IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 4, April 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904057 |

Distributed Framework of a Big Data Analysis and Rule Mining for IoT Big Data

L.Yamuna.,[1] Mrs. K.K.Kavitha[2] , Mr. C. Senthil Raja[3]

Research Scholar, Department of Computer Science, Selvamm Arts and Science College (Autonomous), TamilNadu, India[1]

Head of the Department, Department of Computer Science, Selvamm Arts and Science College (Autonomous), TamilNadu, India[2]

Assistant Professor, Department of Computer Science, Selvamm Arts and Science College (Autonomous), TamilNadu, India[3]

ABSTRACT: Data Mining aims to examination for implicit, previously recognized and potentially valuable in order from data. Big Data Mining is the ability of getting helpful data from large datasets or stream of data. The presented system attempts to seek the pattern of interest from probabilistic database. If the output formed is not matching the user condition then the effort put to create that line is complete waste. In several real-life applications, users might look for a little part of this enormous search space for Big Data Mining. This requested data is most probably doubtful to the system. Another challenge to representation the secret information from the data sets which creates results less useful the proposed method reduces the search space to a greater extent as it concentrates extra on the conditions by using the Map Reduce form. The users are known whole freedom to state their interests by specifying their own conditions. Besides categorization and clustering, anomaly recognition, frequent pattern mining and association rule mining are integrated as the latter two analyze precious data and helps the producer by discovery the interesting or popular patterns that make known customer purchase behavior. The probabilistic database refers to the database where necessary result will be present. The existential probability recommends to the probability of data being present in the searched database. The algorithm provides here greatly decreases the search space for Big Data mining of doubtful data, returning only those patterns that are motivating to the users for Big Data analytics.

KEYWORDS: Big Data, Data Mining, Analytics, Data Input, Analysis, Framework

I. INTRODUCTION

As the information technology spreads fast, most of the data were born digital as well as exchanged on internet today. According to the estimation of Lyman and Varian , the new data stored in digital media devices have already been more than 92 % in 2002, while the size of these new data was also more than five Exabyte's. In fact, the problems of analyzing the large scale data were not suddenly occurred but have been there for several years because the creation of data is usually much easier than finding useful things from the data. Even though computer systems today are much faster than those in the 1930s, the large scale data is a strain to analyze by the computers we have today.

In response to the problems of analyzing large-scale data, quite a few efficient methods [2], such as sampling, data condensation, density-based approaches, grid-based approaches, divide and conquer, incremental learning, and distributed computing, have been presented. Of course, these methods are constantly used to improve the performance of the operators of data analytics process. The results of these methods illustrate that with the efficient methods at hand, we may be able to analyze the large-scale data in a reasonable time.The term 'Big Data' describes innovative techniques and technologies to capture, store, distribute, manage and analyze petabyte-or larger-sized datasets with high-velocity and different structures. Big data can be structured, unstructured or semi-structured, resulting in incapability of conventional data management methods. Data is generated from various different sources and can arrive in the system at various rates. In order to process these large amounts of data in an inexpensive and efficient way, parallelism is used. Big Data is a data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. At the same time, cloud technology is emerging as an infrastructure appropriate for building large and complex systems. Storage and compute resources provisioned from converged infrastructure and distributed

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904057 |

resource pools present a cost-effective different to the traditional in-house data center. The cloud offers new levels of scalability, flexibility and availability, and permits easy access to data from several locations and any device. In addition, the cloud representations a data model of objects that contain data integrated among its user-defined and system-defined metadata as a single part.

II. RELATED WORK

Image-based recommendations through clustering play a vital role in many application areas, and profound research works are done mostly on the medical images for various analyses using artificial intelligence and other machine learning techniques. Hancer et al. [22] proposed an artificial bee colony approach for the three benchmarks Lena, Remote Sensing, and Brain MRI images for the clustering approach compared with the particle swarm optimization, and K-means cluster approaches. M. Gong et al. [23] had come with a new clustering approach using the Kernal metric for the fuzzy c-mean clustering process and tested on the various image datasets like synthetic, natural, and medical images for the performance evaluation. Karthikeyan and Aruna [24] proposed a probabilistic text and image-based semi-supervised clustering approach. They used the topic modeling, comparing image features such as color sets and image block signature computing the similarity distance measure. Their method further compared with the K-means and Dbscan unsupervised clustering algorithms.

