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/IJSRCSEIT

Enhancement of Performance of Clustering Technique during Data Mining To Investigating Sentiment Analysis Using KDD Process

Dr Kapil Kumar Kaswan[1], Monika[2]

[1]Assistant Professor, Department of CSE, CDLU, Sirsa, Haryana, India

[2]M.Tech. Scholar, Department of CSE, CDLU, Sirsa, Haryana, India

ARTICLEINFO

Article History:

Accepted : 02 April 2025
Published: 04 April 2025

Publication Issue
Volume 11, Issue 2
March-April-2025

Page Number
3027-3036

ABSTRACT

Data mining is the act of searching through big data sets to find patterns and correlations that, when analyzed, might assist solve issues faced by businesses. The methodologies and tools of data mining provide businesses with the ability to forecast future trends and make better educated business choices. Finding unique groupings, or "clusters," within a data collection is the objective of the clustering technique. Using an algorithm written in machine language, the tool produces groups in which the individual objects in each group will, in most cases, share characteristics with the other members of the group. The major challenge to big data processing is management of unmanaged data. Map reduce function is used to get the frequency of unmanaged data and makes it manageable. Moreover soft computing mechanism might be used to improve the performance of clustering operations. Present research is focused on enhancement of performance of clustering techniques that are used in data mining. In order to gauge public opinion, researchers are analyzing tweets and user comments using sentiment analysis. As a result of technological development, the globe is altering at a breakneck pace. Having Internet connectivity is crucial in today's society. People are increasingly using social network applications to voice their opinions on current events. When trying to sell a product or improve a government service, collecting and analyzing customer feedback is essential. Data mining, also known as sentiment analysis, is often done in advance of a discussion in which the attitudes behind different points of view are to be found. The use of sentiment analysis to gauge consumer sentiment has exploded in recent years. The tweets of Twitter users are analyzed using neural networks in the latest study. There has been a rise in the use of Twitter data for survey research in recent years, and researchers are more interested in "tweets" (comments) and the content of these expressions. Accordingly, this research aims to evaluate the efficacy of several approaches to sentiment analysis applied to Twitter data. Sentiment analysis academics have been looking at how people feel about a wide range of things, like movies, commercial goods, and everyday social problems. Twitter is a very popular micro blog where customers can talk about what they think. Opinion research using Twitter data has been getting a lot of attention in the last decade. Because there is a lot of interest in sentiment analysis, the proposed work used an RNN model to predict sentiment based on text and graphic sentiments. These thoughts have been taken into account from Elon Musk's tweets. Research is using filters that people can set up to classify and remove useless content before training. The user-defined classification and filtering system has cut down on the amount of time it takes to learn. The accuracy of predictions has gone up because useless things have been removed. The proposed work used RNNs to come up with a more reliable and smart way to do things. This work has been flexible, scalable, and efficient when it comes to twitter sentiment analysis.

Keyword: Big Data, Data mining, clustering, Map Reduce, Performance

Introduction

1.1 Big data:

Big data is a term for massive data sets having large, more varied and complex structure with the difficulties of storing, analyzing and visualizing for further processes or results. The process of research into massive amounts of data to reveal hidden patterns and secret correlations named as big data analytics. These useful information's for companies or organizations with the help of gaining richer and deeper insights and getting an advantage over the competition. For this reason, big data implementations need to be analyzed and executed as accurately as possible. This paper presents an overview of big data's content, scope, samples, methods, advantages and challenges and discusses privacy concern on it. Since the early 1990s, people have been discussing the idea of "big data." It gained widespread recognition and esteem, and its future importance is certain to increase. These days, no company can succeed without mastering the art of managing massive amounts of data. According to MGI, Big Data is only a collection of datasets. Data of this magnitude would need sophisticated database software for recording, tracking, and analyzing. The world's information vaults are continuously being depleted. The dispute is stoked by people's usage of digital, social media, and other online forums. Acquiring new information happens at a lightning pace. The present business climate is ripe with opportunity, and the influx of data from a wide range of sources offers a treasure trove of knowledge that might prove crucial. Working with large datasets is complicated by the fact that data within a group tend to be more similar to one another than data within other groups or clusters. Many industries, from telecommunications to healthcare to banking to insurance to marketing to biology to online document categorization to city planning to seismic research to transportation, all employ big data applications.

