A Survey on Different Techniques for Mining Frequent Itemsets

Anbumalar Smilin.V[1], S.P.Siddique Ibrahim[2]

Abstract— **Data mining faces a lot of challenges in the big data era. Association rule mining algorithm is not sufficient to process large data sets. Apriori algorithm has limitations like the high I/O load and low performance. The FP-Growth algorithm also has certain limitations like less internal memory. Mining the frequent itemset in the dynamic scenarios is a challenging task. A parallelized approach using the mapreduce framework is also used to process large data sets. The various techniques for mining the frequent itemsets have been discussed.**

***Index Terms*— Apriori algorithm, Big data, Data mining, Frequent itemset mining.**

I. INTRODUCTION

Data mining faces a lot of challenges in this big data era. The term big data refers to the voluminous amount of data which is difficult to store, analyze and process. The Big data includes various technologies to obtain useful information from the huge amount of data. The mining of big data is a difficult process. One of the main challenge of the data mining is the finding the frequent itemset. Many different approaches are used for obtaining the frequent itemset. The frequent itemsets are used to obtain the association rules. The association rules are used for obtaining the regularities in the products being bought. It is the basis for the decision making in the business analysis. The data mining and the big data has lead to the emergence of the business intelligence. The data is gathered and analyzed to improve the profit of the organizations by adopting many techniques.

The big data has certain unique characteristics like the volume, velocity, variety, veracity and value [8]. The big data has applications in many fields like the healthcare, governmental organizations, manufacturing industries, media, retail banking and research. The big data requires many technologies to obtain useful information. Big data has become the important area of research today. With the use of the social media like the facebook, twitter and many other social platforms the data is growing in a rapid manner. The fast growing data becomes the basis for the big data.

***1.Anbumalar Smilin.V**,CSE Department,Kumaraguru College of Technology,Coimbatore,India,MobileNo:9585510132,*
***2.S.P.Siddique Ibrahim**,CSE Department,Kumaraguru College of Technology,Coimbatore,India,MobileNo:9894442939,*

II. CHARACTERISTICS OF BIG DATA

A. Volume

It refers to the vast amount of data generated every second. The data is growing in a rapid manner by the use of social media like the facebook and the twitter. In the facebook a lots of messages and the pictures are shared every second. Many emails are used for the communication purposes. These all lead to the vast amount of data generation.

B. Velocity

It refers to the speed at which the new data is generated and the speed at which the new data moves around. One of the best examples is that we can easily detect the credit card fraudulent within seconds. Another example is that messages in the facebook go viral within seconds.

C. Variety

It refers to the different types of data that can be used. The big data can handle both the structured and the unstructured data. Messages, photos, audio and many other different types of data can be processed.

D. Veracity

It refers to the trustworthiness of the data. With rapid growing of data and its huge volume quality of the data is very much less. The big data technology helps to deal with these kinds of data.

E. Value

It is one the most important aspect of big data. The value must be extracted from the data. The value plays an important role in analysis and decision making.

III. APPLICATIONS OF BIG DATA

A. Healthcare

With the advent of the big data the diagnosis of the diseases has become simpler. Diseases can be diagnosed at a very

early stage. The epidemics can be predicted earlier and the living standards of the people are improved.

B. Governmental organizations

Many of the governmental organizations use the big data as a tool to overcome many problems faced by the government. The big data is much useful in the election campaigns.

C. Manufacturing Industries

The use of the big data in the manufacturing industries helps in improving the quality of the product. A huge amount of the sensory data and the historical data makes up the big data which helps in the manufacturing.

D. Media

Big data plays an important role in the media. It provides more technologies to gather much information from the targeted consumers which mainly help the advertisers for the marketing purposes.

IV. CURRENT RESEARCH IN BIG DATA

The American society of engineering education demonstrated the encrypted search and the cluster formation in big data in March 2014. The security of big data is one of the latest research topics. The White house has announced more than 200 million dollars for the big data research. The European commission also has announced funds for the big data public private forum to discuss the issues related to the big data. Research is also done in the sampling of the big data.

V. DATA MINING

The main goal of the data mining is to extract useful information from the large datasets. Many hidden patterns are obtained from the data sets. The data mining involves many tasks like the anomaly detection, Association rule learning, clustering, classification, regression and summarization.

VI. ASSOCIATION RULE MINING

It is an important data mining model used to find the interesting relationship between the data in the database. It is mainly used for the Market basket analysis to help improve the business activity. Association rules are obtained using two main criteria the support and the confidence.
The support indicates how frequently the items appear in the database. The confidence refers to the number of times the if/then statements have found to be true.

