International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307

OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT24104105

Clustering Social Networking Data With K-Means Algorithm Using R Language

Sujeet Kumar Sahani[1], Dr. Sonam Singh[2]

[1]Research Scholar, SHEAT Group of Institutions, Babatpur, Varanasi, Uttar Pradesh, India

[1]SHEAT Group of Institutions, Babatpur, Varanasi, Uttar Pradesh, India

ARTICLEINFO

Article History:

Accepted : 20 June 2024

Published: 01 July 2024

Publication Issue

Volume 10, Issue 4

July-August-2024

Page Number

23-30

ABSTRACT

The main objectives of this research work are to report detailed empirical studies on sequential and parallel algorithms for diverse clustering tasks executed on very large social network datasets using memory efficient out-of-core approaches. We evaluate the spark implementation for R on Cloudera using the data from social media review datasets like k-means and hierarchical clustering to rank these algorithms. This implementation leverages the YouTube dataset from UCI Machine Learning Repository. Our goal is to compare a few algorithms, so we can know exactly how accurately these models are performing. Ultimately we want to deal with testing and ranking clustering method, and mining and finally clustering massive amounts of unstructured data.

Keywords : Social Networking, Clustering, R Programming, Cloudera, K-Means, Big Data

I. INTRODUCTION

When you have a large datasets that can not be managed by usual computer manner, it is called "big data". This consists of an extensive mixture of information made with the aid emails, logs, postings, pics, songs movies might posts searches social networking programs and on-line transactions Over the years, big data has evolved into an entire host of tools and techniques; it is not just about massive datasets anymore.

When dealing with large volumes of data, "big data," it requires new thoughts and models to derive meaningful information from this massive piece of datasets. The existing database management systems are not equipped to maintain these huge datasets and the reason could be attributed as their increasing volume. The data is sharing, storing, searching analyzing along with capturing the full structure of big-data. The big data attributes such as volume, velocity, diversity and value with complexity impose equally difficult challenges to handle.

The practice of "big data mining" refers to the set of techniques used for searching large volume information or patterns among an enormous database. It is often a mix of structured and unstructured data in this methodology.

Structured data in full form: This is something that comes with all type related to a known format such as relational database system.

For this unstructured kind of data from sources such as Facebook, twitter or YouTube whom we can tell them modern information gathering paradigms.

Semi-structured data - It is schema-less or self-describing dataset, no need to have predefined schema. Similar to text or numeric datasets, you can often achieve improved results grouping similar things together. Clustering implementations will typically represent items in the group as vectors within an n-dimensional space; similar distances between these vectors result clusters generally indicating that they are close together.

One of the questions addressed in this article is how to save hundreds megs TBs PBs of unstructured data on local storage server / repo?

Some common questions are as follows:

1. Big Data - How to Map and Analyze Unstructured Data?
2. How can we analyze social media like Twitter or live data? We all know Hadoop big data analysis tools but how they help in this case?
3. What can we do with big data to get some knowledge
4. How Clustering Works With Big Data?
5. Which way is an ideal structure for unstructured data?

II. LITERATURE REVIEW

It is simpler, for example to handle data if the scale of data was smaller and as S. Vikram Phaneendra et al. [1]. people have resorted to RDBMS in past years. Recently, RDBMS techniques have not been able to handle large volumes of data (Big Data). They explained that volume, velocity, variety and complexity are the four main characteristics of big data apart from value. Current big data is, however referred Dnvsl Indira and Kiran Kumara Reddi [2] for structured semi-structured and unstructured, Homogeneous(data type same) Heterogeneous (hybird/hetrogeneous structure). They proposed a model that uses free, or low-demand times-where there is lots of bandwidth available-to shift around huge volumes of data on networks.

Big data mining is the task of extracting useful information from the large scale collections of datasets or data streams, which could not previously be done because of its volume, variability and velocity according to Albert Bifet Wei Fan [3]. As discussed by Albert Bifet et al[5]. Cascading analysis is becoming the preferred way of gaining grounding-based insights that help businesses respond to problems or improve their performance quickly. Bernice Purcell et al. The immense datasets of big data contain unstructured, semi-structured, and structured information , making them difficult to process using standard systems. Additional information on object-based storage and clustered network-attached storage as methods for storing lots of data was also shared.

Sameer Agarwal [6] and some co-authors introduced Blink DB, an approximation query engine for concurrent interactive SQL queries on big datasets.. Blink DB is based on two key principles: 1) a selective sampling strategy that computes an appropriate sample size according to the desired accuracy of queries or query response time, and 2) a dynamic optimization framework which continually creates and maintains multi-dimensional samples from original data.