1.2 Data Mining:

Data mining is the process of discovering and extracting patterns from large data sets utilizing methods from ML, statistics, and DBMS. Information is extracted from a data collection and translated into an understandable structure for future use in data mining, an interdisciplinary field at the confluence of computer science and statistics. Data mining and other analytic methods are used to study KDD. Data management, model and inference concerns, interestingness measures, complexity considerations, post-processing of detected structures, visualization, and live updates are all part of the process. The term "data mining" is misleading since the goal is not the collection of raw data but the discovery of useful patterns and insights within large datasets. Commonly used as a synonym for AI and BI, as well as for any kind of massive data or information processing, "big data" has become something of a catchall term in recent years. The book now known as Data mining: Practical machine learning tools and techniques with Java was originally intended to be named Practical machine learning, but the term data mining was added for marketing reasons. When referring to particular methods, the terms AI and ML are frequently more precise.

Data mining, in practise, refers to the semi- or fully-automated study of large datasets with the goal of discovering hidden, useful patterns, such as groupings of records, outliers, and interdependencies. Database techniques, such as geographical indexes, are often used for this function. These regularities may be seen as a condensed version of the original data and used in follow-up studies or in predictive analytics and machine learning applications. For instance, if the data mining step discovers multiple categories in the data, the decision support system may be able to provide more trustworthy prediction results. Although they are integral to the KDD process as a whole, they are not carried out during the data mining stage. However, data analysis is used to test models and hypotheses on the dataset, such as determining the success of a marketing campaign. However, data mining uses statistical and machine learning algorithms to discover previously unknown relationships within a large dataset. Spying on data, going fishing for data, and dredging for data are all synonyms for the same thing: employing data mining methods to infer meaning from data samples that are too small to be statistically significant. On the other hand, these methods may be utilised to come up with new theories that can be put to the test on larger data sets.

1.3 Clustering

In a computer cluster, two or more computers, known as nodes, collaborate to achieve a common goal. This allows for large, parallelizable tasks to be dispersed throughout the computer nodes in the cluster, which improves overall performance [7]. Performance is enhanced because the combined memory and processing power of each machine may aid in a wide variety of tasks. An internode network is required for the nodes of a computer cluster to communicate with one another. In order to group nodes together, specialised software is needed. It's possible that each node will utilise its own local storage device, or they may all share a single storage system. A cluster's primary entry point is usually a node inside the cluster called the "leader node." This node may, for instance, be in charge of assigning tasks to subordinates, collecting outcomes, and reporting them to an outside entity. Additionally, latency and bottlenecks may be reduced by optimising a cluster's inter-node communication [8]. Cluster computing may be broken down into various categories. HP clusters use computer clusters and supercomputers to tackle complex computational problems. The employment of nodes for communication is commonplace in the tasks they are used to doing. The throughput is increased by a dispersed group of nodes working together.

i. Load-balancing clusters: a group of computers working together to balance the workload of several users accessing the same or similar data or applications [9, 10]. As a result, no one node will be slowed down by an excessive workload. Host computers often use DFS, or a distributed file system.

ii. HA Clusters are built with spare nodes in mind, so that they can take over seamlessly in the event of a breakdown. Some examples of always-on computer services include business processes, complex databases, and consumer services like websites and file-sharing networks. Customer data is available around-the-clock, which is a major selling point.

