International Journal of
Advances in Scientific Research and Engineering (ijasre)

E-ISSN : 2454-8006

DOI: http://dx.doi.org/10.7324/IJASRE.2018.32584

Volume 4, Issue 1
January-2018

A Conceptual Review on Data Mining in Big Data

Tamizharasi Thirugnanam and R. Jaya subalakshmi

School of computer science and engineering

VIT University, Vellore.

India

ABSTRACT

Big data is a term used for massive collection of complex data sets. Social networking, easily available technology and devices, etc. lead to unprecedented Big Data generation in past few years. Data comes from everywhere like posts on social networking website (example Facebook, Twitter), digital photos and videos. From this large collection of data sets, useful extraction of data can be done by using data mining. Proper handling of Big Data helps to discover facts, patterns and new models, data mining techniques are used for this purpose. In this article HACE theorem has been mentioned, that portrays the features of big data and how Big data can be processed, taking data mining as a prospect. This paper presents an overview of different Data Mining techniques used to analyze Big Data. It contains reviews of thirty different research paper on Data mining techniques, we have tried to explain the uniqueness, limitations and how well the techniques is used to resolve the challenges and issues in Big Data.

Keywords: Big data, Data mining, Algorithm.

I. INTRODUCTION

The 'Big Data' term first appeared as title in academic paper in 2000 by Diebold. Big Data represents large and complex growing data sets. The traditional methods for capture, data processing, storage, analyzing, sharing, transfer, querying and updating cannot handle it. Big Data is composed of structured and / or unstructured data. Social networking, easily available technology and devices, etc. lead to unprecedented Big Data generation in past few years. About 90% of data was produced in past two year. The amount of webpages handled by Google indexing service was about one billion in 1998. In next decade, the number increased rapidly and Google had already indexed one trillion webpages in 2008.

Proper handling of Big Data helps to discover facts, patterns and discover new models. It can be also used to forecast. Big Data has been used for getting insights of customer's behavior, to win elections, intelligent algorithm are being used to gather information from photo and video, etc. E-commerce firms like Amazon, Flipkart, Goibibo, eBay, Aliaba, etc. has been using Big Data. Flipkart takes data-driven decision by analysing 25 million rows of inventory data each day. Daily 30%-40% of orders are generated in other e-commerce's firms like Snapdeal using the Big Data tools. Face book and Google records huge amount of Big Data by tracking user's behavior. It is used for personalization of online advertisements, to improve user experience and user interface, to develop products like Trending news by Facebook, etc. Even automotive companies like Tesla and Ford are using tools like Apache Hadoop® cluster to collect and analyzes sensor's data received from customer's cars. This is used in research and development department, to notify drivers for car maintenance, to improve car performance and to increase customer satisfaction.

1.1 Characteristics of Big Data

Big Data has few important characteristics. 3 V's can be used to characterize Big Data.

1.1.1 Variety of Big Data:

Data is generated in various ways. Based on variety of data, Big Data can be classified as follows:

Structured data: The data which is organized and can be easily categorized and analyzed is called structured data. The data from sensors, electronic devices like televisions, smartphones and data stored in database are examples of structured data.

Unstructured: Unorganized data which can't be analyzed directly and contains complex information is called unstructured data. The data from social networking sites, documents, emails, audios, videos, photographs, stock data, customer reviews, etc. are examples of unstructured data.

Semi-structured: The classification of data in unstructured and unstructured type is not well defined. Some data follow organizational structure or carries tags. This make them easier to organize and can be categorized and analyzed as they are more accessible.

1.1.2 Velocity of Big Data: Velocity of Big Data deals with rate of generation of Big Data from sources. Big Data can also be categorized based on rate of data creation. Data may be created in batch or in streams, real time or near-time.

1.1.3 Volume of Big Data: Big Data have huge data sets. They can be also classified based on the volume of storage required used to store data like Terabytes, Records, Transactions, Tables, Files, etc.

1.2 HACE theorem

Hace theorem is used to characterize Big data. This helps in understanding underlying arrangement of data as well how well the data is arranged within a cluster. Big data starts with large volume, heterogeneous , autonomous and tries to analyze complex and evolving relationships with data.

This feature makes it hard to extract useful information from the large data sets. Let us assume, three blind men are given a task to size up a giant statue, which will be the Big data in this example. Now each guy has a goal to draw the picture of the statue with the data they gathered in the process. The information of each men is limited to their own local region. Everyone will have a different feeling about the statue; someone might feel it as a rope, or a wall depending upon their limited area. Now to make this problem a little bit complex some more conditions are included a) Size of statue is increasing rapidly b) pose is changing constantly c) and each person is allowed to discuss their information with other two persons. In this scenario analysing Big data would be collecting information from each blind men and then draw the picture in the best possible way. It would be much better if we collect information from each blind men and then get an expert to draw the picture using the collective information, in regard that each guy speaks different language and has some privacy issue about the message they discuss in the exchange process.

1.2.1 Large data sets with heterogeneous and diverse dimensionality:

One of the main features of Big data is that large set of data is represented by heterogeneous and diverse dimensionalities. Every information collector has their own way to represent a data, they can select whatever features they want to represent a particular data. For example, an individual is having accounts on three different social networking websites. Now, this single individual is being represented in three different ways by these websites. So here heterogeneous feature refers to different types of representation of a single individual and diverse features that are used to represent him. Now in this case every website has their own schemata to represent each individual, now if we try to aggregate data form different sources, heterogeneity and diverse dimensionality of data becomes the major challenge.

1.2.2 Autonomous sources with decentralized control:

This is the main characteristics of Big data application. By being autonomous, data sources can collect information without depending or relying on a centralized source for information. For example, in World Wide Web (WWW), there are different web servers that contains information and doesn't have to depend on any centralized source. And also keeping one centralized source makes the application more vulnerable to attacks or malfunctions. Big companies like Facebook, Google, Twitter and Walmart have deployed several different web servers all over the

world to provide non-stop and quick services to local market. Every country has their own rules and regulation, so by being autonomous companies can deploy there servers according to the rules of a particular area. For example Indian markets of Amazon are different than American markets, in terms of costumer behaviour, top sell items and seasonal promotions.

1.2.3 Complexity and Evolving relationships:

As the volume of Big Data increases, the complexity and the relationships underneath the data also increases. Earlier when the data was centralized, the focus was to find best features that can represent an observation. Like using set of features such as age, gender, salary, etc. to represent a single individual. This type of data representation is known as sample feature representation where each individual is treated as a separate entity, without taking social connections into consideration which is a very important part of a human life. People are making social connections based on their hobbies, likes and dislikes or are connected biologically. These connections in a huge set of data is making it more and more complex. For example, social networking sites such as Facebook and twitter are using social functions such as friends (Facebook) and followers (Twitter) to connect people. In sample feature representation, two entities are similar if there features have same values, while in sample feature relationship representation two people can be connected even though none of their feature values matches. In dynamic world, the values of features and the social connections between people can evolve rapidly based on spatial, temporal and other factors. Such a complication is becoming a reality of Big data applications, where the key is to take different (many to many) relationships and evolving changes into consideration to find useful patterns in Big Data Collecti

1.3 Data Mining for Big Data

Data mining is the process of extraction of hidden predictive information, analysing data from various perspective and summarize it in meaningful information. It is used to predict future trends and behaviors, take knowledge driven decisions. Data mining tools can answer questions which were too time consuming to resolve using traditional methods.

