International Journal of Latest Trends in Engineering and Technology
Special Issue IC3NS 2018, pp. 005-008
e-ISSN:2278-621X

CLUSTERING OF UNSTRUCTURED DATA IN BIG DATA WITH HADOOP

Navneet Kaur[1], Niranjan Lal[2]

Abstract—We live in the data and information area. The amount of unstructured data is used linearly increases over time. The heterogeneity of data on Social networking sites are increasing day by day on Facebook, Twitter, and analysis and discovered the growth of data that will be unmanageable and uncontrollable in coming future for organizations due to increased heterogeneity of data. Currently, most of the information stored in the organization is unstructured models. More than 80% of all potentially useful business data is unstructured data like sensor readings, console records, etc. The Internet is overflow with unstructured content, almost 95% of existing data is unstructured, much of which is in the form of text, streamed data, videos, and images. There are three varieties of data known as structured formatted data, unformatted unstructured and partial formatted semi-structured data, unstructured and semi-structured are a new area for researchers in Big Data, as we know, the unstructured data is increasing double in two years. Clustering is an important technique used to group numerical, text and image datasets. Clustering makes image retrieval work easy by finding similar images. The images are grouped in a certain number of clusters. The image data sets are clustered according to some characteristics like color, shape, etc. This paper describes clustering analysis of images using K-means clustering algorithm in Big data.
Keywords—Big Data mining, Unstructured data, clustering, Image Clustering, k-mean clustering.

1. INTRODUCTION

Big data is a process of evaluating a big volume of data and information to reveal hidden forms and interrelations between dissimilar nodes. This information is developed by online transactions, videos, emails, images, records, social network [1] provides several benefits, such as more targeted marketing, direct business information, market sources, and sales recognition. In the current world of Big Data, most of the data is not structured and it is estimated to represent more than 95% of all data generated. In contrast, unstructured data refers to data that does not fit perfectly with the traditional column and row structure of relational databases. Examples of unstructured data include e-mail, video, audio files, web pages, images and messages on social media.

Apache Hadoop is right the tool for analyzing the unstructured data. Hadoop is an open source framework written in Java to process large structured and unstructured data sets. Hadoop basically has two main components distributed system of Hadoop (HDFS) and Map-Reduce [5]. The complete technology stack contains common utilities, a distributed file system, storage and data analysis platforms, and an application layer that handles distributed processing, parallel computing, workflow, and management. Hadoop is more efficient to handle large unstructured data sets than conventional approaches provides massive scalability and speed.

The SparkR project was initially launched in AMP Lab as an effort to discover dissimilar methods to integrate the usability of R with Spark's scalability. It is an R package that provides alightweight interface to use Apache Spark R. SparkR provides an implementation of Data Frame that supports distributed operations such as selection, aggregation, filtering, etc. SparkR Data Frames presents an API similar to the local R data frames but can climb on large datasets using Spark distributed processing support. SparkR can read data from various sources, including JSON files, Hive tables, etc.

The most important challenge for Big Data applications is to explore huge volumes of unstructured data and extract useful knowledge. In many situations, the process of extracting knowledge must be very efficient and near real-time because storing all the observed data is almost impossible. The Big Data system, which integrates software and hardware components, is hardly available without the support of the main industrial shareholders. In fact, for decades, corporations made business decisions based on transactional data stored in relational databases. Big Data mining offers the opportunity to go beyond their relational databases to have less structured data: weblogs, social sites, email, and pictures that can be extracted to obtain useful data.

