Impact Factor (SJIF): 3.632

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444
Volume 3, Issue 5, May-2016

Survey on Big Data as Frequent Itemset Mining Technique

Tamanna Jethava, Rahul Joshi

Department of Information & Technology
Parul Institute of Engineering & Technology, Baroda, Gujarat, India

ABSTRACT

Abstract - Frequent Item sets plays an essential role in many Data Mining tasks that try to find interesting patterns from database. Typically it refers to a set of items that frequently appear together in transaction dataset. There are several Mining algorithm are being used for frequent item set mining ,yet most do not scale to the type of data we are presented with today, so called “BIG DATA”. Big Data is collection of large data sets. Our approach is to work on frequent item set mining over the large dataset with scalable and speedy way. Big Data basically work with Map Reduce along with HDFS is used to find out frequent item sets from Big Data on large cluster. This paper focus on using pre-processing & mining algorithm as hybrid approach for big data over Hadoop platform. Also survey on big data mining with New algorithm which Provide scalability & effectiveness .Also increases it’s time complexity with the help of parallel processing over cluster on Hadoop.

Keywords – Mining of frequent itemset, Big Data, MapReduce, Hadoop platform, HDFS.

I. INTRODUCTION

The data mining algorithm has been widely used in different fields. As the algorithm needs a lot of information as the basis for the exploration, quickly finding the useful information from a huge dataset is important. But when a single host cannot afford large amounts of computation, using grid computing, cloud computing the Hadoop and MapReduce computing frameworks, etc., the use of multiple computers to compute the mining algorithm has been put forward[7]. Social networking is used widely as basic communication media with exponential growth. Sites such as Twitter, Facebook, Linked In and My Space are frequently used by people. Facebook had more than 845 million active users in February 2012 . Research is done on Big Data which shows that it can create significant value in the word’s economy improving productivity of companies and public sector Storing huge amount of data won’t have any value without KDD (Knowledge discovery in Database) which is process of finding information from database and extracted knowledge can be used for making effective business decisions Discovery of association rules from large database is one of the problems in KDD (Knowledge Discovery in Database). One of the most important areas of data mining is association rule mining; it is a task to find all items or subsets of items which frequently occur and the relationship between them by using two main steps: finding frequent item sets and generating association rules. Frequent Item set Mining (FIM) tries to discover information from database based on frequent occurrences of an event according to the minimum frequency threshold provided by user [1]. Basically for small amount of data it is easy with simple mining algorithm like Apriori, FP-Growth etc. but for Big Data mining it is complex to work on this basic algorithms for better efficiency of large amount of data we need algorithm which manages mining of Big Data[2].

This paper contain introduction to hybrid approach of basic data mining algorithm for big data using MapReduce.

II. Related Terms

A. Frequent Item set mining

Frequent sets play a vital role in Data mining basic which try to find out interesting pattern from database like association rule, correlations, clusters, sequences. Among these ARM is popular problem, because discovery of the sets of items, products, symptoms and characteristics which occur together in the given database can be seen as one of the most basic task in data mining [18]. Let I be the set of items.

- A transaction over I is a couple T = (tid, I), where tid is the transaction identifier & I is the set of items from I.

- A transaction T = (tid, I) is said to support a set X.
- The cover of a set X in D consists of the set of transaction identifiers of transactions in D that support X.
- The support of a set X in D is the number of transactions in the cover of X in D.
- The frequency of a set X in D is the probability that X occurs in a transaction, or in other Words, the support of X divided by the total number of transactions in the database.
- A set is called frequent if its support is no less than a given absolute minimal support threshold min_sup with 0> min_sup_{abs} >|D|.

FRM (frequent itemset mining) is the first step of ARM (association rule mining), after having all the frequent itemsets, for every single frequent itemset, it enumerate all the possible association rules.

B. Big Data

Big Data concern large-volume, complex, growing data sets with multiple, sources. The Big Data is nothing but a data, available at heterogeneous, autonomous sources, in extreme large amount, which get updated in fractions of seconds. Big Data spam deal with three properties.

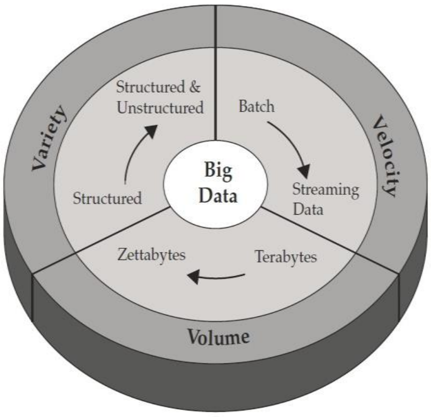


Fig 1: Big Data Properties [4]

Volume

The quantity of generated data is important in this context. The size of the data determines the value and potential of the data under consideration, and whether it can actually be considered big data or not. The name ‘big data’ itself contains a term related to size, and hence the characteristic.

Variety

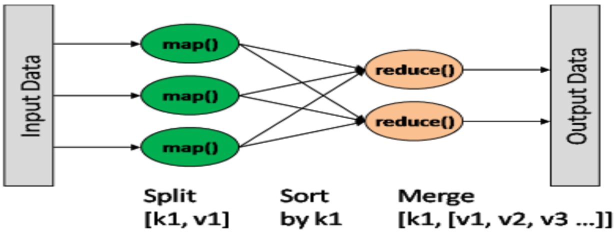
The type of content, and an essential fact that data analysts must know. This helps people who are associated with and analyse the data to effectively use the data to their advantage and thus uphold its importance.

Velocity

In this context, the speed at which the data is generated and processed to meet the demands and the challenges that lie in the path of growth and development.

C. MapReduce

MapReduce is a parallel programming model for writing distributed applications devised at Google for efficient processing of large amounts of data (multi-terabyte data-sets), on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. The MapReduce program runs on Hadoop which is an Apache open-source framework [6].



MapReduce framework has two phases, Map phase and Reduce phase. Map and reduce functions are used for large parallel computations specified by users. Map function takes chunk of data from HDFS in (key, value) pair format and generates a set of (key', value') intermediate (key, value) pairs. MapReduce framework collects all intermediate values which are bind to same intermediate key and same are passed to reduce function; it is formalized as, map :: (key, value) → (key' , value'); Value of map function is used by reduce function.

Intermediate key details are received by reduce function, that are merged together. The intermediate values are provided to reduce function through iterator, by using which too large values fit in memory, formalized as, reduce :: (key', list (value')) → (key'', value'')

Output can have one or more output files which are written on HDFS. Examples such as Inverted Index, Term Vector per host Distributed Sort, Distributed Grep, count of URL access frequency can be completed through MapReduce framework.

D. Hadoop

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models[5]. The Hadoop framework application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.

Hadoop Architecture

At its core, Hadoop has two major layers namely:

- Processing/Computation layer (MapReduce), and
- Storage layer (Hadoop Distributed File System).

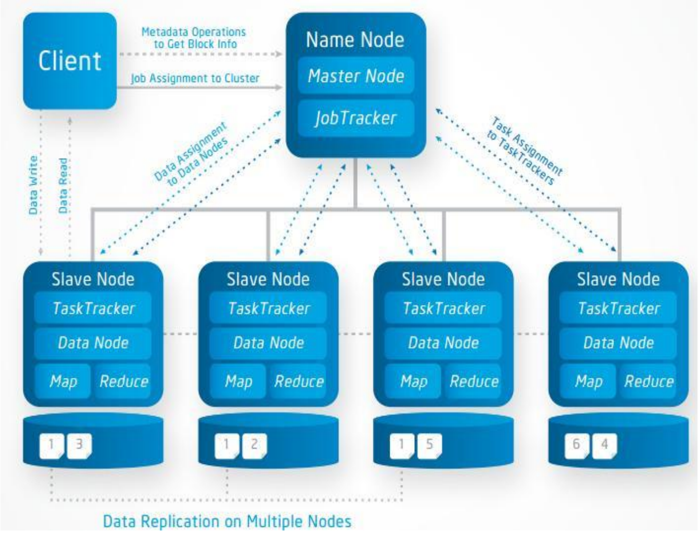

Client
Metadata Operations to Get Block Info
Job Assignment to Cluster
Name Node
Master Node
JobTracker
Data Write
Data Read
Data Assignment to Data Nodes
Task Assignment to TaskTrackers
Slave Node
TaskTracker
Data Node
Map
Reduce
1 3
1 2
1 5
6 4
Data Replication on Multiple Nodes

Fig 2: Hadoop Architecture [5]

Hadoop Distributed File System

The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. It is highly fault-tolerant and is designed to be deployed on low-cost hardware. It provides high throughput access to application data and is suitable for applications having large datasets[5].

Apart from the above-mentioned two core components, Hadoop framework also includes the following two modules:

- Hadoop Common : These are Java libraries andutilities required by other Hadoop modules.
- Hadoop YARN : This is a framework for job scheduling and cluster resource management.

