INTERNATIONAL JOURNAL FOR TRENDS IN ENGINEERING & TECHNOLOGY
VOLUME 4 ISSUE 1 – APRIL 2015 - ISSN: 2349 - 9303

Characterizing and Processing of Big Data Using Data Mining Techniques

S. Vishnu Priya[1]
Student,II Year ME,
Department of Information Technology
[1]National Engineering College,
priyavp91@email.com

V. Manimaran M.E[2]
Assistant Professor,
Department of Information technology,
[2] National Engineering College,
jeyamaran2002@yahoo.co.in

Abstract— Big data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured. It concerns Large-Volume, Complex and growing data sets in both multiple and autonomous sources. Not only in science and engineering big data are now rapidly expanding in all domains like physical, bio logical etc...The main objective of this paper is to characterize the features of big data. Here the HACE theorem, that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective, is used. The aggregation of mining, analysis, information sources, user interest modeling, privacy and security are involved in this model. To explore and extract the large volumes of data and useful information or knowledge respectively is the most fundamental challenge in Big Data. So we should have a tendency to analyze these problems and knowledge revolution.

Index Terms — Big Data, Data Mining, HACE Theorem, Subset Selection.
.

1 INTRODUCTION

Big data is the term for a collection of data sets so large and complex that it becomes difficult to process or describe the exponential growth and availability of data, both structured and unstructured. In short, the term Big data apply to information that can't be processed or analyzed using traditional processes. Data set that exceeds the boundaries and sizes of normal processing capabilities are termed as big data. Big Data may be as important to business – and society – as the Internet has become. There are 3 V's in Big Data management:

- Volume: there is large data than ever before, its size continues to increase, but not the percent of data that our tools can process [2].
- Variety: there are variety of data, as text, sensor data, audio, video, graph, and more [2].
- Velocity: data are arriving continuously as streams of data, and we are interested in obtaining useful information from it in real time [2].

Nowadays, there are two more V's:

- Variability: there are modifications in the structure of the data and how the users want to interpret that data[2]
- Value: business value that gives the organization a compelling advantages, due to the ability of taking decisions based on answering questions that were previously considered beyond reach[2].

In the definition of Big Data in 2012 is a high volume, velocity and variety of information assets that demand cost-effective, innovative forms of information are processing for enhanced insight and decision making. Key enablers for the growth of "Big Data" are: Storage capacities increase, Processing power Increase and Data availability.

2 DATA MINING TECHNIQUES USED TO PROCESS BIG DATA

2.1 Clustering

This is also called unsupervised learning. Here, the given a database of objects that are usually without any predefined categories or classes. It is required to partition the objects into subsets so elements of that groups share a common property. Moreover the partition should be the similarity between members of the same group should be high and the similarity between members of different groups should be low [7].Clustering can be said as identification of similar objects. By using the clustering techniques we can further identify dense and sparse regions in object space and can discover overall pattern and correlations among the attributes of data. Classification approach can also be used for effective means for distinguishing groups or classes of object due to its high cost the clustering can be used as preprocessing step for subset selection and classification. For example, to form a group of customers based on their purchasing patterns, to categories blood groups with same functionality [7].

2.2 Classification

Classification is one of the most commonly applied techniques of data mining, which contains a set of pre-classified examples to develop a model that can classify the records at large instant. Fraud detection and credit card risk applications are the best examples that suites to this type of analysis [4]. This type frequently contains decision tree or neural network-based classification algorithms. The process of data classification involves learning and classification. In Learning process the classification algorithm analyze the training data. In classification process, to estimate the accuracy of the classification rules test data are used [7]. If the accuracy is

acceptable the rules can applied to the new data. For an application of fraud detection, this would contain complete information of both fraudulent and valid activities that determined on the record-by-record basis. The classifier-training algorithm makes use of these pre-classified examples to examine the set of parameters required for proper discrimination [7].

2.3 Prediction

Regression can be adapted for prediction technique. Regression analysis can be used to define the relationship among or between one or more independent and dependent variables. In data mining, the independent variables are the attributes that we already known and response variables are the one that we want to predict. Moreover, many real-world problems are not simply based on prediction [7]. For examples, volumes of sales, stocks price, and rate of the product failure are all very difficult to predict, because they depends on complex interactions of predicted variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be needed to predict future values. The same types of model can often be used for both the regression and the classification. For example, the Classification and decision tree algorithms can be used to build both classification (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks can also create both classification and regression models [7].

