Journal of Advances in Science and Technology
Vol. 20, Issue No. 2, September-2023, ISSN 2230-9659

An Overview of Characteristics and Application of Big Data

Rajendra Mahto[1]*, Dr. Nidhi Mishra[2]

[1] Research Scholar, Kalinga University, Raipur, Chhattisgarh, India

Email: rmahto2250@gmail.com

[2] Assistant Professor, Dpt. of Computer Science, Kalinga University, Raipur, Chhattisgarh, India

Abstract- Big data is huge volume, heterogeneous, appropriated data. Big data applications where data gathering has developed ceaselessly, it is costly to oversee, catch or concentrate and procedure data utilizing existing programming apparatuses. Clustering Technique for unstructured for big data is a primary errand of exploratory data investigation and data mining applications. This phenomenon, commonly referred to as Big Data, presents both opportunities and challenges across various domains. This paper provides a comprehensive overview of the characteristics and applications of Big Data. it delves into the characteristics of Big Data, including its volume, velocity, variety, veracity, and value. Understanding these characteristics is crucial for harnessing the full potential of Big Data analytics. the paper explores the diverse applications of Big Data across industries such as healthcare, finance, retail, manufacturing, and transportation.

Keywords- Data, Big data, Datamining, Clustering, HADOOP

---x---

INTRODUCTION

Big data has real time data-intensive processing that runs on high performance clusters. Big data communication is used for distributing the data across various locations. This process is very expensive and it needs a large memory space to hold the data for computing. Big data contains the transaction and interaction datasets based on size and complexity that exceed regular technical capability in capturing, organizing and processing data at a reasonable cost in cloud environment. The big data computation and information sharing are carried out effectively using data preprocessing in cloud environment. In big data applications, data collection process has grown exponentially. Big data applications involve the process of collecting and sharing the information with larger memory consumption. The main problem in big data applications is to analyze the large volumes of data and extract useful information or knowledge for future activities. The unnecessary noises, obtained from various sources present in the data are removed with the help of preprocessing, which minimizes the time taken for computation and improves the information sharing. The distributed data mining on huge amount of cloud data needs minimum computational overhead and communication costs. Big data is frequently described by its quantity which exceeds the typical range of databases. In big data classification, the challenge comes from evaluating and understanding uniqueness of huge datasets by retrieving valuable geometric and statistical patterns. Due to the availability of extensive information and benefits related to data processing, big data has gained great research significance. Big data applications are processed with the scalable nature of data via MapReduce programming model.

DATA MINING

The Information mining is a various are that incorporates colossal instruments acquiring different answers for the raised issues and delays its wings in multidisciplinary. As of late, information mining assumes a urgent part in data industry and society because of wide accessibility of tremendous measures of information. This huge measure of information is helpful when it was changed over into data, so there is the need of breaking down the information. It enables clients to think about information from various measurements to shape the connection between the information and to characterize the information. Information mining is exceptionally useful in numerous applications in this present reality jump at the chance to discovery of misrepresentation, examination of market, client maintenance, organic, metrological, and other logical applications. Real-world data contain inconsistent, noisy, and missing data because it was collected from heterogeneous sources. To enhance effectiveness and simplify the extracting method, we can apply the pre-processing techniques on the real-time data.

Data cleaning: It removes the dirty data by smoothing noisy data, filling in missing values with most probable values and removal of inconsistencies, outliers, finally obtained the accurate data.

Data selection: Here applicable information are recovered from the database for the investigation assignment. Information diminishment may likewise be performed to get the first information with littler portrayal.

Data integration: Which blends information from numerous information sources as data cubes, databases, flat files into a solitary information store as information warehousing. These prompts enhance the precision and speed through the information mining process by the evacuation of redundancies and irregularities in the subsequent informational collection.

Data's transformation: Data are transformed into various forms by using normalization, generalization, aggregation and attribute selection because, which are suitable for mining.

Data mining: It is essential process here shrewd techniques are connected thinking the final objectives to extricate designs. To this progression, fascinating examples are exhibited to the client, and this new information might be put away in a learning base.

Pattern evaluation: Based on interestingness measures knowledge was represented, which is mainly to identify the truly interesting patterns.