Data mining as a term used for the comprehensive classes of common six actions or tasks as follows:

1.3.1 Classification:

Classification is process of generalization of data and organizing into categories to be use it efficiently and effectively. Classification algorithms in data mining are of several major kinds like Apriori, Naive Bayes, k-nearest neighbor classifier, Decision tree and Ada Boost.

1.3.2 Estimation

Estimation deals with continuous outcomes. Based on data supplied to algorithm, estimation is used to discover values for certain unidentified variables like height, age, location, weight, salary, etc.

1.3.3 Prediction

Prediction is process of forecasting and giving statement about how things that will happen in future. It is often dependent on knowledge and experience. Some expected outcome may be part of statement in prediction.

1.3.4 Association Rules

An association rules are if/then statements which infers some certain associative relationships between seemingly unrelated data collections of objects in a relational database.

1.3.5 Clustering

Clustering is one of the most important unsupervised learning problem. It is used for discovering a structure in a set of unlabelled data.

Table 1: Difference among Big data and Data mining

Big data refers to large data set.	Data mining is an activity to discover important information among big data sets.
Big data is a resource	Data mining is the handler which delivers valuable outcome.
Big data varies based on the limitation of the organization handling the data set and on the abilities of the applications that are used for processing and analysing data.	Data mining is a procedure which involves comparatively complex search task.

This paper presents an overview of different Data Mining techniques used to analyze Big Data. Section II contains reviews of thirty different research paper on Data mining techniques, we have tried to explain the uniqueness, limitations and how well the techniques is used to resolve the challenges in Big Data.

II. LITERATURE REVIEW

Yahya et. Al., paper discussed the Aprori algorithm for classification for Big Data. Apriori algorithm is used for finding frequent item sets from a transactional dataset and scans the data repeatedly to generate as much candidate item sets as possible. In order to reduce the time required parallel and distributed computing are used. Map Reduce Apriori (MRApriori) algorithm is implemented using Hadoop. The algorithm has two phase to compute all repeated k-item sets. In first phase, the map function takes whole split as value parameter [1]. The traditional Apriori algorithm processes the split with partial minimum support count.

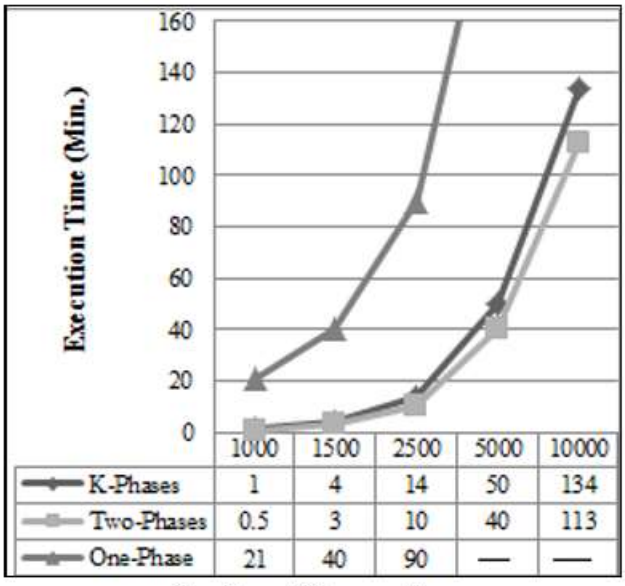

Figure 1: Aprori algorithm

Execution time in minutes was measured for different amount of transactions. The use of Map Reduce introduces computational overhead.

A Haque et. al., describes speeding up the process labelling of instance in Data Streams and scale it for huge data sets. MapReduce is used for large amount of data streams to classify data using ADABOOOST algorithm [2]. The MapReduce is used to implement multi-tiered ensemble with help of Hadoop and Heterogeneous Hierarchical Ensemble (HHE) approach. Pipeline processing procedure is used to rapidly put on anticipated labels to data occurrence in data stream and similarly classifiers are maintained. Time required processing the chuck i.e. data sets using parallel computing environment is measured. If number of maps is doubled from 5 to 10 the Average Execution time (per chunk) for 30000 data sets decreases by 31.1 second of Forest Cover data set and 7.1 seconds for PAMAP2

data set. The implementation adapts finely with data streams. The decision boundaries are estimated more tightly by using this approach of the target classes than other alternative methods. The approach doesn't scale well for all type of data sets.

In the Liu, B., Blasch, E., et. al., paper, the implementation of Naive Bayes classifier (NBC) which is efficient and scalable in parallel computing environment is described. It accurately discover useful information from large data sets[3]. The MapReduce framework is utilized to perform classification on large datasets. The operation is divided in three steps: Pre-processing Raw Dataset, Preparing Input Datasets, Sentiment Classification Using Hadoop and Result Collection. Algorithms to implement operations are mentioned. Accuracy with respect to dataset size is used to evaluate efficiency and scalability of implementation. Processing time for processing per ten thousand (seconds/10K reviews) movie reviews of different data set size was used to measure scalability. For smaller dataset the accuracy is unstable but for larger database the accuracy reaches 80% to 82% and remains stable. Time required to process 400000 dataset was 4.24 seconds/10K reviews, for 1200000 dataset the time reduces to 2.77 seconds/10K reviews and for 2000000 datasets were only 2.33 seconds/10K reviews indicating good scalability. The accuracy can be further improved using filtering methods

Support Vector Machines (SVM) is used for classification in data mining. SVM is implemented to process large dataset of image using MapReduce. Alham et al [4] proposed MRSMO, an efficient algorithm based on SVM is used for image explanation. MRSMO is based on SMO algorithm. MRSMO for parallel computing environment uses Hadoop which is an implementation MapReduce. Average accuracy level of classification was used to detect accuracy of implementation. The average accuracy level of 93% was achieved using MRSMO. The MRSMO algorithm reduced the training time. So, the MRSMO is scalable and can process large number efficiently. The MRSMO has less accuracy than standard SMO. Also the overhead is greater for MRSMO then standard SMO implementation.