VII. FREQUENT ITEMSET MINING

It is mainly used for market basket analysis. The regularities in the shopping behavior of the customers can be found using the frequent itemset mining. The products which are bought together can be found using the frequent itemset mining.

VIII. RARE ITEMSET MINING

Rare itemset mining refers to the mining of itemsets that do not occur frequently in the database. The rare itemsets are also referred to as the contrasting frequent itemsets. Rare itemset mining refers to the mining of contradicting beliefs and exceptions mainly in the field of biology or medicine.

IX. DYNAMIC THRESHOLD VALUE

The threshold value refers to the values like the minimum support count which changes dynamically in the incremental databases. The minimum support count is used for obtaining the frequent itemsets. The threshold values play an important role in the pattern mining.

X. APRIORI ALGORITHM

The Apriori algorithm is mainly used to find the frequent itemsets. The frequent itemsets found are used to form the association rules. The Apriori algorithm finds the frequent individual items in the database and extends them to larger and larger itemsets. It uses a uniform minimum support threshold.

XI. FP-GROWTH ALGORITHM

FP-Growth algorithm is mainly used to find the frequent itemset without candidate itemset generation. Two steps are followed in the FP-growth algorithm. In the first step, the FP-tree is constructed. In the second step the frequent itemset are extracted from the FP-tree.

XII. HADOOP

Hadoop is used for processing large data sets. It works in a distributed environment. Hadoop uses the map reduce framework to process the data in parallel across different CPU nodes. The hadoop framework is written in the java language.
The Hadoop framework consists of four modules:

A. Hadoop Common

It consists of the java libraries and utilities. The libraries provide the java files and the scripts to start the hadoop.

B. Hadoop YARN

Job scheduling and cluster resource management is done by this module.

C. Hadoop Distributed File System

It is a distributed file system used to run on large clusters of machines.

D. Hadoop Mapreduce

It is mainly used for parallel processing of large amounts of data. It includes two main tasks the map task and the reduce task.

XIII. HADOOP DISTRIBUTED FILESYSTEM

HDFS is mainly used for storage and computation of large amount of data across various clusters of servers. The HDFS uses master/slave architecture. The metadata is stored on a server called the namenode which acts as the master. The datanode is a server which stores the application data. There is more than one datanode which acts as the slave.

XIV. MAPREDUCE

Hadoop Mapreduce framework is used for parallel processing of large amounts of data. It involves two main tasks, the map task and the reduce task.

A. Map task

It takes the input data and produces the key/value pairs. The key/value pair is the output of the map task. The map task is performed by the mapper.

B. Reduce task

The reducer obtains the output data from the mapper and process the data. The processed data is stored in the HDFS. Reducer has three main tasks which are the shuffling, sorting and reducing.

XV. LITERATURE SURVEY

A. Mining Frequent Patterns without Candidate Generation:

JW.Han, J. Pei and YW.Yin have proposed a new method called as the Frequent Pattern tree method. The frequent pattern tree stores the compressed information in an extended prefix tree structure. The frequent patterns are stored in a compressed form. A FP-tree based mining method known as the FP-growth is developed. The proposed algorithm helps in mining the frequent itemsets without the candidate set generation.

Three techniques were employed to achieve the efficiency of mining:

1) A large database is converted into a small data structure to avoid the repeated database scans which is said to be costly.

2) It adopts a pattern frequent growth method to avoid generating large candidate sets which is very costly.

3) The mining tasks are divided into smaller task which is very useful in reducing the search space.

The FP-tree based mining also has many research issues like the SQL-based FP-tree structure with high scalability, mining frequent patterns with constraints and using FP-tree structure for mining sequential patterns.

B. PFP: Parallel FP-growth for query recommendation

H. Li, Y. Wang, D. Zhang, M. Zhang, and E. Chang have proposed a parallel FP-Growth algorithm. In parallel FP-growth algorithm the mining task is divided into a number of partitions. Each of the partitions is provided to the different machines and each partition is computed independently. To overcome the challenges faced by the FP-growth algorithm like the storage, distribution of computation and highly expensive computation parallel FP-growth algorithm is proposed. The PFP algorithm consists of five steps. In the first step, the database is divided into small parts. In the second step the mapper and the reducer are used to do the parallel counting. In the third step the frequent items are grouped. In the Fourth step the FP-tree is constructed and the frequent itemsets are mined. In the fifth step the local frequent itemsets are aggregated. The PFP algorithm is effective in mining tag-tag associations and WebPage-WebPage associations which are used in query recommendation or any other search.