The notion of the outlier has various interpretations in different scientific communities. In statistics [32], [8], outliers are data points that deviate from a specific model assumption. The notion of distance-based outliers was recently introduced in databases [23], [24], [33]. According to this notion, a point, P, in a multidimensional data set is an outlier if there are less than (a user-specified constant) p points from the data in the -neighborhood of P. A related notion of outliers was introduced in [6]. Deviant points [21] are special forms of outliers. Sampling [39], [40] is a well-recognized and widely used statistical technique. In the context of clustering in databases, both partitional [28] and hierarchical [17] clustering techniques employ uniform random sampling to pick a set of points and guide the subsequent clustering effort. In data mining, Toivonen [38] examined the problem of using sampling during the discovery of association rules. Sampling has also been recently successfully applied in query optimization [11], [10], as well as approximate query answering [16], [2].incorporate substances and the connections between them shapes the connections.

III. HETEROGENEITY AND INCOMPLETENESS

When humans consume information, a great deal of heterogeneity is comfortably tolerated. In fact, the nuance and richness of natural language can provide valuable depth. However, machine analysis algorithms expect homogeneous data, and cannot understand nuance. In consequence, data must be carefully structured as a first step in (or prior to) data analysis. Computer systems work most efficiently if they can store multiple items that are all identical in size and structure. Efficient representation, access, and analysis of semi-structured data require further work.

SCALE

Of course, the first thing anyone thinks of with big data is its size. After all, the word “big” is there in the very name. Managing large and rapidly increasing volumes of data has been a challenging issue for many decades. In the past, this challenge was mitigated by processors getting faster, following Moore’s law, to provide us with the resources needed to cope with increasing volumes of data. But, there is a fundamental shift underway now: data volume is scaling faster than compute resources, and CPU speeds are static.

TIMELINESS

The flip side of size is speed. The larger the data set to be processed, the longer it will take to analyze. The design of a system that effectively deals with size is likely also to result in a system that can process a given size of data set faster. However, it is not just this speed that is usually meant when one speaks of Velocity in the context of big data. Rather, there is an acquisition rate challenge

PRIVACY

The privacy of data is another huge concern, and one that increases in the context of big data. For electronic health records, there are strict laws governing what can and cannot be done. For other data, regulations, particularly in the US, are less forceful. However, there is great public fear regarding the inappropriate use of personal data, particularly through

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904057 |

linking of data from multiple sources. Managing privacy is effectively both a technical and a sociological problem, which must be addressed jointly from both perspectives to realize the promise of big data.

HUMAN COLLABORATION

In spite of the tremendous advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a hard time finding. Ideally, analytics for big data will not be all computational rather it will be designed explicitly to have a human in the loop. The new sub-field of visual analytics is attempting to do this, at least with respect to the modeling and analysis phase in the pipeline. In today's complex world, it often takes multiple experts from different domains to really understand what is going on. A big data analysis system must support input from multiple human experts, and shared exploration of results together in one room. The data system has to accept this distributed expert input, and support their collaboration.

3.1 K-MEAN (++) CLUSTERING APPROACH

Clustering is an unsupervised learning technique to find out the K-Patterns in the given image dataset. For our image similarity approach, we have taken a Fashion image dataset with 40,000 images for analysis. We have computed K-Mean clustering by computing ten iterative times with different centroid positions, and the K-Means++ initialization method has been used for better convergence. An earlier version of K-Means consists of the initialization method, assignment of data points, updating the centroid, and repeats the stages until it finds out the convergence. Hence, choosing the random k-centroids is also known as the Initialization sensitivity procedure. To overcome the initialization sensitivity procedure in our approach, we have adapted the K-Means ++ approach for finding out the better initial selection of K-centroids for PCA transformed images using the SVD approach on our e-commerce 40K dataset.

