



Data Mining with Big Data Analysis Algorithm Tools, Application and Challenges

Dr. Shalini Lamba^a, Shivam chaurasiya^b

^a Head of Department, Department of Computer Science, National Post Graduate College, Lucknow, India

^b Student, Department of Computer Science, National Post Graduate College, Lucknow, India

drshalinilamba@gmail.com, shivamchaurasiya08738@gmail.com,

KEYWORD

Big data,
Data Mining,
Big Data Mining
Algorithm, Big-Data
challenges, Big data
Tools

ABSTRACT

Organizations now days gather and store vast volumes of data in the hopes that it may be valuable later. Big data can be utilized by enterprises to accomplish a range of goals where success depends on astute analysis, in addition to seeking greater insights for enhancing the quality of their services and profit. Measure mistakes, noise accumulation, spurious correlation, scalability or storage bottlenecks, and other special computing or statistical problems are brought about by big data. These problems are unique and need for a modern statistical and computational framework. This paper presents the literature criticism about the Big data Mining and the problems and challenges including emphasis on the distinguished features of Big Data. It also covers a few strategies for working with large amounts of data. An overview of big data, including its type, source, and features, is provided in this work. This research also includes an assessment of several large data mining platforms, techniques, and problems. However, big data also brings with it a number of obstacles, including those related to data storage, analysis, visualization, and capture. The purpose of this article is to provide an in-depth understanding of big data, including its applications, opportunities, and challenges. It also aims to showcase the cutting-edge approaches and technology that we now use to address big data issues.

1. Introduction

We must now fiercely confront and resist the exabytes period of history as the petabyte era has come to an end. Millions of individuals have benefited from the technological revolution as a result of the creation of huge/extreme data through the ever-increasing use of digital equipment, particularly remote sensors that produce continuous streams of digital data, or "big data." The fact that massive volumes of data are being created on a never-before-seen and ever-increasing scale is a confirmed major occurrence or large event. For instance, a survey indicates that over 2 million inquiries are sent to Google every day, over 2 million pieces of content are shared on Facebook, over 2 million hours of video are uploaded on YouTube, etc. Gathering meaningful information from this massive volume of data is our primary task. Numerous technologies are emerging to meet these needs, such as cloud computing—that is, Google's paradigm. Reduce Map, etc. From the perspective of data mining, one of our biggest challenges is extracting information from Big Data. Predicting future trends and making a variety of business decisions can both benefit from the retrieved knowledge. Decisions made by organizations can be informed by knowledge. There are many different data mining techniques available for extracting knowledge from databases. To increase performance, these techniques are frequently used in conjunction with distributed storage systems and parallel processing architectures.

Corresponding Author: Dr. Shalini Lamba, Department of Computer Science, National P.G. College, Lucknow, India
Email: drshalinilamba@gmail.com

2. Big Data

The collection of quantities and variables that are somewhat connected and somewhat dissimilar is known as data. The sizes of databases have, however, increased dramatically recently. Big data is described as data that is overwhelming in terms of amount, variety, velocity, and connections to other data, making it challenging to handle using conventional database management systems or tools. When a dataset is thought to be able to perform security, curation, analysis, and perception using modern advances, it can be referred to as enormous information. The organization's new information management architecture and systems are the result of innovation moving forward. Analytically produced awareness will increasingly drive decision-making efficacy. Thus, a (human or computerized) leader requires a greater chance of predicting what modifications need to be made in order to link with the organization's final aim the more accurate and convenient these are. Most importantly, big data is a multidisciplinary and developing combination of new innovations in a mix with new measurements in data handling and stockpiling (volume and speed), another phase of a variety of data sources, and the difficulty of adequately managing information quality (veracity), according to On the other hand, has defined big data's attributes using the five Vs, which are summed up as follows:







Value		Clinically relevant data Longitudinal studies
Volume		High-throughput technologies Continuous monitoring of vital signs
Velocity		High-speed processing for fast clinical decision support Increasing data generation rate by the health infrastructure
Variety		Heterogeneous and unstructured data sources Differences in frequencies and taxonomies
Veracity		Data quality is unreliable Data coming from uncontrolled environments
Variability		Seasonal health effects and disease evolution Non-deterministic models of illness and health

