A Survey Paper on Hybrid Algorithm with Map Reduce Framework to Mine Distributed Association Rules from Big Data

Jaini A. Doshi[1] Rakesh Shah[2] Jaini A. Doshi[3]
[1]Student [2]Professor [3]Assistant Professor
[1,2,3]Department of Computer Engineering
[1,2,3]Grow More Faculty of Engineering, Himmatnagar, India

Abstract— When its big data, it's extremely large data sets. These datasets are analyzed computationally to reveal patterns, trends, and associations. That could be related to human behavior and interactions and product reordering ratio or understanding the symptoms or related signs of a disease. Many data analysis techniques are already defined and used by researchers. Results are still showing the scope of improvement. Based on the volume, variety, and velocity of data, the techniques are needed to be used or improved. Association rule mining is one of the technique to solve issues of accuracy in retrieved results. They are used to detect changes in customer behavior, buying trends and reasons that affect such process. Researches till date has proven the results are better than the earlier one. Though several methods have been suggested for the extraction of association rules, problems arise when data is in growing pattern with large volume. To overcome such issue, we propose, in this paper, a hybrid approach based on ARM techniques with Map Reduce framework, modified for processing large volumes of data in an increasing manner. Furthermore, because real life databases lead to a huge number of rules' including many redundant rules, our algorithm proposes to mine a compact set of rules with no loss of information. The results of experiments tested on large real world datasets highlight the relevance of mined data. Additionally in this research, the experiments are performed in continuous growing data which still yields comparative results.

Key words: Hybrid Algorithm, Map Reduce Framework, Big Data

I. INTRODUCTION

Big data means really a big data, it is a collection of large datasets that cannot be processed using traditional computing techniques. Big data is not only a data, relatively it has become a complete subject, which involves various tools, techniques and frameworks. Big Data includes massive volume, high velocity, and extensible variety of data. The data in big data will be of three types.

- Structured data: Relational data.
- Semi Structured data: XML data.
- Unstructured data: Word, PDF, Text, Media Logs.

Association rule mining is a methodology that is used to discover unknown relationships hidden in big data. It is used for personalized marketing promotions, smarter inventory management, product placement strategies in stores, and a better customer relationship management. The most known algorithm is the Apriori algorithm, but the FP Growth algorithm is often used.

MapReduce is a programming paradigm or model used to process large datasets with a parallel distributed algorithm on a cluster. In Big Data Analytics, MapReduce plays a critical role. When it is combined with HDFS we can use Map Reduce to handle Big Data.

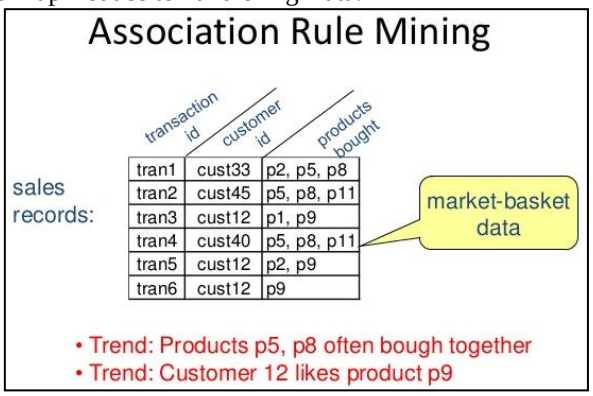

Fig 1. Association Rule Mining

The basic unit of information used by MapReduce is a key-value pair. All the data whether structured or unstructured needs to be translated to the key-value pair before it is passed through the MapReduce model.

II. LITERATURE SURVEY

[1] When the data is distributed and grows large, the efficiency decreases since the communication cost and the memory usage become improper. Researchers are focusing on accelerate the processing of the ARM algorithms by parallelizing and executing them on multiple clusters. To conquer the above drawbacks, we introduce a distributed algorithm for mining generic basis of ARM based on the cloud computing framework MapReduce. Indeed, generic basis offer a compact set of informative and no redundant rules, with the guarantee of no loss of information.

[2] This paper presents a parallel and distributed solution to the problem of extracting frequent item sets from Big Weighted Datasets. The proposed system, running on a Hadoop cluster, overcomes the limitations of state-of the art approaches in coping with datasets enriched with item weights.

[3] Frequent Item set Mining is one of the most popular techniques to extract knowledge from data. However, these mining methods become more difficult when they are applied to Big Data. Fortunately, current improvements in the field of parallel programming provide many tools to tackle this problem. Hadoop technology can provides a better platform to overcome the shortcomings mining techniques.

