International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-4, Issue-6)

Research Article

June 2015

K-Means Clustering Method for Big Data Mining

[1]Pravin Anil Tak, [2]Dr. S. V. Gumaste, [3]Prof. S. A. Kahate
[1]M.E. II[nd] Year Student, [2]Professor, [3]Assistant Professor
[1, 2, 3] Computer Engineering Department, Sharadchandra Pawar College of Engineering,
Dumberwadi, Otur, Tal- Junnar, Dist- Pune, Maharashtra, India

Abstract—

The Big Data is an important term now days since various decisions related to business, social, political, educational etc. are taken by analyzing such huge data. Clustering is an efficient way to distribute data among various groups by specifying attributes. The elements in the cluster organize in such a way that elements in a same cluster shows similarity but dissimilar with elements of other cluster.

Keywords— Cluster; privacy; partitioning; decision making;

I. INTRODUCTION

A word "Big Data" takes a several forms since concentrating from different view of researcher. Now in current age an everyone interacting with network to fulfil needs as it comes, but clean observation indicates that correct information not get replied from various service providers. Actually Data plays an important role to take different decisions that are deals with business, educational, social, politic al, personal, and geographical etc. Data mining is not a new concept as information technology deals with very early from but pervious data structure somewhat unique shows structured, limited in size and not variant but as time passes everyone generating data in his own schemata since the unstructured data generated which is large size such data called as Big Data. Due to this attributes existing database management systems are not able to process data and not gives efficient results within specified time.

Big data having various attributes like heterogeneous, large in size, coming from different diverse sources, speedy growth in data. To handle big data efficiently, the understanding of these attributes are important since through these attributes data management can possible and result oriented approach comes out. An unstructured data defines the property of heterogeneous since data not following specific structure due to this data becomes complex in nature and such data are not get handle by existing database management systems. Data warehouse has been utilized there to handle attribute heterogeneous in well manner.

As big data comes from various sources that are located in the network follows their own schemata and approaching data at central site (at Server) becomes complex. As time passes data generation from different sources are enhancing continuously, here time factor shows how data is growing fast and it's don't having any structure.

This Paper consist of five sections such as section I introduces agenda of Big Data and explains various terms related to big data, section II shows background work of various authors was they had done, section III defines efficient factors that are needed to mine big data within specified time, section IV security measures required to enhance the compatibility regards to big data mining and finally section V concludes the paper.

II. BACKGROUND

The Literature survey shows the work done of an authors which had carried out in past. Data Management is not a new term comes to know but the factors are changed related to data. Since data get enhancing continuously existing systems are not capable to handle big data, here few of back grounded working has explained to define knowledge and approach towards Big data Mining.

Hui Chen et all was shows ways to find parallel frequent patterns on big transitional data. They was defined in big data era, data size has risen from TB-level to PB-level. Traditional algorithm cannot able to fulfill the needs of big data computing. They had design a parallel algorithm for mining frequent pattern over big transactional data based on an extended MapReduce Frame. In which, the mass data file is firstly split into many data sub files, the patterns in each sub file can be quickly located based on bitmap computation by scanning the data only once. And the computing results of all sub files are merged for mining the frequent patterns in the whole big data. In order to improve the performance of the proposed method, the insignificant patterns are pruned by a statistic analysis method when the data sub files are processed. The experimental results show that the method is efficient, strong in scalability, and can be used to efficiently mine frequent patterns in big data.

Yang Song et all was explains the concept of storage mining due to this information technology meets the big data analytics. They was defined The emerging paradigm shift to cloud based data center infrastructures imposes remarkable challenges to IT management operations, e.g., due to virtualization techniques and more stringent requirements for cost and efficiency. On one hand, the voluminous data generated by daily IT operations such as logs and performance measurements contain abundant information and insights which can be leveraged to assist the IT management.

On the other hand, traditional IT management solutions cannot consume and exploit the rich information contained in the data due to the daunting volume, velocity variety, as well as the lack of scalable data mining and machine learning frameworks to extract insights from such raw data. As an example, here introduce our project of Storage Mining, which exploits big data analytics techniques to facilitate storage cloud management.