N. Lal et al. [11, 17, 18], "A system for heterogeneous data: Digital audio, video and photo management from industries like healthcare and retail," in. They say there are about 30 billion web pages on the World Wide Web (WWW). As same as the programs above, HaLoop is a modified version of Hadoop MapReduce and it supports an iterative program from Yinji Bu et al. [9].

HaLoop provides loop-aware scheduling, loop-invariant data caching and powerful fixed-point verification algorithms.

Osama Abu Abbas [10] compared some clustering algorithms such as the K-means algorithm, hierarchical clustering algorithm, expectation-maximization (EM)algorithm and self-organizing map. Conclusion was with the performance, quality and accuracy of these algorithms after going through evaluation on dataset resize, type or number of clusters created in this case based programming.

A part of the task obligations was tackled by Bao Rong Chang et al., they debated that utilizing a couple platforms to combining there computational power and accomplish better performance results is quite challenge just as desirable (Chang et al. [14]. They introduced new ways to both big data platforms R Hadoop and SparkR so as to build powerful big data analytics using the programming language called R as part of business intelligence (BI) for fast retrieval and analysis of huge volumes of structured or semi-structured datasets.

Simon Mulwa Kiio et al. [22]. created a forensic tool based on Apache Spark. To collect and analyze social media data for hate speech & cyber-bullying on twitter. We used a Naïve Bayes model with the Spark ML API to automatically split hate speech and cyber-bullying.

The most critical problem with K-means clustering was also solved by Agnivesh et al. [16] it often converges to the local - not global optima, which can lead to suboptimal results The performance of the algorithm depends largely on initial centers which a particular concern for large data.

III. Clustering of Unstructured Data

Generalized adoption of new technology gave the birth to multiple data, or big data. As people are connecting more online and through social groups, info is spreading faster than it has previously. Some experts argue that effective management and exploitation of this large data is fast becoming a scientific problem itself - one whose solution will be essential to further progress in science (and perhaps also the wider economy).

A. Clustering

The widespread adoption of new technology has led to an explosion in diverse data, or "big data." Information is growing faster than ever before because people are meeting more online and through social networks. If only we could control and make good use of all that data, the masses claim (by which they probably mean certain experts), it will be critical to advancing science - not to mention "the economy" too in some expert-speak quarters.

There are two main types of clustering: supervised and unsupervised clustering

Supervised Clustering: This method will utilize a few sample data for us to describe the rest of the data points. It attempts to learn how instances can be assigned into one of a number of predefined classes; its is known as Classification. Data elements are at the same cluster level comparable to one another and share certain properties.

Unsupervised clustering: So Process, the first step in machine learning is similar as well - it groups some examples together to understand an entity (Data Set) of a Machine Learning System. The process of grouping unlabeled examples is called clustering. And since these are unlabeled examples this is an example of unsupervised machine learning used for clustering.

K-means is options the better choice but on smaller data sets because it iterates over all of the points. This in turn will end taking more time on classifying data points if there is high amount of such amongst the dataset.

Clustering algorithms are varied as their characteristics:

Hierarchical Clustering: This approach does not split the input data to form a number of clusters directly. Instead, it a series of more and larger combining Of the Data until number is reached.

Non-hierarchical/Iterative Clustering: Here, you have to define a number of clusters where the cluster formation starts. Afterwards, we come back to our groups and reassign data items in order to improve the quality of grouping.

Hard and soft clustering: Is the type of association direct or probabilistic

Disjunctive clustering: An item may be in many clusters

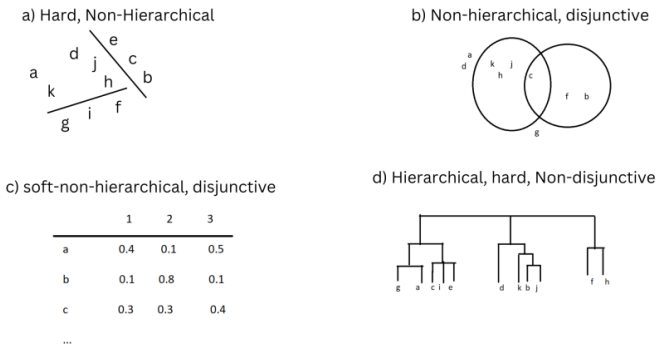

Figure 1. An Example of Grouping Properties

Clustering (in above Figure) is the method of organizing data into groups with similar things. These clusters are defined by two distinct characteristics:

1. High Intra-cluster similarity: Shows a high degree of the same data items belonging to one group.

2. Low Inter-cluster Similarity: Alternatively, data pieces from different of clusters are dissimilar.

Clustering belongs to unsupervised learning, and is divided into the following two major categories:

- Single Data Mining Clustering
- Multi Data Mining Clustering

The single data mining clustering type analyses one dataset to discover patterns and relationships within that same group of datasets. On the other hand, few data mining clustering involves combining and analyzing multiple datasets concurrently to give an idea how different sources are associated which is actually known as multi-data Mining Clustering .