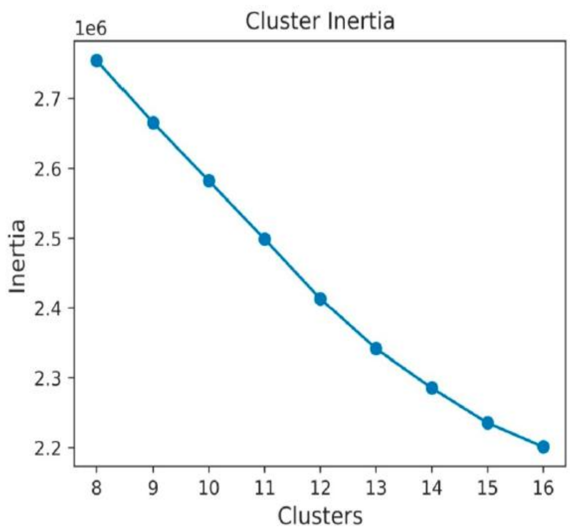

Figure 3.1 K-Mean (++) Clustering Approach

3.2 DENSITY ESTIMATION TECHNIQUES

A nonparametric density estimator technique attempts to define a function that approximates the data distribution. It is imperative, for performance reasons, to be able to derive the approximate solution quickly. Thus, description of the function must be kept in memory. The success of the various density estimator techniques hinges upon several parameters, which include the simplicity of the function, the time it takes to find the function parameters, and the number of parameters we need to store for a given approximation quality.

3.3 CLUSTERING

We run clustering algorithms on the density-biased sample to evaluate the accuracy of the proposed technique. To cluster the sample points, we use a hierarchical clustering algorithm. The algorithm starts with each point in a separate cluster and at each step merges the closest clusters together. The technique is similar to CURE in that it keeps a set of points as a representative of each cluster of points. The hierarchical clustering approach has several advantages over K-

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904057 |

means or K-medoids algorithms. It has been shown experimentally that it can be used to discover clusters of arbitrary shape, unlike K-means or K-medoids techniques, which typically perform best when the clusters are spherical. It has also been shown that a hierarchical clustering approach is insensitive to the size of the cluster. On the other hand, K-means or K-medoids algorithms tend to split up very large clusters and merge the pieces with smaller clusters. The main disadvantage of the hierarchical approach is that the runtime complexity is quadratic. Running a quadratic algorithm even on a few thousand data points can be problematic.

3.4 CLUSTERING EXPERIMENTS

A viable approach to overcome this difficulty is to reduce the size of the data set in a manner that preserves the information of interest to the analysis without harming the overall technique. A hierarchical clustering algorithm that is running off the biased sample is in effect an approximation algorithm for the clustering problem on the original data set. In this respect, our technique is analogous to the new results on approximation clustering algorithms since these algorithms also run on a (uniform random) sample to efficiently obtain the approximate clusterings. The techniques are not directly comparable, however, because these algorithms are approximating the optimal K-medoids partitioning, which can be very different from the hierarchical clusterings we find. We note here that we can also use biased sampling in conjunction with K-means or K-medoids algorithms if this is desirable. However, K-means,or K-medoids algorithms optimize a criterion that puts equal weight to each point in the data set. For example, for the K-means clustering.

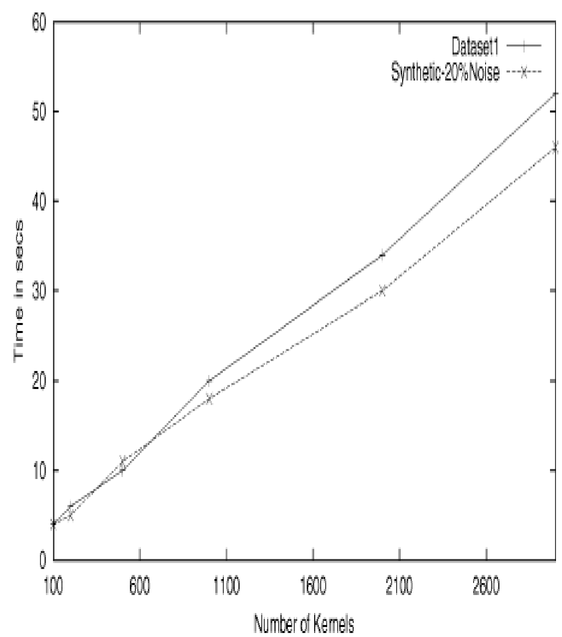

Figure 3.4 Clustering Experiments

3.5 TECHNOLOGY DEVELOPMENT

The big data technology is still in its infancy. Many key technical problems, such as cloud computing, grid computing, stream computing, parallel computing, big data architecture, big datamodel, and software systems supporting big data, etc. should be fully investigated. –

Format conversion of big data: Due to wide and diverse data sources, heterogeneity is always a characteristic of big data, as well as a key factor which restricts the efficiency of data format conversion. If such format conversion can be made more efficient, the application of big data may create more values. Mobile Netw Appl (2014) 19:171–209 203 –

Big data transfer: Big data transfer involves big data generation, acquisition, transmission, storage, and other data transformations in the spatial domain. As discussed, big data transfer usually incurs high costs, which is the bottleneck for

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904057 |

big data computing. However, data transfer is inevitable in big data applications. Improving the transfer efficiency of big data is a key factor to improve big data computing.

