Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Comprehensive Expansion in Big Data: Innovation and Technology

Suhiar Mohammed Zeki Abd Alsammed

Computer Science Department, University of Technology, Iraq.Email:110121@uotechnology.edu.iq

ARTICLE INFO

Article history:
Received: 19 /05/2019
Rrevised form: 03 /06/2019
Accepted : 09 /06/2019
Available online: 17 /06/2019

Keywords:

Big data, IT companies, Energy Management, Confusion Matrix.

ABSTRACT

In recent year, the use of the internet and cloud-based application has increased tremendously. With this lot of internet apps and social network have been developed. This apps and networks are growing very rapidly. Data on this network belongs to different data sizes and structures. This data in simple language is called big data. With the increasing demand for services, all the big data services are now automated, which stores and retrieve consumer's data. This data is in a different structure which makes it difficult to handle. This can include an example of road traffic data, different vehicle or person or location or atmosphere can produce a different set of data. Sometimes this data is structured while some times it may contain repeated, null or noisy data. This data may contain images and videos. The traditional database cannot handle this type of unstructured data. The best to solve this dimensionality problem is by using a dimensionality reduction technique. This paper provides the concept of big data collection, its analysis, storage and handling issues, security challenges, and other data handling technique.

1 . Introduction

Today, the use of the internet has increased very much, because of this increased rate of internet usages, a lot of data is getting generated every second. This data size is present in EB and PB that is Exabytes and Petabytes. It is predicted that by 2050, the size of the data will be much more than the brain power of the entire world. This data is growing because of the ease of the internet, satellite, sensors, communication channels. With increases size of data, lot many storage devices are also produced which can store a lot amount of data. Roger Magoulas is a person who gives the name 'Big Data' for introducing this technology [6].

A lot of social networks has evolved in the last decade; ease of internet brings almost everybody on social media which is creating a web of data. It is not just social media which is developing this data but many other sectors like the stock market, GPS sensors, Data collected by satellite, this technology has evolved to a great extent which is producing tremendous amount of data each second. This data is stored in a storage device because it contains a lot

Corresponding authorSuhiar Mohammed Zeki Abd Alsammed

Email addresses: 110121@uotechnology.edu.iq

Communicated by Qusuay Hatim Egaar

of information and can be very much useful in prediction. Historical data can be used along with some algorithm to predict future data. This data has proven a good track record of prediction using machine learning technology which includes stock market, Agriculture, Biology, Marketing, Weather Forecast and almost each and every sector [7].
Data Mining technology and Artificial Intelligence Algorithm, mines this data and finds the relations between all the data set, using this relation one model can be built which is called as machine learning model or Artificial intelligence algorithm. This Technology is useful in Big data analysis. Using these techniques two different datasets can be compared with each other to find similarities between them. This big data technology helps to understand any concept using dataset containing that subject. It also helps to find the hidden patterns present in the data and also help to determine missing values. Big data technology has some challenges too, which includes storage of data, as the volume of data is too high. This dataset is growing very fast, because of that handling and storing of this data has become a big challenge. Sometimes this data contains a redundant value which is not useful in manner, to understand this value it is necessary to visualize this data, and this rapidly growing data is becoming difficult to handle and visualize. This data visualization is a very important part of big data analysis.
Computer science is improving a lot with new technologies. this technology generates and gathers user data very smoothly. YouTube is the best example for understanding this thing. Many YouTube users upload their video on YouTube, which generates data of MP4 format. People watch this, and it gets stored in XML format in YouTube servers. Some users Likes and comment on the video which is stored in text format. In this way, a single source creates a huge amount of data with a different type. This is some most important big data challenges as the number of sources increases, the format, speed, and structure of data also change which create difficulty in storage and handling of data.
With respect to social media, now cloud computing and IoT (Internet of Things) both of these computer science technologies are growing very fast. In cloud computing, all the data, as well as services and functions, are stored on the cloud for ease of access. This layer is called an enterprise layer. While in the case of Inter of Things, sensors and physical devices are present which collects the data and sent it to cloud for analysis of this data. This sensor generates data even for a minor change. It is capable of detecting minor changes and sending it to underlying servers. Because of this huge amount of data gets generates which exceed the limit of capability of computers to handle it. This increasing data has become a major challenge to store, analyze and retrieve the heterogeneous form of data [2].
This big data analysis survey briefly explains each aspect of big data analysis. The flow of this literature survey is as below. Chapter II is an explanation of big data analysis concepts and Application of big data analysis. Chapter III consist of different technologies present to handle big data and to extract information from it. Chapter 4 Contains the challenges present in big data analysis and Chapter V contains big data and machine learning algorithm for analysis of big data, the conclusion of the literature survey and future work.

2. Overview

Financial Institutes, Engineering, Government, agriculture and every filed in using big data technology to analyze datasets. Data has now become an important aspect of every field. With a good big data analysis technique, one can expand their business to the right clients or can do their research on specific points instead of wasting their time on other aspects of technology. This data is too important but strong this data is also a big problem, it is not possible to store this data in Traditional SQL based database, as they appear in a different format, it becomes very much difficult to store in a SQL database. Data coming from mobile text messages, YouTube, Sensors, and GPS are entirely in a different format from each other, along with that velocity of speed is also different, some data get generated very often while other is continually getting generated. This also impacts the handling of big data. This data can be in incomplete, unstructured, semi-structured, heterogeneous or multi-dimensional.