This clustering approach helps in the categorization of complex datasets to enable better analysis and decision making across a diverse set of domains such research, corporate intelligence or scientific discovery.

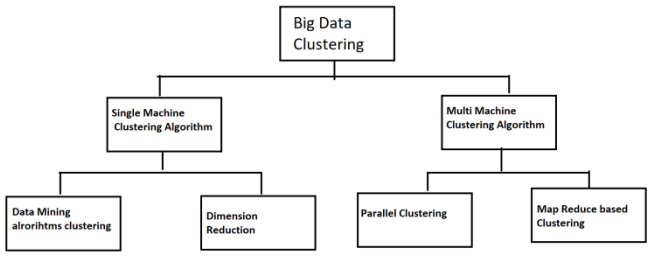

Figure 2. Clustering Categorization

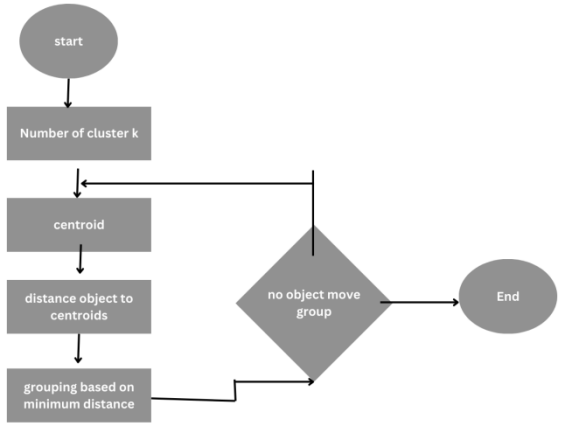

Figure 3. Design Model for the clustering Algorithm

R Language

Environment R is an environment and programming language, created for statistics assist in graphics. R is a free implementation of the S programming language, which was developed by Bell Laboratories (part of AT&T) adapted to the World Wide Web. It has a very flexibility for data analysis and visualization applications comparison to other languages because so many statistical techniques, graphing tools libraries are provided within this. The language (basis of the programming environment R) is widely adopted for creating and evaluating new methods in statistical research. In addition, the open source R makes it possible to cooperate with numerous people in the

statistical industry so that more creative and advanced statistics becomes feasible.

Spark

Spark was essentially created in 2009 as a research project by the AMP Lab at University of California, Berkeley. And since then, it has grown to be one of Apache Software Foundation's top-level project. The widely supported big data processing framework Spark, was indeed designed around speed, usability and advanced analytics. This is implemented as an open source project.

Apache Spark outshines related large data and map reduce platforms like Hadoop, Storm etc. on multiple accounts. It is suitable for many applications and environments because it supports multiple programming languages such as Python, R, Scale...etc. While Spark is close connected with Hadoop's core storage architecture, it supplies additional functionality and reduces the time complexity in data processing over large clusters.

With its fast in-memory processing and efficient handling of iterative algorithms, Spark is a particularly good fit for Toptal's real-time data analytics, machine learning and interactive searches over large datasets. With its ability to scale and flexibility, it is ranked high in terms of the list for best solutions for businesses with big data age.

SparkR

Research out of the AMP Lab also begat SparkR - a new R package designed to let users easily write and execute code in R that utilizes Apache Spark, (combining supposedly the best aspects of both estimates about Python with visualization interfaces for debugging purposes) The upshot here is that an implementation facilitates a subset of Data-Frame functionality including filtering, aggregation and selection. Like a local R data frame, these SparkR Data-Frames have an API but are built on top of large scale distributed computing with Apache Spark to handle very large datasets.

SparkR reads data from JSON files and Hive tables well. In a Spark cluster, all data frame operations in SparkR could be done distributed automatically on both computers and cores. With its scalability, SparkR can crunch Terabytes of data and fit well on clusters with thousand servers running in tandem. For data scientists and analysts working with big data settings, SparkR is an indispensable tool as it offers a simplified route to perform complex R-based data analyses over large-scale datasets.