Lu, W., Shen, Y et al., discusses the problem in performing classification of large data sets. An efficient k nearest neighbor and kNN join operation implementation for classification is described [5]. The kNN algorithm is implemented using MapReduce is proposed. The kNN join operation is performed on by using mappers to cluster objects in group and the reducer to do kNN join for each collection of objects. To decrease costs for the shuffling and computational pruning rules for distance filtering are used. Two approximate algorithms to reduce the amount of duplications are suggested for minimizing shuffling cost. The partitioning and grouping strategy is used in PGBJ algorithm. Instead of grouping PBJ employs the same framework used in H-BRJ. The running time, computation selectivity and shuffling cost is measured with respect to data size for all algorithms and are compared. The PGBJ has lowest runtime, selectivity and shuffling cost for given data sets.

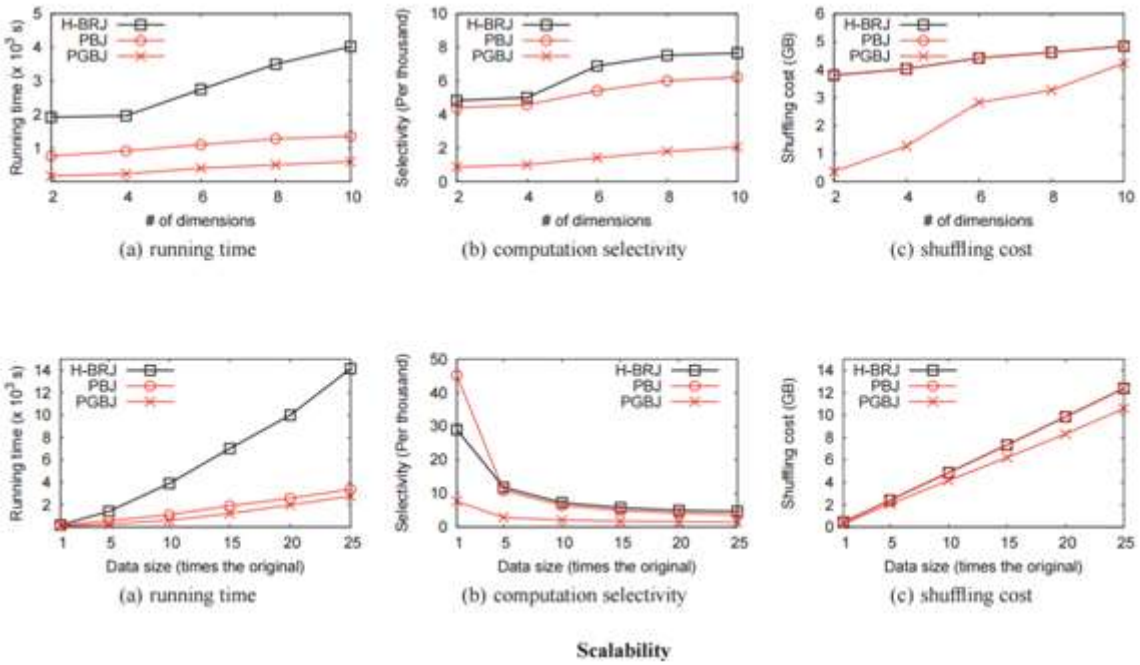

Figure 2: Effect of dimensionality

Bayesian Networks is a popular classification algorithm in Data Mining. Scalable Bayesian Networks is implemented for parametric learning is required to handle Big Data effectively. Aniruddha Basak et. al., paper [6] describes implementation using Hadoop, an implementation of MapReduce algorithm. It is used to perform calculation in the Bayesian Network. The operation is divided into following algorithms: Map-Reduced Bayesian Update (MRBU), Sequential EM (SEM), Map-Reduced EM (MREM). Hadoop is an implementation of MapReduce algorithm. It is used to perform calculation in the Bayesian Network. The operation is divided into following algorithms: Map-Reduced Bayesian Update (MRBU), Sequential EM (SEM), Map-Reduced EM (MREM). The Speed Up ratio increases as number of data record increases significantly. Speed Up ratio for Bayesian Network with large junction trees increased significantly for 20 or even smaller datasets. The parallel execution time for processing various data sets halved if number of nodes are doubled. The Bayesian Network with smaller junction trees have lower Speed Up ratio.

The neural network implemented using traditional technologies can't handle huge amount of data and it affects both accuracy and effectiveness. The redundant pieces of data in large data sets also affect the system performance. Kairan et al., describe usage of cloud computing platform for running deep learning algorithm for processing large data sets [7]. A Handwriting character recognizer is designed using MapReduce. Training and testing of cases uses RBM algorithm and back-propagation algorithm. Runtime (minutes) wither respect to number of MapReduce slaves is calculated for 60000 training cases. The running time decreases for processing as the number of nodes increases. The implementation is scalable and robust. Deep learning requires training for long period of time.

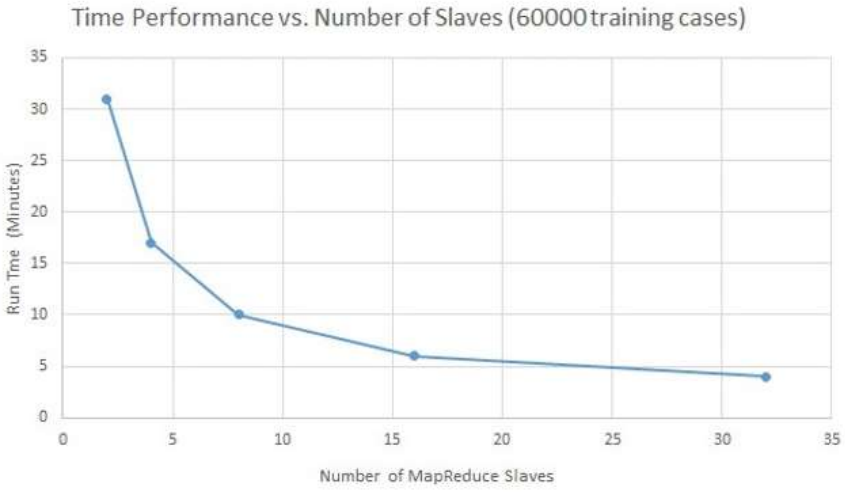

Figure 3: RBM algorithm

Jesus Maillo et. al., describes the k-nearest Neighbour method used for classification of data in Data Mining [8]. It is effective and simple. Traditional way to use this method on large volume of data fails. MapReduce-based approach is used implement k-Nearest neighbour classification. In map phase, arranging the calculation of correspondence between test examples and training set is splinted among a cluster of computing nodes. For each map, the k nearest neighbours along with respective distance values will be supplied to reduce stage. Further the reduce phase decides the final k nearest neighbours. Metric used for evaluation is ratio of Reference time to Parallel time called Speedup. *Reference time* and *Parallel time* is the runtime spent with the sequential version and improved version of algorithm respectively.