1.4 Clustering in Data Mining

Clustering, which is based on unsupervised machine learning and is used in the area of data mining, goes by a few distinct names. Clustering arranges data points such that related things may be located next to one another. Cluster analysis is a technique for classifying data into distinct groups. In order to make sense of the information mined, data mining uses both classification and clustering techniques. Data has been tagged as a consequence of the classification process. Data samples with similar characteristics may be grouped together using a technique called clustering.

The following arguments provide insight on why clustering is crucial in data mining:

1. The following justifications illuminate the significance of clustering in data mining:
2. The most effective algorithms will be able to process a broad range of data types, such as those that are interval-based, category-based, and binary.
3. The clustering technique must be adaptable enough to recognize clusters of varied sizes and forms. They need not be restricted to only the distance measurements used to find the smallest spherical clusters.
4. Both low- and high-dimensional data must be handled by the clustering technique.

Data in databases is often incomplete, inaccurate, or noisy.

1.5 Data mining for sentiment analysis:

as an alternative research technique for collecting and analyzing textual data on the internet. Sentiment analysis is a data mining technique that systematically evaluates textual content using machine learning techniques. As a research method in marketing, sentiment analysis presents an efficient and effective evaluation of consumer opinions in real time. It allows data collection and analysis from a very large sample without hindrances, obstructions and time delays. Through sentiment analysis, marketers collect rich data on attitudes and opinion in real time, without compromising reliability, validity and generalizability. Marketers also gather feedback on attitudes and opinions as they occur without having to invest in lengthy and costly market research activities. The paper proposes sentiment analysis as an alternative technique capable of triangulating qualitative and quantitative methods through innovative real time data collection and analysis[22]. Origins of sentiment analysis are rooted in the disciplines of psychology, sociology and anthropology and flow from the theory of affective stance and appraisal theory which focus on emotions in shaping cognitions. Emotions are feelings generated from both conscious and unconscious processing. An emotional assessment of a situation is a general evaluation of that situation (whether positive or negative) that manifest in mental and bodily responses. The role of emotions in marketing is not new. To the marketer, customer emotions are indirect motivators of purchase behavior. Consumer resources such as social media posts, review and survey responses, and health materials may be analyzed using sentiment analysis for variety of purposes, from marketing to medical and customer service. World is fast changing as a result of contemporary technological breakthroughs. The Internet has become a necessity for everyone in today's environment. Because of the rapid expansion of these platforms, many people are turning to social networking applications to air their grievances and express their opinions on current events. People's reactions to utilizing public services, buying a product, and so on must be gathered and analyzed. In discussion preparation, sentiment analysis is used to identify the sentiments that underlay various viewpoints in a range of texts. There have been several applications of sentiment analysis that are making use of machine learning techniques **[1].** Recently, specialists in sentiment analysis have been concentrating their efforts on determining people's perspectives on a wide range of topics, such as movies and consumer goods. Sentiment analysis has found applications in a variety of fields.

It shapes brand saliency, influences attitudes, beliefs, opinions and perceptions. Sentiment analysis is a systematic analysis of online expressions. Specifically, sentiment analysis focuses on evaluating attitudes and opinions on a topic of interest using machine learning techniques

Literature Review

Dutta Niham*&* **Laura** ***Elle*** *(2023)analyzes extensive data libraries and how they benefit when completed. This research has results, namely library data can be called big data if the data fulfills the three characteristics of big data: volume, speed, and variety. Data mining can process big data, for example, by associations that create links between objects. If an extensive data library contains borrowed data, this means being able to establish connections between borrowed collections. Librarians, in particular, can use the results of this comprehensive data work to decide how to proceed. Therefore, it is essential to play a role in improving the library. This research aims to increase discussion between readers and librarians about using Big Data, especially in libraries. The benefit of this research is to broaden the knowledge of readers and librarians about how to utilize big data. [1]*

