Big Data Mining: Challenges, Technologies, Tools and Applications

Asha M. PAWAR
Assistant Professor, SKNCOE, Computer Engineering Dept.
aashapawar20@gmail.com

Big data is a data with large size means it has large volume, velocity and variety. Now a day's big data is expanding in a various science and engineering fields. And so there are many challenges to manage and analyse big data using various tools. This paper introduces the big data and its Characteristic concepts and Next section elaborates about the Challenges in Big data. In Particular, wed discuss about the technologies used in big data Analysis and Which Tools are mainly used to analyse the data. As big data is growing day by day there are lot of application areas where we need to use any of the technology and tools discussed in paper. Mainly this paper focuses on the Challenges, Technologies, Tools and Applications used for big data Analysis.

Keywords: *big data, big data analysis, mining, heterogeneous data.*

1 Introduction

I) Data:

Data is unprocessed raw material that may or may not specify meaning. Data gives the general values to the user. E.g. Financial data, medical data and so on. It gives the collection of the things at one place. But it is very difficult to find out some meaning from those raw figures so that we need to process this data to get some meaningful information. Data may be in the form of fact, text and numbers.

II) Information:

Data can be represented in information to perform some analysis. Information gives us some meaningful information about the raw data. It is processed and interpreted in some specific format.

III) Knowledge:

Knowledge is the processed information that may give us the typical structure or value. With information it adds some experience and gives the correct analysis.

Data ⟶ Information ⟶ Knowledge

E.g. in Financial Services like Bank the data is all collected information from all stakeholders of bank, Information is separate department data which is represented in some specific format and Knowledge is withdraw process by each customer with previous experience.

Data can be represented in different kinds like Relational Data, Transaction Data and Data Cubes.

IV) Data Mining

Data Mining is one important way to analyse the data in some proper format. Data Mining is a process in which data is analysed on different criteria and summarize it for further use. In other words Data Mining is extract information from large set of data values. That means mining knowledge from large data values is Data Mining also referred as Knowledge Discovery. Data Mining can be useful in different areas like fraud detection, Market analysis, Target Analysis.

V) Big Data and its characteristic

Big Data is typically a data itself with large size and it is difficult to maintain, collect and manage by any analysis tool or user. E.g. Data stored on Facebook, twitter are Big data. The volume and velocity of this kind of data is too large. Big data includes video, photo, audio, and simulation and 3D models [3].

There are some important characteristic of big data like Volume, Velocity, Variety, Veracity, and Value. These 5V's are main characteristics of the big data [1], [2] :

1. Volume represents the size of the database in terabyte or petabyte.
2. Velocity is the speed of data in communication in a real world.
3. Veracity describes Incompleteness data.
4. Variety gives us the structure of different data types used in real world.
5. Value represents the important information from the data source.

VI) Big Data Mining

Data mining is the process to gather information in same domain. The purpose of data mining is to collect information for any particular issue which will helpful to data analyst for classification and prediction. So now todays challenge is to mine this large volume of data so that it can be analysed by any mining tool.

2 Challenges of Big Data Mining

a. Volume and Scalability.

It is the biggest challenge to deal with the size of data. As Twitter generates 7 + Terabytes of data and Facebook generate 10 + Terabytes of data every year so it becomes difficult to manage and analyse. As we are moving from Terabytes to Petabytes and from Petabytes to Zeta bytes of data it's the important task to analyse this Big data by some methodology. Scale the data in proper way is the important issue in big data mining.

b. Miss-Handling of Big Data

Data handling mainly depends on the scalability of data. And scalability depends on data size, hardware size, and concurrency. Day by day data size is increasing and format to store data is also changing and not fixed in future so it's the task of data analyst to overcome such challenge so mishandling of data by different users.

c. Privacy and Security

In Big data, data size and format are not fixed so it's difficult to maintain privacy of one user from another. And because of this volume of data security algorithms are not fixed. When size of data changes or format changes then we need to apply new security algorithms. Ones we define the security or privacy algorithms to it cannot be applicable to upgraded data. E.g. In hospital the data collected and it may upgrade daily and it may be in different format, so it becomes difficult to analyse and secure the newly added data.

As data is linked with so many formats and users it's a fear to keep privacy of data and hence it's a big challenge in data mining.

d. Speed and Velocity

Velocity refers to unique speed with timely manner. But in many cases it is difficult to maintain unique speed because of variety and size of data.

e. Heterogeneity of Data

Data analysis has first step that data must be structured in a well format. Some errors and confusion in data may lead to miss-classification of data. Machine analysis algorithm only understands homogeneous or structured data. Hence to make the data in homogeneous format is a big challenge in big data mining.