C. MREclat: an Algorithm for Parallel Mining Frequent Itemsets

Zhigang Zhang, Genlin Ji, Mengmeng Tang have proposed a parallel algorithm MREclat based on Map/Reduce framework. In the vertical layout algorithm the frequent patterns are mined using the algorithm Eclat. The algorithms for mining frequent patterns in horizontal layout databases are different from the algorithms for mining vertical databases like the Eclat. A parallel algorithm MREclat which uses a mapreduce framework has been proposed to obtain the frequent itemsets from the massive datasets.

Algorithm MREclat consists of three steps. In the initial step, all frequent 2-itemsets and their tid-lists are obtained from transaction database. The second step is the balanced group step, where frequent 1-itemsets are partitioned into groups. The third step is the parallel mining step, where the data got in the first step is redistributed to different computing nodes. Each node runs an improved Eclat to mine frequent itemsets. Finally, MREclat collects all the output from each computing node and formats the final result.

MREclat uses the improved Eclat to process data with the same prefix. It has been proved that MREclat has high scalability and good speedup ratio.

D. Parallel Mining Frequent Patterns over Big Transactional Data in Extended MapReduce

Hui Chen, Tsau Young Lin, Zhibing Zhan and Jie Zhong have proposed a parallel algorithm for mining frequent pattern in large transactional data. It uses an extended MapReduce Framework. A number of subfiles are obtained by splitting the mass data file. The bitmap computation is performed on each subfile to obtain the frequent patterns.

The frequent pattern of the overall mass data file is obtained by integrating the results of all subfiles. A statistic analysis method is used to prune the insignificant patterns when processing each subfile. It has been proved that the method is scalable and efficient in mining frequent patterns in big data.

E. An Improved Parallel Association Rules Algorithm Based on MapReduce Framework for Big Data

Xinhao Zhou and Yongfeng Huang have proposed an improved parallel Apriori algorithm. The mapreduce framework is used to find the count and the candidate set generation. The proposed algorithm is compared with the existing traditional apriori algorithm. The time complexity of both the algorithms has been used to compare the performance of the algorithms. It has been proved that the proposed algorithm is more efficient compared to the traditional algorithm.

F. MRPrePost-A parallel algorithm adapted for mining big data

Jinggui Liao, Yuelong Zhao and Saiqin Long have proposed a MRPrePost algorithm.It is a parallel algorithm which is implemented using the Hadoop platform. The MRPrePost is an improved PrePost algorithm which uses the mapreduce framework. The MRPrePost algorithm is used to find the association rules by mining the large datasets.

The MRPrePost algorithm has three steps. In the first step the database is divided into the data blocks called the shards which are allocated to each worker node. In the second step the FP-tree is constructed. In the final step the FP-tree is mined to obtain the frequent itemsets. Experimental results have proved that the MRPrePost algorithm is the fastest.

G. Frequent Itemset Mining for Big Data in social media using ClustBigFIM algorithm

Sheela Gole and Bharat Tidke have proposed a new method, ClustBigFIM. Large datasets are mined using the Mapreduce framework in the proposed algorithm. BigFIM algorithm is modified to obtain the ClustBigFIM algorithm. ClustBigFIM algorithm provides scalability and speed which are used to obtain useful information from large datasets. The useful information can be used to make better decisions in the business activity. The proposed ClustBigFIM algorithm has four main steps. In the first step the proposed algorithm uses K-means algorithm to generate the clusters. In the second step the frequent itemsets are mined from the clusters. By constructing the prefix tree the global TID list are obtained. The subtrees of the prefix tree are mined to obtain the frequent itemsets. The proposed ClustBigFIM algorithm is proved to be more efficient compared to the BigFIM algorithm.

H. Distributed FP-ARMH Algorithm in Hadoop MapReduce Framework

Surendar Natarajan and Sountharrajan Sehar have proposed a new algorithm named Association rule mining based on Hadoop (ARMH). The proposed algorithm utilizes the clusters effectively and helps in mining frequent pattern from large databases. The workload among the clusters is managed using the hadoop distributed framework. The hadoop distributed file system stores the large database. Three mapreduce jobs have been used to mine the frequent patterns. The FP-tree is produced in the first mapreduce job. The FP-tree is stored in the array data structure format. The FP-tree array data structure is given as the input for the second mapreduce job. The second mapreduce job produces condition pattern base as output for all the item sets. The third map reduce job takes the condition pattern base as input and produce frequent pattern corresponding to the item set to which the conditional pattern base has been created. In third map-reduce program, the map job would produce the conditional FP-tree for conditional pattern base and reduce job would produce frequent pattern from the corresponding conditional FP-tree. The conditional FP-tree is also stored in array data structure. From the conditional FP-tree the frequent patterns are obtained. The proposed ARMH algorithm utilizes the hadoop cluster effectively to obtain the frequent pattern from large databases.