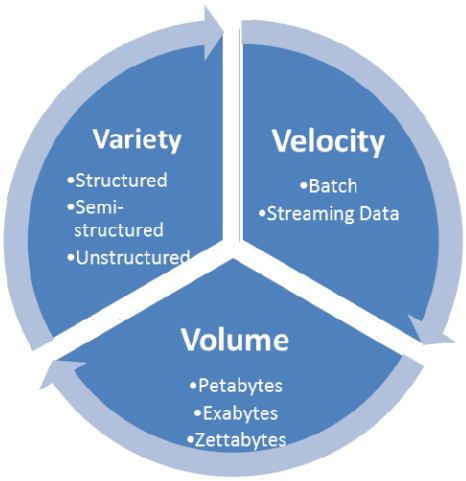

Figure 1: V in Big Data

Doug Laney is a big data scientist. he always uses 3 V's in his big data analysis. In 2000 he introduces this 3 V [6].

1) Volume (size of data): with improvements in computer science technology and the internet, people are most likely to use cloud platform and internet-based services. Sensors, social media, financial markets are producing a large amount of data on each moment, which is a cause of increased data volume.
2) Velocity: Speed of data is a major factor while handling it. Streams of data appear in different speed and it is completely depending on the source which is cresting this data. The sensor creates data on event triggers which make the velocity different for each data stream, some mechanical devices have the facility to handle this problem by using a timer, but it is not possible in every case.
3) Variety: Big data has different suture or format has it comes from different sources it is quite acceptable. Data from Social media, form Video Camera, from sensors are different in nature, data can be in text format, can be video or audio format, data from sensors are in different formats, or it can be the encrypted format. this makes data heterogeneous and this is called as a variety by Doug Laney.

All major research centers, IT companies, and Research firms have already started using big data approaches in their research and marketing strategies. To extract the meaningful information from big data, Processing power is required, a good machine learning algorithm and Skiles are also an important aspect while extracting information from big data. That is why it is necessary to handle 3 V's of big data which are velocity, volume, and variety.

The process of extracting meaningful information from datasets is known as KDD (Knowledge discovery in databases.) Fayyad and his teammates have introduced the KDD process. It contains several steps for extracting Knowledge or information from data. These steps include data collection, data selection, data pre-processing, data transformation, model creation, and interpretation. This steps collectively are able to collect the data from different sources and extract hidden data or knowledge from the given datasets. It is also possible to visualize raw data, pre-proceed data, and model formed by datasets.

Data processing in nothing but the collection, modification, and processing of data to form a machine learning model which will be able to produce a prediction for new inputs. Karmasphere is a big data analyst who splits the task of big data analysis into 4 A's. This is Acquisition, Analyses, assembly, action.

1) Acquisition: This is the first step of big data analysis, in this step data is collected from different sources. We have seen earlier that data is present in different forms and in a different structure. That is why a proper data collection technique is important for web data, the crawler is one technique which is used. Other techniques may include user own responsibility to submit data in form, or sensors or even web camera is also a way to acquire data for storage.
2) Assembly: In this stage of data mining, data is already collected from a different source, now the storage is a problem in data analysis, that is why it is necessary to remove unusable and redundant data. This all task has been done in this stage, this stage also includes data transformation in which data in transformed into the different structure so that all variance of the data comes into a singularity, which reduces the size of data, also it helps to visualize data and to create data mining model.

3) Analyze: In this stage, data should be trustworthy and appropriate. This is the most important step in KDD. In this stage, a different algorithm is applied to big data. This algorithm can include classification, clustering, regression. After applying this data to the algorithm, it gives a better inside of data. An algorithm which is giving better result is selected as the best algorithm for the given data. so that with the underlying algorithm, a new training model is constructed which can create relationships among the data as well as it can predict the missing values for new input. The construed model is called a data mining model.
4) Action: In this stage, new inputs are applied to a trained model so that missing values or missing classes from that dataset is predicted. This stage depends on user requirement, whether the user wants to mine new information or whether the user wants to find the relationship present among the current data. Before using any trained model for data analysis, it is very important to find and compare its accuracy with other trained model based on a different algorithm and different pre-processed data.
5) Privacy: Privacy is an important aspect while data collection. Because if data mining needs consistency then it should acquire all user data for predicting better and real-time accuracy. but this may invade user privacy. So that it is equally important to maintain user privacy and create the best training model.