T. Sajana, C. M. Sheela Rani and K. V. Narayana(2023)presents the survey of clustering techniques defined with 4 V's of Big Data characteristics - Volume, Variety, Velocity and Value. Volume is the basic characteristic of Big Data which deals with data size, dimensionality of the data set and outlier's detection. Variety is deals with type of attributes of data set like numerical, categorical, continuous, ordinal and ratio. Velocity deals with algorithm analysis for computation of various attributes to process data. Finally Value deals with the parameters which are used for processing. In the present paper Introduction to Big Data is discussed in section1, Architecture of Big Data in section2, Description of clustering algorithms in section3 and finally in section4 comparison of different clustering algorithms is presented.[2]

Chonghui Guo &Jingfeng Chen (2023) discusses the research background—big data analytics in healthcare, the research framework of big data analytics in healthcare, analysis of the medical process, and the literature summary of diagnosis-treatment pattern mining. Then the challenges for data-driven typical diagnosis-treatment pattern mining are highlighted, including similarity measures between diagnosis and treatment records, typical diagnosis-treatment pattern extraction, prediction, evaluation, and recommendation, when considering the rich temporal and heterogeneous medical information in EMRs. Furthermore, a data-driven unifying diagnosis identification and prediction method (UDIPM) embedding the disease ontology structure is proposed from EMRs to assist in better coding integration of diagnosis. Three categories of typical treatment patterns are mined from doctor order content, duration, and sequence view respectively, which can provide a data-driven guideline to achieve the "5R" goal for rational drug use and clinical pathways.[3]

Rustam Mussabayev [a]**, Nenad Mladenovic** [a]**, Bassem Jarbouidiscuss**(2022)K-means plays a vital role in data mining and is the simplest and most widely used algorithm under the Euclidean Minimum Sum-of-Squares Clustering (MSSC) model. However, its performance drastically drops when applied to vast amounts of data. Therefore, it is crucial to improve K-means by scaling it to big data using as few of the following computational resources as possible: data, time, and algorithmic ingredients. We propose a new parallel scheme of using K-means and K-means++ algorithms for big data clustering that satisfies the properties of a “true big data” algorithm and outperforms the classical and recent state-of-the-art MSSC approaches in terms of solution quality and runtime. The new approach naturally implements global search by decomposing the MSSC problem without using additional metaheuristics. This work shows that data decomposition is the basic approach to solve the big data clustering problem. The empirical success of the new algorithm allowed us to challenge the common belief that more data is required to obtain a good clustering solution. Moreover, the present work questions the established trend that more sophisticated hybrid approaches and algorithms are required to obtain a better clustering solution.[4]

Wiley, Hoboken(2022)In the context of big data, many scientific communities aim to provide efficient approaches to accommodate large-scale datasets. This is the case of the machine-learning community, and more generally, the artificial intelligence community. The aim of this article is to explain how data mining problems can be considered as combinatorial optimization problems, and how metaheuristics can be used to address them. Four primary data mining tasks are presented: clustering, association rules, classification, and feature selection.[5]

Safanaz Heidari et al. presented a mapreduce-based method for density-dependent grouping of massive datasets (2019). The DBSCAN approach stands out among density-based clustering algorithms because to its superior sensitivity to noisy data and clusters of varying sizes and forms. Using the MapReduce architecture, the authors of this study want to provide a novel approach to the problem of clustering large datasets of varying densities on the Hadoop platform.[6]

P. Praveen et al. (2019) investigated large data clustering, which necessitates adapting standard data mining approaches for use with massive datasets. In this study, we provided a high-level evaluation of both classic clustering methodologies and new clustering model advancements for big data processing, with the goal of improving the management and analysis of today's massive datasets. Research into the clustering of massive datasets is a burgeoning area with plenty of room for new ideas. [7]

Ahmed Ismail et al. discussed the use of intelligent big data analytics to the study of healthcare, including their experiences, their plans for the future, and the current challenges they face (2019). They provide healthcare analytics with techniques, programmes, and software that may help with real-world issues. Our strategy necessitates the installation of middleware between the various data sources and the MapReduce Hadoop cluster so that we may integrate the data. The approach dealt with the ineffectiveness of combining information from several sources. Taking cues from computer models of bee hives, S. Sudhakar Ilango et al.(2019) has presented a method for clustering large datasets for optimization [8].