I. Incremental FP-Growth Mining Strategy for Dynamic Threshold Value and Database Based on MapReduce

Xiaoting Wei, Yunlong Ma, Feng Zhang, Min Liu and Weiming Shen have proposed a parallelized incremental FP-Growth mining algorithm. The large scale data is processed using a mapreduce framework. The proposed incremental algorithm is effective when threshold value and original database change at the same time.

The parallel incremental FP growth algorithm comprises of seven stages. There are four mapreduce phases. In the first step the database is divided into small chunks and the frequent list is obtained. In the second step the FP-tree is constructed and the local frequent itemsets are obtained. The third step is to aggregate the frequent itemsets mined. In the fourth step the database is updated. The fifth step is to update the frequent item list. In the sixth step the FP-tree is constructed and the local frequent itemsets are mined. In the final step the local frequent itemsets are aggregated. The algorithm proves to be more effective in the incremental databases.

XVI. CONCLUSION

The mining of frequent itemset is an important research area in the field of data mining. The association rules are formed using the frequent itemset mined. Many different methods have been proposed for mining the frequent itemsets. The literature survey illustrates the different approaches for mining frequent itemsets. Each method has its own advantages and disadvantages.

REFERENCES

[1]JW.Han, J.Pei and YW.Yin, “Mining Frequent Patterns without Candidate Generation”, International Conference on Management of Data, vol. 29(2), 2000, pp. 1-12.

[2] H. Li, Y. Wang, D. Zhang, M. Zhang, and E. Chang, "PFP: Parallel FP-growth for query recommendation", Proceedings of the 2008 ACM Conference on Recommender Systems, 2008, pp. 107-114.
[3] Zhigang Zhang, Genlin Ji, Mengmeng Tang, "MREclat: an Algorithm for Parallel Mining Frequent Itemsets", 2013 International Conference on Advanced Cloud and Big Data.
[4] Hui Chen, Tsau Young Lin, Zhibing Zhang and Jie Zhong, "Parallel Mining Frequent Patterns over Big Transactional Data in Extended MapReduce", 2013 IEEE International Conference on Granular Computing.
[5] Xinhao Zhou, Yongfeng Huang, "An Improved Parallel Association Rules Algorithm Based on MapReduce Framework for Big Data", 2014 11th International Conference on Fuzzy Systems and Knowledge Discovery.
[6]Jinggui Liao, Yuelong Zhao, Saiqin Long, "MRPrePost-A parallel algorithm adapted for mining big data", 2014 IEEE Workshop on Electronics, Computer and Applications.
[7] Sheela Gole, Bharat Tidke, " Frequent Itemset Mining for Big Data in social media using ClustBigFIM algorithm", International Conference on Pervasive Computing.
[8] Siddique Ibrahim S P, "Extract data in Large Database with Hadoop", International Journal of Advances in Engineering and Scientific Research", Vol.1 2014, pp. 5-9.
[9] Surendar Natarajan, Sountharrajan Sehar, "Distributed FP-ARMH Algorithm in Hadoop Map Reduce Framework", 2013 IEEE.
[10] Xiaoting Wei, Yunlong Ma , Feng Zhang, Min Liu, Weiming Shen , "Incremental FP-Growth Mining Strategy for Dynamic Threshold Value and Database Based on MapReduce", Proceedings of the 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design.
[11] Siddique Ibrahim S P, Priyanka R, "A Survey on Infrequent Weighted Itemset Mining Approaches" , 2015, IJARCET, Vol.4, pp. 199-203.

Anbumalar Smilin V was born in Coimbatore, India. She received B.E-computer science and engineering from KGISL institute of technology.Currently she is pursuing M.E-computer science and engineering in Kumaraguru college of technology, Coimbatore.

S.P.Siddique Ibrahim was born in India**.** He received M.E-computer science and engineering form Bannari Amman Institute of Technology. Currently he is pursuing his PhD in VIT University. He is Assistant Professor in Kumaraguru College of Technology.