Carson Kai-Sang Leung et all introduces the concept that help for Reducing the Search Space for Big Data Mining for Interesting Patterns from Uncertain Data. An existing data mining algorithms search interesting patterns from a set transactional database of precise data. However, there are situations in which data are uncertain. Items in each transaction of these probabilistic databases of uncertain data are associated with existential probabilities, which express the likelihood of these items to be present in the transaction. When compared with mining from precise data, the search space for mining from uncertain data is much larger due to the presence of the existential probabilities. This problem is worsened as moving to the era of big data. Furthermore, in many real-life applications, users may be interested in a tiny portion of this large search space for big data mining.

III. CLUSTERING METHOD TO MINE BIG DATA

The studies of different factors of big data are important to mine big data properly. Clustering is the process of grouping a set of data elements into multiple groups or clusters so that objects/elements within a cluster have high similarity, but are very dissimilar to objects in others clusters. Dissimilarities and similarities are accessed based on the attribute values describing the objects and often involve distance measures. Clustering as an efficient data mining tool has its roots in many application areas such as biology, security, business intelligence, and Web search. Cluster analysis is the process of partitioning a set of data objects into subsets. Each subset is a cluster, such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters. In business intelligence, clustering takes an important role used to organize a large number of customers into groups, where customers within a group share various similar characteristics.

As a Big data mining function, cluster analysis can be used as a standalone tool to gain insight into the distribution of data, to observe the attributes of each cluster, and to focus on a particular set of clusters for further mining process. Clustering also called as data segmentation because clustering partitions big/large data sets into groups according to their similarity. There are various requirements of clustering in big data mining such as scalability, heterogeneity, discovery of clusters with arbitrary shape, requirements for domain knowledge to determine input parameters, ability to deal with noisy data, capability of clustering high dimensionality data, constraint based clustering, interpretability and usability.

The highly scalable clustering algorithms are needed to mine large data set; it may lead to generate biased results. Many algorithms are designed to cluster numeric data. However, applications may require clustering data types, such as binary, nominal and ordinal data types. Recently, more and more applications need clustering techniques for complex data types such as graphs, sequences, images, and documents. Most real-world data sets contain missing; unknown, erroneous data are often noisy. Real world applications may need to perform clustering under various kinds of constraints. Users want clustering results to be interpretable, comprehensible and usable. That is clustering may need to be tied in with specific semantic interpretations and applications.

Basically there are four methods of clustering such as partitioning method, hierarchical method, density based method, and grid based methods. In partitioning method a large data set has been partitions into number of subsets called as clusters but they are defined by following specific criteria. A hierarchical method creates a hierarchical decomposition of the given set of data objects. In density based methods clusters are formed based on distance between objects but such methods can find only spherical-shaped clusters and encounter difficulty in discovering clusters of arbitrary shapes.

A PARTITIONS METHOD: K-MEANS ALGORITHM

The simplest and most fundamental version of cluster analysis is partitioning, which organizes the objects of a set into several exclusive groups or clusters. Formally, given a data set, D of n objects, and k, the number of clusters to form, a partition algorithm organizes the objects into k partitions ($k \leq n$), where each partition represents a cluster. K-Means is a one of partitioning algorithm which defines efficient way to cluster analysis.

The K-means algorithm is simplest unsupervised learning algorithm that solves the well known clustering problem. The procedure defines a simple and easy way to classify a given data set through a certain number of clusters (consider/assume k clusters) fixed a priori. The main idea is to evaluate k centroids, one for each cluste and such centroids must be placed in a cunning way because of different location causes different outcomes. So, the better way is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

Let consider a data set, D contains n objects in Euclidean space. Partitioning methods distribute the objects in D into k clusters, C1, C2,……..Ck, that is $Ci < D$ and $Ci \cap Cj = \emptyset$ for ($1 \leq i, j \leq k$). An objective function is used to assess the partitioning quality so that objects within a cluster are similar to one another but dissimilar to objects in other clusters. This is the objective function aims foe high intracluster similarity and low intercluster similarity. A centroid based partitioning technique uses the centroid of a cluster, Ci , to represent that cluster. Conceptually, the centroid of a

cluster is its center point. The centroid can be defined in various ways such as by the mean or medoid of the objects or points assigned to the cluster. The difference between an object p & ci, the representative of the cluster, is measured by dist(p,ci), where dist(x,y) is the Euclidean distance between two points x and y. The quality of cluster ci can be measured within cluster variation, which is the sum of squared error between all objects in ci and the centroid ci, defined as

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, c_i)^2$$

Where E is the sum of the squared error for all objects in the data set; p is the point in space representing a given object; and ci is the centroid of cluster Ci here both p and ci are multidimensional. In other words, for each cluster, the distance from the object to its cluster center is squared, and the distances are summed. This objective function tries to make the resulting k clusters as compact and as separate as possible.