[1] Mody University of Science and Technology,Lakshmangarh,Sikar,Rajasthan, India
[2] Mody University of Science and Technology,Lakshmangarh,Sikar,Rajasthan, India

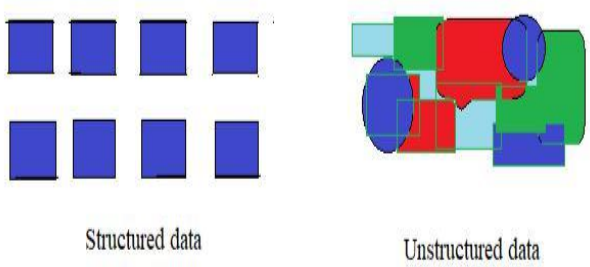

Fig.1 Structured data v/s Unstructured data

Clustering is an essential tool for data mining and the analysis of large volumes of data [3]. There are problems in applying clustering techniques duo of big data to new challenges with big data. The clustering of a series of images into meaningful classes has various applications in the group of image databases and in the design of their human interface. Clustering of images can also be used to segment a movie, and to facilitate image extraction [6].

In this paper, we introduce the most popular Big Data's clustering techniques and clustering of an image set into groups of similar images.

The paper is organized as follows: the second section represents the related work which shows a different state of the papers. The third section describes the challenges of Big Data. The fourth section provides a global view of the various clustering techniques dealing with images.

2. RELATED WORK

Zomaya et al. [1] present a survey of current clustering algorithms of different categories, such as partition and hierarchy based. In their work they established a comparison of five classes with their most representative algorithm, their goal was to find the greatest performance for huge data.

In [2] the author attention is on the best and most used algorithms in the literature like k-means, they present some comparative work of clustering algorithms.

Sherin et al. [3] present an overview of big data mining algorithms and platforms that can be used in the Big Data field discussing different characteristics and challenges. Dr. Meenu Dave et al. [6] discussed different applications and the significance of clustering approaches. To inspect the giant volume of data, clustering algorithms help provide a powerful tool. Numerous grouping techniques in reference to huge data sets with their cons and pros are discussed in thepaper.

Researchers in [7] present a review of some algorithms that can handle large image data sets such as searching for the nearest neighbors, in this paper, then describe a scalable type of a search algorithm for the nearest neighbor and analyze how it can be used to find duplicates in more than a billion images.

In Herawan et al.[8] discuss various clustering techniques including MapReduce, parallel classification using MapReduce. They present an overview of different data mining clustering approaches.

3. CHALLENGES OF ANALYZING BIG DATA

Large volumes of data refer to data groups that are too bulky to be managed using management tools existing databases. Big Data presents a major challenge for the database and analysis of research data. To take full advantage of the data it will be necessary to address several challenges related to Big Data. We are listing the main challenges facing Big Data Analysis.

Dealing with data storage- The most recognizable challenge of Big Data is simply analyze and store all the information. Companies are developing at a very fast speed. In addition, the evolution of companies rapidly growths a large amount of data created. Storing this data is becoming a task for all. We are turning to a number of diverse technologies to deal with big data. With respect to storage, converged and defined software storage can help companies scale their hardware [9].

Heterogeneous data- Much knowledge is needed to obtain data in the right shape for the data analysis. Like if data comes from different social network sites, you necessary to know who the operator is in a general sense, for example, a client using a specific set of objects and recognize what they are trying to show the data. Make sure people understand that analyze the data in depth where the data came from, what type of data is and how to interpret that information.

Speed- Satisfy the need for speed. Companies must not only analyze relevant data they need, but they have to catch them speedily. The challenge is to go through the large volumes of data and access the required level at high speed. The Hardware is a possible solution. Some providers use more memory and massively parallel processing to process big data very quickly. One more technique is to put the data in memory and using a grid computing method, where several machines are used to crack difficulties [11]. Both methods allow companies to explore huge volumes of data and get sales information in almost real time. Security of data- Keep that large data safe is another big challenge for the big data. Security is also a major concern for organizations with large data warehouses. After all, some large data file may be attractive targets for hackers. Most of the data gathered by businesses for strategic purposes is personal data openly from user accounts. The use of this data is then directly related to the connection of faith between the organizations and its clients. Thus, these data security is a crucial issue

for the future of Big Data. Meaningful results-Graphical analysis becomes problematic when handling giant amounts of knowledge. One way to solve this problem is to group the data into a higher-level view in which the smaller datasets become visible. By gathering data or groups, you can view data more efficiently [10].
To handle growing data requests, we need to increase the volume and performance of methods and tools. Big Data needs different solutions to improve the capacity and efficient treatment to exploit data functionally without necessarily taking on new resources.
In fact, with the growth of data, existing data mining algorithms have not been able to meet the significant data processing needs. Therefore, to exploit this large volume of unstructured data, an effective computing model with a reasonable computational cost of this heterogeneous data is needed.