IV. PROPOSED APPROACH

Extracting insight from written language text such as emails, Facebook status updates, Twitter tweets and YouTube comments is called as Text Mining in big data. Text analytics is the application of text mining techniques to business problems. Text mining is especially challenging because it processes unstructured data from so many different sources.

Cluster Before going ahead lets understand some basics about clustering in text mining-Clustering, Clustering is a process of grouping things that are similar to each other based on how many objects resemble both compared to those not. The IBM Knowledge center ,in above Figure 6. There are a number of data mining approaches to mine and make the information discoveries with massive, multifaceted datasets in relevant ways taking into account unique characteristics about big data.

Paying special heed to this approach, organizations and academicians become enabled with the power of uncovering crucial patterns, trends as well as connections embedded within colossal text data records critical for shaping their strategies or informed decision making.

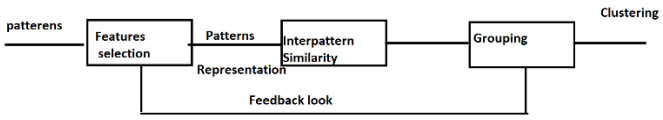

Figure 4. Overview of Clustering

In the last post we learned about two clustering algorithms that have been added and how to used them, from now on let us learn some real world research methodologies in which these algorithm are already provided and evaluate with a specific YouTube social media dataset.

The purpose this part of the hands-on research is to implement use these clustering algorithms and determine how well they can be used in practice, toward understanding social media data. In this study, we intend to analyze how such algorithms behave on instructive YouTube datasets combining lists from roughly 5000 videos gathered and already preprocessed.

This study aims to assess the algorithms capability of clustering similar but loosely connected users or content for enhanced processing and understanding social media data. This empirical tool is particularly useful for large-scale data with diverse forms, which are commonly seen on social platforms and provides an evidence-based validation of the versatility & robustness of algorithms in real-world applications.

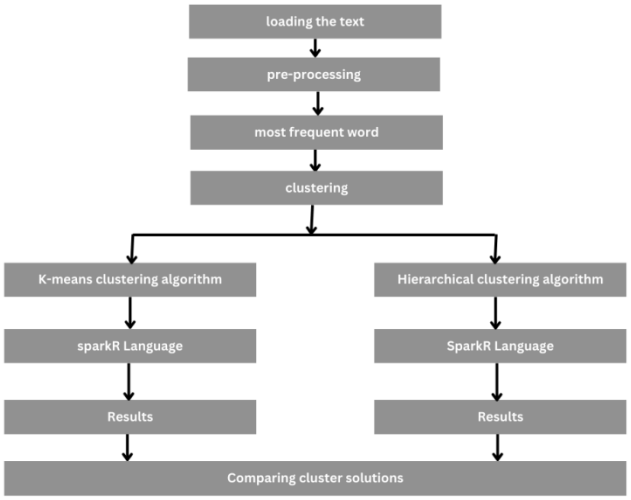

Figure 5. Flow Chart of proposed model

Paper Suggests Method for Comparing Two Clustering Algorithms

In above Figure, Contrasting the partitioned based clustering and hierarchical clustering methods in contrasting our suggested method.

Our approach is to retrieve unstructured information from a given online storage source or local directory. After we load the data; this is where our technique comes into place. Cleaning and organizing data are the activities to be done in this preprocessing step, before the analysis.

Next, the algorithm searches for words that occur commonly in unstructured data. To keep the clustering process focus on meaningful and distinctive features of data, we then remove these most frequent words

Next We use K-means clustering and Hierarchical clustering which are famous algorithms to perform the clusters. These methods are executed on Big Data infrastructure with the support of Hadoop over Cloudera distribution and through a Windows operating system using SparkR programming.

We test our approach in the context of a YouTube data set which we downloaded from UCI Machine Repository as well. We use this dataset as the benchmark for evaluating and comparing how effective clustering algorithms are in dealing with large scale of unstructured social media data, which is not explored by previous work.

To offer some valuable insights about the strengths and slackness of each clustering algorithm in dealing with complex real-world data scenarios, this effort applies (as you might say) evaluates the extent to which it arranges or organizes itself a better way categorizing rather simply speaking for considering YouTube Data set median.