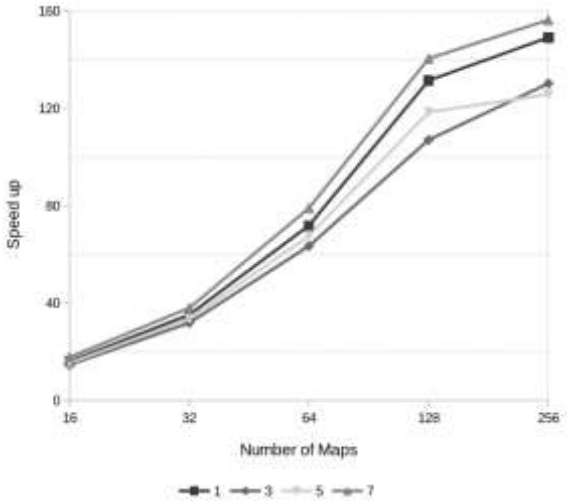

Figure 4: k-nearest Neighbour method

The faster performance can be achieved by using technologies of Spark

Dai et. al., describes the process of constructing decision trees for big data sets while ensuring the accuracy of Decision tree's classification results [9]. The traditional algorithm is transformed in a sequence of Mapping and reducing actions. Operation is separated in four phases: data preparation, selection, update and tree growing. Data structures to minimize communication cost incurred in parallel computing environment are also used. Metric used for evaluation is ratio of Reference time to Parallel time called Speedup.

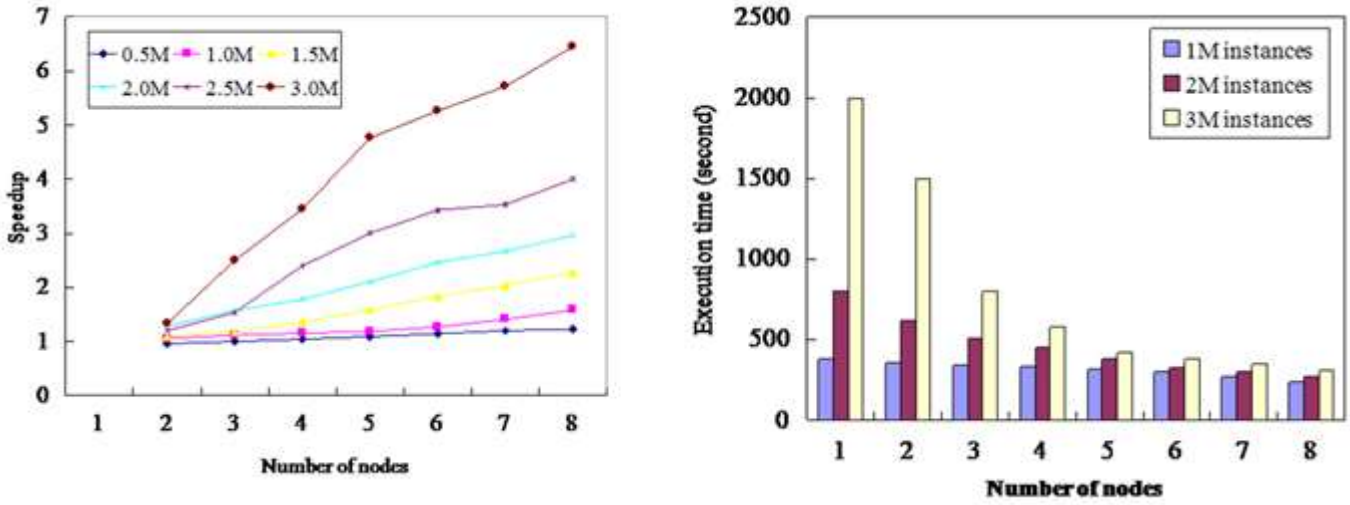

Figure 5: Decision tree algorithm using MapReduce

Reference time and *Parallel time* is the runtime spent with the sequential version and improved version of algorithm respectively. MapReduce implementation of decision tree has better time efficiency and is scalable. Execution time for classification of data sets reduces as number of node and number of instance of training dataset increases.

Estimation of large graphs is ubiquitous. Computation of Diameter, radius of each node, etc. on graphs of large size is impossible by traditional methods. HADI algorithm is proposed for quicker approximation of diameter and for data mining in huge graphs [10]. It is implemented with Hadoop which is based on MapReduce algorithm. Various methods like Bit Shuffle Encoding, Pruning Stable Nodes and Check pointing are used for optimization of HADI algorithm. The runtime in hours is calculated for different number of machines with respect to number of edges in million. The throughput $1/T_M$ is calculated for different number of machines.

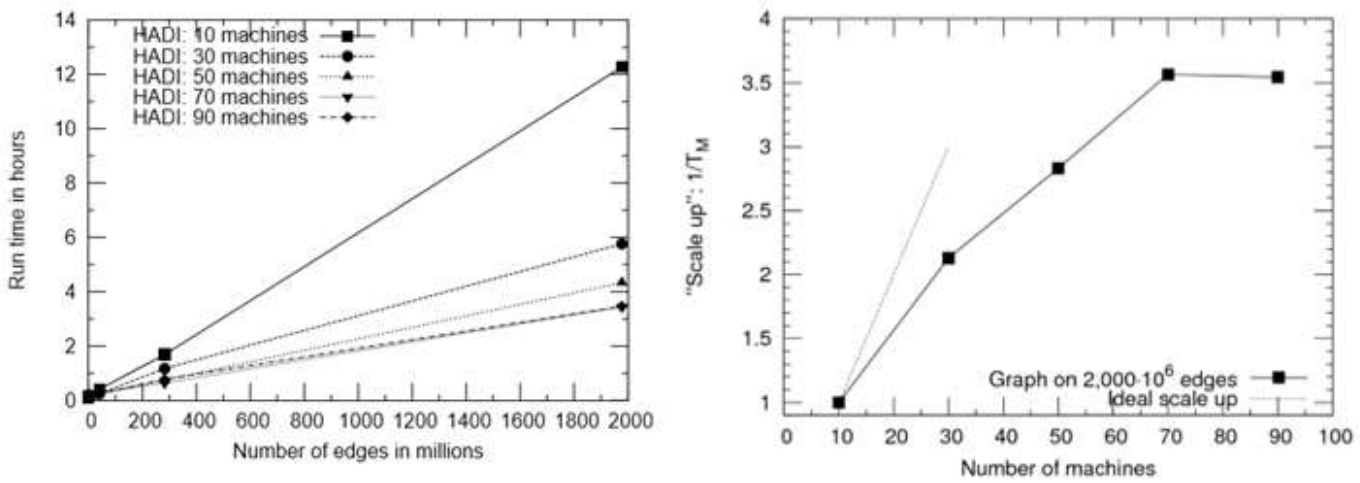

Figure 6: HADI algorithm

The paper [11] contains an overview of various clustering methods along with their explanation, properties and drawbacks. Paper also provides the comparison between all the five methods. The clustering methods are partitioning method, Hierarchical method, Density based , Grid based and Model based. In partioning method two algorithms are explained K-means and K-medoids. According to the paper K-means is the simplest clustering algorithm. The main drawback for this method is that it only works for numeric data. In hierarchical method clustering tree is formed, it can be done by using top down or bottom up approach i.e either you can start with a single cluster and then start joining more clusters or you can start with different clusters and start breaking them into more clusters. The drawback in this method is if splitting is performed, no backtracking can be done. Density based method is based on the density boundary and connectivity. The clusters are formed based on density of the data points in the particular region. In this method there is no need to mention the number of clusters it works automatically. But this method doesn't work on high dimensional data sets. In Grid based method more attention is paid on spatial data. STING (Statistical Information Grid based) and Wave Cluster are examples of grid based clustering. Model based clustering method enhance the fit between the provide information and some (predefined) mathematical model. It involves EM algorithm, which provides excellent performance with respect to the cluster quality but even it can't work on high dimensional data sets.

