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

ISSN : 2456-3307

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/IJSRCSEIT

Performance of Clustering Technique During Data Mining to Analyze Big Data

Dr Kapil Kumar Kaswan[1], Preeti[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: 03 July 2023

Published: 20 July 2023

Publication Issue

Volume 9, Issue 4

July-August-2023

Page Number

124-130

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.

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

I. INTRODUCTION

1.1 Big data

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 analysing. 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.

10 V's of Big Data

The list of 10 V now includes seven additional characteristics that are essential for all users to grasp. This chapter is perfect for introducing the 10 V of big data [2], since all of these properties begin with V.

- **Volume:** It was thought that the sheer quantity of big data was its most distinguishing attribute. Every minute, four thousand hours of video are added to YouTube. In 2016, it was estimated that monthly mobile traffic worldwide will average 6.2 Exabytes, or 6.2 billion gigabytes. Problems with timeliness and precision are due to the sheer amount of data being generated.

- **Velocity:** When we speak about how quickly anything happens, we're referring to the pace at which new data is generated and updated. Awesomely, Facebook's data centre may be able to hold four terabytes of data.

- **Variety:** Utilizing organised data is ideal when dealing with massive volumes of data. Sometimes it's necessary to work with data that's only kind of in order.

- **Variability:** The intrinsic diversity of this theme allows for the expression of a wide range of possibilities. Since there is a great deal of incoherence in the database, an external detection technique is essential for obtaining useful findings. The dimensions of an attribute allow for inferences to be made across a broad range of possible data sources and formats, which is made possible by the attribute's intrinsic variability as a consequence of data's inherent heterogeneity.

- **Veracity:** To be deemed "veracity," data & its sources must be trusted and validated.

- **Validity:** When we speak about validity in relation to large amounts of data, we are referring to data accuracy with adjustments.

- **Vulnerability:** Since a compromise of such massive volumes of data might have far-reaching repercussions, the security risk is high. The CRN reports that hackers once again offered stolen data for sale on the dark web in May of 2016. Those 360 million email addresses and passwords were stolen, including some that belonged to MySpace, resulting in a total of 197 million compromised user accounts.

- **Volatility:** A longer retention period is an indication of data's instability. There are a variety of applications for this kind of data.

- **Visualization:** In addition, data may be easily analyzed thanks to visualization, which is a second unique aspect. Technical issues include memory technology limitations, inadequate scalability, functionality, and reaction speed, to name a few.

- **Value:** Ultimately, the value it provides is the most important attribute. The rest of big data's characteristics haven't proven to be helpful either. If the real value or economic potential of the data can be determined. We now know that huge data may be very valuable. More consumer research is needed to better understand their needs and simplify services accordingly. Our focus here is on both productivity increases and overall company success.

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

II. Literature Review

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. [1]

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. [2]

In their presentation, 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 [3].

Comparison of dataset sizes to processing times is shown. The results of the ABC algorithm's observer and worker phases are used to calibrate the percentage of classification error for a range of fitness and probability values. [4]

T. Ramalingeswara Rao et al. analysed the big data system's constituent parts (2019). In this research, we focus on the characteristics of the data management process and analyse the properties of numerous MapReduce-compatible distributed file systems and NoSQL databases. In addition, we provide a variety of crucial cloud-based ML tools used throughout the data

model construction, development, and deployment processes. [5]

In order to better understand how cloud-based data storage & retrieval works, Somnath Mazumdar et al. (2019) did a thorough literature study. On this post, we examine the newest methods for deploying and storing Big Data in the cloud. For the sake of Big Data management, an effort is being made to bring attention to the real connection between the two. [6]

Li Zhu1 et al. conducted research on how to optimise parallel collaborative filtering techniques for large data mining (2019). In this paper, we analyse the execution flow of a standard parallel collaborative filtering algorithm, discuss its limitations, and outline the steps for improving it, beginning with the creation of node scoring vectors and continuing all the way through the retrieval of neighbouring nodes and the formation of recommendation information. Rabindra Kumar Barik et al. [7] explored the pros and cons of using a hybrid mist-cloud for massively parallel spatial analysis & processing (2019). In this study, the author explains why and how mist computing has recently surged in popularity for use in geospatial analysis. Additionally, MistGIS, a mist computing framework, was developed to aid in mining geographic large data for mining-related insights. [8]

R. Joseph Manoj et al(2019) ACO-based ANN feature selection approach is tailored to massive datasets. Feature selection is a method for streamlining a large dataset by selecting individual items that have common characteristics. It can help reduce the massive data set. These findings show that it is possible to create a feature selection method for text classification by merging ACO and ANN algorithms. The simulations they ran using Reuter's data set provided conclusive evidence in favour of this hybrid method. [9]