Algorithm K-Means-

Input:

k: the number of clusters,

D: a data set containing n objects.

Output:

A set of k clusters.

Method:

1) Arbitrarily choose k objects from D as the initial centers;

2) repeat

3) assign/reassign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;

4) update the cluster means, that is calculate the mean value of the objects for each cluster;

5) until no change;

According to above defined algorithm clusters are formed in such that objects in the different cluster shows the attribute of heterogeneity and objects in the same cluster shows the property of homogeneity. If the iterative procedure goes successfully then clusters get formed are compact and consistent, it gives better results.

As all knows that, big data shows the attributes like heterogeneous, large in size, complexity and most important is it comes from autonomous sources from the network. To mine efficient data from such huge data clustering is an appropriate way to handle to big data.

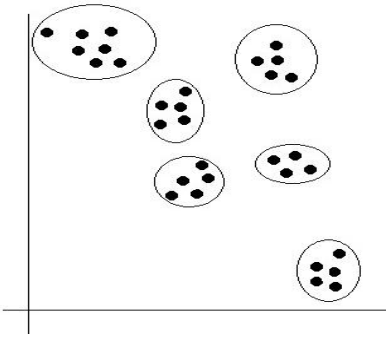

Fig-Clustering using k-means

IV. SECURITY MEASURES OF BIG DATA MINING

Security concerns are most important if anyone going to deal with database and tools. The security aspects deal many things for the big data mining applications. The human related errors and mishandling is also a security concern for the big data mining since data comes in large amount and it generated by different organization or peoples. These type security measures are based on the characteristics of big data mining.

1. Privacy-

This is compulsory for the each individual who operates the big data mining tools. Privacy is concerned with individual user. The individual duties are to keep the data items undisclosed to other peoples. The organization should have to educate the employees about the privacy and its related aspect time to time according to attacks and breaches of current scenarios and past scenarios. Data privacy internally should be maintained with the help of different types of integrity constraints.

2. Data Correctness-

Data correctness is more important thing for the big data mining. If a database or data warehouse contains incorrect data then mining tools will produce incorrect result. Thus, there would be a filter that filter out the data and correct the data which is not correct. Data correctness should be ensured before entry into the database. Correct data items always produces the correct output by extracting data by data mining tools or by any other tools.

3. Data Integrity-

Integrity of data is also a security aspect because huge data comes at system. If data numeric field is in mode of character then it produces the incorrect result of mathematical operations during data mining. Integrity of data under database is managed by the help of various different types of integrity constraints of databases. Once an integrity constraint is enforced on data items then user should not have to right about removal of that integrity constraint.

4. Correction of Mistaken Data-

The data and information stored at different storage medium are not correct completely since the format of storage get change according to respective medium. Thus, there should be a technique that finds the mistaken and incorrect data to be corrected before the storing into the large and consistent databases. The correction must be automated not manual. Correction of large mistaken data requires algorithms having considered integrity and availability. Manual correction takes too much time and there would be threat for disclosure of sensitive data. A proper mechanism should be implemented on behalf of the company policy to handle the correctness of data if manual procedure is applied for that.

Again there are several attributes are considered while measuring security of large data. As business organization knows big historical data is needed while important decision are taken related to critical issues in organization. Therefore now a day's an efficient algorithm and effective data mining tools are having demand since all have to mine data which gives real and accurate informative results.

V. CONCLUSIONS

The conclusion tends to define clustering is a major factor needed while big data management has been carried out since k means algorithm basically a way to form efficient clusters. Through clustering large amount of data can be easily handled and with the help of that heterogeneous huge data can be manage and finally gives a better results.

ACKNOWLEDGMENT

All faith and honor to Lord Shri Ganesh for his grace and inspiration. I wish to express my sincere thanks to all the departmental staff members for their support. Last but not the least; I would like to thank all my Friends and Family members who have always been there to support and helped me to complete this paper work in time.