The paper [12] addresses the problems with existing clustering algorithms which either works on spherical clusters or are fragile in presence of outliers. It also contains the solution for this problem i.e. CURE. CURE is based on hierarchical clustering model, basically this algorithm chooses some random multiple points in a cluster, that are know as the representatives of that cluster. The selected points are moved towards the center by some fraction. If two different cluster's representative points values are nearby then those two clusters are merged. This method helps in decreasing the number of outliers. In large data sets, two methods are used. First they are partitioned into k number of clusters then random sampling is done i.e. some random points are chosen for clustering process. According to their experimental results, CURE is found out to be much better algorithm for clustering than the existing ones. Furthermore, partitioning and random sampling helps CURE to perform well with large data sets without worrying about clustering quality.

K-means algorithm is the simplest clustering algorithm. The main advantage of this algorithm is that it works for large data sets, but it is limited to numeric data. This paper [13] presents a new algorithm i.e. k-modes to overcome this drawback. K-modes algorithms works on categorical data. The paper also explains about categorical data, its domain and attributes and categorical objects. Some modifications have been done on k-means to transform into k-modes. In k mode algorithm initially modes are selected for each cluster. Dissimalirty measure in k mode is different, the number of mismatches between two objects determines dissimalirty measure. Instead of number, frequencies are used as an attribute for modes. The paper demonstrates experimental results done on soybean disease data sets, it shows the new algorithm performed very well with large and complex data set. Even the clustering quality was not compromised.

The paper [14] addresses an approximate algorithm based on k-means. For analysis of big data this method is very fast, expandable and have high efficiency. It fixes the number of iterations which is big drawback of k-means algorithm. The paper shows an overview of the new proposed algorithm. In a data set there are large set of attributes for a single data point. The approximate algorithm only focuses on the attributes in which user is more interested and operates on only those attributes, it helps the algorithm to reduce the complexity in k-means. The number of iterations in approximate algorithm depends on the number of clustering thus helps in overcoming the drawback of k-means algorithm. The paper contains the implementation results that shows 1) increase in number of dimensions directly increases the run-time 2) Increasing number of clusters directly increases run time. The paper demonstrates the efficacy and precision of the algorithm on real and synthetic data. In most of the tests the efficiency of the proposed algorithm is more than k-means, also cluster recovery is better in the new algorithm because it does not reject any data points. The main drawback of new algorithm is that it only works on numerical data same as k-means, and for large data sets hierarchical clustering can be used alongside approximate algorithm.

The research paper [15] addresses various methods used to solve problems of big data using map reduce framework over Hadoop Distributed File system. MapReduce is a programming tool that is used to process large volume data sets in computer clusters. It uses two main functions i.e. map() and reduce(). This paper thoroughly explains the working of MapReduce and the techniques that are used for it. As every other tool map reduce also has a limitation, the algorithms that depends on shared global state during processing faces difficulties in implementing on map reduce. Therefore, implementing online algorithms on map reduce is problematic.

The paper [16] contains an overview of all the techniques that are being used for clustering big data. Author has written in detail mainly about two clustering algorithms i.e. k-means and Self organising maps. A lot of modifications have been done on these two algorithms. The author has mentioned all the algorithms who are either an extended version of these two or are based on the concept of these two algorithms. Other than this, in the paper other popular clustering algorithms are also mentioned. Technologies that are being used for big data are mentioned in the paper with the purpose they are serving for big data. During applying a technique to cluster big data different strategies must be followed, the author have described two systems KNIME and WEKA that helps in telling that how much amount of data can be processed using limited resources.

As time passes, volume of big data is also increasing exponentially. And therefore the need of extracting useful patterns in the data is also increasing. To analyze this amount of big data, parallel clustering algorithms with iterative applications would be a better way to do it. In this paper, an overview of different parallel clustering algorithms has been presented. The paper have explained how the algorithm is solving the issues of big data, and have also mentioned the limitations of that algorithm. The paper paid more attention on MapReduce framework, as it is really easy to program on that and its fault tolerant property. But even MapReduce framework doesn't work for iterative algorithms. The main finding of the paper is that still no framework exists that supports iterative algorithms. In future to handle the amount of data we have to come up with a new framework on which we can apply parallel algorithms with iterative applications.

The paper [18] focuses on evaluating different hierarchical clustering algorithms and it also presents a comparison between partitional clustering and agglomerative clustering. They have demonstrated experimental evaluation, which shows that partitional algorithm performs much better than agglomerative clustering, which suggests that partitional algorithm is well suited for clustering not just due its low computational cost but also the clustering quality is better than agglomerative clustering. This paper presents a new clustering algorithm i.e. Constrained agglomerative clustering that combines the both partitional and agglomerative features . There experimental results have shown that this new algorithm is consistently showing better results than partitioned or agglomerative algorithms alone.

In this paper [19], they have presented a new algorithm known as DBSCAN, which works on density based clustering method and can be used to find arbitrary shaped clusters. It works on only one parameter which is determined by the user. They have demonstrated an experimental evaluation on DBSCAN to check its efficiency and effectiveness using synthetic and real data of the SEQUOIA 2000 benchmark. There results have shown following things: 1) DBSCAN can find better arbitrary shaped clusters than well-known algorithm CLARANS 2) DBSCAN efficiency is more than 100 times better than CLARANS. The only drawback in existing DBSCAN algorithm is that it

works only for point objects, in future this algorithm must be extended so that it can also work on spatial data that contains polygon objects.

The paper [20] presents a new grid based algorithm which uses adaptive mesh refinement technique which is capable of applying higher resolutions to high density regions. Instead of using a single uniform grid, the AMR clustering algorithm uses different resolution grids depending on the regional density, for example the algorithm places high resolution grid where high computational cost is required. For highly irregular data, single uniform grid based algorithm cannot efficiently provide good clustering quality, while AMR can dynamically discover nested clusters which is useful for highly irregular data. In the paper they have shown some experimental results that compares the efficiency and effectiveness of AMR algorithm over single uniform grid based algorithm. For experimental evaluation they have used the star particle data-sets generated from ENZO and there results clearly shows that AMR algorithm performs better over single uniform mesh.