3. Big Data Mining Techniques

a. ANN

In general ANN is called as "Neural Network". NN is a non-linear statistical data modelling approach and used to manage complex relationships between I/P and O/P. As dataset used in ANN grows massively we need to analyse it automatically. It is also helpful to recognize the pattern from which it belongs. Classification, Prediction, Clustering & Association Rules are the steps of data mining and are useful in neural network to identify patterns. [8]

b. Decision Tree

It is one of most powerful tool in data mining process. Decision tree is originally implemented in decision theory and

statistics. It is used for nominal and numeric data values. Decision analysis is performed with the help of tree shaped structure. Three main components used in decision tree are 1) Square represents decision node 2) Circle represents chance node and 3) Triangle represents end node.

c. Genetic Algorithm

Genetic algorithm involves three steps selection, crossover and mutation for every gene. It follows survival of fittest law. Genetic algorithm has found applications in phylogenetic, computational science, engineering, economics and many more. GA can be used as a classifier in many areas also it can be used for prediction analysis purpose. GA uses many techniques like K-Nearest Neighbour (KNN) and Rule Induction. [9] [12].

d. Classification

Classification is a method of assigning a label to unclassified data. There are different methods in classification like Supervised and Unsupervised classification methods. There are many techniques used for classification like Bayesian Classification, ANN and Support Vector Machine (SVM).

e. Clustering

Clustering is the process of making the groups of similar objects together. While performing clustering we need to group the items on data similarity and assign the label to each group. It is more advantageous than classification because it helps to find useful features that represent different groups. Clustering has many application areas like market research, pattern recognition, data analysis, and image processing. Clustering deals with the High Dimensional Data. It has many techniques to make groups like Partitioning Method, Hierarchical Method, Density-based Method, Grid-Based Method, Model-Based Method, and Constraint-based Method.

4. Applications of Big Data

a. Healthcare and medicine

Big data creates the link between patients, doctors, diagnosis and predicting the surgery and pharmaceutical companies. Also Big data haves the computing power too high to decode DNA sequence and predict the disease or required pattern. Big data can also be helpful to monitor premature babies and sick baby unit. By analysing every heartbeat now a days it can be possible to identify the disease before its actual symptoms. [6]

b. Banking

In financial sector it will also help for predictive area in customer behaviour. Different applications in banking like credit risk analysis, customer changes, marketing policy, and historical transaction are handled by many big data tools. Business Intelligence with respect to financial banking can be handled by Big data tools.[7]

c. Telecommunication

Nowadays in telecommunication field call analysis, pricing, prediction of funds, customer loyalty and customer ratio all these can be handled by Big data mining tools.
Mobile user data mining tool is one example.

d. Marketing

Marketers and retailers uses the big data mining tools to decide whether advertisement is in correct way or not, based on customer facial recognition software. Because marketers want that advertisement is so viral so that they can be useful to improve sell of product. Also big data mining tools are helpful to prepare customer report and shopping cart analysis. [7]

e. Industry

Many Big data mining tools are helpful to prepare report on production management like quality of the product, process

optimization, store inventory and employee management on various products.
It also helpful to optimize staffing through data, reduce fraud and timely analysis of inventory.

f. Social Media

Now a day, social sites generates tremendous data every day, so many tools are used to analyse and manage it properly.

5. Tools used in Big Data

Big data deals with many data types like structure, unstructured and also volume of data is too large so to analyse these data we need some tool like Hadoop, NoSQL, MapReduce, R-Language, RapidMiner, WEKA and KNIME.

a. Hadoop

Hadoop is an open source software, that stores and process big data. It uses cluster approach to manage big data. It is java based framework allows to use inbuilt library functions and different tools. Hadoop can execute large, complex data set on different operating system like Windows, LINUX, and UNIX and so on. GOOGLE and YAHOO use Hadoop Framework to analyse their data efficiently.

b. NoSQL

NoSQL refers to not only SQL. NoSQL allows using structured data as well as un-structured data. Means data can be accessed by SQL queries as well as some new techniques can be used. NoSQL is open source software and uses DaaS (Database as a Service). As Graph databases are more popular solution for big data NoSQL is the right application who manages it efficiently.

c. MapReduce

It is processing model for distributed environment on parallel architecture. It is an open source tool to implement process and manage large data efficiently. MapReduce can be categorized in two parts Map converts the data into another type of data and divides into data tuples like key/value pair and Reduce takes input from Map and combines data tuples into small set of tuples.
MapReduce is used to sort Petabytes of data in an hour.

d. R-Language

developed by Bell Labs. R uses S Programming to handle large volume of data.
R is programming language for statistical values, complex data and graphical information. Effective data handling and storage can be done by R-Language. R provides graphical facilities for data analyst and has many tools to perform data analysis

6. Conclusions

The basic purpose of this paper is to introduce basic concepts about big data and its mining techniques. It's a survey about the big data analysis methods. Hence we introduce in first part all basic about the big data mining then we introduce the different challenges to handle and manage big data. Lateral part gives information about various techniques used in data mining to improve the accuracy and performance. Than we discuss about the tools used for the big data analysis for different techniques of big data. And finally we discuss about the applications of big data in various domains. In future, Big data and Cloud are nearer terms so we can further expand big data analysis using cloud technology.