Real-time performance of big data: The real-time performance of big data is also a key problem in many application scenarios. Effective means to define the life cycle of data, compute the rate of depreciation of data, and build computing models of real-time and online applications, will influence the analysis results of big data.

Processing of big data: As big data research is advanced, new problems on big data processing arise from the traditional data analysis, including (i) data re-utilization, with the increase of data scale, more values may be mined from re-utilization of existing data; (ii) data re-organization, datasets in different businesses can be re-organized, which can be mined more value; (iii) data exhaust, which means wrong data during acquisition. In big data, not only the correct data but also the wrong data should be utilized to generate more value.

3.6 TECHNIQUES FOR BIG DATA ANALYTICS

Supervised, unsupervised, and hybrid machine learning approaches are the most widely used tools for descriptive and predictive analytics on big data. Apart from that, various techniques from mathematics have been used in big data analytics. The problem of big data volume can be somewhat minimized by dimensionality reduction. Linear mapping methods, such as principal component analysis (PCA) and singular value decomposition, as well as non-linear mapping methods, such as Sammon's mapping, kernel principal component analysis, and laplacian eigenmaps, have been widely used for dimensionality reduction. Another important tool used in big data analytics is mathematical optimization. Subfields of optimization, such as constraint satisfaction programming, dynamic programming, and heuristics & metaheuristics are widely used in AI and machine learning problems. Other important optimization methods include multi-objective and multi-modal optimization methods, such as pareto optimization and evolutionary algorithms, respectively.

Statistics is considered as a counterpart to machine learning; differentiated by data model versus algorithmic model respectively. The two fields have subsumed ideas from each other. Statistical concepts, such as expectation-maximization and PCA, are widely adopted in machine learning problems. Similarly, machine learning techniques, such as probably approximately correct learning are used in applied statistics. However, both of these tools have been heavily used for big data analytics. Big data analytics has a close proximity to data mining approaches. Mining big data is more challenging than traditional data mining due to massive data volume. The common practice is to extend the existing data mining algorithms to cope with massive datasets, by executing on samples of big data and then merging the sample results. This kind of clustering algorithms include CLARA (Clustering Large Applications) and BIRCH (Balanced Iterative Reducing using Cluster Hierarchies). Researchers have also emphasized on the reduction of computational complexity of data mining algorithms. For example, spectral regression discriminant analysis significantly reduces the time and space complexity by simplifying discriminant analysis to a set of regularized least squares problems. Similarly, Shi et al. reduce the space complexity of non-linear discriminant analysis from O(n2) to O(n), to minimize computation and storage problem on large-scale datasets.

Nevertheless, time and space complexity of most of the machine learning and statistical methods are very high to be effective for real time analysis on large-scale dataset. In the recent years, distributed and parallel computing technologies have emerged as the prime solution to large scale computing problems, due to their scalability, performance, and reliability. Therefore, efforts have been made to perform big data analytics using distributed computing, under strict performance and reliability constraints. Consequently, distributed data analytics algorithms have been proposed in the literature.

Mining of distributed data in itself has emerged as a new paradigm of data analytics. It should be noted that, to be effective, the nodes should perform the computations independently, i.e., without constantly sharing intermediate data with peer nodes. Park and Kargupta discuss the distributed algorithms for classifier learning, association rule mining, and clustering. Rana et al. propose a component-based system, designated as PaDDMAS, for developing distributed data mining applications. Similar systems for distributed machine learning methods are proposed, such as MLbase. Further, cloud computing infrastructure-based systems are also proposed for performing distributed machine learning, such as the Distributed GraphLab framework that emphasizes on consistency and fault-tolerance in distributed analytics. The main driving force for big data analytics has been the industry researches for massive-scale commercial applications. Although cluster and grid computing have existed for long, they are designed specifically for particular applications and require high cost and expertise.

Therefore, the technologies for big data analytics did not evolve significantly in that period. When cloud computing infrastructure and distributed processing platforms, such as mab reduce, and their open source implementations became widely available, the research on big data analytics escalated. Iterative graph processing systems, for solving large scale practical computing problems, have also been proposed. The proprietary graph processing architecture developed at

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904057 |

Google, known as Pregel, addresses distributed processing of large scale real-life graphs. An opensource counterpart of Pregel is Apache Giraph14, which provides additional features, such as edge oriented input and out-of-core computation. Moreover, the rising data volume has contributed to the increasing demand for big data analytics. In the recent years, distributed file system technologies, such as HDFS and QFS as well as NoSQL databases for unstructured data, such as MongoDB15 and CouchDB16 have been widely used for big data analytics.