Knowledge presentation: Mined knowledge present to the user using knowledge representation and visualization technique. Data mining techniques are efficient and scalable. Because these algorithms reduce running time of mining process of course, the size of data is very large. Above mentioned steps are very crucial to reduce data size by considering only relevant data, and also reduce the repentance of data. Data mining system finally gives the requested knowledge or information to the user according to their query.

The pre-processed data can be used by the data mining process which was accumulated in large different data storages. We know function of mining the data is to discover interesting knowledge from above components as explained in the structural design of data mining represented as Fig. 1.

- **Dataware house,** Database, Other Information Repository and World Wide Web: Which may include spread sheets, set of data bases, and different kinds of information repositories. Preprocessing methods applied to these database systems for mining of data effectively.
- **Data Cleaning:** It removes the dirty data by smoothing noisy data, filling in missing values with most probable values and removal of inconsistencies, outliers, finally obtained.
- **Data Warehouse Server:** Dependent upon the users request data warehouse server or database fetches the relevant data to them.
- **Knowledge Base:** This is connected to both data mining engine and pattern evaluation component. It evaluates or searches the interestingness of resulting patterns by the guidance of domain knowledge as of the user beliefs. Domain knowledge is like thresholds, metadata, and interesting constraints. Knowledge base describes data from several heterogeneous sources.
- **Data Mining Engine:** It is an extremely fundamental segment inside the information mining framework since it comprises of an arrangement of functionalities as connection and affiliation, portrayal, expectation, order, development investigation, group examination, and exception investigation.
- **Pattern Development Module:** This module connects with the information mining module and UI modules to remove just intriguing examples by considering intriguing quality measures.
- **User Interface:** This module enables the client to communicate inside the information mining framework by indicating an errand or an information mining inquiry. Clients get suitable data by the information mining comes about. This segment furthermore enables clients to see information distribution center diagrams, peruses through information bases, assess mined examples and diverse types of examples can be envisioned.

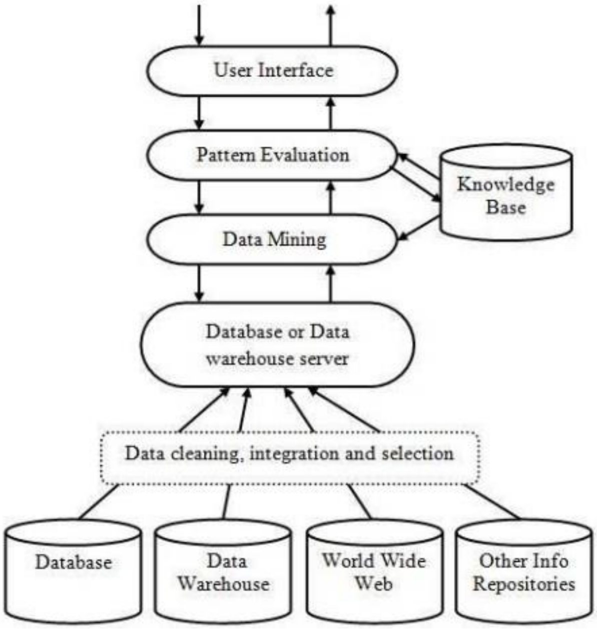


Figure 1: Data mining Basic Architecture

Data mining system is an advanced stage of the On-Line Analytical Processing (OLAP) in the perspective of a data warehouse. It has a wide range of dealing out within the data storage system using latest methods to perform analysis. Architecture of data warehousing consists of warehouse database server, OLAP server and front-end client layer. Warehouse database server is used to feed data from operational databases and external sources. It also was having the utilities and tools to perform cleaning, extraction, transformation, load, and refresh functions in the data. Next layer is OLAP server to implement standard relational operations and multi-dimensional data. The last and important layer is a front-end client. Here we have been reporting tools, analysis tools, and queries. Models of the data warehouse include the virtual warehouse, the enterprise warehouse, and the data mart.

BIG DATA

Big data innovations characterize another age of advances and models, structured exclusively to financially extricate valuable data's from extremely substantial volumes of a wide assortment of data, by allowing high speed catch, disclosure, and analysis. O'Reilly characterizes big data is the data that surpasses the preparing limit of traditional database frameworks. He likewise clarifies that the data is extremely big, moves exceptionally quick, or doesn't fit into customary database models. Further he has stretched out that to pick up an incentive from this data, one needs to pick an elective method to process it.