2.1 The infrastructure of big data analysis

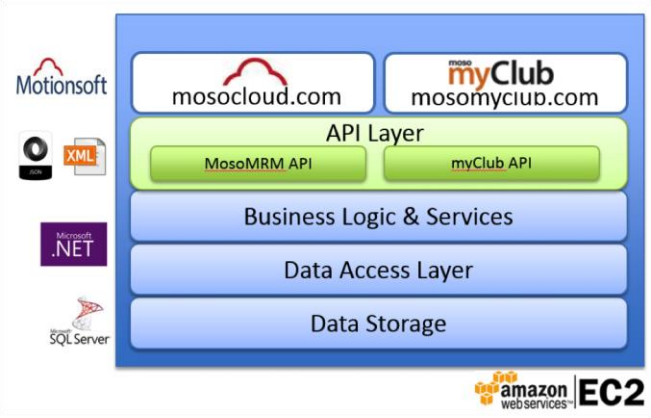

Figure 2: Infrastructure of big data analysis

1)Data Layer: Big data present in a different format, so it not always possible to store it in tradition SQL databases. Proper data handling techniques are required for storing data. There are different databases present which can handle unstructured, semi-structured or heterogeneous data at the same time. This may include Mango DB and Cassandra. This data is collected from the web, social media or sensors. Hadoop technology like Map reduce supports data layer. Along with that some software like HBase, Hive, the spark is also present in the data layer.
2) Analytics Layer: This layer contains some services which can create a model from big data. Input to this layer is clean data, and output is a machine learning model. This Layer consist of services like Python and R, which are capable of producing a machine learning model. This is the backbone of entire big data analysis.
3) Integration Layer: This is a connecting layer. This layer takes new input from the user, send it to analytics layer where a machine learning model is present, based on the algorithm it produces the result and sends back to the upper layer, which is Decision layer. This layer acts as a service to the user application. This can also be termed as API Layer where user app can connect with big data analysis system.
4) Decision Layer: This is the topmost layer in big data analysis. This is the point where the user actually gives input to the system. This may include Desktop Applications, Web Services, Websites or Mobile Application. Integration layer plays an important role in connecting Decision layer with the system.

2.2 Big Data Application

Big data has so many applications. With the ease of big data analysis, its Applications are also increasing. few of the big data application is present below:

1) Financial Market: Financial markets like the Stock market daily produces a lot of data which contain stock prices and company fundamentals, this data can be collected easily using exchange API. Many researchers have found that this data is capable of producing a machine learning algorithm and hence it can be very much useful in determining future stock market prices[5]. It has become very simple to find stocks with the same chart or data using big data analysis tools present in the market [15].

2) Agriculture: For the farmer, it is a difficult task to determine the correct level of water, temperature, humidity or chemical fertilizers require for a specific plant. If they store this record and collect it from another successful farmer who is planting that plant. By collecting this data, it becomes easy to understand the best possible environmental condition require for plantation of that plant [3]. This helps farmer for farming with improved performance.

3) Social media analysis: There is a lot of user data present on social media. This data is very useful in the analysis of a certain trend in publication. Privacy preservation is also an important factor here. User shares their like and dislikes on social media, this information is useful in finding people's choice about certain things, like movie, Game or anything trending [16].

4) Advertisements and Marketing: Data present on websites, or on social media are useful to find clients for a specific product. If the data is present, it becomes possible to find one's likes and dislikes, by which it becomes easy to decide wheatear that person is appropriate to target for that product or not. Big data analysis has proven the successful implementation of a system which finds potential target [17].

3. Big Data Technologies

Big data technologies involved collecting, analyzing and prediction data. For these purposes, there are different tools present for big data analysis. This tool can collect data, perform pre-processing, it can storage and find hidden patterns present in it. This section contains different tools which can be useful for this task. Some of these tools are described below:

A. Data Analysis

1) Hadoop: Hadoop is an open source. Hadoop has interactive tools which make data analysis task very easy. Data can be collected and visualize on the front end of Hadoop, which makes data visualization further easy as a user don't need to write long code for that. The different machine learning algorithm is also present in Hadoop which are useful in creating a machine learning model for finding hidden patterns [18].

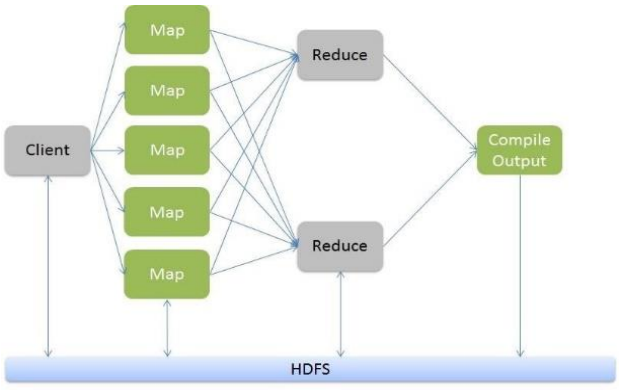

Figure 3: Hadoop – Map Reduce Architecture

2) Map Reduce: Map-reduce is a part of Hadoop where uses can program to convert a large task into small one so that they can speed up the task. Map reduce has 2 important parts including mapping and reduce. Mapping converts the data into value. In reduce task, the dataset is converted into reduces tuples to minimize the time required for that task [12].
3) Hive: Hive is much similar to SQL frontend. In Hive, users can write query just like SQL databases. This query can work as program and provides the required output to users. Hive is also an open source project. SQL query like structure makes it simple for using to users [13].
4) PIG: It is much similar to Hive. PIG programming is much similar to Perl. Users can execute queries on Hadoop through PIG.

B. Big Data Storage

Volume and velocity of data is a big concern. For storing heterogeneous data different type of storage technologies are used. These technologies have evolved the functionality of data compression and complete visualization of data. Few of these techniques are present below:

Data Warehouse Vs Data Lake

There are different storage forms of big data. Mainly this is known as Data warehouse or DWH. Similar to the data warehouse, a data lake is another approach to store data.
Although both of them are utilized to store big data, they have many different approaches to storing and retrieving data. In the case of data warehouses, a relational database is used.
All these databases have rows and columns as it is present in table format. Data warehouses also contain schema similar to SQL Databases. Data is being stored in the way defined in the schema,which makes it very simple to store and retrieve the data. When it comes to the analysis part, Data warehouse is a little bit more time consuming as data lakes, because it is bound to schema [28].
Alternatively, a data warehouse is more flexible as there is no schema or structure, data can be easily analyzed. Scalability of data lakes is much better than Data warehouse. Data lakes have no structure, storing data in data lakes are similar to throwing all the data without any form. This ends up with bad practice storing data. Retrieving this data becomes difficult and this problem is called data swamp [25].Data present in data lakes is like miscellaneous data. because of less structured form, a data warehouse is preferred over data lakes.
1) HBase: Heterogenous nature of big data makes it not possible to store data in traditional SQL. HBase is a NoSQL database. It can store multi-dimensional data as well. HBase is an open source database, where user can store the data and they have read and write access to stored data. It is mostly used with Hadoop with YARN engine [11].
2) Sky Tree: Sky Tree is much similar to HBase, along with that it has machine learning tools, which makes big data analysis task much smoother. Big data volume is too high because it is not possible to do all big data analysis steps manually. Sky Tree helps in automated data collection and Machine learning.
3)No SQL: Heterogenous nature of big data make it not possible to store data in traditional SQL. No SQL support heterogeneous data. It is also called a Not Only SQL database. It is built on Cassandra platform. No SQL is also an open source database.