V. CONCLUSION

The outcomes of the experiment reveal that big data has devastated every industry to such a degree that retrieving one piece of useful information from large databases seems an insurmountable goal. K-means and hierarchical clustering helps to group similar data points together which will eventually be beneficial for the purpose of a effective Data Analysis.

In the subsequent, these clustering algorithms could be considered more precisely suitable for the falsely clustered and various data varieties as such heterogeneous datasets. Advanced algorithms are to be designed to inject the heterogeneous, semi-structured, structured and unstructured data into big data area.

The field of clustering algorithms primarily still demands the development of more efficient ones, so that analyses can be automated for different types forms at once and reliable in practice. This will make it more adaptable to the big data analytic skills and develop meaningful knowledge which could spontaneously enhance decision-making across layers.

VI. REFERENCES

[1]. S.Vikram Phaneendra & E.Madhusudhan Reddy "Big Data- solutions for RDBMS problems- A survey" In 12th IEEE/IFIP Network Operations & Management Symposium (NOMS 2010) (Osaka, Japan, Apr 19, 2013).

[2]. Kiran kumara Reddi & Dnvsl Indira "Different Technique to Transfer Big Data: survey" IEEE Transactions on 52(8) (Aug.2013).

[3]. Vaithiyanathan, V., Rajeswari, K., Tajane, K., & Pitale, R.,"comparison of different classification techniques", International Journal of Advances in Engineering & Technology, May 2013. ISSN: 2231-1963, 6(2), 764–768.

[4]. Albert Bifet "Mining Big Data In Real Time", Informatics, 37 (2013) 15–20 DEC 2012.

[5]. Bernice Purcell "The emergence of "big data" technology and analytics" Journal of Technology Research 2013.

[6]. Sameer Agarwal, Barzan MozafariX, Aurojit Panda, Henry Milner, Samuel MaddenX, Ion Stoica "BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data", ACM , 978-1-4503-1994 2/13/04.

[7]. K. A. Abdul Nazeer & M. P. Sebastian" Improving the Accuracy and Efficiency of the K Means Clustering Algorithm" .Proceedings of the World Congress on Engineering, 2009 Vol I WCE 2009, London, U.K, July 1 – 3.

[8]. Niranjan Lal & Bhagyashree Pathak "mining of unstructured data with clustering approach", International journal of engineering research & Management Technology, 2016.

[9]. Yingyi Bu _ Bill Howe _ Magdalena Balazinska _ Michael D. Ernst "The HaLoop Approach to Large-Scale Iterative Data Analysis", VLDB, 2010 paper.

[10]. Osama Abu Abbas "Comparison between Data Clustering Algorithms", the International Arab Journal of Information Technology, Volume 5, July 2008.

[11]. Jeffrey Dean and Sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", OSDI, 2010.

[12]. D. Napoleon & P. Ganga lakshmi, "An Efficient K-Means Clustering Algorithm for Reducing Time Complexity using Uniform Distribution Data Points", IEEE, 2010.

[13]. Monika kalra, Niranjan lal, & samimul Qamar, (2017)"K-mean Clustering algorithm for data Mining of Heterogeneous Data", International and Communication Technology for Sustainable Development , pp61-70.

[14]. Bao Rong Chang, Yun-Da Lee, and Po-Hao Liao "Development of Multiple Big Data Analytics Platforms with Rapid Response" Scientific Programming Volume 2017, Article ID 6972461, https://doi.org/10.1155/2017/6972461

[15]. Simon Mulwa Kiio, Elisha O. Abade " Apache Spark based Big Data Analytics for Social Network Cybercrime Forensics " International Journal of Computer Applications (0975 – 8887) Volume 179 – No.8, December 2017.

[16]. Agnivesh, Rajiv Pandey, Amarjeet Singh" Enhancing K-means for Multidimensional Big Data Clustering using R on Cloud " International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue-7, May 2019.

[17]. Kaur N., Lal N. (2018) Clustering of Social Networking Data Using SparkR in Big Data. In: Singh M., Gupta P., Tyagi V., Flusser J., Ören T. (eds) Advances in Computing and Data Sciences. ICACDS 2018. Communications in Computer and Information Science, vol 906. Springer, Singapore. https://doi.org/10.1007/978-981-13-1813-9_22

[18]. N. Lal, M. Singh, S. Pandey and A. Solanki, "A Proposed Ranked Clustering Approach for Unstructured Data from Dataspace using VSM," 2020 20th International Conference on Computational Science and Its Applications (ICCSA), Cagliari, Italy, 2020, pp. 80-86, doi: 10.1109/ICCSA50381.2020.00024.