3 CHARACTERIZATION OF BIG DATA

To know about big data we need to analysis its characters. The HACE theorem is used to analysis the characteristics of big data. HACE theorem stands for Heterogeneous, Autonomous, Complex and Evolving ie.) Big Data starts with large-volume, heterogeneous and autonomous sources with distributed and decentralized control, that needs to explore complex and evolving relationships among such data [1].

A. Huge with heterogeneous and diverse data sources

Heterogeneous Data is a data from much number of sources, that are largely unknown and unlimited, and also in many varying formats. One of the fundamental characteristics of the Big Data is its huge volume represented by heterogeneous and diverse dimensionality. This huge volume of data comes from various sources like Twitter, Myspace, Orkut and LinkedIn, etc.[1].

B. Decentralized control

Autonomous data sources with distributed and decentralized controls are the main characteristics of Big Data applications. Being autonomous, without any involvement (or relying on) of the centralized control each data source is able to generate and collect information. This is similar to the World Wide Web (WWW) settings where each web server provides a certain amount of information and each server is able to fully function without necessarily relying on the other servers [1].

C. Complex data and knowledge associations

Multistructure, multisource data is a complex data, examples of complex data types are materials bills, documents of word processed, maps, time-series, images and videos. Such combined characteristics suggest that Big Data requires a “big mind” to consolidate data to get maximum values [1].

4 PROPOSED WORK

The dataset is formed into a cluster using the connectivity based clustering. Then the subset selection search method is used to select the important feature by removing the irrelevant features that is it takes the relevant feature in account to provide the search result in the appropriate time. The Fig 1. Show the Big Data Processing model.

4.1 Data Set

Here the large data set is imported and the many of the databases are connected together to form a larger set of the data. The data set import process has two major steps. At first data is fetched into the system and kept in a temporary table. From there it get processed and taken into the main database [5]. These two step process helps to prevent errors in the data which affecting the main database. The first step to import the data requires a definition and information about the data that need to be imported and to put in the temporary tables respectively. This can be done by an Import File Loader and an Import Loader Format. In computing, extract, transforms, and load refers to a process in usage of database and especially in data warehousing that extracts data from the outside sources and transforms it to fit operational needs, which can include quality levels. Loads it into the end target (database, more specifically, operational data store, data mart, or data warehouse) the systems are commonly used to integrate the data from the multiple applications, typically developed and supported by different vendors that are hosted on separate computer hardware.

The disparate systems that containing the original content are frequently managed and operated by various employees. For example a cost accounting system may combine data from payroll, sales and purchasing. A healthy Big Data environment begins with an investment in data storage, but must lead to payoffs via aggressive data usage.

The path from storage to usage goes directly through data transformation. The progression is straightforward:

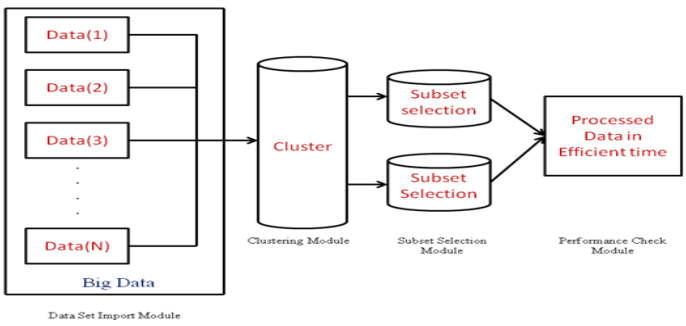

Fig 1. Big Data Processing

1. A shared watering hole of format-agnostic Big Data storage makes data capture easy, which attracts users across the organizations.

2. The capability to easily capture data attracts more—and more interesting data—over time. Inevitably, that data comes in a wide variety of formats.
3. Data transformation is required to wrangle, that variety of data into structured formats and features for analysis.

Data can be characterized using HACE Theorem: Big Data starts with large-volume, heterogeneous and autonomous sources with distributed and decentralized control that to explore complex and evolving relationships among data[1]. These characteristics make it an extreme challenge for discovering useful knowledge from the Big Data.

Before processing with data it need to be pre processed. Here it preprocesses all the queries that are related to the next surveillance of the further datasets. Data pre-processing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable in data mining and also in machine learning projects. Data-gathering methods are often controlled, resulting in out-of-range values (e.g., Income: −100), impossible data combinations (e.g., Sex: Male, Pregnant: Yes), missing data, etc.. Analyzing the data that has not been carefully screened for such problems can be lead to misleading results.