Diverse areas of data mining for user-generated product evaluations have been the subject of scholarly works.

Ronan Collobert and his coworkers have employed convolution networks. With a single neural network and learning approach, the author shows how to do part-of-speech tagging, chunking, named entity recognition, and semantic role labelling. [9]

The goal of this study was to collect and classify user feedback using K. Dave's product reviews. In order to perform feature extraction and scoring, the classifier makes use of information retrieval techniques, hence the results may be different depending on the setting. Traditional machine learning isn't your sole choice when looking for novel ways to learn. Because of their sheer number and sometimes vague nature, internet search results made it impossible to execute operations on individual phrases. L. Maria Soledad Elli

[3] was swayed by customer feedback. The corporation has reviewed the data and adapted its approach accordingly. The book's main argument is that the author's proposed technology is more exact than existing methods. Scientists have used multinomial Naive Bayesian analysis. Here, classifiers do the heavy lifting. Mechanisms are also used to support vector machines. Machine learning methods were also used to detect fake trends and reviews. M. S. Hota and S. Pathak[4] used the Knn classifier to do multi-class sentiment analysis on Twitter data. Emotions are now simply called "feelings," as opposed to the outdated word "sentiment," which formerly described human emotions. Sentiment analysis, also known as Data mining, is a subfield of data mining that involves the examination of user-generated content (UGC) to ascertain public opinion on topics such as recent events, organisations, products, and brands. UGC includes but is not limited to microblogging sites, social media, online news, and user reviews. In order to classify participants' emotions, this study used a multidimensional framework. If we compare the innovative approach with the current one using industry-standard evaluation criteria, we find that it performs better. Surveys were conducted by L Zhang and K. Liu for Data mining and Sentiment Analysis. Sentiment analysis, often known as Data mining, is a technique used to analyse the general public's feelings on certain topics. A lot of research on NLP ideas and models was reviewed in Rain [6]. In their studies, the researchers used decision list classifiers and naive Bayesian classifiers. A review may be categorised in these ways. It's possible that this is either a good or a terrible review. There has been a rise in the use of deep neural networks to scientific inquiry. Recent years have seen a rise in the use of neural networks for the analysis of sentiment. When customers received their orders, they were prompted to provide evaluations on Amazon. M. R. Socher [7] pioneered the use of recursive neural networks, which are now being put to use to better comprehend compositionality in areas such as sentiment identification. Naive Bayesian, K-nearest neighbour, and Supporting Vector Machine are just a few of the classic techniques that will be used in this study, along with more modern deep learning strategies. N. Xu Yun et al. [8] used computer science techniques from Stanford University, including the perceptron, naive bayes, and supporting vector machines. [10]

Seventy percent of it was really put to use here. N. KumarRavi and VadlamaniRavi analysed the tasks, methods, and applications of Data mining and sentiment analysis. In the past ten years, researchers have published over a hundred publications that explore different sentiment analysis methods, techniques, and applications. The methods are examined thoroughly in this research. The article includes a summary of over a hundred research and a list of open questions. [11]

In 2017, aspect-based sentiment analysis on Twitter was used to classify hybrid sentiment, an idea first proposed by O. Zainuddin. In order to dig further into the data, our study used Twitter's aspect-based sentiment analysis. [12]

In this research, we explain how to use a feature selection technique for sentiment classification on Twitter. According to N. M. Z. Asghar], a hybrid classification algorithm may be used to examine Twitter sentiment. The researchers in this investigation used a hybrid method of categorization to address these issues. [13

O. Alsaeedi, Abdullah, and Mohammad Khan just completed a research of methods for analysing sentiment in Twitter data. They believed that rapid global change was being brought about by advances in modern technology. [14]

P. Shathik and Anvar put out the idea of sentiment analysis using machine learning techniques, with the two authors providing an overview of the existing literature on the topic. [15]