C. SQL vs NoSQL Databases

Table 1: SQL vs NoSQL Databases.

	SQL	NoSQL
Structure of data	SQL databases are relational (RDBMS)	NoSQL databases are distributed or non-relational
Form	It is in the form of Table.	It is in the form of documents, tree, graph or key-value pair
Format of database	They Have a specific format	NoSQL databases have no Format [29].
Schema Type	SQL has a pre-defined schema	No Predefined schema, it has Dynamic schema type.
Query	SQL query focused on Table and its relations.	NoSQL query is complex in structed, and it varies from databases to databases.
Scalability of System	Scalability can be increased by hardware capacity	Scalability is increased by increasing Server pools.
Data Manipulation Technique	SQL uses structed Language for Data manipulation.	UnQL((Unstructured Query Language) is used in case of NoSQL.
Handling hierarchical data	SQL databases can't store hierarchical data	NoSQL database can store, manipulate and retrieve hierarchical data [30].
Example	MySQL, Oracle, SQLite, Postgres and MS-SQL	MongoDB, Redis, Cassandra, Hbase and Raven DB [31].

D. DATA Visualization Tools

It is a good habit to visualize data before doing pre-processing and after pre-processing. It gives a better understanding of the data. There are plenty of Data visualization tools are available.

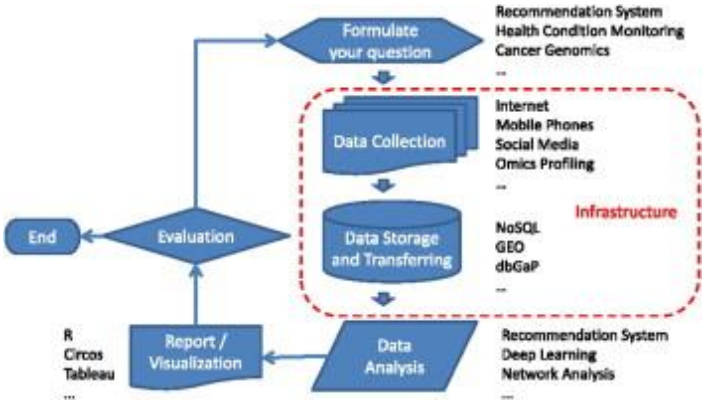

Figure 4: Big Data Analysis

1) R Tool: R is an open source project which is much similar to python. R is basically made for mathematical complexity reduction in big data analysis. Using R, the researcher can focus on the concept of data instead of wasting their time on math presents behind the data. It has many open source libraries presents which makes the computational task much simple. R library is present for data collection, pre-processing, Calculations and data mining. So, it has all the functionality needed for big data analysis.
2) Weka: Weka is based on Java. It is also an open source project. What makes Weka much popular is its simplicity. Even the person with zero machine learning knowledge can also work on Weka. It has complete user interface tools so there is no need to write any sort of code for data analysis. As the user interface doesn't allow to write many scripts, complex calculations are not possible on Weka, this is its limitation.
3) Tableau: a Tableau is a desktop software, it supports data collection and data handling. In Tableau panel, the user can visualize the data. For more simplicity data visualization is available in chart, bar, graphs and maps format. User can select any preferred option. This is good for small scale data analysis project.
4) Infogram: Infogram is much similar to Weka tool. It makes data analysis very simple. User can do an analysis of data in 3 simple steps. Like Tableau users can visualize the data in the form of charts, bar, graphs or maps. With that, they can share their project on Infogram platform itself. It has different channel present where students and researches can publish their project so that other people can understand and enhance that project.
5) ChartBlocks: ChartBlock is online data analysis tools. It involves no complex coding. User can use this tool for free of cost. It has the functionality to extract datasets from the database, excel, and spreadsheets.
6) Ember Charts: It is based on a JavaScript framework. Ember.js is used to build Ember Charts platform. Like Tableau users can visualize the data in the form of charts, pie, graphs and time series data. It is very simple to use.
7) Tangle: This support scripting. So, if the user wants to write their own script for complex calculations based on their project demand, then Tangle is the best alternative to R. As it has less community than R, many new types of research use R for better support.