Thus, the representation and the quality of a data is a first and foremost that before running taking an analysis[3]. If there is much irrelevant and redundant information present on the noisy and unreliable data, then the knowledge discovery during the training phase is more difficult. Data preparation and filtering steps that can take a considerable amount of processing time. Data pre-processing step includes data cleaning, normalization, data transformation, feature extraction, feature selection, etc.. The products of the data pre-processing is the final training set.

4.2 Clustering and Connectivity Based Clustering

The clustering has been used to cluster the words into a groups based either on their participation in particular grammatical relations with other words or on the class of labels associated with each word.

- *Cluster analysis or clustering:* Clustering is the task of grouping a set of objects in such a way that objects are in the same groups (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task in exploratory data mining, and also a common technique for statistical data analysis, that are used in many fields, including machine learning, pattern recognition, image analytics, information retrieval, and bioinformatics.

Cluster analysis itself is not one specific algorithm, but the general task to be solved [10]. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of a clustering includes group with small distances within the cluster members, dense areas of a data space, intervals or particular statistical distributions. Clustering can therefore be tabulated as a multi-objective optimization problems. The most appropriate clustering algorithm and its parameter settings (including values such as the distance function, a density threshold or the number of the expected clusters) will depend on the individual data set and intended use of the results[11]. Cluster analysis as such not an automatic task, instead an iterative process of knowledge discovery or interactive multi-objective optimization that involves in trial and failure. It will often be necessary to modify the data preprocessing and model parameters until the result achieves the desired properties.

- *Connectivity based clustering:* Connectivity based clustering, also known as *hierarchical clustering*, is based on the core idea of objects being more related to nearby objects than to objects farther away. These algorithms connect "objects" to form "clusters" based on their distance. A cluster can be described largely by the maximum distance needed to connect parts of the cluster. At different distances, different clusters will form, which can be represented using a dendrogram, which explains where the common name "hierarchical clustering" comes from: these algorithms do not provide a single partitioning of the data set, but instead provide an extensive hierarchy of clusters that merge with each other at certain distances[8][9]. In a dendrogram, the y-axis marks the distance at which the clusters merge, while the objects are placed along the x-axis such that the clusters don't mix .

Connectivity based clustering is a whole family of methods that can be differ by the way of distances that are computed. Apart from the usual selection of distance functions, the user also needs to take decision on the linkage criterion (since a cluster consists of multiple objects, and there are multiple candidates to compute the distance to) to use. Popular selections are known as single-linkage clustering (the minimum of object distances), complete linkage clustering (the maximum of object distances) or "Unweighted Paired Group Method", also called as an average linkage cluster. Moreover, hierarchical clustering can be agglomerate (starting with a single elements and aggregating them into a clusters) or divisive (starting with the complete data set and dividing it into partitions).

4.3Subset Selection

To efficiently retrieve the data from the large data set the subset search method is used. Information retrieval is an activity to obtaining the information resources relevant to an information needed from a collection of information resources. Searches can be done based on metadata or on full-text (or other content-based) indexing. An information retrieval process begins when a user enters a query into the system [2]. Queries are the formal statements of information. In information retrieval a query does not identify a single object in the collection. Instead, several objects can match the query, with the different degrees of relevance. An instance is an entity that is represented by an information in a database. User queries are matched with the information present in the database. Depending on the application s the data objects may be, for example, text documents, images, audio, word files,videos,etc,.

- *Minimum Spanning Tree* : Given a connected graph, a spanning tree of that graph is a sub graph that is a tree which connects all the vertices together. A single graph may have many different spanning trees. Assign a weight to each edge and used to assign a weight to a spanning tree by computing the sum of the weights of the

edges in that spanning tree[6]. A minimum spanning tree (MST) or minimum weight spanning tree is then a spanning tree with weight less than or equal to the weight of every other spanning tree. More generally, any undirected graph (not necessarily connected) has a minimum spanning forest, which is the union of minimum spanning tree for its connected components [12].

4.4 Performance Measure

A performance is a graphical representation of progress over time of some an entity, such as an enterprise, an employee or a business unit, towards some specified goal. Performance scorecards are widely used in many industries both in the public and private sectors. The performance is an essential component of the balanced scorecard methodology. Performance scorecards are also used independently of the balanced scorecard methodology to monitor the progress of any organizational goal.