Acknowledgment

I am really thankful to my Husband Mr. Mohan Pawar for guiding me. Also I am thankful to all my friends who encourage me all the time.

References

[1]. A Survey on Big Data, Mining: (Tools, Techniques, Applications and Notable Uses) Nour E. Oweis 1,4 , Suhail S. wais 2 , Waseem George 1 , Mona G. Suliman 3 , Václav Snášel 1,4 Springer Publication.

[2]. Kudyba, S. (2014). Big Data, Mining, and Analytics: Components of Strategic Decision Making. CRC Press.

[3]. Gupta, R. (2014). Journey from Data Mining to Web Mining to Big Data. arXiv preprint arXiv:1404.4140.

[4]. Big Data Challenges Alexandru Adrian T OLE Romanian American University ,Bucharest, Romania

[5]. A Survey on Big Data, Mining: (Tools, Techniques, Applications and Notable Uses) Nour E. Oweis 1,4 , Suhail S. Owais 2 , Waseem George 1 , Mona G. Suliman 3 , Václav Snášel 1,4 Springer Publications.

[6]. Big Data Challenges: Data Analysis Perspective Riya Lodha , Harshil Jain and Lakshmi Kurup Computer Engineering Department, D.J.Sanghvi College of Engineering, Mumbai University, Mumbai, India Accepted 10 Sept 2014, Available online 01 Oct 2014, Vol.4, No.5 (Oct 2014) International Journal of Current Engineering and Technology E-ISSN 2277 – 4106, P-ISSN 2347 – 5161 ® ©2014 INPRESSCO , All Rights Reserved

[7]. W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential", Health Inf Sci Syst, vol. 2, no. 1, p. 3, 2014.

[8]. Applications of Big data in Various Fields Kuchipu di Sr avanthi, Tatireddy Subba Reddy, Assistant Professor, CSE Department, Assistant Professor, CSE Department, QIS Institute of Technology,. VVIT ,NAMBUR, Vengamukkapalem, Ongole, India, 523272 GUNTUR, Andhra Pradesh,

[9]. Journal of Theoretical and Applied Information Technology © 2005 - 2009 JATIT. All rights reserved. www.jatit.org

[10]. NEURAL NETWORKS IN DATA MINING DR. YASHPAL SINGH, ALOK SINGH CHAUHAN Reader, Bundelkhand Institute of Engineering & Technology, Jhansi, India Lecturer, United Institute of Management, Allahabad, India E-mail: yash_biet@yahoo.co.in , alok_sc@yahoo.co.in

[11]. International Journal of Information Management 35 (2015) 137–144 Contents lists available at ScienceDirect International Journal of Information Management jou rnal h om epage: www.elsevier.com/locate/ijinfomgt

[12]. Beyond the hype: Big data concepts, methods, and analytics Amir Gandomi ∗ , Murtaza Haider Ted Rogers School of Management, Ryerson University, Toronto, Ontario M5B 2K3, Canada

[13]. Lahoti, A. A., & Ramteke, P. L. (2014). Data Mining Technique its Needs and Using Ap-plications. IJCSMC, Vol. 3, Issue. 4, April 2014, pg.572 – 579.

[14]. http://www.ibm.com/developerworks/library/ba-data-mining-techniques/

[15]. http://thenewstack.io/six-of-the-best-open-source-data-mining-tools/

[16]. The Four V's of Big Data – IBM http://www.ibmbigdatahub.com/infographic/four-vs-big- data (last seen 05–April–2015).

[17]. Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data.Knowledge and Data Engineering, IEEE Transactions on, 26(1), 97-107.

[18]. Jain, N., & Srivastava, V. (2013). DATA MINING TECHNIQUES: A SURVEY PAPER. IJRET: International Journal of Research in Engineering and Technology.

[19]. Saed Sayad, Data Mining Map, An Introduction to Data Mining, http://www.saedsayad.com/ (2012). (Last seen 05–April–2015).

Mrs. Asha M. Pawar graduated from BAMU University in 2005 from Information Technology field. She has completed her M.Tech Computer Science & Engineering from VTU Belgaum in 2012. She has total 11 Yrs of Experience as an Assistant Professor in Pune University. She has published total 6 papers in various international journals.