E. Challenges in Big Data

There are many challenges when it comes to collect, store and analysis of big data. While Big data analysis it is very much important to handle this problem. Many data scientist has observed the following challenges in big data.
A. Storage: As described earlier, the volume of big data is increasing rapidly. It has become very difficult to store this data in servers and databases. This data is currently present in Terabytes. This is really a huge amount of data; current computational power is not capable of handling and storing this data. It is expected to grow further with the increased use of the internet. Databases were able to handle a large amount of data, but that was in homogenous form. But in the case of big data, it is present in the heterogonous form [23]. No SQL databases are able to handle this heterogeneous form of data. Mongo DB and Cassandra are databases which are able to store heterogeneous data [4].
B. Data representation: In data visualize tools sections we have seen various visualization tools and its importance in big data analysis. This data visualization tools are capable of handling data of different format, but not all. Videos and encrypted data handling are not possible for data visualization tools. While doing data analysis it is important to have in deep knowledge of data, for that reason data representation is an important step. Heterogenous data representation is not possible from current data visualization tools.
C. Discarding outdated Data: Big data analysis has its first step to collect data, and then data analysis. but many times, previously collected data becomes outdated and that sentiments changes with time. This data should be replaced or discarded as they will increase the overhead of data storage and data handling. This data also decreases the accuracy of the algorithm. Finding this outdated data is another challenge in big data [4].
D. Analysis: Big data analysis is done by different machine learning algorithm like classification, clustering, regression. After data pre-processing it is not an easy task to select a perfect machine learning algorithm to select for that data. Because of this time required for data analysis increases. Also, it is not always the case that data scientist picks the correct machine learning algorithm. This process is handled by data visualization task. As a researcher comes to know the structure of data, it becomes flexible to find a probable algorithm which can give better results [8].

F. Reporting: Output of the data analysis is represented to the users in the form of charts, graphs or values. This process is called as Reporting. In the case of User, test data is too high, then it becomes a challenge to present all outputs for the test data to the user, as it is difficult to understand to users. In this kind of cases, outputs can be represented in terms of graphs or charts [8].

G. Redundant Data: In data collection step of big data, data is collected several times. This may be possible that repeated data is collected over time, because of this repeated data, the volume of dataset grows very rapidly and this is not very much useful in several cases. This data also increases the overhead on the machine learning algorithm while creating machine learning model. This redundant data should not be collected or if it is present in databases then it should be removed to save storage and processing time.

G. Energy Management: Large storage devices and high computational power increases the demand for electricity requires for big data management. Big data technology is rapidly growing, and it will grow further which will require a lot more energy. This is necessary to create energy efficient big data technology to conserve energy.

H. Data Privacy: In the data collection stage, data is collected from different sources, this may include forum, personal blog or social media channel. Few users don't want to share their data to data collection tool [19]. This is a privacy issue and it should be handled by data collection tool. Private data of any person should be safe, or this will become illegal or unethical. It is the responsibility of the system that it should conserve the user's privacy [20].

J. Scalability: Big data has many dimensional and volume of data may be large or small. The velocity of data may be fast or slow. It is necessary that the system should be able to handle these situations. A scalable and expendable system is very much necessary [10].

K. Dimensionality Reduction: Visualization is an important step in big data analysis. Dimensionality reduction is required for clear understating of data and data visualization. Dimensionality reduction involves removing redundant data. There is a certain problem in Dimensionality reduction in big data analysis. In big data, there are lot many datasets which are redundant or unrelated with respect to other approach or they are uncorrelated with data, this data set should be removed from big data, this procedure is known as dimensionality reduction in big data analysis [9].

After Pre-processing of data, the next step is to select the feature and extract it from data. After removing redundant data, Visualization of data becomes simple and handy to end users. Another benefit is storage reduction; as redundant data is removed storage require also gets reduced. Machine learning algorithm also performs the operation with better time complexity thereby reducing time in data analysis and machine learning step. There are different dimensionality reduction methods:

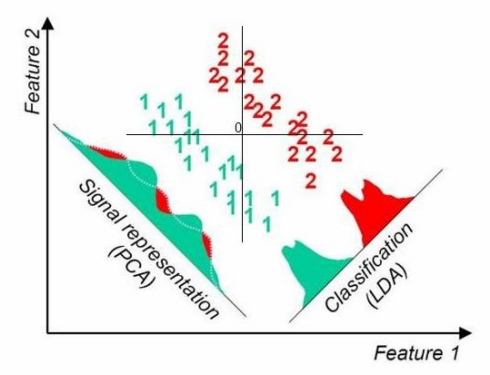

Figure 5: Dimensionality Reduction

1) Eliminating data column of Null Values: some databases have many formats of data, some of which appears very often, and they do not have significance in data analysis. This dimension often has null values or missing values. This

column or features should be pulled out and eliminated from the dataset. Process of removing column or dimension with null value is a good solution for dimensionality reduction.

2) Low variance mesh: It is useful for dimensions with numeric values. In this method system scan for the entire column and find out the values which have low variance. It is computed by a threshold value. All such dimensions who have less variance are removed from data.

3) Reduce highly resembled dimension: In data visualization step it is observed that two or many columns have identical values. It is quite considerable to remove such a column which has an identical column. As both have the same values, one column's value should be enough for data analysis. Association is a technique to find out relations between two data columns, if the association is more than the threshold then it can be considered as a resembled column. This is also known as Linear correlation node. All such nodes are reduced to save storage utilization.

4) Principal component analysis (PCA): Principal component analysis is a data transformation technique which is very popular in dimensionality reduction. In PCA original y coordinate of the dataset is copied to new y' coordinate called as principal coordinate [24].

5) Tree ensemble method: In the data analysis task, classification tress is prepared which contains all the data present in datasets. It is found that it creates a problem of data overfitting. This can be controlled by the ensemble method, in which the use of each attribute present in the dataset is calculated, and based on this statistic, attributes which have less influence in analysis task is removed.