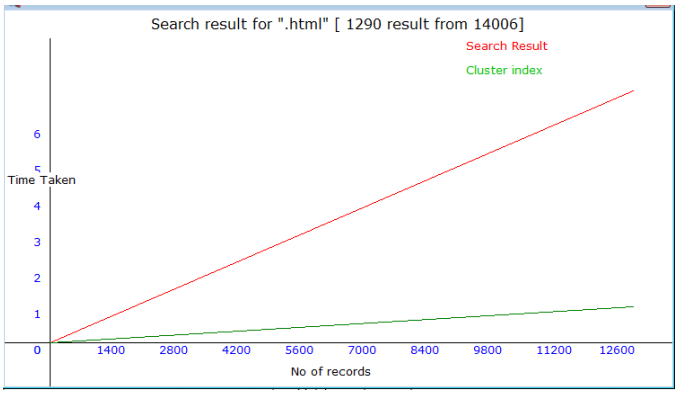

Fig 2. Performance Measure

Fig.2.shows that the performance metrics is calculated between subset selection search and deterministic search method with respect to time. The subset selection method uses the relevant cluster for its search purpose. Where the deterministic method uses the irrelevant cluster for its search purpose. As the result the subset selection algorithm shows the better performance than the deterministic search.

5 CHALLENGES IN BIG DATA

Meeting the challenges of big data will be difficult. The volume of data is already high and increasing every day. The velocity of its growth is increasing, driven in part by the internet connected devices [1]. Furthermore, the variety of data being generated is also expanding, and organizations capability to capture and process this data is limited [7]. Current technologies, architectures and analysis approaches are unable to work with the flood of data, and the organizations want to change the way they think about, plan, govern, process and report on the data to know the potential of big data.

- Privacy, security and trust.
- Data management and sharing.
- Technology and analytical systems.

6 CONCLUSION

We have entered into an era of Big Data. Through better analysis of the large amount of data, there is the chance for making the data faster disciplines and improving the profitability and success of enterprises. Hence, to characterize the characteristics of Big Data are important, here the Big Data was characterized using the HACE theorem. Subset selection is used to select the features in the dataset to provide the best result while processing with Big Data. The challenges include heterogeneity, security and trust, lack of structure, error-handling, privacy, timeliness, Data management, provenance, visualization and sharing that leads to result interpretation. Therefore, these challenges will need transformative solutions. Hence we must support and encourage fundamental research towards addressing these technical challenges to achieve the benefits of Big Data.

REFERENCES

[1] Xindong Wu, Xingquan Zhu,Gong-Qing Wu, and Wei Ding, "Data Mining with Big Data" Ieee Transactions On Knowledge And Data Engineering, Vol. 26, No. 1, January 2014.
[2] I. Kopanas, N. Avouris, and S. Daskalaki, "The Role of Domain Knowledge in a Large Scale Data Mining Project," Proc. Second Hellenic Conf. AI: Methods and Applications of Artificial Intelligence, I.P. Vlahavas, C.D. Spyropoulos, eds., pp. 288-299, 2002.
[3] M.H. Alam, J.W. Ha, and S.K. Lee, "Novel Approaches to Crawling Important Pages Early" Knowledge and Information Systems, vol. 33, no. 3, pp 707-734, Dec. 2012.
[4] R. Ahmed and G. Karypis, "Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks," Knowledge and Information Systems, vol. 33, no. 3, pp. 603-630, Dec. 2012.
[5] Sharayu S. Sangekar, Pranjali P. Deshmukh,"Data Mining of Complex Data with Multiple, Autonomous Sources," International Journal of Pure and Applied Research in Engineering and Technology ijpret, 2014; volume 2 (9): 793-799.
[6] Y.-C. Chen, W.-C. Peng, and S.-Y. Lee, "Efficient Algorithms for Influence Maximization in Social Networks," Knowledge and Information Systems, vol. 33, no. 3, pp. 577-601, Dec. 2012.
[7] Bharti Thakur, Manish Mann "Data Mining for Big Data: A Review"International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 5, May 2014.
[8] Amandeep Kaur Mann, Navneet Kaur "Survey Paper on Clustering Techniques" International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, April 2013.
[9] Anoop Kumar Jain, Satyam Maheswari "Survey of Recent Clustering Techniques in Data Mining" International Archive of Applied Sciences and Technology Volume 3 [2] June 2012: 68 – 75.
[10] K.Kameshwaran, K.Malarvizhi "Survey on Clustering Techniques in Data Mining" International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2272-2276
[11] http://www.cc.gatech.edu/~isbell/reading/papers/berkhin02survey.pd.
[12] https://www.ics.uci.edu/~eppstein/161/960206.html.
[13] ftp://ftp.cs.ubc.ca/.snapshot/sv_weekly.2/local/techreports/1994/TR-94-13.pdf.