4. CLUSTERING OF IMAGE DATA

In current days there has been a growing interest in developing effective methods for the recovery of images based on images. The categorization of clusters and images is a means for the high-level description of the content of the image. The goal is to map clustered file images so that the group of classes provides essentially the similar knowledge about the picture file of the entire collection of the picture. The created classes provide a visualization and summary of the content of the image that can be applied for dissimilar tasks related to the management of the image database [7]. The clustering of images allows the execution of efficient recovery algorithms and design of a user-friendly interface for the database. A common approach to grouping images includesthe following problems:
Image features – Represent the image.
Organization of data – Organize the feature data.
Classifier – Classify an image to a cluster.

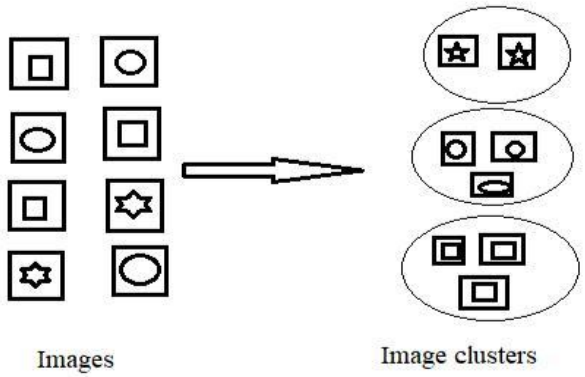

Fig.2 A diagram of the image clustering

In image clustering, there is three way to match between images and cluster centers:
in the low-level representation, the correspondence between the cluster centers and the images is established on a measure of the distance between matching pixels. The computational work is least in phase representation, with the computational cost of the matching process. 2) Anotheroption is to move to a representation of the high-level, in which every picture is labeled as belonging to a class such as nature and animal. A substantial computational power is required in the representation phase, such as the use oftechniques as supervised learning to images classification, and this allows a very simple pairing phase to group images of similar content by the category tags [8].
There is a representation of a middle level that balances the two previous ones, in it, a transition is made from the pixels to the characteristics. The feature vectors are used to define the image content and clusters in a compact way. The correspondence phase produces the correspondence of the functionalities. Distance metrics and Similarity measures are necessary to match the clusters in the spaces of the chosen characteristics. Most work on cluster images uses mid-level representation, including the histogram methods most frequently used.

5. CLUSTERING ALGORITHMS

The k-means is a simple and easy algorithm and uses a squared error criterion. It begins with a random initial partition and continues to reassign models to cluster centers based on the matching between the design and the cluster centers until a convergence criterion is met. In k-means clustering approaches, a measure of distance between data points and between cluster center and a data point is given a priori as aportion of the problematic configuration. The grouping consists of finding limited classes with lowest intra-class variability. However, in several grouping difficulties e.g. clustering of images, the items that we want to categorize have a complex three-dimensional structure and the choice of the correct distance measure is not a simple task. The choice of a specific distance measure can influence the results of the grouping. The user must also specify the number of clusters. The Linde Buzo Gray (LBG) clustering algorithm is thegrowth of the k-means algorithm to

overwhelm the problem of selecting initial clusters. When grouping images, the user cannot predict how many clusters are needed [12].

Clustering will be more beneficial to reduce the time it takes to find images in the database. Fuzzy C-means is one of the clustering approaches that allows part of the data to belong to more than one clusters. In this grouping, each point has a degree of cluster membership, as in fuzzy logic, instead of completely belonging to a single cluster. Therefore, the points on the edge of a cluster may be in the cluster to a less extent than the points in the middle of the cluster.