Wenzhun Huang et al. looked into a new method of cluster computing based on Hadoop for grouping signals and creating an analytical hierarchy model (2019). High-performance computing on massive data in the cloud is investigated here using the Hadoop data model. The researchers compared the suggested approach to existing literature. [10]

Kai Peng et al.(2018) established intrusion detection across massive data in a mobile cloud environment using hierarchies and principal component analysis (pbirch). In this study, the authors describe a new clustering method they've dubbed PBirch for tackling this issue. The results show that, contrary to PMBKM, PBirch has the ability to provide a decent clustering result, and that this result may be further enhanced by optimization of the necessary parameters. The processing time of PBirch drops dramatically when more clusters are added. [11]

Samiya Khan, et al., presented their work on cloud-based analytics, along with a literature assessment and suggestions for the field's future (2018). Since the dawn of the information era, the volume of data in all its forms has increased dramatically. It is anticipated that most information will be stored on the cloud by 2016. Organizations can only expect to get insights from this data if they have a system in place to collect, clean, and analyse it. Because of this, research into cloud-based analytics may progress. [12]

III. Problem statement

Several research projects have examined clustering, Big data, and data mining. That mountain of information may be kept in a public, private, or hybrid cloud. Numerous clustering strategies have been used in data mining. Data management, however, presents certain challenges. The performance of data transmission via the mining mechanism is degraded when big data solutions are implemented. In addition, earlier research has only provided a slice of security for massive data.

IV. 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.

V. 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.

VI. 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.

VII. REFERENCES

[1]. 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

[2]. 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

[3]. 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

[4]. 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

[5]. 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

[6]. 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. 6, Issue 1). Springer International Publishing. https://doi.org/10.1186/s40537-019-0178-3

[7]. Zhu, L., Li, H., & Feng, Y. (2019). Research on big data mining based on improved parallel collaborative filtering algorithm. Cluster Computing, 22, 3595–3604. https://doi.org/10.1007/s10586-018-2209-9

[8]. Barik, R. K., Misra, C., Lenka, R. K., Dubey, H., & Mankodiya, K. (2019). Hybrid mist-cloud systems

for large scale geospatial big data analytics and processing: opportunities and challenges. Arabian Journal of Geosciences, 12(2). https://doi.org/10.1007/s12517-018-4104-3

[9]. Joseph Manoj, R., Anto Praveena, M. D., & Vijayakumar, K. (2019). An ACO–ANN based feature selection algorithm for big data. Cluster Computing, 22, 3953–3960. https://doi.org/10.1007/s10586-018-2550-z

[10]. Huang, W., Wang, H., Zhang, Y., & Zhang, S. (2019). A novel cluster computing technique based on signal clustering and analytic hierarchy model using hadoop. Cluster Computing, 22, 13077–13084. https://doi.org/10.1007/s10586-017-1205-9

[11]. Khan, S., Shakil, K. A., & Alam, M. (2018). Cloud-based big data analytics—a survey of current research and future directions. Advances in Intelligent Systems and Computing, 654, 595–604. https://doi.org/10.1007/978-981-10-6620-7_57

[12]. Peng, K., Zheng, L., Xu, X., Lin, T., & Leung, V. C. M. (2018). Balanced iterative reducing and clustering using hierarchies with principal component analysis (PBirch) for intrusion detection over big data in mobile cloud environment. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 11342 LNCS. Springer International Publishing. https://doi.org/10.1007/978-3-030-05345-1_14

[13]. Skourletopoulos, G., Mavromoustakis, C. X., Mastorakis, G., Batalla, J. M., Dobre, C., Panagiotakis, S., & Pallis, E. (2017). Towards Mobile Cloud Computing in 5G Mobile Networks: Applications, Big Data Services and Future Opportunities. 43–62. https://doi.org/10.1007/978-3-319-45145-9_3

[14]. Ularu, Elena Geanina, “Perspectives on Big Data and Big Data Analytics.“Journal of DBSJ, DBSJ (Database Systems Journal) 2012.

[15]. NirmalKaur, Gurpinder Singh,” A Review Paper On Data Mining And Big Data”, ISSN No. 0976-5697, Jalandhar, Punjab, India, 2017

[16]. Bandara, I., Ioras, F., Maher, K.: Cybersecurity concerns in e-learning education. In: ICERI2014 Conference, 728-734 (2014).

[17]. Meslhy, E.: Data Security Model for Cloud Computing. Journal of Communication and Computer 10, 1047-1062, (2013).

[18]. Yang, H., Tate, M.: A Descriptive Literature Review and Classification of Cloud Computing Research. Communication Association Info System 31, (2012).

[19]. Kumar, A.: Secure Storage and Access of Data in Cloud Computing. In: International Conference on ICT Convergence, 15-17 (2012).

Cite this article as :

Dr. Kapil Kumar Kaswan, Preeti, "Performance of Clustering Technique During Data Mining to Analyze Big Data", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 4, pp.124-130, July-August-2023.