Data Mining of Big Data: The Survey and Review

ReshmaNarawade
Department of Computer Engineering
M.G.M.C.I.T, New Mumbai, Maharashtra, India

Prof. Vilas Jadhav
Department of Computer Engineering
M.G.M.C.I.T, New Mumbai, Maharashtra, India

Abstract- Big Data word is defined for a collection of large and complex data in big volume. Big Data can be collected from available in structure and non-structure format. .It can born in different field such as Engineering, Social Media sites, videos , Metallurgy, Geographical area, from various research and many more resources etc. This is the only part of Big Data. In data mining concept useful data is collected by various technology which are invented to manage and analyze such as Hadoop , Storm ,Apache S4 etc. Data mining is a technique for discovering various useful pattern as well as descriptive, understandable models from big volume data. In this paper we analyze the types of Big Data, techniques to analyze data and challenges in Big Data for future as well Technique use for it.

Keywords : – Big Data, Data mining Techniques, HACE theorem, 5V's,Privacy

I. INTRODUCTION

Data Mining is the technology to extract the knowledge from the data. It is used to explore and analyze the same. The data to be mined varies from a small data set to a large data set i.e. Big Data. Big Data mining was very relevant from the beginning, as the first book mentioning 'Big Data' is a data mining book that appeared also in 1998 by Weiss and Indrukya. However, the first academic paper with the words 'Big Data' in the title appeared a bit later in 2000 in a paper by Diebold .The origin of the term 'Big Data' is due to the fact that we are creating a huge amount of data every day. Usama Fayyad in his invited talk at the KDD BigMine‟ 12Workshop presented amazing data numbers about internet usage, among them the following: each day Google has more than 1 billion queries per day, Twitter has more than 250 million tweets per day, Facebook has more than 800 million updates per day, and YouTube has more than 4 billion views per day. The data produced nowadays is estimated in the order of zettabytes, and it is growing around 40% every year. A new large source of data is going to be generated from mobile devices and big companies as Google, Apple, Facebook, Yahoo are starting to look carefully to this data to find useful patterns to improve user experience.

"Big Data" is pervasive, and yet still the notion engenders confusion. Big Data has been used to convey all sorts of concepts, including: huge quantities of data, social media analytics, next generation data management capabilities, real-time data, and much more. Whatever the label, organizations are starting to understand and explore how to process and analyze a vast array of information in new ways. In doing so, a small, but growing group of pioneers is achieving breakthrough business outcomes. In industries throughout the world, executives recognize the need to learn more about how to exploit Big Data. But despite what seems like unrelenting media attention, it can be hard to find in-depth information on what organizations are really doing. So, we sought to better understand how organizations view Big Data – and to what extent they are currently using it to benefit their businesses.

The rest of the paper is organized as follows. Data Mining of Big Data is explained in section II. Types of Big Data and Sources III. HACE Theorem IV. Data Mining for Data V. Data Mining for Big Data VI. New Techniques of Data Mining VII. Estimation of the Future. VIII. Conclusion

II. TYPES OF BIG DATA AND SOURCES

There are two types of Big Data: Structured and Unstructured.

A. Structured Data –

Structured Data can be categorized under words and number easily. Source of these data is like network sensors embedded in electronic devices, smartphones, and global positioning system (GPS) devices. Structured data also consist of things like account figure, transaction data, sales figures etc.

B. Unstructured Data –

Unstructured data include critical information, such as photos, multimedia, reviews from commercial websites, various comment on social networking sites. Separation of such a data is not an easy task to convert into categories or analyzed numerically. "Unstructured Big Data is the things that humans are saying," says Big Data consulting firm vice president Tony Jewitt of Plano, Texas. "It uses natural language." Analysis of unstructured data relies on keywords, which allow users to filter the data based on searchable terms. The explosive growth of the Internet in recent years means that the variety and amount of Big Data continue to grow. Much of that growth comes from unstructured data.