[4] To mitigate high communication and reduce computing cost in MapReduce-based FIM algorithms, they developed FiDoop-DP, which exploits correlation among transactions to partition a large dataset across data nodes in a Hadoop cluster. FiDoop-DP is capable to (1) partition transactions with high similarity together and

(2) group highly correlated frequent items into a list. One of the salient features of FiDoop-DP lies in its capability of lowering network traffic and computing load through reducing the number of redundant transactions, which are transmitted among Hadoop nodes.

[5] Pattern discovery is the important part of knowledge discovery in Database, comes under Data mining. To find out useful patterns, association rule mining is one of the most popularized and revealing technique in data mining. Association rule mining plays a key role in decision making by discovering useful relations between attributes in the database. For this, first frequent itemsets need to calculate followed by Candidate itemset. While generating frequent itemsets, frequent-1 itemset can be generated easily. But frequent 2-itemsets suffered from both time and space complexity experimental results shows the time required to access the items for pairing with each other and to prune the unnecessary itemsets using defined min_sup get minimizes through HBase with a different number of nodes. From this, we reached the conclusion that HBase is very effective to random read/write access.

[6] A service recommender method in light of user Preferencees has huge applications particularly into the era of Big Data. Thus it's been chosen to have propose an service recommendation method, based on preferences of user to create proper recommendation. Preferences of end-user and preferences of past users is considered for recommendation to service. Our customized service recommendation system is giving more accurate service to the users. Likewise, a mapreduce framework in hadoop has been utilized to enhance the efficiency and scalability in massive data surroundings.

[7] In this paper, we mainly address the challenge of using the MapReduce model to parallelize Apriori. Hadoop, the fundamental cloud computing framework, can handle many tough problems, including parallelization, concurrency control, network communication and fault tolerance. The novelty of the algorithm lies in the ability of parallel Apriori to take full advantage of what Hadoop can provide. It can be easily applied to many commodity machines to deal with mass data without consider the synchronization problem.

[8] MapReduce is very lucrative for parallel processing of big data on large cluster of commodity computers. In this paper, we mostly focus on the parallelization of Apriori algorithm on MapReduce framework. The MapReduce computing model is well resemble to the computation of frequent itemsets in Apriori algorithm. We reviewed various planned approaches to parallelize Apriori on Hadoop distributed framework. They are categorized on the basis of Map and Reduce functions used to implement them e.g. 1-phase vs. k-phase, I/O of Mapper, Combiner and Reducer and using functionality of Combiner inside Mapper etc. Scheduling invocations and waiting time overheads are major bottleneck in performance of algorithms and it is addressed by FPC and DPC techniques. 1 and 2-itemsets are in huge number among all k-candidates so we can handle it separately and input it to the third phase of mapreduce. For this triangular matrix data arrangement is used to count the support of 1 and 2-itemsets in one step. All these techniques may not be equally exclusive and some of them can be integrated to increase the performance of resulting algorithm.

III. CONCLUSION

In this paper, we have introduced a hybrid approach for distributed mining of association rules from big data, based on the Map Reduce framework.

We proposed to mine generic basis of association rules. Indeed, generic basis offer a compact set of informative and non-redundant rules, with the guarantee of no loss of information. The obtained affirm the efficiency of our algorithm in ARM from large datasets.

REFERENCES

[1] Marwa Bouraoui, Ines Bouzouita, Amel Grissa Touzi, "Hadoop based Mining of Distributed Association Rules from Big Datas",IEEE, December 21-23, 2017.

[2] Divya.M.G, Nandini.K, Priyanka.K.T, Vandana.B," Weighted Itemset Mining from Bigdata using Hadoop", ICICN16, CSE, RRCE.

[3] Tushar M. Chaure, Kavita R. Singh," Frequent Itemset Mining Techniques - A Technical Review", WCFTR 2016.

[4] Yaling Xun, Jifu Zhang, Xiao Qin, Senior Member," FiDoop-DP: Data Partitioning in Frequent Itemset, International Conference on Innovations in Computing & Networking (ICICN16), CSE, RRCE Mining on Hadoop Clusters", IEEE 2016.

[5] Ashwini A. Pandagale, Anil R," Hadoop-HBase for Finding Association Rules using Apriori MapReduce Algorithm",IEEE, May 20-21, 2016,

[6] Vijay M Bande, Ganesh K Pakle,"CSRS: Customized Service Recommendation System for Big Data Analysis using Map Reduce", Conference: 2016 International Conference on Inventive Computation Technologies (ICICT)

[7] Xin Yue Yang, Zhen Liu, Yan Fu," MapReduce as a Programming Model for Association Rules Algorithm on Hadoop" , IEEE - august-2010

[8] Sudhakar Singh, Rakhi Garg, P K Mishra, "Review of Apriori Based Algorithms on MapReduce Framework", ICC-2014.