The paper[21] discusses a very important field in big data mining which is finding frequently occurring item sets and also discusses the most established algorithm for them in a transactional database which is called Apriori algorithm and improved it using top-k rule.Apriori algorithm generate most of the candidate itemsets by scanning the dataset many times. In that paper, they have implemented an efficient Apriori algorithm based on Hadoop-Map Reduce model which conducts 2 phases to get most frequently occuring k-itemsets. Unfortunately, in the old algorithm when the dataset size is massive, the memory use and computational cost becomes expensive.It could have also happened that the earlier algorithm that we use skipped steps and missed important information and failed. To solve the above problem they proposed TopkRules, an algorithm that discovers the top-k rules which have high support value, and where the k is defined by the user. It uses several optimizations techniques and includes rule expansions that improve its performance. They have compared their proposed MapReduce Apriori algorithm with the current two existing k association rule algorithms which need either one or k phases to find the same frequent n-itemsets.

Data mining in big data is a process involves analysis of a big, complicated, and multi-dimensional dataset to discover the patterns and hidden knowledge of the data set. However, the traditional discovery Association rule mining methods cannot manage massive amounts of data. In this paper[22] an Association rule algo based on Lift interestingness (MRLAR) is developed using MapReduce framework for which manage many datasets with a huge number of nodes. High performance tests and parallel processing was used to test the performance and completeness of the proposed algorithm which was based on MapReduce framework. The outcome of the experiments conducted in this paper give evidence that MRLAR performs very efficiently, the associations among itemsets was detected, by involving the uses of Lift interestingness and MapReduce measured to determine the correlation in RightHandSide and LeftHandSide in the Association rule algo instead of using confidence. While there are benefits of MapReduce method, it also has some drawbacks which cannot be ignored. These limitations are concerned with MRLAR such as extracting the single dimensional asoosiation rules and another limitation is that it operates only on data structures of type pairs which is based on MapReduce , so all the input data must be converted into such datastructure.

In a large database of sales transactions mining for association rules between itemsets has been described as very important database mining problem. In the twenty third paper [23] a fundamentally different algorithm for mining association rules is used that is efficient from known algorithms like apriori and aprioriTid. The proposed algo is called as partition Algorithm. Almost similar itemset generation technique are used by the partition algorithm and apriori algo . The only reason partition is better is that it has better count generating function. Apriori and AprioriTid algorithm were used to compare the performance of partition algorithm. All experiments were done on synthetic data. Apriori's performance was superior to that of AprioriTid's performance as per the results so AprioriTid was not compared to partition algorithm. Compared to previous algorithms, the discussed algorithm not only has lower CPU overhead almost by a factor of 4 for almost all cases but also reduces the I/O overhead significantly. Hence the discussed algorithm is really very suitable for huge sized databases.

In this paper[24] a discussion of some promising research directions and brief analysis and review of the current conditions of frequent pattern mining is done. Additionally, this paper described approaches and includes a comparative study between their performances. In this paper they have discussed many kinds of frequently seen pattern mining methods and Association Rule mining. Then they also reviewed the different ways that can be used to

find frequent item sets using the association rule mining like Apriori, DHP, GSP, etc. The main basis on which different approaches were analyzed were on their performance and memory perception. In the end they highlighted the direction in which further research regarding frequent pattern mining methods and Association rule mining can be done.

In this paper [25], on the MapReduce platform ,they have inspected the validity of FIM techniques. They have proposed two new methods for mining massive datasets: BigFIM is extremely efficient to run on really massive datasets while Dist-Eclat's main aim is to improve speed. They also have conducted experiments to show the scalability of their methods. They studied and tried to implement 2 FIM algorithms for MapReduce in this paper namely Dist-Eclat and BigFIM. BigFIM, utilizes a hybrid approach and the mainly targets massive databases. A second algo, Dist-Eclat uses a really simple load balancing technique based on k-frequent item sets which focuses on speed. The results indicate that using Round-Robin allocation scheme in combination with 3-frequent items results in a better workload distribution. At last, the experiments shows that applying the methods on Big Data using MapReduce outperform previously used FIM methods. Further research can be done in assignment methods which will result in better workload distribution.

This paper[26] discusses about the current status; challenges to forecast the future and controversy in a broad overview. And on challenging side, one of the big issue is Securely Manage Big Data with today's threat spectrum. This paper introduces Big Data mining and its applications in predicting the future. The paper contains 4 such applications big graph mining, netix, etc. and also discuss about Big Data problems. The paper also discusses about some of the main Big Data controversy that are being going on. They also point out the significance of open-source software tools that have been used and give some challenges in forecasting the future. This paper uses illustrations with applications from various areas that have been discussed. Finally this paper discuss about the privacy of big data as an essential issues and how to securely manage it.

This paper [27] introduces "heterogeneous mixture learning", which is considered one of the most advanced data analysis technology which can be applied to a heterogeneous mixture, and also includes details of some actual applications. It also that the data that has previously been collected can be utilized without any specific aim. This paper talks about using heterogeneous mixture learning tech in big data analysis which is considered an advance tech. In the paper, the difficulties that are inherent in heterogeneous mixture data analysis have been introduced, the basic concept of the topic and the results of an experiment that deals with electricity demand predictions has been demonstrated. Data mining technology in heterogeneous mixtures also plays a significant role in the market, as the Big Data analysis increases in its importance. In the coming future, the application of heterogeneous mixture learning is expected to expand exponentially.

This paper [28] illustrates an approach of data mining techniques, using big data and visualization used to analyze and predict traffic. The goal was to find Traffic patterns in Dublin City and understand them. To identify unusual traffic patterns the prediction model of big data was used as a tool of estimation. Red lines and green lines were used to show real time traffic with the help of tweets. The generic model was designed with the help of multivariate regression algorithms, data mining techniques and ARIMA. Visual correlation with real-time traffic tweets were also used to make the algorithm. The result was a really powerful and efficient web application containing over five hundred million traffic observations that provide traffic prediction and analysis through analytical dashboard.

This paper [29] proposes the use of different data mining tools to detect wasted parts and accuracy of these tools was also compared. A combination of physical and thermal characteristics have been used and the algos had been implemented on Ahanpishegan company's current data to deduce the durability of its products. Investigation on each algorithm's accuracy was done. Various prediction algorithms were used in the investigation that were conducted like CHAID, QUEST and C&R. Managers and employees were interviewed firstly to find points that were mindboggling about proportion and features of defective parts. Then by identifying outliers and using integrated database, they cleaned redundant data and excluded the constant variables, they also have applied different prediction algos and compared the final results. The paper's main aim is to increase the quality of industrial products like durability, maintainability and reliability. A recommendation of use of data stream mining tools was made by the researchers to achieve more accurate and quicker results. They also have suggested to conduct such researches in food

industries to test the food quality made by the company. They also mentioned future works can be done on failure frequency, damage done due to the failure and decrease the consequence of such failures by minimizing it.