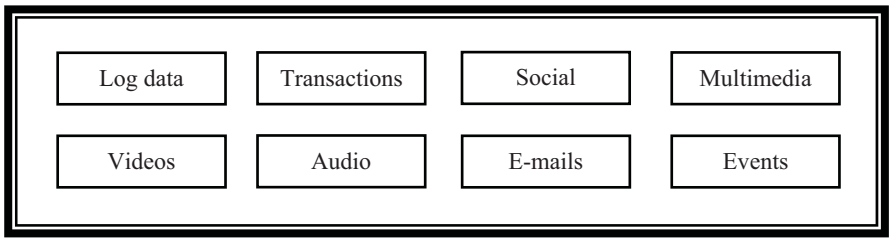

Figure 1. Sources of Big Data

III. HACE THEOREM

Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data. These characteristics make it an extreme challenge for discovering useful knowledge from the Big Data. In a naïve sense, we can imagine that a number of blind men are trying to size up a giant Camel, which will be the Big Data in this context. The goal of each blind man is to draw a picture (or conclusion) of the Camel according to the part of information he collects during the process. Because each person's view is limited to his local region, it is not surprising that the blind men will each conclude independently that the camel "feels" like a rope, a hose, or a wall, depending on the region each of them is limited to. To make the problem even more complicated, let us assume that the camel is growing rapidly and its pose changes constantly, and each blind man may have his own (possible unreliable and inaccurate) information sources that tell him about biased knowledge about the camel (e.g. one blind man may exchange his feeling about the camel with another blind man, where the exchanged knowledge is inherently biased). Exploring the Big Data in this scenario is equivalent to aggregating heterogeneous information from different sources (blind men) to help draw a best possible picture to reveal the genuine gesture of the camel in a real-time fashion. Indeed, this task is not as simple as asking each blind man to describe his feelings about the camel and then getting an expert to draw one single picture with a combined view, concerning that each individual may speak a different language (heterogeneous and diverse information sources) and they may even have privacy concerns about the messages they deliberate in the information exchange process. The term Big Data literally concerns about data volumes, HACE theorem suggests that the key characteristics of the Big Data are:

A. Huge with heterogeneous and diverse data sources: - One of the fundamental characteristics of the Big Data is the huge volume of data represented by heterogeneous and diverse dimensionalities. This huge volume of data comes from various sites like Twitter, MySpace, Orkut and LinkedIn etc.

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

C. Complex data and knowledge associations:- Multistructure, Multisource data is complex data, Examples of complex data types are bills of materials, word processing documents, maps, time-series, images and

video. Such combined characteristics suggest that Big Data require a “big mind” to consolidate data for maximum values.

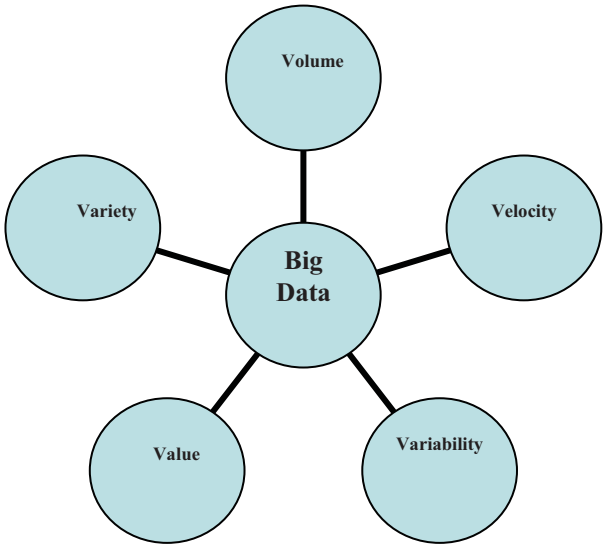

Fig 3.1 Five V’s in Big Data

Doug Laney was the first one talking about 3V’s in Big Data Management.
Volume: The amount of data. Perhaps the characteristic most associated with Big Data, volume refers to the mass quantities of data that organizations are trying to harness to improve decision-making across the enterprise. Data volumes continue to increase at an unprecedented rate.
Variety: Different types of data and data sources. Variety is about managing the complexity of multiple data types, including structured, semi-structured and unstructured data. Organizations need to integrate and analyze data from a complex array of both traditional and non-traditional information sources, from within and outside the enterprise. With the explosion of sensors, smart devices and social collaboration technologies, data is being generated in countless forms, including: text, web data, tweets, audio, video, log files and more.
Velocity: Data in motion. The speed at which data is created, processed and analyzed continues to accelerate.
Now a day’s there are two more V‟s
Variability: - There are changes in the structure of the data and how users want to interpret that data.
Value: - Business value that gives organization a compelling advantage, due to the ability of making decisions based in answering questions that were previously considered beyond reach.

Table 3.1 Difference between Big Data and Data mining

Data Mining	Big Data
Data mining refers to the collecting useful and important information from Big Data	Large data set term represent Big Data
Data mining is the controller which is helpful information	Big Data is the asset
Data mining refers to the operation that involve relatively complicated search operation sometimes known as Knowledge discovery	Big data" differ based on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyze the data.

IV. DATA MINING FOR BIG DATA

Generally, Data Mining is the process of analyzing data from different view and gathers it into valuable information that can be used to increase revenue, cuts costs, or both. Technically, data mining is the process of finding relation between or patterns among multiple fields in huge relational database. Data mining as a term used

for the specific classes of six activities or tasks as follows:
1. Distribution
2. Estimation
3. Prediction
4. Association rules
5. Clustering
6. Description

A. Classification

Classification is a process of generalizing the data according to different instances. Several major kinds of classification algorithms in data mining are Decision tree, k-nearest neighbor classifier, Naive Bayes, Apriori and AdaBoost. Classification consists of examining the features of a newly presented object and assigning to it a predefined class. The classification task is characterized by the well-defined classes, and a training set consisting of reclassified examples.