6) Feature Elimination: it is based on a machine learning algorithm. In backward feature elimination process, training is performed on all the inputs, then its accuracy and error are calculated, then one feature from the whole list is removed. Again, the whole training procedure is carried out on all feature except that one, this step is followed for all the features present in datasets. The error rate for each iteration is calculated for all the loops, feature with a value less than the threshold is Removed [14]. After this step whole procedure is carried out for remaining attributes from the data. In this way, a minimum number of features required for maintaining the performance of the training model is kept and another feature is removed. This is algorithm-based feature elimination technique.

4.Data Mining Algorithms

Data mining is one of the important steps in big data analysis. Data mining is the task of selecting a feature and creating a training model based on that feature, this training model is capable of creating an algorithm based on the correlation of data present in the datasets. There are different approaches used in data analysis, which involves classification, clustering or regression. Which technique should be used is depends on the data type and its correlation [21]. Data visualization provides meaningful information about algorithm selection. Multiple algorithms are applied to the datasets. Accuracy and error rate of each Algorithm is summarized to select best suitable data mining algorithm. Few machine learning Algorithm is as follow:

1. Clustering: Clustering is the machine learning method in which data with similar features form a cluster. This process is called clustering. There are different types of clustering based on the method of forming clusters. Few of them are Centroid-based clustering, density-based clustering, and distribution-based clustering. K-means algorithm is a centroid-based clustering algorithm. In K- means, K numbers are cluster are created based on feature selection. The distance measure is used in the machine learning process. This distance measure is Euclidean distance, F-measure and Jaccard index [1].

A. Euclidean distance measure: In Euclidean distance measure the distance from each point from one cluster to another point in other cluster is found out, square root of summation of their distances is the way to calculate Euclidean distance.

Formula:

$$\text{dij} = \sqrt{(xi1 - xj1)^2 + (xi2 - xi2)^2 + \cdots + (xip - xip)^2}$$

B. F- measure: It uses to balance the condition of false negative votes. It uses precision and recalls in its calculation for finding measure.

C. Jaccard index: It is a numeric value. Range of Jaccard index is from 0 to 1. Jaccard index is used to find out the similarity of two datasets. Jaccard index is calculated by the following formula, Jaccard index 0 means there is nothing common in two datasets, while if Jaccard index is 1, it indicates that both the datasets are identical to each other. Jacccab Index of two datasets A and B is calculated by:

$$J(A, B) = (A \cap B) / (A \cup B) = TP / TP + FP + FN$$

TP - True Positive rate, FP - False Positive rate, FN- False Negative rate.

K-means Algorithm

Most Popular cluster algorithm is a K-means algorithm. In K-means, K number of clusters are formed by using Euclidean distance measure. This process is iterated for all the points present in the dataset until distance becomes less than threshold values.

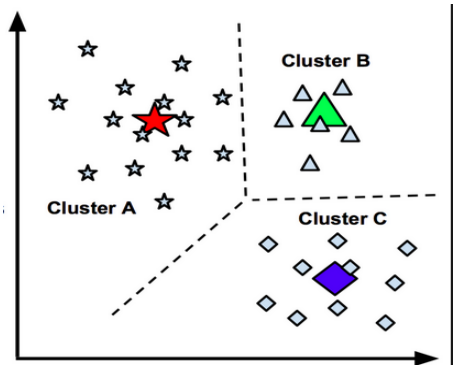

Figure 6: Clustering.

2.Classification: Classification is also a machine learning technique; Classification can be supervised or unsupervised. The supervised classifier has trainer, while unsupervised classifier doesn't require trainer unsupervised classification, output class is defined by classification Algorithm. Classification algorithm contains decision trees, Naive Based or K- nearest neighbour classifiers.

Decision Tree: Decision tree takes data as input and divides it's in the form of tree, node as a most important feature and then follow the iterative step until all the nodes are used to create a decision tree. A C4.5 decision tree is the most common decision tree; it is also known as a J48 decision tree. It has a problem of data overfitting and extended tree size as all nodes are used in the construction of the tree. Tree length is reduced by tree pruning technology in this all the leaf or similar node which has no significance in training procedure are removed. The problem of data overfitting is improved in Random forest algorithm. In Random forest algorithm, many decision trees are created and voting taking is used to build one decision tree which does not have data overfitting issue.

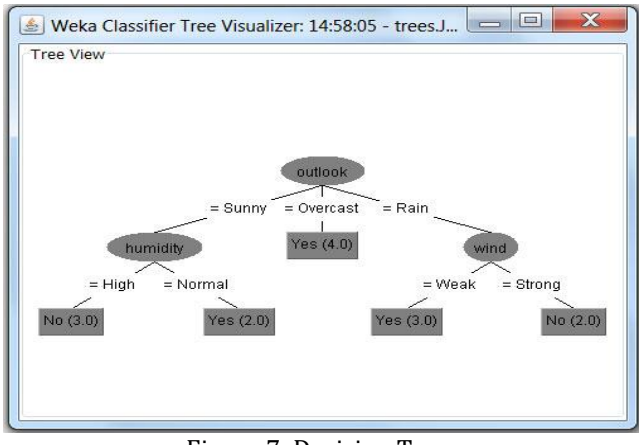

Figure 7: Decision Tree.

Boosting and Bagging: It is a technique in which dataset is distribution in different bags of data, few bags are used as a training sample and rest bags are used for testing purpose. accuracy for such bags are measured and bags with a high accuracy rate is considered for training purpose. Bagging and boosting help to remove the effect of data overfitting.

Table 2: Enhancements in Classification Techniques.