According to Merceron, A et al., the main argument is that cosine and similar values can be interpreted by the teachers easily and belong to educational data. The researchers argued that first the interestingness of the cosines must be calculated first, and then if they rate it not interesting then lift should be tried on cosines. If both measures do not agree, teachers should go by their intuition to decide whether or not to dismiss the association rule. They provided a case study with data from a Learning Management Systems. The case study presented in the paper depicts a standard situation: a Learning Management System that provides additional resources to the students in opposition to the face-to-face teaching context. The main conclusion of the paper is that association rules are useful in Educational Data Mining for analysis of learning data. Another conclusion of this work is that common Learning Management System are far from being data mining friendly. Test data and Log data concerning access to resources are not stored the same way [30].

III. CONCLUSION

In the below table we have mentioned all the techniques that are being used in above thirty literature reviews on Big data papers. With the technique we have also mentioned the Pros and cons of that particular technique.

Table 2: Different Technique Pros and Cons

SL .No	Technique Used	Pros	Cons
1	Apriori algorithm for classification on MapReduce model.	Improves efficiency by calculating all the frequent k-item sets in two phases.	Introduces computational overhead.
2	Heterogeneous Hierarchical Ensemble approach with Pipeline processing.	• Speeds up process labelling. • Average execution time decreases significantly • Fine adaptation with data streams. • Tight estimation of decision boundaries.	Does not scale well for all types of data sets.
3	Naïve Bayes Classifier implemented on MapReduce model.	• Efficient and scalable in parallel computing environments. • High accuracy for large databases • Good scalability.	Unstable accuracy for small databases.
4	Support Vector Machines for classification on MapReduce model.	• Efficient for image expansion. • Very high accuracy. • Reduction of training time. • High scalability. • Can process large numbers efficiently.	Lesser accuracy than the standard SMO algorithm.
5	k Nearest Neighbour algorithm for classification on MapReduce framework.	• Decreases shuffling costs. • Decreases costs for computational pruning. • Reduces duplications. • Lowest runtime, selectivity and shuffling costs for given data sets.	
6	MRBU, SEM and MREM are used for parametric learning and classification on MapReduce model.	• Speed up ratio increased with significant increase in data records. • Parallel execution time halved. • Speed up ratio for Bayesian Networks with large junction trees increased	Bayesian Network with smaller junction trees had lower Speed up ratios.

		significantly.	
7	Use of cloud as a platform to run deep running algorithms for processing large data sets.	• Running time decreases with increase in nodes. • Highly scalable. • Robust.	Deep learning algorithms require training for long periods of time.
8	K nearest neighbour classification which is implemented through MapReduce based approach	Effective and simple	Takes time.
9	Classification using Decision trees	Execution time reduces as number of instance of training dataset increases, scalable.	Communication cost is still high.
10	HADI algorithm in huge graphs, Bit shuffle encoding, Pruning stable nodes for optimization.	Quicker than traditional algorithms	Still takes a lot of time for calculation of runtime and throughput.
11	Comparison between Partitioning method(PM), Hierarchical method(HM), Density based(DB) , Grid based(GB) and Model based(MB)	PM- simplest clustering algorithm DB- no need to mention the number of clusters it works automatically MB-excellent performance with respect to the cluster quality	PM-only works for numeric data HM- no backtracking can be done DB-doesn't work on high dimensional data sets MB-doesn't work on high dimensional data sets
12	CURE- hierarchical clustering model	1. Much better algorithm for clustering than the existing ones 2. Perform well with large data sets	Time complexity $O(s^2)$ for low computational data.
13	K-means algorithm	It works for large data sets	Limited to numeric data
14	Approximate algorithm based on k-means	Very fast, expandable and have high efficiency	Only works on numerical data
15	Map reduce framework over Hadoop Distributed File system	Process large volume data sets	Implementing online algorithms on map reduce is problematic.
16	K-means and Self organising maps	Easy and can process complex data sets.	Limited to numeric data
17	Parallel clustering algorithms(MapReduceframe	Fault tolerant property	Doesn't work for iterative algorithms

	work)		
18	Partitional clustering(PC) and agglomerative clustering(AC)	PC- Low computational cost	AC- Errors may be introduced during the initial merging decisions
19	DBSCAN	1. DBSCAN can find better arbitrary shaped clusters than well-known algorithm CLARANS 2. DBSCAN efficiency is more than 100 times better than CLARANS.	It works only for point objects, not for spatial data.
20	Grid based algorithm which uses Mesh refinement	Efficient and effective than single uniform mesh	Costly
21	Apriori algorithm	Efficient and effective	Takes time to compute the frequent itemsets, expensive.
22	Association rule algorithm based on lift interestingness(MRLAR)	Performs efficiently in high performance test and parallel processing, shows completeness	Extracting the single dimensional association rules is difficult and another limitation is that it operates only on data structures of type pairs which is based on MapReduce, so all the input data must be converted into such Data Structures.
23	Partition algorithm	Better count generation function, better than apriori algorithm, lower CPU overhead, reduces the I/O overhead significantly.	Not discussed.
24	Frequently seen pattern mining methods and Association Rule mining are used.	Easy implementation	FP growth cannot be used for interactive mining and incremental mining
25	Implementation of 2 FIM algorithms for MapReduce namely Dist-Eclat and BigFIM.	BigFIM and Dist-Eclat methods outperform than other FIM methods on Big Data using MapReduce.	Dist-Eclat Algorithm nearly all of the mapper need entire dataset in memory and need to communicate with entire dataset which is not feasible with the network infrastructure which is preferred in the paper.
26	This paper introduces Big Data mining and its applications in predicting the future.	Eventually contributed to a lot of cost savings	It requires special computer power
27	Heterogeneous mixture learning tech in big data analysis which is considered an advance technology.	Sharing only information about prediction models among computers, not massive data samples	Costly
28	Multivariate regression	Efficient and effective	Takes time to compute the frequent

	algorithms, data mining techniques and ARIMA		itemsets, expensive.
29	Various prediction algorithms were used in the investigation that were conducted like CHAID, QUEST and C&R	CHAID generate a wider range of tree which doesn't follow the binary tree split up method. I works for all types of inputs, and it accepts both case weights and frequency variables.	Since multiple splits fragment the variable's range into smaller subranges, the algorithm requires larger quantities of data to get dependable results.
30	Data from a Learning Management Systems	Reduce Expenses On Training.	The main drawback of using a LMS is that it cannot update themselves by online learning.

