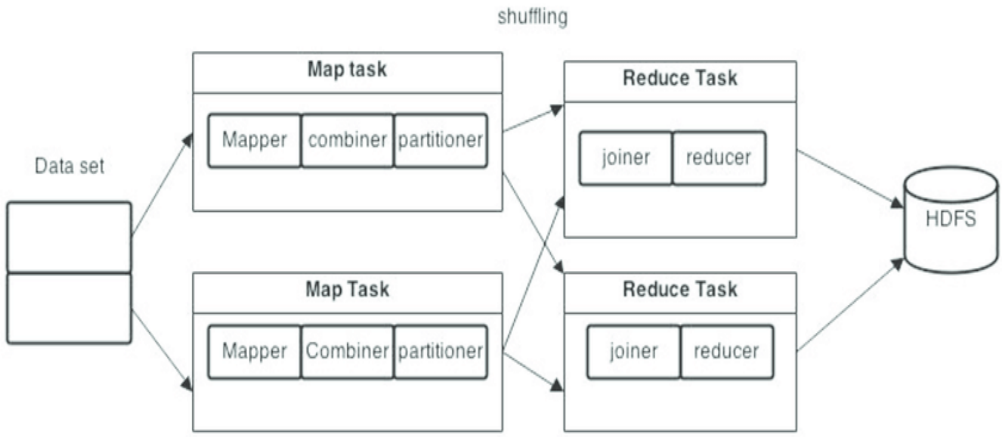

Fig 1 MapReduce Job Execution Flow

1.2 Spark

Spark[4] is based on Hadoop MapReduce and it extends the MapReduce model to efficiently use it for more types of computations, which includes interactive queries and stream processing. The main feature of Spark is its in-memory cluster computing that increases the processing speed of an application. It is designed to cover a wide range of workloads such as batch applications, iterative algorithms, interactive queries and streaming. Spark is an open-source cluster computing framework that supports in-memory computation with help of RDDs(Resilient Distributed Dataset) and Dataframe. Since it supports in memory processing the execution time becomes very less. It provides reusability, Fault Tolerance and real-time stream processing. It can easily integrate with hadoop. Apache spark is 100 times faster in memory than hadoop.

1.3 Association Rule Mining

Association Rule Mining (ARM) is the technique that is used to describe the association relationship among different items. There are many algorithm that are used to find the frequently used item. Apriori is one such algorithm that is widely used identifying all frequent patterns. It is mostly used in market basket analysis. In addition to this Apriori is used in text analysis and also in medical field. The traditional Apriori algorithm is very efficient. But when the data size is very huge the traditional Frequent Item Set Mining algorithms running on a single machine suffers performance degradation and were not able to scale. To overcome this problem association rule mining algorithm are designed for parallel execution on distributed network. Here Hadoop MapReduce and Spark is analyzed.

Implementing Apriori algorithm on map reduce framework improves the performance, by reducing the execution time and good memory utilization. Today the data is increasing in exponential form. The data could be structured and unstructured data. Hadoop map reduce solves that problem. The measures that are used in Apriori Algorithm are support confidence, support and lift.

2. Survey on Apriori algorithms for Association rule mining

In hadoop the performance of Apriori algorithm lacks when clusters are heterogeneous in nature. To improve the performance of Apriori algorithm in [5] authors proposed Parallel Mining of Frequent Itemset algorithms based on Map Reduced Framework on Heterogeneous Hadoop cluster. Frequent item sets are found in many ways using various rules and also various algorithms. When it comes to large volume of data, algorithms and techniques were not able to satisfy any one of the mechanisms like Load Balancing, data distribution or fault tolerance

Yaling Xun et al [6] proposed an algorithm called FiDoop-DP to boost the performance of parallel Frequent Itemset Mining(FIM) on Hadoop clusters. For finding FIM, many approaches focus more on load balancing which brings lack of correlation within the partition. So the overhead cost increased, it required more network. so it brought down the performance. In the FiDoop-DP all relevant transaction are moved into a data partition. By doing this the number of redundant transactions are reduced. FiDoop uses Voronoi diagram-based data partitioning technique which incorporate Locality Sensitive Hashing technique. This algorithm places highly similar transactions into a data partition to improve locality without creating an excessive number of redundant transactions. Here the network traffic and load is reduced as there are less

number of transactions. The drawback of FiDoop is when data-placement mechanism in HDFS on heterogeneous clusters is used the performance degrades.

In [7] to include enhancements to the apriori algorithm authors proposed improved filtered transaction Weighted based Apriori and Hash Tree based Apriori. In their model transaction filtering is performed to improve the performance in terms of execution time. FTWeightedHashT Apriori algorithm contains three main phases: Item Prediction phase, Filtration phase, and Tree Construction phase. The first phase scans the datasets for counting the frequency of the itemsets. In second phase weight is assigned to each items. The last phase performs the relevancy computation and reachability verification using weighted HastTree. The drawback is that the infrequent itemsets are completely removed from the dataset which also may an important itemset.