ID 3 Algorithm	C 4.5 Decision Tree	Random Forest
Invented by Ross Quinlan	Invented by Ross Quinlan	Invented by Leo and Adele Cutler.
It uses Entropy and Information Gain for classification.	C 4.5 is based on the same principal of ID3.	The random forest creates n-number of trees.
Size of the Decision tree is Large	Size of the Decision tree is Small	Size of the Decision tree is Small
Bagging and boosting are not supported.	Boosting feature is added in C4.5	Bagging and Boosting is base of random Forest
ID3 is slower than c 4.5 because of tree length	C4.5 is faster than ID3 and Random Forest	The random forest creates n number of trees, which reduces performance and speed.
It is basic Decision tree based on Information gain calculation.	Tree pruning and boosting techniques are added	Creates multiple trees to reduce the problem of data overfitting.
Accuracy is measured by Confusion Matrix	Accuracy is measured by Confusion Matrix	Accuracy is measured by Out Of Bag Error rate.

3.Association Rule Mining: In Association rule mining, data correlation is found out. Different and similarity measures are important aspects of association rule mining technique. Most popular association technique is the Apriori algorithm. Apriori Algorithm is used for databases which have very large datasets. It is used to find out the association and correlation of variables present in the databases [22].

Finding the Accuracy of big data analysis

After the creation of a machine learning algorithm, it is necessary to find out its accuracy of other dataset and on the real-time dataset. There are different techniques used for finding the accuracy of the trained model. This includes confusion matrix, out of bag error, True positive and false positive rate, and testing data on a real-time scenario. a real-time case study is very important to test case, as confusion matrix and other indices may give false test result because of data overfitting. but real-time tests are always a true test result.

1. True positive, false positive rate and confusion matrix:True positive rate is also known as sensitivity; it is the actual number of correct output to the number of all output. Precision and recall are summarized by True positive rate. Just like TP rate, False positive rate is an incorrectly classified instance to the number of instances present in the complete datasets.Confusion Matrix is a table format structure with all test result of the classification. Confusion matrix has a matrix of n number, where n is a number of the feature present in the datasets. A confusion matrix is used to calculate True positive, false positive, True negative and a false negative rate of test cases. Thereby it becomes simple to find sensitivity, precision, and recall of test cases.

Precision = | {relevant document} ∩ {retrieved documents} | / | {retrieved documents} |

Recall = | {relevant document} ∩ {retrieved documents} | / | {relevant documents} |

2. Out of bag error estimation: Data overfitting issue of the decision tree is handled by ensemble technique, bagging and boosting. A random forest decision tree is an example of ensemble decision trees. In all this technique K-number of the decision tree is constructed, where k is defined by the user. a confusion matrix is not capable of finding the accuracy of the ensemble decision tree. Out of bag estimation is a technique to find out the error rate of

ensemble classification techniques [15]. In a certain bag of data, variables which are producing the wrong output is measured as an error in OOB estimation.

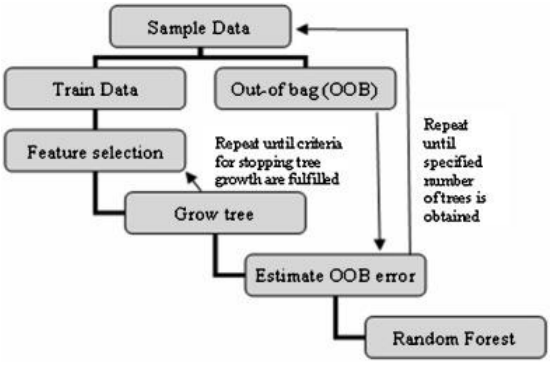

Figure 8: OOB Estimation

3. Test with real-time data: Out of bag estimation and confusion matrix uses some data set from training data, to test the result and find out the accuracy of the training model. As the test data is part of training data, such accuracy may get changed when test data is different. That is why it is better to practice to find the accuracy of the training model with respect to real-time data, before deploying any service to clients. Data has variance in its nature. data format gets change any time so that real-time accuracy is considered as best accuracy rate. In the case of many algorithm precision and recall comes very positive but when it comes to real-time datasets, the accuracy of the model is not up to the mark. Real-time data should be used for testing, if the accuracy is less than expected then algorithm and pre-processing technique should be changed.

Conclusion

In this literature survey, big data analysis is explained with data collection, data pre-processing, data analysis and accuracy calculation. Big data technology is evolving with huge volume. This dataset is a very good tool to analyze the relationship among the data and they're by predicting the future. Current computer technology is not completely capable of handling the challenges of big data analysis. New technology and algorithms are developing which has built in such a way that they can handle a large amount of data. Many schedule algorithms are presented which handle big data and analyze the structure and relationship among the variable present in the data. Data privacy is a big concern in the data collection task, which is described in the paper. Data privacy should be maintained so that people can put their views on the internet without any issue. Some data analysis techniques like clustering and classification are presented for increasing the accuracy of the training model and reducing the processing time for the creation of a machine learning algorithm. Finally, methods of accuracy calculation are followed to find the error rate in the training model.