Big data is a term including various sorts of confounded and substantial datasets that is difficult to process with the regular data handling frameworks. Various difficulties are set up with big data like stockpiling, change, perception, seeking, examination, security and protection infringement and sharing. The exponential development of data in all fields requests the progressive estimates required for overseeing and getting to such data. The creators have featured the requirement for the examination in big data, so as to deal with the online bio-intelligent data road. They have predicted the significance of big data in the natural and biomedical research. It has detonated so that it has minimized an administrative mapping for actually recognizable data. This is conceivable by breaking down the meta data and by utilizing the prescient, amassed discoveries along these lines consolidating the past discrete datasets.

BIG DATA CHARACTERISTICS

Big data investigation has gotten a great deal of attention as of late, and in light of current circumstances. To turn into the piece of this development one needs to think about big data examination. This pushing the envelope on investigation is an energizing part of the big data examination development. Organizations are eager to have the capacity to get to and investigate data that they've been gathering or need to pick up knowledge from, yet do not have the capacity to oversee viably. It may include envisioning enormous measures of different data, or it may include progressed broke down gushing at you continuously. It is both transformative and progressive.

Data: An introduction

Data might be characterized as an accumulation of qualities including subjective or quantitative factors. Data may establish crude certainties, figures, numbers and so forth. Extensively data can be partitioned into three prime classifications. These are given underneath.

Types of Data

Any data can be portrayed by the manner in which it is sorted out in the capacity frameworks. Basically, there are three general classifications of data.

1. Organized data
2. Semi-Structured data
3. Un-Structured data

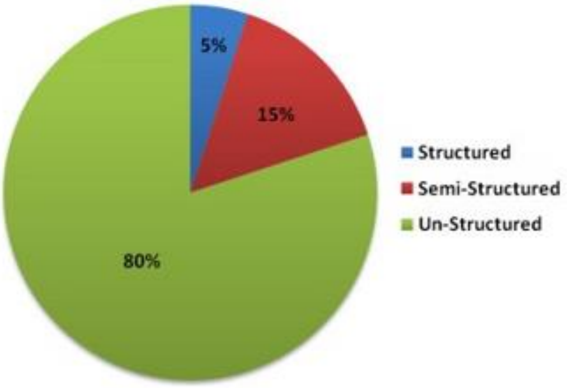


Figure 2: Approximate Percentage composition of different kinds of

Data

Structured Data

These are those sorts of data that adjust to a formal structure of a data-model or outline. They can be effectively mapped into predefined fields in the memory stockpiling frameworks like social database frameworks (RDBMS). The essential preferred standpoint of organized data is the simplicity of capacity, controls (expansion, cancellation, updation) and examination. It turns out to be extremely simple to perform activity on such sort of data since they can be effectively confined to a compartment with some predefined stockpiling structures and instruments. Each organized data has a few confinements related with it like sort, size and arrangements (numbers, characters, alpha-numeric and soon.).

Semi-Structured Data

They are those datasets which don't have an inflexible and severe organization or mapping related with them. In any case, they contain some sort of metadata or semantics connected to them, which help is their distinguishing proof and associations. These data are not put away as lines and segments of a table. They are commonly composed in a progressive system. It gives a genuinely adaptable portrayal of data which can be effectively changed over in any ideal arrangements acclimating inflexible configurations and blueprints for performing investigation and preparing. Scarcely any instances of semi-organized data are: XML records, JSON reports, Email, EDI and so forth.

Unstructured Data

Most of data (over 80%) that we have available to us today is generally unstructured data. These sorts of data don't have any inflexible configurations or outlines related with them. On account of this one-of-a-kind element, these datasets can't be put away in customary database frameworks and along these lines can't be handled and dissected like the conventional organized datasets. Hardly any instances of unstructured data are: pictures, recordings, sounds, web logs, long range informal communication data, Quick Response (QR) codes and geospatial datasets.

Unstructured and Complex nature of Data:

A large portion of the big data created from various sources are to a great extent unstructured in nature. For instance, messages, twitter channels, YouTube recordings, and social systems administration locales data and so on. This implies these data can't be prepared like the traditional data sets. Uncommon instruments and techniques are required for preparing these data.