Research Methodology:

Large datasets have been shielded from clustering for the duration of the intended research. Acceleration and improved efficiency are the results of a modern data mining approach that is driven by data. Data mining methods reduce the length of the material, and cutting-edge big data procedures are applied to the information. There have been several research in existence that are related to sentiment analysis, machine learning and accuracy. Issues faced in previous research are accuracy, reliability and lack of flexibility. Thus novel approach has been proposed in order to perform sentiment analysis on textual tweets. Finally accuracy and f-score is evaluated. Researchers are using sentiment analysis on a dataset culled from Twitter, which contains both visual and textual user comments, to get a sense of user attitude. Rapid transformation is occurring all around the globe as a consequence of recent scientific and technical developments.

The Internet is a necessity in today's society. As the popularity of social media platforms continues to grow, more and more people are utilizing them to share their views on important topics of discussion. It's important to get people's feedback before making any major changes to a product, service, etc. In order to be well-prepared for a conversation, many people engage in Data mining (also known as sentiment analysis), a process in which the emotions behind different points of view are sought to be identified. Consumer sentiment research has become more popular in recent years.

Research in this area focuses mostly on user-posted tweets, both textual and graphical. For visual stuff, we apply a CNN-based approach, whereas neural networks are used for the textual variety. In recent years, opinion research based on Twitter data has paid a lot more attention to "tweets" (comments) and the substance of these statements. Therefore, the purpose of this study is to compare the results of different sentiment analysis applied to Twitter data. The proposed study employs a machine learning strategy based on neural networks to identify emotional states from images. Preprocessing the data retrieved from Twitter allows for the extraction of the most relevant information. Take into account the following set of facts, which will serve as the foundation for your predictions. After then, the text and images in this data are analysed to determine their respective categories. To study visual data, the CNN model is used. When everything is said and done, the current model's precision and efficiency are weighed against those of its predecessors.

In this section, we'll show you how to use the research in a number of different ways.

I. Exploratory Research, which lays out and finds new problems.
II. Constructive Research, which comes up with ways to solve a problem.
III. Empirical Research, which tests the feasibility of a solution based on real-world evidence. This is the third type of research.

The accuracy, f-score, and precision of LSTM-based training and testing improve when Nave Bayes and KNN classifiers are used with them. This work is better at being reliable and adaptable when compared to other work. This means that when the proposed work is compared to previous research, it is found that having a hidden layer and classifier makes it faster and less likely that an error will happen when making predictions. The classification process lets the LSTM system predict in less time and with more accuracy. Cutting down the number of people in a batch makes them more accurate, but it also slows things down. Another way to do this is

Need of research:

Effective clustering mechanisms are used for managing massive data sets. After aggregating all of the raw data, it was determined that. To make it more manageable, we've broken everything down into bite-sized chunks of information and modeled the actual processing of requests in real time. There was also a reduction in the amount of time needed to get a prescription. Cluster sizes have been read by inspecting file sizes. For Big Data Analysis, clustering stands out as a popular unsupervised strategy that is also crucial. Clustering may be used as a statistical tool to find relevant patterns within a dataset, or as a pre-processing step to decrease data dimensionality before executing the learning algorithm It has been observed that there is need to consider the choice of customer during inventory management. If the taste of customer is considered, then it becomes easy to manage the stock. Semantic analysis is assisting in considering customer choice. Machine learning system allows management to predict the demand of products. If right products are available at right time then the scalability and efficiency of business gets increased. Moreover it reduces the probability of losses that are faced by commercial organization. Thus it is required to enhance the capability of semantic analysis by making use of machine learning mechanisms such as LSTM.

