© 2018 IJSRSET | Volume 4 | Issue 9 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

A Review on Leads from Data Mining To Big Data

Urvashi Saraswat[1], Nisha Singh[2]

[1]Department of Computer Science and Engineering, Faculty of Engineering Technology, Rama University, Kanpur, India

[2]Department of Information Technology BBDNITM, Dr. APJ Abdul Kalam Technical University, Lucknow, India

ABSTRACT

This paper aims towards the advancements in the field of Data. This depicts the transition of technologies and market need of data. Data mining was initially the prior technique for knowledge extraction. From this area how the knowledge extraction changed its course to Big Data is the essence of this survey work. The paper also aims in driving the attention towards the advantages, tools and future aspects of Big Data.

Keywords : Data Mining, Big Data, Big Data Analytics, Clustering.

I. INTRODUCTION

Data mining- means discovering the hidden values from data repository. This technique suggests ways and methods to extract the required core information from the large databases. It is a powerful technology that helps many companies to collect important data for their analysis. There are several data mining tools that contribute in the data extraction and filtration. These tools also suggest future trends and behavior, which allows the business to become proactive in the current industry.

Data mining is the investigation and study of enormous information sets, so as to get purposeful pattern and rules. The key plan is to search out effective thanks to mix the computer's power to method the information with the human eye's ability to discover patterns [1] . The definition of data mining is nearly related to another generally used term i.e. Knowledge Discovery [2] .

Data mining has two primary objectives of prediction and outline. Prediction involves the victimization of some variables in information sets, so as to predict unknown values of different relevant variables (e.g. classification, regression, and anomaly detection). Description involves discovering human comprehensible patterns and trends within the information (e.g. clustering, association rule learning, and minimizing data) [4,3].

ISSUES IN DATA MINIG

Data mining is not considered to be an easy task. There are certain algorithms that can get very complex and it is not necessary that data is always available at one place. There are certain issues in the data mining like- the methodology of mining, performance-based issues and diverse data type issues. One of the key problems raised by data processing technology isn't a business or technological one, however a social one. It is the difficulty of individual privacy. Data processing makes its potential to research routine business transactions and reap a big quantity of data regarding people shopping for habits and preferences. Data mining has evolved into a very important and agile space of analysis due to the hypothetical challenges and sensible applications related to the matter of locating fascinating and

antecedently unknown data from real-world databases [3,4].

The prime challenges and issues of Data mining are-

- Poor data quality i.e. full of noise, missing values, inadequate data size, etc.
- Data Redundancy from different forms and sources.
- Unavailability or difficulty in finding the data
- Difficulty in dealing with enormous data sets.
- Dealing with unbalanced and variable data

BIG DATA-a big library

The notion of massive knowledge has been endemic among engineering since the earliest days of computing. "Big Data" originally meant the quantity of knowledge that would not be processed by ancient information ways and tools. Whenever a replacement data-storage medium was fabricated, the quantity of knowledge accessible exploded as a result of it may be simply accessed [4].

The term "Big Data" became noticeable in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey, which had a title of "Big Data and the NextWave of InfraStress". The Big Data mining was considered quite relevant from the start, because the initially, the book mentioned 'Big Data' as a data mining book that also appeared in 1998 by Weiss and Indrukya [5].

II. CLASSIFICATION OF BIG DATA

Big Data is a set of datasets and is quite massive and complicated that is on the far side of the ability of typical software package tools to capture, store, manage and method the information at intervals a tolerable period [6].

Big Data possesses two categories of data. The Figure1. below shows the representation of the two categories.

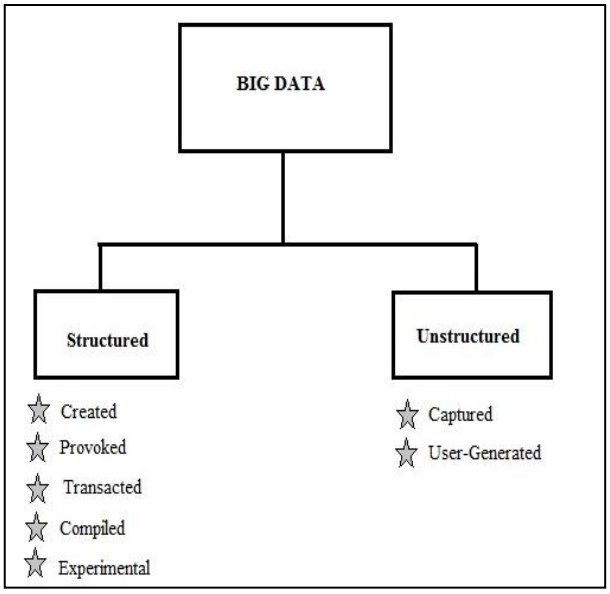

Figure 1. Categorization of Big Data

3.1 Structured Data : This type of data can be easily analyzed because it exists in numerical form,. A structured data is easily Created, Provoked, Transacted, Compiled and is Experimental.

3.2 Unstructured Data: Unstructured data is a bit less familiar as compared to the Structured Data. But, most of the data which is created in the present technology is unstructured. As per its name, this typed of data is features no structure. It can be captured easily in form of clicks, purchases, etc. Basically, Unstructured data is of two types- Captured and User-generated.

III. IMPORTANT V's of BIG DATA

Doug Laney was the first person to introduce the three V's in Big Data Management [7]. The data here, should be of specific or common activity or business that can be either public or private. And, this Big Data is defined by 3 V's [8].

- Volume- is the factor describing the amount of data present. It relates the mass quantity of data. It describes the amount of data. It refers to mass quantities of data.

- Variety – Variety is the various species of data and sources that include the structured, semi-structured and the unstructured data.
- Velocity – This factor connotes the locomotion of data. It states that a data can be generated rapidly and then processed and analyzed [9].

IV. BIG DATA ANALYTICS

Big data analytics refers to the process of collecting, organizing and analyzing large sets of data ("Big Data") to discover patterns and other useful information [10]. Therefore, Big Data Analytics can be defined as the examination of large data sets that uncover the hidden patterns, unknown association, customer priorities and other business-related data.

V. WHY IS BIG DATA ANALYTICS IMPORTANT

Big Data Analytics plays a major role in the current industry whether it is business or any technology emerging organization. The prime aim of Big Data Analytics is to help the organizations in handling their data and explore it for more opportunities. The key factors that show the importance of Big Data Analytics are- cost reduction, fast decision making and emerging services.

High performance analytics allow us to perform us standard operations, that were never imagined to possibly exist, because of bulk data size.
We all talk about Big Data, but how far are we aware of its use in the organization? Let us have a brief look over the brighter side of Big Data involved in the organizations:

- It allows the businesses to spot errors, fraudulent faster and easier.
- The real time data analysis allows the businesses to upgrade much effectual strategies towards the competitors in a much less time.
- The Big Data is a growing platform and, the Big Data Analytics tools are efficiently capable in handling massive data.

VI. ADVANTAGES OF BIG DATA ANALYTICS

Now-a-days, the business cluster businesses are looking for more actionable insights. They require the core data and with this, many data related projects have originated. This data runs the procedure for improved operations. Many advantages are of Big Data are stated as-

- Big Data is timeline oriented: As per the business analysis surveys, it has been observed that, 60% of each working days is consumed by knowledge workers, where they spend their time in managing the data.
- Unlimited Storage: There is the availability of almost unlimited storage for huge volumes of data.
- Accessibility of Big Data: You can access the Big Data from anywhere, at anytime, on any device with cloud service.
- Transmission speed: The speed of transmission is very high and owes to cutting edge technologies.

VII. THE THREE TIERS OF BIG DATA

There are three main challenges that are found in Big Data. These functions are- data ingress and computation of arithmetic computing procedures, semantics and other domain knowledge. The above mentioned factors are for different Big Data applications and the complications are raised by the Big Data volumes, the distribution of distributed data and also by complex and dynamic characteristics. This Big data framework can be categorized into three tiers as shown in Figure2. [11]. As the figure suggests, These different tiers build up a framework which assists in overcoming data availability and accessibility challenges.

Tier1- This is responsible for focusing over the data access and computation of arithmetic procedures.

Because, there is a huge amount of data which is stored at different levels , so it is crucial to focus on the rapid growth of information. Therefore, for handling the computation of large scale data there are different capable computing platforms like Hadoop.

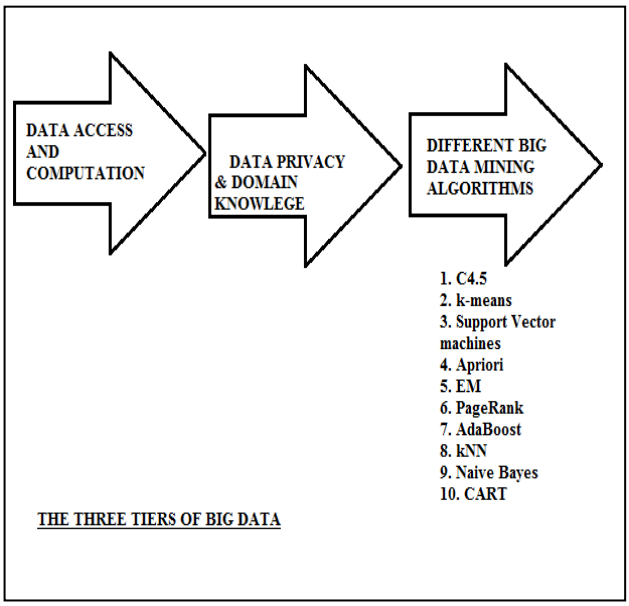

Figure 2. The three tiers of BIG DATA

Tier2- This tier focuses on the privacy of the data i.e. the semantic knowledge and domain knowledge for different Big Data applications.

Tier3- Tier3 uses the different standard algorithms to evaluate and analyze data. These algorithms used in Big Data Analytics are the Data Mining Algorithms on the huge volumes of data[12].

VIII. AREAS OF APPLICATION

Who uses Big Data? What are the areas where the Big Data is required? The answer is: Big Data is open and generalized. This means that Big Data is open to use anytime and anywhere. Every industry or organization is in a need of Big Data Analytics. In other words, we can say that Big Data has become a need for every industrial or organizational aspect. And this need of Big Data has pushed the organizations to adopt Big Data Analytics because of the various types of users:-

- The data scientist: These are the people who perform all the complex analysis on the bulk volume of data. They are familiar with the different models and are capable of applying new models and study structure dependencies of the data.
- The Business Analysts: These people are more proactive with the data analytics tools and are efficient in extracting the information out of the complex existing information. They work on the collective information and with the predictive analytical techniques, process on the raw data.
- Business Manager: They focus on the conclusions and comprehend the models of the existing data.
- IT Developers: These people are responsible basically to list all the prior industry requirements at user end. This is done by the IT developers so as to design a user friendly support end for the analysts and managers.

From the user end to the consumer industry, Big Data is a part of all areas. These areas which utilize the Big Data Analytics and function under are many like: Hospitality, Healthcare, Retail sector, Government and many more.

IX. THE KEY TECHNOLOGIES AND HOW THEY WORK

In the present technology biased industry, there is no organization devoid of the involvement of Big Data. Can you even imagine an industry existing without the involvement of data? Then, to manage this data, the technical standards have moved on to the advanced analytics. Some of the key role players in this are:-

- Data Management: Our data needs to be highly assessed in terms of quality standards. The data should be well manages and reliably analyzed. Once the data is highly reliable and meets the quality benchmarks, the organizations a master data management program and runs the complete enterprise on the same page.

- Data Mining: The mining technology assists in analyzing the huge volumes of data and with this different patterns can be identified and examined. The mined information can be utilized further for future analysis.
- Hadoop: This is the key factor that is an open source framework, that works on the Big Data Analytics and by simply analyzing the data from the system memory, it can derive immediate insights from it. This technology not only saves time, but also yields quick and quality results.
- Predictive Analysis: It uses data, statistic-based algorithms and machine –learning techniques to diagnose the future results of the complex data analysis. Most common applications of the predictive analysis is the fraud detection, marketing and operations, risk analysis, etc.

The above mentioned are the common technologies on which the Big Data Analytics focuses its operations and analysis. Apart from these there are text mining techniques, in-memory analytics and more.

X. MINING TECHNQUES INVOLVED IN BIG DATA

There are various techniques of analysis that can be performed in order to extract information from Big Data. Each of these designed analysis have a different effect or result [13]. Data mining works for the Big Data with the process of finding associations and huge volumes of fields in a large relational database.

Data mining as a term utilized for the specific class of six activities as follows:

- Classification
- Estimation
- Prediction
- Association rules
- Clustering
- Description

With the application of Data Mining techniques, Big Data Analysis is a great help to organizations, and data scientists. The more is the data, higher are the chances of a better model using the data mining techniques.

XI. CONCLUSION

In this survey we have analyzed the existing importance of Data Mining techniques in the Big Data. The different tools and techniques of Big Data are the core of the industry standards, thus, how the data mining yields its impact in the Big Data Analytics are the prime factors to study. This paper also summarizes the application areas and the challenges faced in the data industry and procedures on how to overcome and meet the industry requirements.

XII. ACKNOWLEDGEMENT

I would like to thank all people who supported me to prepare this paper. I would like to thank my guide who helped me with proper suggestions. Finally, I would like to thank all the journal papers which I have referred in the compilation process of my paper successfully.

XIII. REFERENCES

[1]. Han, J., Kamber, M., Data Mining Concepts and Techniques, Morgan Kaufmann Publisher, 2001.

[2]. Pavel Berkhin, A Survey of Clustering Data Mining Techniques, pp.25-71, 2002.

[3]. SaurkarAnand V, Bhujade Vaibhav, Bhagat Priti, Kharpade Amit, "A Review Paper on Various Data Mining Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 4, April 2014.

[4]. Arjun K, Dr. Jabasheela L.," Big Data: Review, Classification and Analysis Survey",in International Journal of Innovative Research in Information Security (IJIRIS), Volume 1 Issue 3 September 2014.

[5]. Thakur Bharti, Mann Manish, "Data Mining for Big Data: A Review", in International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014.

[6]. Mohammed GH. AL Zamil, "The Application of Semantic-based Classification on Big Data," International Conference on Information and Communi cation Systems (ICICS) 978-1-4799-3023 4/14, 2014, IEEE.

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

[8]. Priya P. Sharma, Chandrakant P. Navdeti, (2014), " Securing Big Data Hadoop: A Review of Security Issues, Threats and Solution", IJCSIT, 5(2), pp2126-2131

[9]. Tiwarkhede Ankita S., Prof. Kakde Vinit, "A Review Paper on Big Data Analytics", in International Journal of Science and Research (IJSR), Volume 4 Issue 4, April 2015

[10]. A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer.MOA: Massive Online Analysis http://moa.cms.waikato.ac.nz/. Journal of Machine Learning Research (JMLR), 2010.

[11]. Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding, "Data Mining with Big Data," Transactions On Knowledge And Data Engineering, Vol. 26, No. 1. 1041-4347/14 January 2014, IEEE

[12]. Mohammed GH. AL Zamil, "The Application of Semantic-based Classification on Big Data," International Conference on Information and Communi cation Systems (ICICS) 978-1-4799-3023 4/14, 2014, IEEE.

[13]. Y. Low, J. Gonzalez, A. Kyrola, D. Bickson,C. Guestrin, and J. M. Hellerstein. Graphlab: A new parallel framework for machine learning. In Conference on Uncertainty in Artificial Intelligence (UAI), Catalina Island, California, July 2010.