BIG DATA FIVE V's

Characteristics are shown as follows

1. **Volume:** Large collection of information may be represented as volume as one of five Vs. Size related with enormous volume thus named as big data. In recent decade the price of the hardware and digital equipment has reduced considerably as compared with the earlier decades. This has led to rise in storage space and may hoard enormous amounts of data at a small price.

2. **Velocity:** This mainly involves high speed rate of data which is accumulated. In trading, banking all transactional events such as should take action as and when they happen. The speed of the arriving data is colossal and continuous. The growing rate of change in the types of information is also called as velocity. From the heterogeneous resources like sensors, networks, mobile, social networks data etc., the data generated are mainly categorized as structured, semi structured and more percentage of unstructured data. Facilitates to deal with the hassle like velocity.

3. **Variety:** This mainly consists of mixed data comes from various sources internally and externally. Data variety is a measure of data representation in different forms. Data formats encompass text, audio, numeric, image, video and logs.

4. **Veracity:** A veracity deal with uncertainty, an inconsistency in information which is exist from various resource, sometimes leads to confusion and disordered in which it is tough to organize.

5. **Value:** Value focuses very important V of all the 5 V's which measures the usefulness of valuable data extracted to advantage knowledge and predicts important decisions. The additional characteristic of variability can also be considered:

6. **Variability:** the changing nature of the data companies seek to capture, manage and analyze – e.g., in sentiment or text analytics, changes in the meaning of key words or phrases

BIG DATA APPLICATIONS

Today the world is flooded with statistics. Large Organizations take advantage of those data for their business growth by analyzing which allows to taker faster and better decisions. By analyzing this data made in different cases as discussed below in Fig. 3:

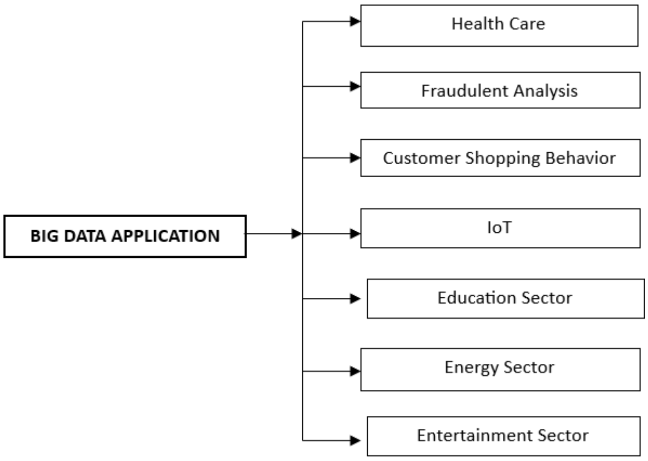


Figure 3: Big Data Applications

- **Health Care:** Big data analysis is widely utilized in healthcare area for collecting, studying a leveraging patient medical and physical information which might be very big or multifaceted implicit by way of conventional method of processing. Nowadays this large volume of data is practiced with the development of technology and tools uses machine learning, AI, and by data scientists. In response to the rise of value-based care due to digitization of healthcare information thus recommended enterprise to apply analytics for effective decision making. Create holistic, 360-degree view of patients, consumers and physicians and improve efficiency and care personalization with comprehensive patient profiles. Provides advanced healthcare support with data about patient, consumer and medical doctor needs and preferences.

- **Fraudulent Analysis:** Bank transactions fraud is a serious concern which has to be handled in proper method by detecting and preventive such threats. It can be used to detect the fraud in financial services by constant monitoring and alerting agencies and thus detect unusual and suspicious behavior, which increases the quality of transactions in business.

- **Tracking customer spending habit, shopping behavior:** In large organization and enterprises, Big data technology helps in concrete decision making, accurate prediction to find out customer sale details, frequency of purchase details, their spending habit etc.

- **IOT:** It collects huge volume of unstructured data generated by IOT devices for analyzing out in real time and accumulate in various storage units. To collect real time operational data the various connected devices via automated intelligent sensors and controls and with other computing capabilities are installed into machines over vast regions and collected and stored in centralized system and efficiently analyzed in real time through big data analytics.

- **Education Sector:** For effective utilization of their data to create opportunities and to find out new fields in their business, big data analyses all these data's and helps to get insights which improve the operational effectiveness of various educational sectors. Educators can reap maximum benefits and they can improve their teaching skills. And also helps the students to obtain many learning experiences in different way and helps them to pursue the career depend on their area of interesting in appreciable and encouraging way.