B. Estimation

Estimation deals with continuously valued outcomes. Given some input data, we use estimation to come up with a value for some unknown continuous variables such as income, height or credit card balance.

C. Prediction

It's a statement about the way things will happen in the future, often but not always based on experience or knowledge. P*rediction* may be a statement in which some outcome is expected.

D. Association Rules

An association rule is a rule which implies certain association relationships among a set of objects such as "occur together" or "one implies the other" in a database.

E. Clustering

Clustering can be considered the most important *unsupervised learning* problem; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data.

- New Techniques of Data Mining :-
  Some New Techniques use for Data mining of Big Data:

a) ***Hadoop*****:** A way to speed up the mining of streaming learners is to distribute the training process onto several machines. Hadoop Map Reduce is a programming model and software framework for writing applications that rapidly process vast amounts of data in parallel on large clusters of compute nodes. A MapReduce job divides the input dataset into independent subsets that are processed by map tasks in parallel. This step of mapping is then followed by a step of reducing tasks. These reduce tasks use the output of the maps to obtain the final result of the job.

b) ***S4****:* Apache S4 is a platform for processing continuous data streams. S4 is designed specifically for managing data streams. S4 apps are designed combining streams and processing elements in real time

c) ***Storm*****:** Storm from Twitter uses a similar approach. Ensemble learning classifiers are easier to scale and parallelize than single classifier methods. They are the first, most obvious, candidate methods to implement using parallel techniques.

VI. ESTIMATION OF THE FUTURE

There are many future important challenges in Big Data management and analytics that arise from the nature of data: large, diverse, and evolving. These are some of the challenges that researchers and practitioners will have to deal during the next years:

A. Analytics Architecture: - It is not clear yet how an optimal architecture of analytics systems should be to deal with historic data and with real-time data at the same time. An interesting proposal is the Lambda architecture of Nathan Marz. The Lambda Architecture solves the problem of computing arbitrary functions on arbitrary data in real time by decomposing the problem into three layers: the batch layer, the serving layer, and the speed layer. It combines in the same system Hadoop for the batch layer, and Storm for the speed layer. The properties of the system are: robust and fault tolerant, scalable, general, and extensible, allows ad hoc queries, minimal maintenance, and debuggable.

B. Statistical significance: - It is important to achieve significant statistical results, and not be fooled by randomness. As E from explains in his book about Large Scale Inference it is easy to go wrong with huge data sets and thousands of questions to answer at once.

C. Distributed mining: - Many data mining techniques are not trivial to paralyze. To have distributed versions of some methods, a lot of research is needed with practical and theoretical analysis to provide new methods.

D. *Hidden Big Data:* - Large quantities of useful data are getting lost since new data is largely untagged file based and unstructured data. The 2012 IDC study on Big Data explains that in 2012, 23% (643 exabytes) of the digital universe would be useful for Big Data if tagged and analyzed. However, currently only 3% of the potentially useful data is tagged, and even less is analyzed.

VII. CONCLUSION

Due to Increase in the amount of data in the field of genomics, meteorology, biology, environmental research, it becomes difficult to handle the data, to find Associations, patterns and to analyze the large data sets. Big Data is the term for a collection of complex data sets, Data mining is an analytic process designed to explore data(usually large amount of data-typically business or market related-also known as "Big Data")in search of consistent patterns and then to validate the findings by applying the detected patterns to new subsets of data. To support Big Data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data. We regard Big Data as an emerging trend and the need for Big Data mining is rising in all science and engineering domains. With Big Data technologies, we will hopefully be able to provide most relevant and most accurate social sensing feedback to better understand our society at real time.

REFERENCES

[1] Alex Berson and Stephen J. Smith Data Warehousing, Data Mining and OLAP edition 2010.

[2] Department of Finance and Deregulation Australian Government Big Data Strategy-Issue Paper March 2013

[3] NASSCOM Big Data Report 2012

[4] Wei Fan and Albert Bifet "Mining Big Data: Current Status and Forecast to the Future",Vol 14,Issue 2,2013

[5] Algorithm and approaches to handle large Data-A Survey,IJCSN Vol 2,Issue 3,2013

[6] Xindong Wu , Gong-Quing Wu and Wei Ding " Data Mining with Big data ", IEEE [7] Transactions on Knowledge and Data Engineering Vol 26 No1 Jan 2014

[7] Xu Y etal, balancing reducer workload for skewed data using sampling based partioning 2013.

[8] X. Niuniu and L. Yuxun, "Review of Decision Trees," IEEE, 2010 .

[9] Decision Trees for Business Intelligence and Data Mining: Using SAS Enterprise Miner "Decision Trees-What Are They?"

[10] Weiss, S.H. and Indurkhya, N. (1998), Predictive Data Mining: A Practical Guide, Morgan Kaufmann Publishers, San Francisco, CA.