REFERENCES

[1] Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE, Gong-Qing Wu, and Wei Ding, Senior Member, IEEE *Data Mining with Big Data* IEEE Transactions on Knowledge and Data Engineering, VOL. 26, NO. 1, JANUARY 2014.

[2] WANG Shuliang, DING Gangyi, ZHONG Ming *Big Spatial Data Mining* IEEE International conference on Big Data 2013.

[3] U Kang and Christos Faloutsos *Big Graph Mining: Algorithms and Discoveries* SIGKDD Explorations Volume 14, Issue 2.

[4] Tak Pravin, Kahate Sandip *An Efficient Approach for Big Data Mining* International Journal of Informative and Futuristic Research 2014.

[5] Hui Chen, Tsau Young Lin, Zhibing Zhang and Jie Zhong *Parallel Mining Frequent Patterns over Big Transactional Data in Extended MapReduce* IEEE International conference on Granular Computing 2013.

[6] Junyu Xuan, Xiangfeng Luo, Jie Lu *Mining Websites Preferences on Web Events in Big Data Environment* 2013 IEEE 16th International Conference on Computational Science and Engineering.

[7] Shuliang WANG Hanning YUAN *Spatial Data Mining in the Context of Big Data* 2013 IEEE International Conference on Parallel and Distributed System.

[8] J. Gerard Wolff, *Big Data and the SP Theory of Intelligence* IEEE Volume 2, 2014.

[9] Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan, And Yong Ren *Information Security in Big Data: Privacy and Data Mining* IEEE Volume 2, 2014.

[10] Rajiv Ranjan *Streaming Big Data Processing in Data center Clouds* IEEE Cloud Computing 2014.

[11] Abdur Rahim Mohammad Forkan, Ibrahim Khalil, Ayman Ibaida, and Zahir Tari *BDCaM: Big Data for Context-aware Monitoring - A Personalized Knowledge Discovery Framework for Assisted Healthcare* IEEE Transactions On Cloud Computing , Vol. X, No. X, February 2015.

[12] SHI Wenhua, ZHANG Xiaohang, GONG Xue, LV Tingjie *Identifying Fake and Potential Corporate Members in Telecommunications Operators* China Communications August 2013

[13] Mohammad Reza Dehghani Zadeh, Mohammad Fathian, Mohammad Reza Gholamian *A New Method for Clustering Based on Development of Imperialist Competitive Algorithm* China Communications December 2014.

[14] FANG Cheng, LIU Jun, LEI Zhenming *Parallelized User Clicks Recognition from Massive HTTP Data Based on Dependency Graph Model* China Communications December 2014

[15] Joseph M. ,Hellerstein,Ron Avnur,Andy Chou,Christian,Hidber,Chris Olston,Vijayshankar Raman, TaliRoth *Interactive Data Analysis: The Control Project* IEEE August 1999.

[16] Katsuya Suto, Hiroki Nishiyama, Nei Kato, Kimihiro Mizutani, Osamu Akashi, And Atsushi Takahara An *Overlay-Based Data Mining Architecture Tolerant to Physical Network Disruptions* IEEE VOLUME 2, No. 3, September 2014.

[17] Jemal H. Abawajy, Andrei Kelarev, and Morshed Chowdhury *Large Iterative Multitier Ensemble Classifiers for Security of Big Data* IEEE Volume 2, NO. 3, September 2014.
[18] Scott W. Cunningham, Wil A. H. Thissen *Three Business and Societal Cases for Big Data: Which of the Three Is True?* IEEE Vol. 42, No. 3, Third Quarter, September 2014.
[19] Xiaochun Cao, Hua Zhang, Xiaojie Guo, Si Liu, and Dan Meng, *SLED: Semantic Label Embedding Dictionary Representation for Multilabel Image Annotation* IEEE Transactions On Image Processing, Vol. 24, No. 9, September 2015.
[20] Shifeng Fang, Li Da Xu, Yunqiang Zhu, Jiaerheng Ahati, Huan Pei, Jianwu Yan, and Zhihui *Liu An Integrated System for Regional Environmental Monitoring and Management Based on Internet of Things* IEEE Transactions On Industrial Informatics, Vol. 10, No. 2, May 2014.