- **Energy Sector:** This analytics of large-scale data performs vital task in energy management on demand side. The energy sector collects huge amount of data on a continuously with various applications like network communication, sensors and cloud computing technologies and wireless transmission. Through big data analytics, energy utilities can optimize power generation and planning and helps in decision making process. It improves the efficiency in energy production hence reduce costs. And also, big data analysis play significant role in renewable energy like solar power. And data forecasting will be more efficient and accurate. Example, weather data analysis, GIS data, energy consumption and energy production data. And also, the efficiency of collaborative operation and asset management can be improved through the analysis.

- **Entertainment Sector:** Bigdata helps the entertainment sector and global media for rapid digital change in which recent methodologies persists along with innovative methods. Examples, It helps in predicting audience interests, also provides insights into customer churn.

HADOOP COMPONENTS

Hadoop Common: This consists of libraries in which Hadoop cluster automatically manages the failures of any hardware.

- **Hadoop Distributed File System:** Hadoop distributed file system mainly used for storage and helps in faster computations, In personal computer system data resides in local file system whereas here all files are stored in HDFS.

- **YARN:** YARN provides resource management. It is called as operating system of Hadoop Eco system takes responsibilities of dealing with and observes workloads and performs batch processing.

- **Map Reduce:** It executes responsibilities in a simultaneous style through providing the information as tiny blocks.

- **Other Hadoop Components:** The various other Hadoop components are shown in Fig. 4, Ambari, Cassandra, Flume, HBase, HCatalog, HadoopSqoop, Hadoop Hive, Hadoop Oozie, Hadoop Pig, Solr, Spark and Hadoop Zookeeper.

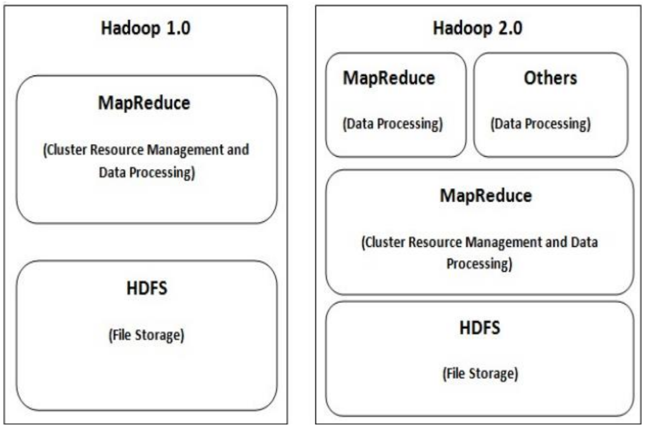


Figure 4: Core Hadoop Components

HADOOP DISTRIBUTED FILE SYSTEM

Refers to main part of Hadoop part developed using distributed file system design that supports large scale of data to be stored across several distributed nodes in a Hadoop cluster and provides easier access. HDFS is a default storage file system used in Hadoop. To access data it offers high performance Master/Slave architecture is followed in HDFS. Name Node and Slave Nodes are the two components of HDFS.

Name node: Name Node is a software program run on commodity hardware plays like master server and keep tracks of the document device name-space processes like entire file operations. This node also normalizes client access to files and acts as metadata and holds the information associated with mapping files to blocks.

Data node: Data Nodes are commodity hardware possessing GNU or Linux operating data node software responsible for data storage. HDFS stores files in blocks with a default size of 64MB whereas in Big Insights it uses of 128MB length and it can also be modified as in per HDFS configuration.

In Data Node the blocks are stored on the underlying file system in a redundant fashion which performs write and read operations as per client request.

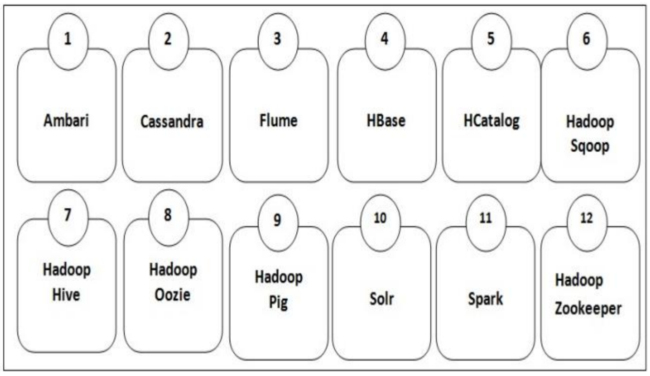


Figure 5: Complementary/Other Hadoop Components

ARCHITECTURE OF HADOOP DISTRIBUTED FILE SYSTEM.