References

1. The survey on approaches to efficient clustering and classification analysis of big data Bhagyashri S. Gandhi ; Leena A. Deshpande Platforms for big data analytics: Trend towards hybrid era AlkaLondhe ; PvrdPrasada Rao 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)
2. A survey of recent technologies and challenges in big data utilizations JieZhang; Xiao Yao;GuangjieHan;YiqiGui 2015 International Conference on Information and Communication Technology Convergence (ICTC)
3. Big Data in smart farming – A review - SjaakWolfertab,LanGea,CorVerdouwab,Marc-Jeroen Bogaardta-www.sciencedirect.com-Volume 153, May 2017
4. A survey on various challenges and aspects in handling big data S. Pradeep; Jagadish S. Kallimani 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)
5. A Survey on Big Data in Financial Sector AnkuJaiswal ;PurrushottamBagale 2017 International Conference on Networking and Network Applications (NaNA): 2017
6. A Survey Paper on Big Data Analytics - Anto Praveen a, B. Bharathi - INTERNATIONAL CONFERENCE ON INFORMATION, COMMUNICATION & EMBEDDED SYSTEMS (ICICES 2017)

7. CheikhKacfahEmani, Nadine Cullot, Christophe Nicolle, Understandable Big Data: A Survey, Computer Science Review, 2015, Vol: 17, pp: 71-80
8. Open research challenges with Big Data — A data-scientist's perspective Sreenivas R. Sukumar 2015 IEEE International Conference on Big Data (Big Data) Year: 2015
9. Jun Yan, Banyu Zhang, Ning Liu SquishingYan , Effective and efficient dimensionality reduction for large scale and streaming data preprocessing, IEEE Transactions on Knowledge and data engineering, 2016, Vol:18, issue:3
10. Toward Scalable Systems for Big Data Analytics: A Technology Tutorial Han Hu;YongkangWen; Tat-Seng Chua;Xuelong Li IEEE Access
11. Hadoop-HBase for large-scale data - Mehul Nalin Vora -Proceedings of 2011 International Conference on Computer Science and Network Technology -2016
12. Big Data Analysis Solutions Using MapReduce Framework - Sara B. Elagib ;Atahur Rahman Najeeb ; Aisha H. Hashim ; Rashidah F. Olanrewaju - 2014 International Conference on Computer and Communication Engineering – 2014
13. A performance evaluation of Hive for scientific data management - TaoyingLiu ; Jing Liu ; Hong Liu ; Wei Li - 2013 IEEE International Conference on Big Data-2013
14. Vallabh Dhoot, ShubhamGawande, Pooja Kanawade and AkankshaLekhwani, Efficient Dimensionality Reduction for Big Data Using Clustering Technique, Imperial Journal of Interdisciplinary Research (IJIR), Vol-2, Issue-5, 2016, ISSN: 2454-1362 3 https://www.thegeekstuff.com/2014/01/sql-vs-nosql-db/?utm_source=tuicool
15. Short and long-term stock trend prediction using decision tree. Rupesh A. Kemble. 2017 International Conference on Intelligent Computing and Control Systems (ICICCS) Year: 2017.
16. Visualization of big data analysis on social media - SrinathVijayaragavan ; Abhishek Anand ; SundarVignesh ; R Arockia Xavier Annie - 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS) - 2017.
17. The power of big data and algorithms for advertising and customer communication - Nico Neumann - 2016 International Workshop on Big Data and Information Security (IWBIS) - 2016.
18. Big Data Analysis Using Apache Hadoop - Shankar Ganesh Manikandan ;Siddarth Ravi - 2014 International Conference on IT Convergence and Security (ICITCS) – 2014
19. Meta-analysis of big data security and privacy: Scholarly literature gaps, Kenneth David Strang;ZhaohaoSun,2016 IEEE International Conference on Big Data (Big Data), Year: 2016
20. Analysis of big data security practices - P Revathy ;RajeswariMukesh - 2017 3rd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT) - 2017.
21. Study of data analysis model based on big data technology, JinhuaChen; Qin Jiang;YuxinWang; Jing Tang,2016 IEEE International Conference on Big Data Analysis (ICBDA),Year: 2016
22. Comparative analysis of association rule mining algorithms - S. Vijayarani ; S. Sharmila - 2016 International Conference on Inventive Computation Technologies (ICICT) - 2016
23. Data optimized computing for heterogeneous big data computing applications, EricaYang ; Derek Ross ; SrikanthNagella ; Martin Turner ; Winfried Kockelmann ; GenovevaBurca ; Federico Montesino Pouzols,2015 IEEE International Conference on Big Data (Big Data),Year: 2015.
24. Big Data Dimension Reduction Using PCA - TonglinZhang ;Baijian Yang - 2016 International Conference on Inventive Computation Technologies (ICICT) - 2016
25. Design and implementation of data warehouse with data model using survey-based services data, BoonKeongSeah; Nor EzamSelan,Fourth edition of the International Conference on the Innovative Computing Technology (INTECH 2014),
26. Big data analysis technology application in agricultural intelligence decision system,Ji-chunZhao ; Jian-xin Guo,2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)
27. Comparing HiveQL and MapReduce methods to process fact data in a data warehouse,Haince Denis Pen ;PrajyotiDsilva ; Sweedle Mascarnes,2017 2nd International Conference on Communication Systems, Computing and IT Applications (CSCITA),Year: 2017.
28. An application of a healthcare data warehouse system,BoonKeongSeah,Third International Conference on Innovative Computing Technology (INTECH 2013),Year: 2013.
29. Security analysis of unstructured data in NOSQL MongoDB database,Jitender Kumar ; Varsha Garg,2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN),Year: 2017
30. Exploring the merits of NoSQL: A study based on mongo dB,BenymolJose ;Sajimon Abraham,2017 International Conference on Networks & Advances in Computational Technologies (NetAct),Year: 2017
31. A Practical Framework for Privacy-Preserving NoSQL Databases, RicardoMacedo;JoãoPaulo;RogérioPontes; Bernardo Portela ; Tiago Oliveira ; Miguel Matos ; Rui Oliveira,2017 IEEE 36th Symposium on Reliable Distributed Systems (SRDS),Year: 2017