Objectives:

1. To handle extensive databases, the clustering algorithm should be scalable. Data should be scalable, if it is not scalable, then we can’t get the appropriate result which would lead to wrong results.
2. To handle high dimensional space along with the data of small size clustering algorithm is suitable.
3. To handle unstructured data and give some structure to the data by organizing it into groups of similar data objects. This makes the job of the data expert easier in order to process the data and discover new patterns.
4. To handle multiple kind of data clustering algorithm is capable.
5. To consider the existing researches in area of sentiment analysis and machine learning.
6. To identify the problem of accuracy and performance in previous research work.
7. To propose integrated classifier to RNN based approaches in order to provide high performance, accurate and efficient solution for sentiment analysis.
8. To compare the performance and accuracy of the proposed model with the conventional mechanism.

Scope of research

A high-availability cluster, also known as a failover cluster, makes use of many systems that are already installed, configured, and connected such that if a problem causes one of the systems to fail, another may be leveraged smoothly to preserve the availability of the service or application. The use of a clustered computing environment has several advantages, including increased availability thanks to fault tolerance and resilience, the capacity to balance and scale workloads, and enhanced performance. Clustering is a method for classifying a set of data items as a cohesive unit based on their shared characteristics. The term "group" is shorthand for "cluster." Cluster analysis is a method for organizing data sets into subsets based on their similarities. In order to gauge public opinion, researchers are analyzing tweets and user comments using sentiment analysis. As a result of technological development, the globe is altering at a breakneck pace. Having Internet connectivity is crucial in today's society. People are increasingly using social network applications to voice their opinions on current events. When trying to sell a product or improve a government service, collecting and analyzing customer feedback is essential. Data mining, also known as sentiment analysis, is often done in advance of a discussion in which the attitudes behind different points of view are to be found

References

[1]. Dutta NihamMonash University Malaysia Laura ElleSoutheast University Chinahttps://doi.org/10.34306/ijcitsm.v3i2.128

[2]. T. Sajana, C. M. Sheela Rani and K. V. Narayan Researchgate.net/profile/SajanaTiruveedhula/publication/298082409_A_Survey_on_Clustering_Techniques_for_Big_Data_Mining/links/5aa39ee145851543e63d7333/A-Survey-on-Clustering-Techniques-for-Big-Data-Mining.pdf

[3]. https://link.springer.com/chapter/10.1007/978-981-99-1075-5_2

[4]. Clarisse Dhaenens & Laetitia Jourdan https://link.springer.com/article/10.1007/s10479-021-04496-0

[5]. Heidari, S., Alborzi, M., Radfar, R., Afsharkazemi, M. A., & Rajabzadeh Ghatari, A. (2019). Big data clustering with varied density based on MapReduce. Journal of Big Data, 6(1). https://doi.org/10.1186/s40537-019-0236-x

[6]. Praveen, P., & Jayanth Babu, C. (2019). Big Data Clustering: Applying Conventional Data Mining Techniques in Big Data Environment. In Lecture Notes in Networks and Systems (Vol. 74). Springer Singapore. https://doi.org/10.1007/978-981-13-7082-3_58

[7]. Ismail, A., Shehab, A., & El-Henawy, I. M. (2019). Healthcare Analysis in Smart Big Data Analytics: Reviews, Challenges and Recommendations. Springer International Publishing. https://doi.org/10.1007/978-3-030-01560-2_2

[8]. Ilango, S. S., Vimal, S., Kaliappan, M., & Subbulakshmi, P. (2019). Optimization using Artificial Bee Colony based clustering approach for big data. Cluster Computing, 22, 12169–12177. https://doi.org/10.1007/s10586-017-1571-3

[9]. Rao, T. R., Mitra, P., Bhatt, R., & Goswami, A. (2019). The big data system, components, tools, and technologies: a survey. In Knowledge and Information Systems (Vol. 60, Issue 3). Springer London. https://doi.org/10.1007/s10115-018-1248-0

[10]. Mazumdar, S., Seybold, D., Kritikos, K., & Verginadis, Y. (2019). A survey on data storage and placement methodologies for Cloud-Big Data ecosystem. In Journal of Big Data (Vol.