The following are the features supported by HDFS are shown in Fig. 6:

- **Scalability:** HDFS is scalable to petabytes or more of data and is flexible to add and/or remove of any number of data nodes in order to store and process huge datasets effectively.
- **Reliability:** When data is stored on HDFS, it divides the data blocks and are stored in data nodes in the Hadoop cluster. Block replication is maintained to provide reliability of information such that even if a particular data node goes down due to power or hardware failures, the data is accessible.
- **Fault tolerant:** One of the essential reasons in Hadoop platform is high degree of fault tolerance. There are probably possibilities of failures at Name Node, Data Node or network components. Quick Detection of faults and automatic recovery is the goal of HDFS. Replication of data makes HDFS reliable and fault-tolerant. By default, the replication factor used in Hadoop is three. Due to this replication, the Hadoop clusters are highly fault-tolerant.
- **Computation cost:** HDFS moves the processing code to the data nodes, which data localization is important concept of Hadoop that brings computations work closer to the node where the data resides and thus reduces network traffic and increase the throughput, making the data processing much faster.
- **Flexible:** Hadoop is a Java-based platform that can run on any operating system. Hadoop enables to easily access various data types includes structured, unstructured and semi-structured data formats. Hadoop can be used for a variety of purposes such as fraud detection, data warehousing, and trade analysis.
- **High throughput:** The amount of data moved successfully from one node to another in a given time period is throughput. Parallel processing of data makes it possible in reducing the time taken to send the data from one node to another and which achieves high throughput.

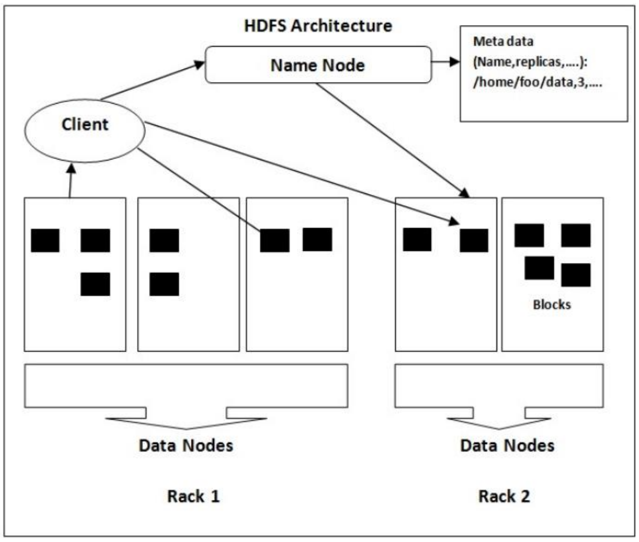


Figure 6: Architecture of Hadoop Distributed File System

PROCESS OF CLUSTERS IN BIG DATA

A gathering of the indistinguishable components firmly together is known as clustering. Data clustering is otherwise called bunch examination or section investigation which sorts out an accumulation of n objects into a segment or a chain of command. The principal point of clustering is to order data into bunches with the end goal that objects are assembled in a similar group when they are "comparable" as indicated by similitude, qualities and conduct. The most ordinarily utilized calculations in clustering are apportioning, various leveled, lattice based, thickness based, and model-based calculations. Apportioning calculations is called as the censored based clustering.

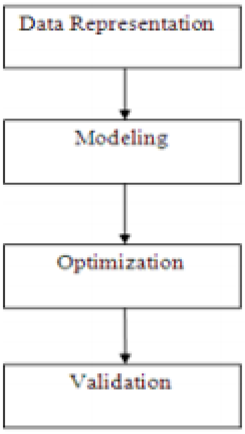


Figure 7: Processes of Data Clustering

CLUSTERING TECHNIQUES

Cluster Analysis: A primary purpose of clustering is to divide objects into groups and creates labels from data. Hence cluster analysis is sometimes called as unsupervised set of rules. And the clustering is typically divided into two principal kinds as partition primarily based and hierarchal based clustering. Requirements of unstructured textual clustering

- Reducing the high dimension of textual documents, an efficient text clustering techniques are required.
- Overlapping must be allowed between clusters because a document can cover numerous topics.
- Labelling a Cluster. Since cluster label shall provide adequate description of clusters
- Clustering algorithm need to discover the quantity of clusters by means of itself.

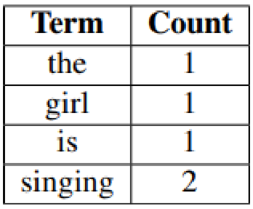
Clustering frequently takes the subsequent steps:

1. **Tokenization**: Tokenization is breaking the textual data into tokens. Generally used tokenization methods encompass N-gram method and Bag-of-words version.
2. **Stemming and lemmatization:** In this step mainly after performing tokenization, many individual terms repeats continuously and this can be avoided using method of stemming where the repeated same terms can be reduced.
3. **Stopword Removal:** After completing tokenization and stemming step, the terms which are of no use, carries no meaning or value, useless words can be removed, then the punctuation symbols can be removed. Then the cluster result produced after elimination of these steps will be better.
4. **Calculating term frequency inverse document frequency:** Once the data is cleaned for pre-processing, from textual document in order to produce features the common method used is finding term frequencies for entire generated tokens. Once frequency is calculated for all tokens it helps to give few clues regarding topic of the textual document. And the term frequency inverse document frequency mainly helps to find the importance of each term in a document from all textual corpus. The weight factor is generally used so that weight is assigned to every tokens. In term frequency, It mainly calculated the number of times the term occurred in the document, for example if we take two textual document d1 and d2, In document d1shown in Table 1,

In document d2 shown in Table 2,

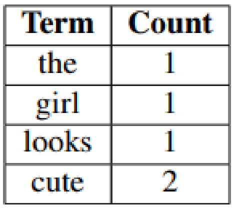
In document d1, the term ‘the” occurs one time, and in the document d2appeared one time, So, term frequency (“the”, document1) = 1 / 5 = 0.2

Table 1: Document d1



Term	Count
the	1
girl	1
is	1
singing	2

Table 2: Document d2



Term	Count
the	1
girl	1
looks	1
cute	2

Term frequency (“the”, document2) = 1 / 6 = 0.16

the term “the” present in entire documents are, idf(“the”,D)= log(2/2) =0, since the term frequency inverse document frequency is 0, thus it give clues that the term “the” is not much useful although it presents in entire textual documents.

tfidf(“the”, document1,D)=0.2*0=0

tfidf(“the”, document2,D)=0.16*0=0

Then we take another term like “cute”,

Term frequency (“cute”, document1) = 0 / 5 = 0

Term frequency (“cute”, document2) = 2 / 6 = 0.33

Inverse document frequency (“cute”, D) = log (2 / 1) = 0.30

So term frequency inverse document frequency (“cute”, document1, D) = 0 * 0.3 = 0

Term frequency inverse document frequency (“cute” document2, D) = 0.33 * 0.30 = 0.14

1. **Clustering**: From the above steps the feature generated are taken as input, based on the similarity all the term are grouped as a cluster using different clustering algorithms. The cluster quality is good if the similarity is high within clusters and less between clusters.

2. **Assessment and visualization:** Once the clusters are evaluated using different measures, It is every so often helpful to visualize the output by using plotting the clusters into low (two) dimensional space. Example Clustering,

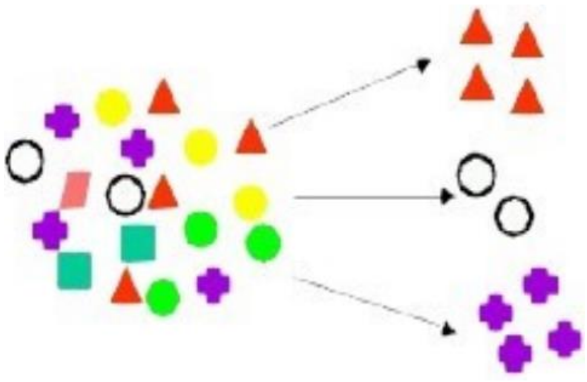


Figure 9: Clustering Example

Clustering is a very important data mining methodologies also called as unsupervised algorithm mainly performs automatic grouping of data which is called as clusters.

Some of the major algorithms for text Clustering are,

- Clustering based on Partition.
- Clustering based on Hierarchical method.
- Clustering based on density.
- Clustering based on Model
- Clustering based on fuzzy technique.

CONCLUSION

The comprehensive exploration of the characteristics and applications of Big Data underscores its pivotal role in today's data-centric landscape. As we navigate through the vast seas of data generated at an unprecedented pace, understanding the intrinsic properties of Big Data—its volume, velocity, variety, veracity, and value—becomes imperative. From healthcare to finance, from retail to manufacturing, the applications of Big Data are manifold and transformative. Predictive analytics, personalized recommendations, operational optimizations, and risk management are just a few examples of how organizations are harnessing Big Data to drive innovation, enhance efficiency, and gain competitive advantage. Overcoming these challenges demands a concerted effort towards developing robust frameworks, implementing stringent data governance policies, and fostering a culture of ethical and responsible data usage. One of the most common aspects of data mining is called clustering. It is employed in a wide number of fields nowadays. Researchers have developed a variety of different techniques for clustering data.

REFERENCES

1. BABU, G.P. and MARTY, M.N. 1994. Clustering with evolution strategies Pattern Recognition, 27, 2,321-329.

2. Fayyad, U. Data Mining and Knowledge Discovery: Making Sense Out of IEEE Expert, v. 11, no. 5, pp. 20-25, October1996.

3. Guo, G, Neagu, D. (2005) Similarity-based Classifier Combination for Decision Making . Proc. Of IEEE International Conference on Systems, Man and Cybernetics, pp.176-181

4. Jyothi Bellary, BhargaviPeyakunta, SekharKonetigari "Hybrid Machine Learning Approach In Data Mining", 2010 Second International Conference on Machine Learning and computing.

5. Oyelade, O. J, Oladipupo, O. O, Obagbuwa, I. C" Application of k- means Clustering algorithm for prediction of Students Academic Performance" (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7,2010.

6. Varun Kumar and Nisha Rathee, ITM University, "Knowledge discovery from database Using an integration of clustering and classification", International Journal of Advanced Computer Science and Applications, Vol. 2, No.3,March 2011.

7. McKinsey Global Institute (2011) Big Data: The next frontier for innovation, competition and productivity.

8. Chen, H., Chaing, R.H.L. and Storey, V.C. (2012) Business Intelligence and Analytics:FromBigDatatoBigImpact,MISQuarterly,36,4,pp.1165-1188.

9. Patel, A.B., Birla, M. and Nair, U. (2012) Addressing Big Data Problem Using HadoopandMapReduce,NIRMAUniversityConferenceonEngineering,pp.1-5.

10. Wu Yuntian, Shaanxi University of Science and Technology, "Based on Machine Learning of Data Mining to Further Explore", 2012 International Conference on Machine Learning Banff, Canada.

11. NeelamadhabPadhy, Dr. Pragnyaban Mishra and Rasmita Panigrahi, "The Survey of Data Mining Applications And Feature Scope", International Journal of Computer Science and Information Processing(CSIP).

12. Gandomi, Amir; Haider, Murtaza (April 2015). "Beyond the hype: Big data concepts, methods, and analytics". International Journal of Information Management. 35 (2): 137–144. doi:10.1016/j.ijinfomgt.2014.10.007. ISSN0268-4012.

13. Joshi, K., Khanduja, M., Kumar, R., Saxena, P., & Sharma, A. (2023). Big data-based clustering algorithm technique: A review analysis. *Automation and Computation*, 397-404.

14. Ahad M. A. & Biswas R. (2019) Handling Small Size Files in Hadoop: Challenges, Opportunities, and Review. In: Nayak J., Abraham A., Krishna B., Chandra Sekhar G., Das A. (eds) Soft Computing in Data Analytics. Advances in Intelligent Systems and Computing, vol 758. (pp. 653-663) Springer, Singapore.

15. Chen, M., Mao, S., Zhang, Y., & Leung, V. C. (2014). Big data: related technologies, challenges and future prospects (pp 1-89). Heidelberg: Springer.

Corresponding Author

Rajendra Mahto*

Research Scholar, Kalinga University, Raipur, Chhattisgarh, India

Email: rmahto2250@gmail.com