Machine learning libraries have been developed for big data analytics. The most notable machine learning library for big data analytics is Apache Mahout, which contains implementations of various machine learning techniques, such as classifiers, clustering, and recommender systems, which are scalable to large scale datasets. MLlib17 is a similar library to perform machine learning on big data on the Apache Spark platform, a MapReduce variant for iterative and fast computations on big data. However, these libraries still lack many important machine learning methods and more contributions are needed from the community.

IV. RESEARCH METHODOLOGY

4.1 BIG DATA ANALYSIS FRAMEWORKS AND PLATFORMS

Various solutions have been presented for the big data analytics which can be divided into Processing/Compute: Hadoop, Nvidia CUDA or Twitter Storm Storage: Titan or HDFS, and Analytics: MLPACK or Mahout. Although there exist commercial products for data analysis, most of the studies on the traditional data analysis are focused on the design and development of efficient and/or effective "ways" to find the useful things from the data. But when we enter the age of big data, most of the current computer systems will not be able to handle the whole dataset all at once; thus, how to design a good data analytics framework or platform3 and how to design analysis methods are both important things for the data analysis process.

4.2 BIG DATA ANALYSIS ALGORITHMS

Because the big data issues have appeared for nearly 10 years, in Fan and Bifet pointed out that the terms "big data" and "big data mining" were first presented in 1998, respectively. The big data and big data mining almost appearing at the same time explained that finding something from big data will be one of the major tasks in this research domain. Data mining algorithms for data analysis also play the vital role in the big data analysis, in terms of the computation cost, memory requirement, and accuracy of the end results. In this section, we will give a brief discussion from the perspective of analysis and search algorithms to explain its importance for big data analytics.

4.3 CLUSTERING ALGORITHMS

In the big data age, traditional clustering algorithms will become even more limited than before because they typically require that all the data be in the same format and be loaded into the same machine so as to find some useful things from the whole data. Although the problem of analyzing large-scale and high-dimensional dataset has attracted many researchers from various disciplines in the last century, and several solutions have been presented in recent years, the characteristics of big data still brought up several new challenges for the data clustering issues. Among them, how to reduce the data complexity is one of the important issues for big data clustering. In divided the big data clustering into two categories: single-machine clustering (i.e., sampling and dimension reduction solutions), and multiple-machine clustering (parallel and Map Reduce solutions). This means that traditional reduction solutions can also be used in the big data age because the complexity and memory space needed for the process of data analysis will be decreased by using sampling and dimension reduction methods. More precisely, sampling can be regarded as reducing the "amount of data" entered into a data analyzing process while dimension reduction can be regarded as "downsizing the whole dataset" because irrelevant dimensions will be discarded before the data analyzing process is carried out Process in parallel. BIRCH and sampling method were used in Cloud Vista to show that it is able to handle large-scale data, e.g., 25 million census records. Using GPU to enhance the performance of a clustering algorithm is another promising solution for big data mining. The multiple species flocking (MSF) was applied to the CUDA platform from NVIDIA to reduce the computation time of clustering algorithm in. The simulation results show that the speedup factor can be increased from 30 up to 60 by using GPU for data clustering. Since most traditional clustering algorithms (e.g., k-means) require a computation that is centralized, how to make them capable of handling big data clustering problems is the major concern of Feldman et al. who use a tree construction for generating the core sets in parallel which is called the "merge-and-reduce" approach. Moreover, Feldman et al. pointed out that by using this solution for clustering, the update time per datum and memory of the traditional clustering algorithms can be significantly reduced.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904057 |

4.4 CLASSIFICATION ALGORITHMS

Similar to the clustering algorithm for big data mining, several studies also attempted to modify the traditional classification algorithms to make them work on a parallel computing environment or to develop new classification algorithms which work naturally on a parallel computing environment. In the design of classification algorithm took into account the input data that are gathered by distributed data sources and they will be processed by a heterogeneous set of learners. In this study, Tekin et al. presented a novel classification algorithm called "classify or send for classification" (CoS). They assumed that each learner can be used to process the input data in two different ways in a distributed data classification system. One is to perform a classification function by itself while the other is to forward the input data to another learner to have them labeled. The information will be exchanged between different learners. In brief, this kind of solutions can be regarded as a cooperative learning to improve the accuracy in solving the big data classification problem. An interesting solution uses the quantum computing to reduce the memory space and computing cost of a classification algorithm. For example, in Rebentrost et al. presented a quantum based support vector machine for big data classification and argued that the classification algorithm they proposed can be implemented with a time complexity O(log NM) where N is the number of dimensions and M is the number of training data. There are bright prospects for big data mining by using quantum-based search algorithm when the hardware of quantum computing has become mature.