III. RELATED WORK

Table 1: Survey of Different Frequent mining Paper

Index	Paper Title	Limitation
1	Frequent Itemset Mining for Big Data,IEEE,2013	Better Workload Distribution for mapreduce
2	A Parallel Algorithm for Approximate Frequent Itemset Mining using Map Reduce,IEEE,2014	Not accurate because it may miss the several frequent item sets which are infrequent in some subset of input dataset
3	IOMRA –A High Efficiency Frequent Itemset Mining Algorithm Based on the Map Reduce Computation Model,IEEE,2014	More amount of memory occupied, more utilization ratio
4	PaWI: Parallel Weighted Itemset mining by means of Map Reduce ,IEEE,2015	Absence of process for discovering combinations of items that were frequently bought together such one can recommended system
5	Reducing the Search Space for Big Data Mining for Inserting Patterns from Uncertain Data,IEEE,2014	More time consuming
6	ClustBigFIM-Frequent Itemset Mining of Big Data Using Pre-Processing Based on Map Reduce Framework,IEEE,2015	Time Consuming at mining level

IV. PROPOSED WORK

For finding frequent items from the given input data Apriori is the most popular algorithm. Even working with Big Data it is also implement on the Map Reduce. But as Apriori works on particular itemset scanning with each step of generating candidate the process of mining FI is more time consuming therefor replacing the Apriori with another algorithm which is more scalable at the stage of Mining in the existing hybrid approach give us speedy frequent items. Also rather to work with whole scanning, modified algorithm gives FI only in one scan of the database [4]. This will work on Map Reduce so first mapper function generate local support This will work on Map Reduce so first mapper function generate local support from the output of clustering process and then global support is calculated by reducer.

Here clustering is used as pre-processing technique. For clustering K-mean algorithm is used. Parallel K-means give approximate result in short time. Even to deal with the vertical database for generating global support for whole item sets .Éclat is being used as FIM method and for resolving memory problem using sub tree mining. Basically improvement in the mining part improve the efficiency of the existing algorithm.

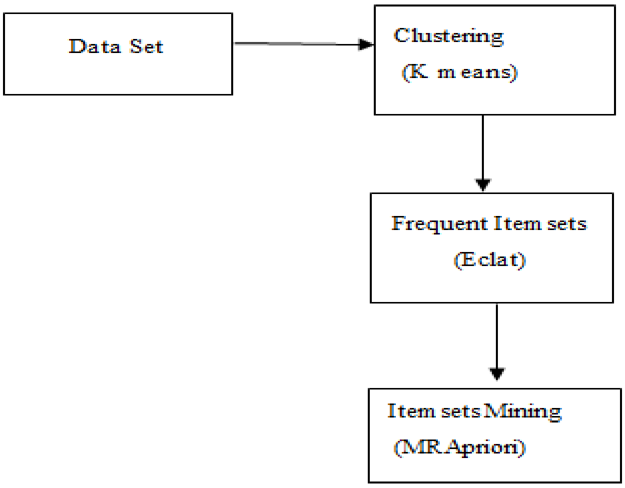


Fig 3: Block Diagram of Hybrid Approach

Proposed algorithm has below four steps which need to be applied on large datasets, steps are Find Clusters, Finding k-FIs, Generate Single Global TID list, Mining of Subtree.
1) Find Clusters
2) Finding k-FIs
3) Generating Single Global TID list
4) Mining of Subtree

V. CONCLUSION

This paper gives brief introduction about the terminology used by big data and the multiple algorithms working over the map reduce framework as a mining process. Also introduction to Hadoop platform for big data is being notifying here. The proposed algorithm has hybrid method for finding frequent item sets using parallel k-means, MRApriori and Eclat algorithm on MapReduce framework. Parallel k-means can give approximate results but in short time; MRApriori finds frequent itemsets having size k; Eclat algorithm finds potential extensions to frequent item sets and subtree mining by resolving memory problem. MapReduce platform can be used extensively for mining Big Data from social media as tradition tool and techniques cannot handle Big Data. Planning to apply frequent item set mining algorithm and MapReduce framework on stream of data which can be real time insights in Big Data.

REFERENCES

[1]Sheela Gole, Bharat Tidke, “Frequent Itemset Mining for big Data in Social media Using ClustBigFIM algorithm”, IEEE Pervasive Computing (ICPC), 2015 International Conference on Jan 2015
[2] Fabio Fumarola, Donato Malerba, “A Parallel Algorithm for Approximate Frequent Itemset Mining using Map Reduce”, IEEE High Performance Computing & Simulation(HPCS) , 2014 International Conference on July 2014.
[3] S. Moens, E. Aksehirli and B. Goethals "Frequent Itemset Mining for Big Data", Big Data, 2013 IEEE International Conference on, pp.111 -118
[4]http://www.planetdata.eu/sites/default/files/presentation/Big_Data_Tutorial_part4
[5]http://www.tutorialspoint.com/hadoop/hadoop_introduction.htm
[6]http://www.tutorialspoint.com/map_reduce/map_reduce_introduction.htm
[7] Sheng, Hui Liu,Shi-Jia Liu,Shi-Xuan Chen ,Kun-Ming Yu,”IOMRA—A High Efficiency Frequent Itemset Mining Algorithm Based on the Map reduce Computation Model”, Computational Science and Engineering(CSE),2014 IEEE 17th International Conference on, Dec 2014