Yassir Imad [8] proposed PrePost algorithm based on the concept of N-lists using Hadoop for Big Data. PrePost algorithm presents a data structure named N-list, which is a modification of the vertical database for storing the information related to Association rule mining. PrePost scans the database twice to construct a PPC-Tree, and make use of PPC-Tree to generate the N-list of frequent 1-itemsets. In the mining process, the database does not require rescanning, only need to intersect the merger N-list. To improve the overall performance HPrePostPlus algorithm is used. It is a scalable Hadoop based method for frequent itemset mining that has no intermediate data, and small network communication

In [9] , authors proposed Kavosh: An effective Map-Reduce based association rule mining method. The problem faced while when the data volume is very high is the data locality, itemset support and skewness. Even Hadoop map reduce framework is ineffective. It does not have iteration and also data locality. Kavosh propose a method in which the input data is converted in a custom format names Kavosh format. The paper propose that these data are distributed over nodes and does not require inputs from other nodes. Here iterations are not required for extracting rules and also the load balance in done. Kavosh converts data into a unified format that helps nodes perform their tasks independently without the need to exchange data with other nodes. In addition, the proposed method compresses input data to facilitate data management. Another advantage is the lack of process skewness because it is possible to allocate a predefined amount of data to each node. This is not suitable for high-dimensional data since it requires many nodes.

Jerry Chun-Wei Lina et al[10] proposed an efficient Particle Swarm Optimization(PSO)-based high-utility itemset mining model(HUIM). They used binary PSO model which has four main process-Preprocessing, Particle encoding, Fitness evaluation and the updating process. The PSO based high utility itemset reduced the overall runtime. They also proposed privacy-preserving utility mining algorithm(PPUM), a genetic based evolutionary algorithm to secure the sensitive high utility itemsets and to reduce the computations. Two proposed models for the applications of HUIM and PPUM, not only generates high quality profitable itemsets according to the user-specified minimum utility threshold, but also enables the capability of privacy preserving for private or secure information (e.g., HUIs) in real-world applications.

Apriori algorithm are used to found Frequent itemset mining where entire data has to be processed to get the result. It is difficult if the data volume is very large. Map reduce can be used but the performance of Map Reduce also sometimes degrades. In [11] the authors used Hybrid Frequent Itemset Mining (HFIM) technique in Spark to optimize the execution time. HFIM uses vertical and horizontal layout of the dataset to find the association. According to the paper this overcomes the challenges that are there in Apriori Algorithm i.e. it needs huge memory and resource for computing. HFIM has two phases. In the first phases it scans the horizontal transaction data and produces k-frequent itemsets. In the second phase, support of candidate items is computed by exploring vertical data on each node. It has k sets of iteration and produces k-frequent itemset

The high utility itemset mining algorithm P-FHM+[12] is modeled to work in parallel environment of Apache Spark framework. A parallel version of FHM+ algorithm named P-FHM+ was developed which is suitable to run big data in the cluster computing environment. In traditional Frequent Itemset Mining, unit profit of items is not taken into consideration. Parallel High Utility Itemset mining discovers itemsets with their utility more than user defined utility threshold. High Utility itemset(HUI) is combination of internal and external utility. Internal utility is frequency of an itemset in a transaction and external utility is unit profit of items. Whole procedure of P-FHM+ are divided into 3 steps: search space division, node data generation and node data mining. The authors used apache Spark framework which provides fast and efficient performance to run big transaction data and it is more scalable.

Sainan Liu and Haoan Pan[13] proposed a rare itemsets mining algorithm based on RP-Tree and Spark framework. In their work in order to solve the problem of scanning the entire datasets, the data are

arranged vertically according to the transaction identifier. Vertical datasets are divided into frequent vertical datasets and rare vertical datasets. Then, it adopts the RP-Tree algorithm to construct the frequent pattern tree that contains rare items and generate rare 1-itemsets. Support of the itemsets is calculated by scanning the two vertical data sets. finally, it used the iterative process to generate rare itemsets. The experimental show that the algorithm can effectively excavate rare itemsets and minimizes the execution time.

Feng Zhang et al [14] proposed an efficient distributed frequent itemset mining algorithm (DFIMA). In their model they used matrix based pruning approach which can significantly reduce the amount of candidate itemsets. The algorithm has been implemented using Spark to further improve the efficiency of iterative computation. Numeric experiment results using standard benchmark datasets by comparing the proposed algorithm with the existing algorithm, parallel FP-growth, show that DFIMA has better efficiency and scalability. In addition, a case study has been carried out to validate the feasibility of DFIMA.