V. CONCLUSION

Big data analytics is trying to take advantage of the excess of information to use it productively. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises. However, many technical challenges described in this paper must be addressed before this potential can be realized fully. Furthermore, these challenges will require transformative solutions, and will not be addressed naturally by the next generation of industrial products. It must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of big data. The intuition here is that the earlier iterations provide some partial clustering information.

VI. FUTURE ENHANCEMENT

This information can potentially be used to construct the tree such that the pruning is more effective. As a vital improvement of the next age of Internet, the Internet of Things pulls in numerous considerations by industry world and scholarly circles. IoT information has numerous qualities, for example, distributed storage, mass temporal and spatial related information, and constrained assets of nodes and so forth. These makes the issue of information mining in IoT turns into a test assignment.

REFERENCES

[1]. Lyman P, Varian H. How much information 2003? Tech. Rep, 2004. [Online]. http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/ printable_ report. pdf.
[2]. Xu R, Wunsch D. Clustering. Hoboken: Wiley-IEEE Press; 2009.
[3]. Ding C, He X. K-means clustering via principal component analysis. In: Proceedings of the twenty-first international conference on machine learning; 2004. pp. 1–9.
[4]. Kollios G, Gunopulos D, Koudas N, Berchtold S. Efficient biased sampling for approximate clustering and outlier detection in large data sets. IEEE Trans Knowl Data Eng. 2003;15 (5):1170–87.
[5]. Fisher D, DeLine R, Czerwinski M, Drucker S. Interactions with big data analytics. Interactions. 2012;19(3):50–9.
[6]. Laney D. 3D data management: controlling data volume, velocity, and variety. META Group, Tech. Rep. 2001. [Online]. http://blogs.gartner.com/doug-laney/files/2012/01/ad9493D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.
[7]. van Rijmenam M. Why the 3v's are not sufficient to describe big data. BigData Startups, Tech. Rep. 2013. [Online]. http://www.bigdata-startups.com/3vs-sufficient-describe-big-data/ .
[8]. Borne K. Top 10 big data challenges a serious look at 10 big data v's. Tech. Rep. 2014. [Online]. https://www.mapr.com/blog/top-10-big-data-challenges-look-10-big-data-v.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904057 |

[9]. M. Cox, D. Ellsworth, Managing Big Data For Scientific Visualization, ACM Siggraph, MRJ/NASA Ames Research Center, 1997.
[10]. J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A.H.Byers, Big data: The next frontier for innovation, competition, and productivity, (2011).
[11]. P. Zikopoulos, K. Parasuraman, T. Deutsch, J. Giles, D. Corrigan, Harness the Power of Big Data The IBM Big Data Platform, McGraw Hill Professional, 2012.
[12]. J.J. Berman, Introduction, in: Principles of Big Data, Morgan Kaufmann, Boston, 2013, xix–xxvi (pp).
[13]. J. Gantz, D. Reinsel, Extracting value from chaos, IDC iView (2011) 1–12.
[14]. J.K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, M. Miettinen, The mobile data challenge: Big data for mobile computing research, Workshop on the Nokia Mobile Data Challenge, in: Proceedings of the Conjunction with the 10th International Conference on Pervasive Computing, 2012, pp. 1–8.
[15]. D.E. O'Leary, Artificial intelligence and big data, IEEE Intell. Syst. 28 (2013) 96–99.
[16]. M. Chen, S. Mao, Y. Liu, Big data: a survey, Mob. Netw. Appl. 19 (2) (2014) 1–39.
[17]. J. Hurwitz, A. Nugent, F. Halper, M. Kaufman, Big data for dummies, For Dummies (2013).
[18]. Russom P. Big data analytics. TDWI: Tech. Rep; 2011.
[19]. Ma C, Zhang HH, Wang X. Machine learning for big data analytics in plants. Trends Plant Sci. 2014;19(12):798–808.
[20]. Boyd D, Crawford K. Critical questions for big data. Inform Commun Soc. 2012; 15(5):662– 79.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING