STUDY ON BIG DATA USING AND DATA MINING TECHNIQUES

Dr.Rakesh Kumar Giri,
Assistant Professor,
Department of Computer Science & Engineering,
Bharath University, Chennai, India

Abstract :In this day and age, when nearly everything is being digitalized, we deal with a vast variety of data types. These three corporations, Google, Microsoft, and Amazon, will each process a very large amount of data. Daily, these companies handled a large amount of data in their operations. Because of this, we need to find a way to tweak the technology so that it can handle all of the data in an effective manner. This presents a challenge for us, but we must find a solution. Big Data is a relatively new concept that defines innovative methods and technology that may be used to analyse large quantities of intricate datasets. The term "big data" was coined in the year 2010. These datasets are being produced at an exponential rate, coming from a wide variety of sources, and doing so at a variety of different rates. The discipline of big data analytics is finding that approaches involving data mining are proving to be of considerable benefit. Because dealing with enormous amounts of data offers considerable challenges for apps, this is something that really needs to be done. The ability to extract information that is useful from very big datasets is at the core of what is known as "Big Data analytics." The objective of this research is to conduct a literature review that covers a variety of topics, including the significance, challenges, and applications of big data in a number of different industries, as well as the numerous methods that are utilised for big data analysis by making use of data mining techniques.

IndexTerms **: Big Data, Data Mining, techniques**

I. INTRODUCTION

In this day and age, when everything is becoming digitalized, analysts have access to a massive amount of data right at their fingertips. Big Data is a collection of datasets that may be unstructured, semi-structured, or structured, and whose volume, complexity, and rate of growth make it difficult to capture, manage, process, or analyse the data utilising the typical database software tools and technologies. These datasets can be referred to collectively as "Big Data." In the year 2010, IBM was the company that first popularised the phrase "Big Data." There are many various sorts of data, some examples of which include text, video, photos, and audio, as well as web page log files, blog entries, tweets, location information, and sensor data. These are just a few examples. Utilizing analytical services, programming tools, and applications that are both clever and scalable is required in order to obtain insights that are of any value from datasets of this scale (Puneet Singh Dugga2013). Data mining is an analytical process that is used in a range of industries to search for significant correlations between variables in enormous data sets. Data mining is also known as Knowledge Discovery in Databases (KDD), which is another name for the technique. When swift and massive volumes of stream data are analysed, it is possible to find potentially helpful new insights and theoretical frameworks. Big data has the potential to aid businesses in boosting their operations and making decisions that are both more timely and smarter. This is one of the many potential benefits of big data. The practise of examining and analysing huge amounts of data with the intention of gleaning usable information from the analysis is known as "data mining." A number of databases store a lot of information, and that information may be extracted and used for a wide variety of purposes. This information can be used to retrieve a lot of data. For instance, a company that operates in the commercial sector may use the transaction data in order to figure out the best way to structure their sales in order to maximise their profits. Because of this, using data mining tactics at a company might potentially result in a rise in that company's level of revenue. The fundamental goal of data mining is to classify or make predictions based on the data that is collected. When it comes to categorization, there are many distinct situations that may be modelled and practised through simulation. For instance, when it comes to an advertisement for a product, it is feasible to ascertain which of the respondents are interested in the ad and which of the respondents are not interested in the ad. From there, ideas concerning the reasons why somebody could be interested in the advertisement might be generated. Data mining is going to be utilised, and it is expected that this information may be forecasted using that method.

II. DATA ATRIBUTE

According to the type of data can be divided into :

Nominal Attributes

The word "nominal" gives the impression of being "connected to names." Symbols or the actual names of the entities themselves most frequently serve as representations for the values of nominal qualities. Because each value indicates a category, code, or condition of some kind, nominal attributes are also frequently referred to as categorical attributes. This is because nominal attributes have values. The arrangement of the data does not follow any visible pattern at any point. In the field of computer science, the values that are employed are referred to as enumerations.

Binary Attributes

Binary characteristics are notional qualities that can only take on one of two potential values, either 0 or 1, and are hence classified as binary. In the vast majority of situations, a value of 0 denotes that the property is absent, whereas a value of 1 indicates that it is present. A binary attribute is deemed to be of the Boolean type if its two possible states correspond, respectively, to the truth value true and the false value false.

Ordinal Attributes

The term "binary attribute" refers to a nominal attribute that can only take on one of two potential values, either 0 or 1. A value of 0 denotes that the characteristic in question does not exist, whereas a value of 1 indicates that it does. If a binary attribute may only have two states, and those states correspond to true and false, then that attribute is regarded to be of the Boolean type.

Big Data

The term "big data" refers to data sets that contain a greater variety of components, are arriving in larger numbers, and are flowing at a rate that is continually accelerating. The term "the three Vs" is commonly used to allude to this idea.

Volume

The amount of data is a crucial consideration. When dealing with big data, it is necessary to process significant quantities of data that have a low data density and are not organised. This can be data that has an unknown value, such as the data feeds on Twitter, the clickstreams on a website or a mobile app, or the data acquired by sensor-enabled equipment. Other examples are tweets and clickstreams. Depending on the size of the company, this might amount to tens or even hundreds of terabytes of data. For some people, it may be hundreds of petabytes, while for others it might be much more.

Velocity

The speed at which data is received and (perhaps) acted upon is referred to as velocity, and it is measured in bits per second (bps). When information is sent directly to memory rather than being written to disc, the data transfer rate is often at its highest and most efficient. Certain internet-enabled smart gadgets perform their functions in real time or very near to real time, which requires the user to evaluate and respond in real time.

Variety

The term "variety" refers to the numerous distinct types of data that one may collect from a given set of sources. The conventional data types were well-organized and did not provide any challenges when it came to being stored in a relational database. The advent of new forms of unstructured data has been brought about as a direct consequence of the development of big data. Text, audio, and video are all examples of formats that may be classified as unstructured or semistructured data, and each of these types of data must undergo further preprocessing in order to derive meaning and supply metadata.

Data Mining

Data mining is a process that involves utilising large data sets to seek for anomalies, patterns, and correlations in order to produce accurate predictions. This approach may be used to make accurate forecasts.

III.DATA MINING TECHNIQUES

It is absolutely necessary to have an understanding of the data that is being processed in order to ensure that the results of data mining will be meaningful. Data mining techniques are frequently hindered by a variety of challenges, the most common of which is noisy data, which may include values such as null as well as values that do not conform to the norm (i.e. outliers). In response to shifts in the properties of the data that are intended to be mined, many enhancements to the data mining process have been created. These extensions include graph mining, which is used for mining data in networks; spatial data mining, which is used for mining spatial data; web usage mining and web content mining, which are used for mining users behaviours and specific topics over the web, respectively; web usage mining, which is used for mining spatial data; and most recently, big data mining, which is an evolved branch of big data analytics to fit different types of data (Richa Gupta,2014) .

Predictive Data Mining:

The predictive task makes use of certain data set variables or values in order to make educated guesses about the unknown or forthcoming values of a wide range of other variables of interest (Washio, Takashi2015). The following is a selection of the many different approaches that have been recommended for use in the process of prediction:

Classification

Finding out which category a freshly collected observation fits into is an important part of the effort involved in data mining. Given a training data set that comprises a large number of characteristics, where the identification of a model is completed as a function of the values of the other attributes, the given scenario is as follows: In order to accomplish this, a training set that contains observations that have been correctly recognised is required. The categorization is used to automatically place records into the many categories that have been set. For instance, it might be used to identify whether or not a credit card transaction was real or fraudulent. Additionally, it could be used to classify news items according to whether or not they relate to finance, entertainment, sports, and so on. There have

been many advancements made to the different classification systems. Nevertheless, decision tree-based methods (Mohammed J. Zaki,2015), neural networks and support vector machines (SVM), naive bayes classifier, and k-nearest neighbor (KNN) are the approaches that have been utilised the most frequently in the process of resolving problems that are encountered in the real world. This is because these methods have been shown to be the most accurate in predicting the outcomes of tests (Arinto Murdopo2016). The information that is anticipated may be used by methods that are based on decision trees to infer applicable rules, which enables the information to be used for the categorization of data. C4.5, which means for Classification and Regression Tree, ID3, which stands for Iterative Dichotomies 3, and CART are the three most frequent algorithms. CART stands for Classification and Regression Tree. The application of neural networks, which are also employed in classification owing to their capacity to extract meaningful information from complex data is utilized in the process of identifying patterns that are deemed to be too intricate for humans to be able to carry out successfully. Neural networks are used for this purpose because of their capacity to derive useful information from large amounts of complicated data.

The "neurons" that are comprised of neural networks have a structure that is comparable to that of the neurons that are located in the human brain. On the other hand, support vector machines, often known as SVMs, discriminate between objects that belong to various classes by defining decision boundaries in line with the concept of decision plans. On the other hand, the naive Bayes classifier is an easy-to-understand probabilistic algorithm that uses Bayes' theorem and makes the assumption that the characteristics have strong independent correlations among themselves. Bayes' theorem is used to make the determination of which characteristics belong to which classes. The k-nearest neighbour approach is yet another categorization procedure that is utilised often and extensively today. Using the votes cast by neighbours in a communal election, this technique chooses which class possesses the function that calculates the shortest distance between two points. The data item is then assigned to that class. In addition to this, there is a technique that is referred to as Repeated Incremental Pruning to Produce Error Reduction (RIPPER), and it is utilised for the classification of objects by utilising a sequence of if... then... rules. This technique was developed in the 1980s. This technique results in the production of a detection model that is composed of resource rules that are built in order to recognise potential occurrences of malicious executables in the future (A Min Tjoa, Iman Paryudi2016) .

Regression

Regression is a supervised mining function for predicting a numerical goal . Predictive data mining's opposite is regression (Anoop Verma2016). During training, the regression model calculates how to assess the target value depending on each data item's predictors. The goal value and predictors are linked in a model that may be used to analyse several data sets with uncertain target values. GLM is a common regression approach. It conducts linear regression on continuous targets (Ozer, Patrick2017). Linear regression, continuous dependent variable. Continuous or discontinuous independent variable(s). Logistic regression categorises binary target values .

Classifier Ensembles

Classifier Ensembles aggregate numerous classifiers to improve their performance. Combining many classifiers can achieve this (Han, Jiawei,2017). These classifiers can be based on several classification methodologies, resulting in varied degrees of effectiveness in categorizing people. Bagging is a bootstrap classifier ensemble. It a method for building an ensemble of models from bootstrap replicates. Random forest is a classifier ensemble consisting of numerous decision trees that output a node for each tree class . It provides an accurate classifier for multiple data sets and can handle large datasets. Rotation forest uses feature extraction to build classifier ensembles. To produce training data for a basic classifier, divide the feature set into k sections and conduct PCA on each. In most situations, key components are omitted to preserve data variability. Consequently, k-axis rounds are used to build the principal classifier's new features.

Descriptive Data Mining

The purpose of descriptive models is to get insight into how to approach future occurrences by analysing what happened in the data in the past. These models are able to comprehend previous performance by mining historical data to investigate the factors that led to either successful or unsuccessful outcomes in the past. It is possible to utilise this to quantify relationships in data in order to categorise things like consumers into assemblies. As a result, it is distinct from the other predictive models, which centre their attention on analysing the actions of a single consumer (Li, Deren, and Shuliang Wang.2014). Following is a list of the various strategies that have been inferred using descriptive models:

Association Rules Mining

It investigates links between variables in huge datasets. When it studies groupings of transactions, it finds rules that can anticipate an item's presence based on other items. Predictive models are rules. It may be used to place things in stores to increase sales, examine web server logs to learn about website users, and study biological data to uncover new correlations. Association rules mining techniques include FP Growth and Apriori . Apriori analyses criteria that, when met, result in high support and confidence levels.

Clustering

Cluster Analysis is unsupervised learning approach. This approach classifies comparable but separate objects into groups. Related publications in emails and proteins and genes with comparable functions are examples. Many clustering techniques have been created, such as the nonexclusive approach, in which data may belong to multiple clusters. Fuzzy clustering provides each cluster a weight between 0 and 1 for each data item. Hierarchical clustering, sometimes called agglomerative clustering, forms tiered, tree-shaped groups. K-means is a popular clustering technique. This technique employs a partitioned approach to split data into clusters with centroids. Similar data items cluster around their centroid. The K-medoids algorithm clusters data points like the K-means approach (S.Vikram Phaneendra2013).

Anomaly Detection

This method identifies outliers, or groups of data points that are significantly different from the rest. This approach finds outliers. Anomaly detection is used to identify credit card fraud, telecom fraud, network intrusions, and faults . Anomaly detection identifies fault-related fraud. It develops a pattern or gathers summary data on average population behaviour to find outliers. Anomaly detection may be graphical. This type of anomaly identification identifies unexpected network components (nodes, edges, and subgraphs) given the graph's structure. Statistical, distance-based, and model-based anomaly detection are also used. In the first two techniques, data is represented as a vector of features, and the model-based approach computes distance between data points. Model-based approaches assume a parametric data model.

Rough Sets Analysis

Rough sets analysis evaluates confusing and fragmentary information (Arinto Murdopo2012). Rough sets are crucial to knowledge discovery. Mathematical computations reveal hidden data patterns. It helps with data reduction, feature selection and extraction, and decision rules.

Optimization Data Mining

Optimization involves finding the most cost-effective or high-performing options within specified constraints. This is done by maximising desired qualities and minimising unwanted ones. Genetic algorithms solve optimization and search difficulties. Using simulated evolution, these methods "breed" computer solutions. A random population starts evolution. During each generation of the optimization process, each member in the population is appraised for selection in the next algorithm iteration. When a given number of generations pass or the population reaches a desirable fitness level, the algorithm stops iterating (Lim, A., L. Breiman2015). Data preparation uses data mining techniques. During this step, clustering and regression are used to remove outliers and smooth noisy data. Data sampling is essential before employing most data mining techniques. Sampling is one kind of necessary statistics. In data mining, sampling is done because processing the entire data set is impracticable and wasteful .

IV. EVOLUTION TO BIG DATA ANALYTICS TECHNIQUES

"Big Data" was originally used in a 1998 presentation deck for Silicon Graphics (SGI) titled "Big Data and the Next Wave of InfraStress" The first book to mention "Big Data" was a data mining book released in 1998. (Weiss and Indrukya 2014). This highlights big data mining's importance. Diebold's 2000 article was the first to utilise "Big Data" in the title. "Big Data" was initially used to characterise the everyday huge volumes of human-generated data. (2012) He presented remarkable internet use figures at the KDD BigMine'12Workshop. Google receives 1 billion inquiries, Twitter 250 million tweets, Facebook 800 million updates, and YouTube 4 billion views daily. Modern data creation is estimated to be in the zettabyte range, growing at 40% yearly. Mobile devices are going to produce a new vast quantity of data, and Google, Apple, Facebook, Yahoo, and Twitter are paying special attention to it to identify fascinating trends that will improve the user experience. When large amounts of data are analysed, analysts, academics, and business users can make better judgments using data that was previously unknown, inaccessible, or inappropriate. The tremendous expansion in data quantities has rendered the well-known data mining methodologies inapplicable to such massive volumes. As a result, many studies are being undertaken on how to improve data mining techniques to handle massive data, which has given rise to big data analytics. Big data analysis techniques include association rules mining and classification tree analysis. In this part, we analyse large data mining jobs. We also explain the innovations provided to accomplish such acceptance and the V dimension of big data handled by these changes. summarizes the studies on the move from data mining to big data analytics .

Grouping approaches by their data mining purpose. In the following table, the state of each strategy, including whether it handles big data analytics, is shown, along with the dimension of big data managed by each technique. The following sections describe how data mining techniques have been improved to handle enormous data and grow into big data analytic approaches. These upgrades made the approaches big data analytic. (XuShunxiang2016)

V. CONCLUSION

The exponential growth in terms of capacity and complexity of data that has happened over the course of the past decade has directly resulted in the considerable amount of research that has been done in the field of big data technology. This research has been done in a significant quantity. In this post, we have made an attempt to present a condensed summary of the most recent literature review in the topic of big data and its analysis utilising a number of different analytical approaches. This was a challenge for us, but we believe we have succeeded. This review is broken down into categories based on the years. Text analytics, which is considered to be the next generation of Big Data and is now much more frequently acknowledged as a mainstream analysis is used to get relevant information out of the millions of opinions that are published on social media. This can be accomplished by analysing the text of the opinions.

The approach of video, audio, and picture analytics has expanded with breakthroughs in machine vision, multi-lingual voice recognition, and rules-based decision engines as a result of the tremendous demand that exists in real time data of rich image and video material. This is as a result of the availability of high-quality picture and video content. They are examples of possible solutions to the economic, political, and social issues that have been discussed .

REFERENCES

[1] Puneet Singh Duggal and Sanchita Paul, Big Data Analysis : Challenges and Solutions.

[2] Wei Fan and Albert Bifet, Mining Big Data: Current Status, and Forecast to the Future, SIGKDD Explorations, Volume 14, Issue 2, 2013.

[3] S.Vikram Phaneendra and E.Madhusudhan(2014) Reddy, Big Data- solutions for RDBMS problems- A survey, IEEE/IFIP Network

[4] Operations & Management Symposium (NOMS 2010),Osaka Japan, Apr 19-23 2013.

[5] Sagiroglu, S. and Sinanc, D.,(2015) Big Data: A Review, International Conference on Collaboration Technologies and Systems (CTS), pp. 42-47, 20-24, May 2015.

[6] Richa Gupta, Sunny Gupta and Anuradha Singhal, Big Data : Overview, IJCTT, Vol 9, Number 5, March 2014.

[7] Ian H. Witten, Eibe Frank, "Data Mining Practical Machine Learning Tools and Techniques", 500 Sansome Street, Suite 400, San Francisco, CA 9411, 2015

[8] Bart van der Sloot, Dennis Broeders, Erik Schrijvers, "Exploring the Boundaries of Big Data", Amsterdam, 2016

[9] Nikita Jain, Vishal Srivastava, "Data Mining Techniques: A Survey Paper", eISSN: 2319-1163 | pISSN: 2321-7308, Rajasthan, India, 2013

[10] Nirmal Kaur, Gurpinder Singh," A Review Paper On Data Mining And Big Data", ISSN No. 0976-5697, Jalandhar, Punjab, India, 2017