6. IMAGE SEGMENTATION

The segmentation of images has attracted significant attention in recent days, because of advances in multidimensional image clustering. The segmentation of images is used to define few characteristics and features of the various images. The segmentation carried out in image segments the images, extracts certain of their important characteristics and combines these characteristics with the image pixels. Efforts have been creating to segment a whole volume using fuzzy clustering. The similar creates a cluster and the related cluster is dissimilar from another cluster.

7. CONCLUSION

This paper describes various algorithms used to manage large image data sets. Show that these algorithms are not enough to address all the challenges posed by Big Data. In fact, there is no clustering algorithm that can be used to answer all huge amount of unstructured data difficulties. Although the parallel classification is potentially very useful for the grouping of Big Data, the complexity of implementing these approaches remains an important challenge. The problem of clustering images is also debated from grouping approaches based on global learning since K-mean and Fuzzy-C mean are recent challenges in this area. Generally, to manage large amounts of data while maintaining the needs of an acceptable resource, we need to improve the clustering algorithms by reducing their time and memory complexity.

8. REFERENCES

[1] Fahad, N. Alshatri, Z. Tari, A. ALAmri, A. Y. Zomaya, I. Khalil, F. Sebti, and A.

[2] Bouras, “A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis,” IEEE transactions on emerging topics in computing, (2014).

[3] A BEN AYED, M.BEN HALIMA and M. ALIMI, “Survey on clustering methods: Towards fuzzy clustering for Big Data,” In Soft Computing and Pattern

[4] Recognition (SoCPaR), 6th International Conference of. IEEE, p. 331-336, (2014).

[5] A Sherin, S. Uma, K.Saranya and M. Saranya Vani ”Survey On Big Data Mining Platforms, Algorithms And Challenges”. International Journal of Computer Science & Engineering Technology,Vol. 5 No, (2014)

[6] S.ARORA, I.CHANA, “A survey of clustering techniques for Big Data analysis,” in Confluence The Next Generation Information Technology Summit (Confluence), 5th International Conference-. IEEE, p. 59-65, (2014).

[7] Swati Sharma,and Manoj Sethi, “Implementing collaborative filtering on large-scale dataset using Hadoop and mahout”, International Research Journal of research and technology(IRJET),04 July 2015

[8] Dr. Meenu Dave and RemantGianey, “Different Clustering Algorithms for Big Data Analytics: A Review”, 5th International Conference on System Modeling & Advancement in Research Trends,IEEE, November 2016.

[9] Ting Liu, Charles Rosenberg, Henry Rowley. “Clustering Billions of Images with Large-Scale Nearest Neighbor Search”, (2007).

[10] Zhang, Bingjing, Judy Qiu, Stefan Lee, and David Crandall, "LargeScale Image Classification using High-Performance Clustering."

[11] A S. Shirkhorshidi, S. Aghabozorgi, T. Y. Wah, and T. Herawan, “Big Data Clustering: A Review,” In Computational Science and Its Applications–ICCSA 2014. Springer International Publishing, p. 707-720. (2014).

[12] Stephen Kaisler, Frank Armour, J. Alberto Espinosa, William Money, “Big data: Issues and Challenges Moving Forward”, 46th Hawaii International Conference on System Science (2013).

[13] Dr. Siddaraju1, Sowmya C L2, Rashmi K3, Rahul M4, “Efficient Analysis of Big data Using Map-Reduce Framework”, International Journal of Recent Development in Engineering and Technology, Volume 2, Issue 6, June 2014.

[14] AKannan1 Dr.V.Mohan2 Dr.N.Anbazhagan3, “Image Retrieval Based on Clustering and NonClustering Techniques using Image Mining”-Int J Engg Techsci Vol 1(1),54-61, (2010).

[15] H. Lin, S. Yang, and S. Midkiff. RABID: A General Distributed R Processing Framework Targeting Large Data-Set Problems. In IEEE Big Data 2013, pages 423–424, June 2013.

[16] L. Yejas, D. Oscar, W. Zhuang, and A. Pannu. Big R: Large-Scale Analytics on Hadoop Using R. In IEEE Big Data 2014, pages 570–577.