Big data is term used for a massive and complicated data. Data mining is a technique used for the analysis of big data, it explores the given data in search of consistent patterns and extraction of useful data. In our paper we have reviewed different data mining techniques that are being used to analyze big data. We have summarized the research summary in the papers and tried to explain the uniqueness, limitation and tried to explain how the techniques that have been used to resolve the challenges that are faced in big data. Although many researches have been done on this topic, but still there are some issues that are yet to be resolved. Hadoop, an implementation of map reduce technique is used for implementing algorithm in parallel computing environment, new technologies of SPARK can be used to improve performance. Data is evolving overtime, these techniques must be improved to detect and adapt the changes. The data mining techniques which we have discussed are trivial and cannot be paralyzed. A lot of research should be done to provide new methods, which are distributed versions of some mentioned methods. Day by day big data is becoming hard to deal with, the quantity and complexity of data is gradually increasing, therefore need of data mining is rising exponentially in all science and engineering fields. So this paper provides a quick review on different techniques that are used in data mining.

REFERENCES

[1] Yahya, O., Hegazy, O., & Ezat, E. (2012). An Efficient Implementation of A-Priori algorithm based on Hadoop-MapReduce model. *International sjournal of Reviews in Computing*, *12*.

[2] Haque, A., Parker, B., & Khan, L. (2013, June). Labeling instances in evolving data streams with mapreduce. In *2013 IEEE International Congress on Big Data* (pp. 387-394). IEEE.

[3] Liu, B., Blasch, E., Chen, Y., Shen, D., & Chen, G. (2013, October). Scalable sentiment classification for big data analysis using Naive Bayes Classifier. In Big Data, 2013 IEEE International Conference on (pp. 99-104). IEEE.

[4] Alham, N. K., Li, M., Liu, Y., & Hammoud, S. (2011). A MapReduce-based distributed SVM algorithm for automatic image annotation. Computers & Mathematics with Applications, 62(7), 2801-2811.

[5] Lu, W., Shen, Y., Chen, S., & Ooi, B. C. (2012). Efficient processing of k nearest neighbor joins using mapreduce. Proceedings of the VLDB Endowment, 5(10), 1016-1027.

[6] Basak, A., Brinster, I., Ma, X., & Mengshoel, O. J. (2012, August). Accelerating Bayesian network parameter learning using Hadoop and MapReduce. In Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications (pp. 101-108). ACM.

[7] Sun, K., Wei, X., Jia, G., Wang, R., & Li, R. (2015). Large-scale Artificial Neural Network: MapReduce-based Deep Learning. arXiv preprint arXiv:1510.02709.

[8] Maillo, J., Triguero, I., & Herrera, F. (2015, August). A MapReduce-Based k-Nearest Neighbor Approach for Big Data Classification. In Trustcom/BigDataSE/ISPA, 2015 IEEE (Vol. 2, pp. 167-172). IEEE.

[9] Dai, Wei, and Wei Ji. "A mapreduce implementation of C4. 5 decision tree algorithm." International Journal of Database Theory and Application 7.1 (2014): 49-60.

[10] Kang, U., Tsourakakis, C., Appel, A. P., Faloutsos, C., & Leskovec, J. (2008). HADI: Fast diameter estimation and mining in massive graphs with Hadoop. ACM Trasactions on Knowledge Discovery from Data (TKDD), 5(2), 8.

[11] Sanse, Keshav, and Meena Sharma. "Clustering methods for Big data analysis."

[12] Guha, S., Rastogi, R., & Shim, K. (1998, June). CURE: an efficient clustering algorithm for large databases. In ACM SIGMOD Record (Vol. 27, No. 2, pp. 73-84). ACM.

[13] Huang, Z. (1997, May). A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. In DMKD (p. 0).

[14] Jain, Mugdha, and Chakradhar Verma. "Adapting k-means for clustering in big data." *International Journal of Computer Applications* 101.1 (2014): 19-24.

[15] Jaseena, K. U., & David, J. M. (2014). Issues, Challenges, and Solutions: Big Data Mining. NeTCoM, CSIT, GRAPH-HOC, SPTM–2014, 131-140.

[16] Kurasova, O., Marcinkevicius, V., Medvedev, V., Rapecka, A., & Stefanovic, P. (2014, November). Strategies for big data clustering. In 2014 IEEE 26th International Conference on Tools with Artificial Intelligence (pp. 740-747). IEEE.

[17] Mohebi, A., Aghabozorgi, S., Ying Wah, T., Herawan, T., & Yahyapour, R. (2016). Iterative big data clustering algorithms: a review. Software: Practice and Experience, 46(1), 107-129.

[18] Zhao, Y., & Karypis, G. (2002, November). Evaluation of hierarchical clustering algorithms for document datasets. In Proceedings of the eleventh international conference on Information and knowledge management (pp. 515-524). ACM.

[19] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd (Vol. 96, No. 34, pp. 226-231).

[20] Liao, W. K., Liu, Y., & Choudhary, A. (2004, April). A grid-based clustering algorithm using adaptive mesh refinement. In 7th Workshop on Mining Scientific and Engineering Datasets of SIAM International Conference on Data Mining (pp. 61-69).

[21]Snehal Ramteke, Association Rule Mining Algorithm Using Big Data Analysis, International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 4 Issue: 5 73 – 75

[22]Oweis, N. E., Fouad, M. M., Oweis, S. R., Owais, S. S., & Snasel, V. A Novel Mapreduce Lift Association Rule Mining Algorithm (MRLAR) for Big Data. *International Journal of Advanced Computer Science & Applications*,*1*(7), 151-157.

[23]Savasere, A., Omiecinski, E. R., & Navathe, S. B. (1995). An efficient algorithm for mining association rules in large databases.

[24]Slimani, T., & Lazzez, A. (2014). Efficient Analysis of Pattern and Association Rule Mining Approaches. *arXiv preprint arXiv:1402.2892*.

[25]Moens, S., Aksehirli, E., & Goethals, B. (2013, October). Frequent itemset mining for big data. In *Big Data, 2013 IEEE International Conference on* (pp. 111-118). IEEE.

[26]Fan, W., & Bifet, A. (2013). Mining big data: current status, and forecast to the future. *ACM sIGKDD Explorations Newsletter*, *14*(2), 1-5.

[27]Fujimaki, R., & Morinaga, S. (2012). The Most Advanced Data Mining of the Big Data Era. *NEC Technical Journal*, *7*, 91.

[28]McHugh, D. (2015). Traffic prediction and analysis using a big data and visualisation approach.

[29]Amooee, G., Minaei-Bidgoli, B., & Bagheri-Dehnavi, M. (2012). A comparison between data mining prediction algorithms for fault detection (Case study: Ahanpishegan co.). *arXiv preprint arXiv:1201.6053*.

[30]Merceron, A., & Yacef, K. (2008, June). Interestingness measures for association rules in educational data. In *Educational Data Mining 2008*.