Data Mining and Big data: Current Scenario, Its Future and Opportunities

Santosh Kumar Singh
University Department of Computer Applications
Vinoba Bhave University, Jharkhand

Dr. Rajiv Kumar Dwivedi
Associate Professor, University Department of Mathematics
Vinoba Bhave University, Jharkhand

ABSTRACT

The purpose of this study is to examine existing position or significance of Data Mining as well as Big Data. In 21st Century the term "Big Data" is no longer just a buzzword, which forced the researchers to think, analyze and be ready for the future challenges related to it, to cope with the evolved nature of data and to develop new analytic techniques. Early this century the rise of the relational databases, web access, use of android phones, and other technologies made the study and management of massive data sets a real and present challenge that needed a name. In July of 2013 the Oxford English Dictionary adopted the phrase, "big data," but it's been around since as early as World War II to apply to working with massive amounts of information. Come 2020, every person in the world will be creating 7 MBs of data every second. We have already created more data in past couple of years than in the entire history of human kind. Big data has taken the world by storm and there are no signs of slowing down. In this paper we will analyze the current scenario and future of Big data on the basis of research.

Keywords — Big Data, Data Mining, IoT, Market Basket Analysis.

INTRODUCTION

Today we are the member of Cyber Village or in a more advanced way we may say that living in an era of digital world. With the rapid development and advancement in the field of automation or digitization various types of data viz. structured, semi structured and unstructured being generated and stored. According to a researcher Usama Fayyad[1] "every day 1 billion queries are there in Google, more than 250 million tweets are there in Twitter, more than 800 million updates are there in Face book, and more than 4 billion views are there in You tube". Each day, 2.5 quintillion bytes of data are generated and 90 percent of the data in the world today were created within the past few years[2]. The data produced nowadays is estimated in the order of zeta bytes, and it is growing around 40% every year. The most fundamental challenge for big data applications is to explore the large volumes of data and extract useful information or knowledge for future actions.[3]

Source: https://www.shutterstock.com

Data mining[4] is the process of finding useful information that is not easily exposed in vast amounts of data. It is a methodology for finding hidden patterns and trends of specific types on the basis of Data Analysis, that extract patterns from data and generate models. A particular data model forms a kind of cluster that describes the relationship between data sets. The most basic analytical tools for handling large-scale structured and unstructured data are Hadoop, NoSQL, and R[5] is used as a tool to analyze the analyzed data with a focus on visualization.
Big Data: Historical Perspective

In fact, the history of using data date back from 7000 years ago when accounting was introduced in ***Mesopotamia*** in order to record the growth of crops and herds. In 1663, ***John Graunt*** recorded and examined all information about mortality roles in London. He wanted to gain an understanding and build a warning system for the ongoing bubonic plague. In the first recorded record of statistical data analysis, he gathered his findings in the book Natural and Political Observations Made upon the Bills of Mortality, which provides great insights into the causes of death in the seventeenth century. Because of his work, ***Graunt can be considered the father of statistics***. The earliest remembrance of modern data is from the 1887 when Herman Hollerith invented a computing machine that could read holes punched into paper cards in order to organize census data.

The 20th Century: The first major data project is created in 1937 and was ordered by the Franklin D. Roosevelt's administration in the USA. The first data-processing machine appeared in 1943 and was developed by the British to decipher Nazi codes during World War II.
In 1952 the National Security Agency (NSA) is created and within 10 years contract more than 12,000 cryptologists. They are confronted with information overload during the Cold War as they start collecting and processing intelligence signals automatically.

In 1965 the United Stated Government decided to build the first data centre to store over 742 million tax returns and 175 million sets of fingerprints by transferring all those records onto magnetic computer tape that had to be stored in a single location. In 1989 British computer scientist Tim Berners-Lee invented eventually the World Wide Web.

The 21st Century : In 2005 Roger Mougalas from O'Reilly Media coined the term ***Big Data*** for the first time. It refers to a large set of data that is almost impossible to manage and process using traditional business intelligence tools.

2005 is also the year that Hadoop was created by Yahoo!. It's goal was to index the entire World Wide Web and nowadays the open-source Hadoop is used by a lot organizations to crunch through huge amounts of data.

In 2010 Eric Schmidt speaks at the Techonomy conference in Lake Tahoe in California and he states that "there were 5 exabytes of information created by the entire world between the dawn of civilization and 2003.Now that same amount is created every two days."

In 2011 the McKinsey report on Big Data: *The next frontier for innovation, competition, and productivity,* states that in 2018 the USA alone will face a shortage of 140,000 – 190,000 data scientist as well as 1.5 million data managers.

Big Data Vs Data Mining

Big data and data mining are two different things. Both of them relate to the use of large data sets to handle the collection or reporting of data that serves businesses or other recipients. However, the two terms are used for two different elements of this kind of operation. Big data is a term for a large data set. Big data sets are those that outgrow the simple kind of database and data handling architectures that were used in earlier times, when big data was more expensive and less feasible. For example, sets of data that are too large to be easily handled in a Microsoft Excel spreadsheet could be referred to as big data sets.

Data mining refers to the activity of going through big data sets to look for relevant or pertinent information. This type of activity is really a good example of the old axiom "looking for a needle in a haystack."

The idea is that businesses collect massive sets of data that may be homogeneous or automatically collected. Decision-makers need access to smaller, more specific pieces of data from those large sets. They use data mining to uncover the pieces of information that will inform leadership and help chart the course for a business.

The big data market is strong and thriving — although it isn't always called "big data" these days. The term "big data" first became part of the tech lexicon in the late 1990s, when people like John Mashey at SGI began using the phrase to describe the enormous and growing stores of enterprise data that were difficult to store and analyze using the technology available at the time.

In 2001, analyst Doug Laney suggested a definition of big data that included three Vs: volume, velocity and variety. Over the next few years, Laney's definition became something of an industry standard, and some people added a fourth V — variability — to the definition. Although big data doesn't refer to any specific quantity, the term is often used when speaking about petabytes and hexabytes of data, much of which cannot be integrated easily.

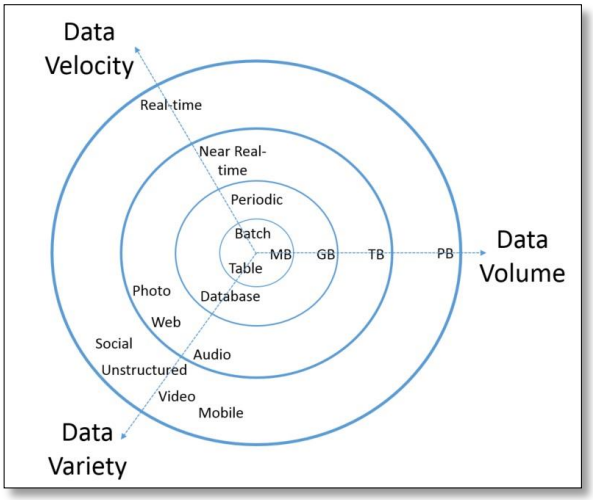

Figure 1: 3 Vs of Big Data

Big Data: Contribution and Future

Big data is no longer just a buzzword. Researchers at Forrester have "found that, in 2016, almost 40 percent of firms are implementing and expanding big data technology adoption. Another 30 percent are planning to adopt big data in the next 12 months." Similarly, the Big Data Executive Survey 2016 from NewVantage Partners found that 62.5 percent of firms now have at least one big data project in production, and only 5.4 percent of organizations have no big data initiatives planned or underway.

Researchers say the adoption of big data technologies is unlikely to slow anytime soon. IDC predicts that the big data and business analytics market will increase from $130.1 billion this year to more than $203 billion in 2020. "The availability of data, a new generation of technology, and a cultural shift toward data-driven decision making continue to drive demand for big data and analytics technology and services," said Dan Vesset, group vice president, analytics and information management. "This market is forecast to grow 11.3 percent in 2016 after revenues reached $122 billion worldwide in 2015 and is expected to continue at a compound annual growth rate (CAGR) of 11.7 percent through 2020." While it's clear that the big data market will grow, *how* organizations will be using their big data is a little less clear. New big data technologies are entering the market, while use of some older technologies continues to grow.

BIG DATA: PRESENT AND FUTURE IN INDIA

Big data has attracted global attention since it's very emergence as a buzzing trend. Currently it is one of the most influential factors in any market. It is rightly said that data is the new oil. Big data can hypothetically be called a system that acquires the crude oil and makes fuel out if it. It is a means of collecting, processing, and analysing data which can be gotten from a wide range of different sources.

The volume of the big data market?

According to a study by NASSCOM the Indian analytics industry is predicted to reach $16 billion mark by 2025. If the prediction comes true, India will have 32% of the global market. Now, we all know that big data is the biggest player in the Indian analytics space. The application of big data in Indian industries is not happening at an all inclusive scale yet the growth is exponential. The industry is expected to grow at a CAGR of 26% up till 2025.

Around 90,000 analytics professionals are currently employed by Indian companies. And there is still a fair bit of skill shortage. There are nearly 600 companies operating in the niche of analytics products and services 400 of these are startups. There is great value added to skills like data mining, predictive analysis, etc.

Big data influencing governance

The Indian administrative body has recognized the power of big data and are Keen enough to put it to good use. A shining example would be the Comptroller and Auditor General or CAG of India. They have drafted a 'Big data management policy'. The main motive behind this is to better audit the humongous amount of data generated by the public sector in the states and the union territories.

Data about power consumption is also being monitored. The data is compared to historical data in order to draw patterns.

Digital space in India

It would be unwise to say that digitization had reached every corner of India – that is still held up for distant future. But what we can say is that more corners of India have come under the umbrella of big data in the last year than ever before.

Currently the internet usage rate in India is next only to USA. Very soon India will have the highest internet using population as well as the highest smart-phone using population. This has changed the big data market forever. The market has expanded through the remote corners of the country.

Turning to Big data

A considerable number of industries have taken a step forward towards big data analytics. The first one that needs mention is the

Finance and banking sector

Big data has emerged as a great way for financial firms to evade risks and minimize losses. Most finance based companies and banks keep a database of taxpayers and borrowers. They analyze historical data to ensure safe returns of their money. While it is understandable that the prediction may not be full proof all the time, finance corps reportedly are happier after implementing big data analytics.

Healthcare has a brighter future

There is a lot more unstructured data in the health sector than we may imagine. It starts with the prescriptions and goes right to the insurance claims and critical diagnostic tests.

This data can be turned into actionable resource with the help of big data analytics. It is unreasonable to think that India struggling to provide efficient rural health support across the country will efficiently employ big data in a large scale to improve health care services right now. But the hopeful thing is that it is happening even if at a slow pace.

- Startups have come up that are using clinical data to improve patient care.
- As automation steps in it will be compulsory to employ analytics.
- The perks of using predictive analysis in diagnosis and prescription is surfacing fast.

The point is that the application of analytics in healthcare will surely make services swifter and more easily available for the mass.

Just like finance, administration and health care, the defence sector is also making good use of data analysis to enhance the security systems.

Building the workforce

The primary concern about the big data analytics space has been an acute shortage of skilled hands. India is dealing with this problem with surprising vigour. There are a considerable number of analytics institutes in India that are providing training for both the fresh and the experienced candidates. The state of the art training facilities are available in NCR and Bangalore. Big data training in Bangalore is producing a large number of impressive analysts on a yearly basis. India has the brains, they just need nurturing to take the industry to new heights

FUTURE SCOPE

Truly keeping track of Big Data trends is like trying to monitor the daily shifts in the wind – the minute you sense a direction, it changes. Open source applications like Apache Hadoop, Spark and others have come to dominate the big data space, and that trend looks likely to continue.

As big data analytics capabilities have progressed, some enterprises have begun investing in machine learning (ML). Machine learning is a branch of artificial intelligence that focuses on allowing computers to learn new things without being explicitly programmed. In other words, it analyzes existing big data stores to come to conclusions which change how the application behaves.

The Internet of Things is also likely to have a sizable impact on big data. One new technology that could help companies deal with their IoT big data is edge computing. In edge computing, the big data analysis happens very close to the IoT devices and sensors instead of in a data center or the cloud.

CONCLUSION

We are living in the big data era where enormous amounts of heterogeneous, semistructured and unstructured data are continually generated at unprecedented scale. Big data discloses the limitations of existing data mining techniques, resulted in a series of new challenges related to big data mining. Big data mining is a promising research area. In spite of the limited work done on big data mining so far, we believe that much work is required to overcome its challenges.

REFERENCES

1. U. Fayyad Big Data Analytics: Applications and Opportunities in On-line Predictive Modeling. http:// big-datamining.org/keynotes/#fayyad 2012

2. “IBM What Is Big Data: Bring Big Data to the Enterprise,” http://www-01.ibm.com/software/data/bigdata/2012.

3. A. Rajaraman and J. Ullman, Mining of Massive Data Sets.Cambridge Univ. Press.2011

4. Witten, Ian H., et al. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.A. Cichocki and R. Unbehaven. Neural Networks for Optimization and Signal Processing, 1st ed. Chichester, U.K.: Wiley, 1993, ch. 2, pp. 45-47.

5. Team, R. Core. "R language definition." Vienna, Austria: R foundation for statistical computing (2000).

6. Puneet Singh Duggal and Sanchita Paul, Big Data Analysis: Challenges and Solutions.

7. Han Hu, Yongyang Nen, Tat Seng Chua, Xuelong Li, Towards Scalable System for Big Data Analytics: A Technology Tutorial, IEEE Access, Volume 2, Page No 653, June 2014.

8. Wei Fan and Albert Bifet, Mining Big Data: Current Status, and Forecast to the Future, SIGKDD Explorations, Volume 14, Issue 2, 2012.
9. S.Vikram Phaneendra and E.Madhusudhan Reddy, Big Data- solutions for RDBMS problems- A survey, IEEE/IFIP Network Operations & Management Symposium (NOMS 2010),Osaka Japan, Apr 19-23 2013.
10. Hardeep Kaur, A Review of Applications of Data Mining in the Field of Education, IJARCCE, Vol. 4, Issue 4, April 2015.
11. Kishor, D., Big Data: The New Challenges in Data Mining, IJIRCST, 1(2), pp. 39-42, 2013.
12. Dheeraj Agarwal, A comprehensive study of data mining and applications, IJARCET, Vol , issue 1, January 2013.
13. Witten IH, Frank E, Hall MA, Pal CJ. Data mining, Fourth Edition: Practical machine learning tools and techniques. 4th ed. San Francisco: Morgan Kaufmann Publishers Inc.; 2016.
14. Kotsiantis SB. Supervised machine learning: a review of classifcation techniques. In: Proceedings of the 2007 conference on emerging artifcial intelligence applications in computer engineering: Real Word AI Systems with applications in eHealth, HCI, Information Retrieval and Pervasive Technologies. IOS Press, Amsterdam, The Netherlands, The Netherlands; 2007. p. 3–24. http://dl.acm.org/citation.cfm?id=1566770.1566773.