When it comes to mine massive data, many algorithms failed to prove efficiency because of the limitation in processing capacity, storage capacity, and main memory constraints. In [15] cloud based approach is used to improve the efficiency. Hadoop-MapReduce is an efficient, scalable, and simplified programming model for massive data processing and it is also available on cloud environment. Cloud computing offers huge computing resources, and capacities to solve big data challenges. They proposed a parallel Transaction Reduction MapReduce Apriori algorithm (TRMR-Apriori) which reduces unnecessary transaction values and transactions from the dataset in parallel manner to overcome above problems. The experiments show that TRMR-Apriori is able to achieve better execution time to discover frequent itemset with different condition on homogeneous computing environment using Hadoop-MapReduce platform in cloud. Overall, the TRMR-Apriori shows the strength to extract the frequent itemset from massive dataset in cloud.

Traditional data analysis algorithms can no longer meet the needs of big data analysis. The rise of Cloud computing provides a new solution to this problem. In [16] Apriori algorithm is used with the Cloud computing technology and big data Hadoop MapReduce. Based on its limitations, it proposes an optimization scheme and introduces the MapReduce model in Cloud computing to achieve parallelization. A MapReduce-based frequent item set mining method is used in their work to improve the efficiency of the algorithm and reduce the overhead required for algorithm execution.

Conclusion

Association Rule Mining is used to find frequent patterns and its associated relationship among item sets. Apriori is one of the association rule mining algorithm. For a large dataset, traditional association rule mining algorithm is not suitable as it requires more scanning of datasets to find the frequent item set. To overcome this problem association rule mining algorithm are designed for parallel execution on distributed network. When used in big data perspective its performance can be improved. Association rule mining can be used in both hadoop framework and spark framework. Hadoop is used for processing the jobs in batches. Spark framework is used for In memory and real time processing.

References

1. J. S. Ward and A. Barker, "Undefined by data: a survey of big data definitions", http://arxiv.org/abs/1309.5821v1.
2. A. Siddiqa, "A survey of big data management: Taxonomy and state-of-the-art", J. Netw. Comput. Appl., vol. 71, pp. 151-166, Aug. 2016.
3. Apache Hadoop, http://hadoop.apache.org
4. Apache Spark [Online]. Available: http://spark.apache.org/
5. Dhanashree Shirke, Deepti Varshney , "Parallel Mining of Frequent Itemsets in Hadoop Cluster Having Heterogeneous Nodes" , International Journal of Advance Research in Computer Science and Management Studies, July 2017.
6. Yaling Xun , Jifu Zhang , Xiao Qin "FiDoop-DP: Data Partitioning in Frequent Itemset Mining on Hadoop Clusters " , IEEE Transactions on Parallel and Distributed Systems, January 2016.
7. Sarem M. Ammar and Fadl M. Ba-Alwi , " Improved FTWeightedHashT Apriori Algorithm for Big Data using Hadoop-MapReduce Model ", Journal of Advances in Mathematics and Computer Science April 2018.
8. Yassir Rochd , Imad Hafidi, "Performance Improvement of PrePost Algorithm Based on Hadoop for Big Data, International Journal of Intelligent Engineering and Systems,", April 24, 2018.
9. Mohammadhossein Barkhordari and Mahdi Niamanesh , "Kavosh: an effective Map-Reduce-based association rule mining method , Journal of BigData" , July 14 2018

10. Jerry Chun-Wei Lina , Wensheng Gana , Philippe Fournier-Viger b, Lu Yang a , Qiankun Liua , Jaroslav Frndac , Lukas Sevcikc , Miroslav Voznak , " High utility-itemset mining and privacy-preserving utility mining , ELSIEVER ScienceDirect " ,10 December 2015
11. Krishan Kumar Sethi, Dharavath Ramesh , " HFIM: a Spark-based hybrid frequent itemset mining algorithm for big data processing ", Springer Science+Business Media New York , Jan 2017
12. Krishan Kumar Sethia , Dharavath Ramesh a, Damodar Reddy Edlab " P-FHM+: Parallel high utility itemset mining algorithm for big data processing, International Conference on Computational Intelligence and Data Science (ICCIDS 2018) " , 2018
13. Sainan Liu, Haoan Pan, "Rare itemsets mining algorithm based on RP-Tree and Spark framework", 6th International Conference on Computer-Aided Design, Manufacturing, Modeling and Simulation (CDMMS 2018), 2018.
14. Feng Zhang, Min Liu, Feng Gui, Weiming Shen, Abdallah Shami, Yunlong Ma," A distributed frequent itemset mining algorithm using Spark for Big Data analytics", Cluster Computing, 18:1493–1501 DOI 10.1007/s10586-015-0477-1, Springer, 2015.
15. Sayeth Saabith, Elankovan Sundararajan Azuraliza Abu Bakar "A Parallel Apriori-Transaction Reduction Algorithm Using Hadoop-Mapreduce in Cloud", Asian Journal of Computer Science and Information Technology, April 2018.
16. Wang Rui, "Research on Apriori Algorithm Optimization of Cloud Computing and Big Data in Software Engineering", 2018 5th International Conference on Electrical & Electronics Engineering and Computer Science (ICEEECS), 2018.