Fig1: Six V's of big data

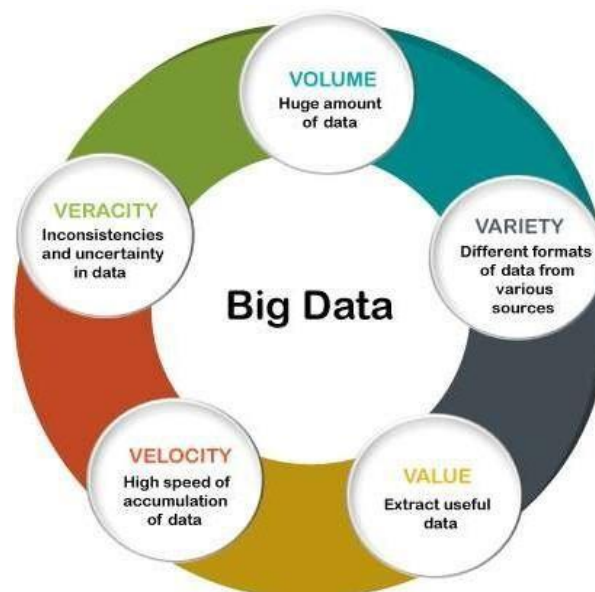


Figure 2: Vs of Big Data

There are many properties associated with big data. The prominent aspects are Volume, Velocity Variety, Veracity and Value.

Volume: The volume of big data is exploding exponentially day to day. The data accumulated through social websites and sensor networks going to cross from petabytes to Zeta bytes.

Velocity: This is a concept which indicates the speed at which the data generated and become historical. Big data is able to handle the incoming and outgoing data rapidly.

Variety: Data produced are from different categories, consists of unstructured, standard, semi structured and raw data which are very difficult to be handled by traditional systems.

Veracity: It describes the amount of variance used in summaries kept within the data bank and refers how they are spread out or closely clustered within the data set.

Value: All enterprises and e-commerce systems are keen in improving the customer relationship by providing value added services. For that, study on customer attitudes and trends in the market are to be analyzed. Moreover, users can also query the data store to find business trends and accordingly they can change their strategies. By making big data open to all, it creates transparency on functional analysis. Supporting real time decisions and experimental analysis in different locations datasets can do wonderful things for enterprises.

Big Data Issues with Challenges

Big data analysis is the act of analyzing enormous data sets using sophisticated analytics and visualization tools to find unknown relationships and hidden patterns that may be used to make wise decisions. Big Data analysis covers several unique processes, such as data recording or collection, data extraction or cleansing, data modeling or analysis, data integration, quantity and representation, query processing, and interpretation. There are difficulties in every one of these stages. Among the difficulties in large data mining include heterogeneity, scale, timeliness, complexity, and privacy.

Big Data and Data Mining

Data stored at the server of Facebook according to that amount is used by people into daily existence the place we upload a number concerning kinds on data as pictures, videos and entire about it data stored regarding the warehouse regarding data at the Facebook servers, we called such big- data due to its complexity. Big-data is nothing but a data available at autonomous and heterogeneous sources of extreme huge amount which gets up to date inside a fraction over second. Another example concerning big- data we can take like analyzing performed from an electronics microscope of the universe. Now the term Data mining be able stay defined namely extraction of useful data from the collected and gathered data or we execute speech extraction of knowledge from database. So big data mining is a close on view that contains a bunch on useful detailed facts on big-data.

BIG DATA TOOLS

Large numbers concerning tools are available after technique big data. In this section, we discuss some current strategies for analyzing big data with emphasis of some important emerging tools namely Cassandra.

➤ Hadoop.

➤ Plotly.

- Bokeh.
- Neo4j.
- Cloud era.
- Open Refine.
- Storm.

DATA MINING TASK:

- Classification (Predictive)
- Clustering (Descriptive)
- Association Rule Discovery (Descriptive)
- Sequential Pattern Discovery (Descriptive).
- Regression (Predictive).
- Deviation Detection (Predictive).
- Collaborative Filter (Predictive).

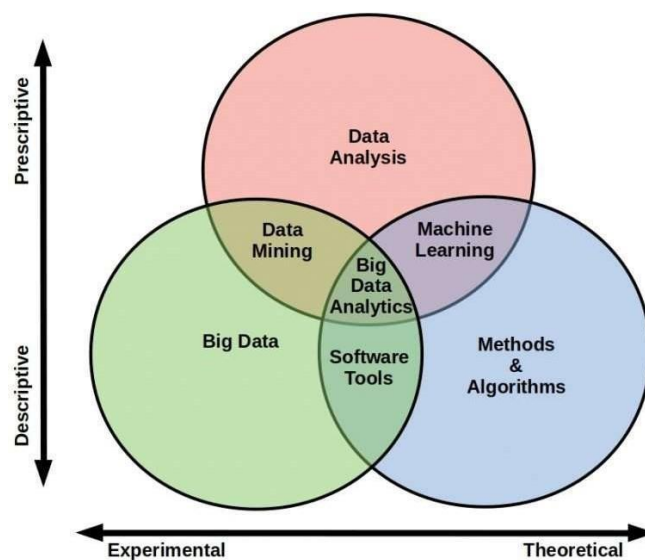
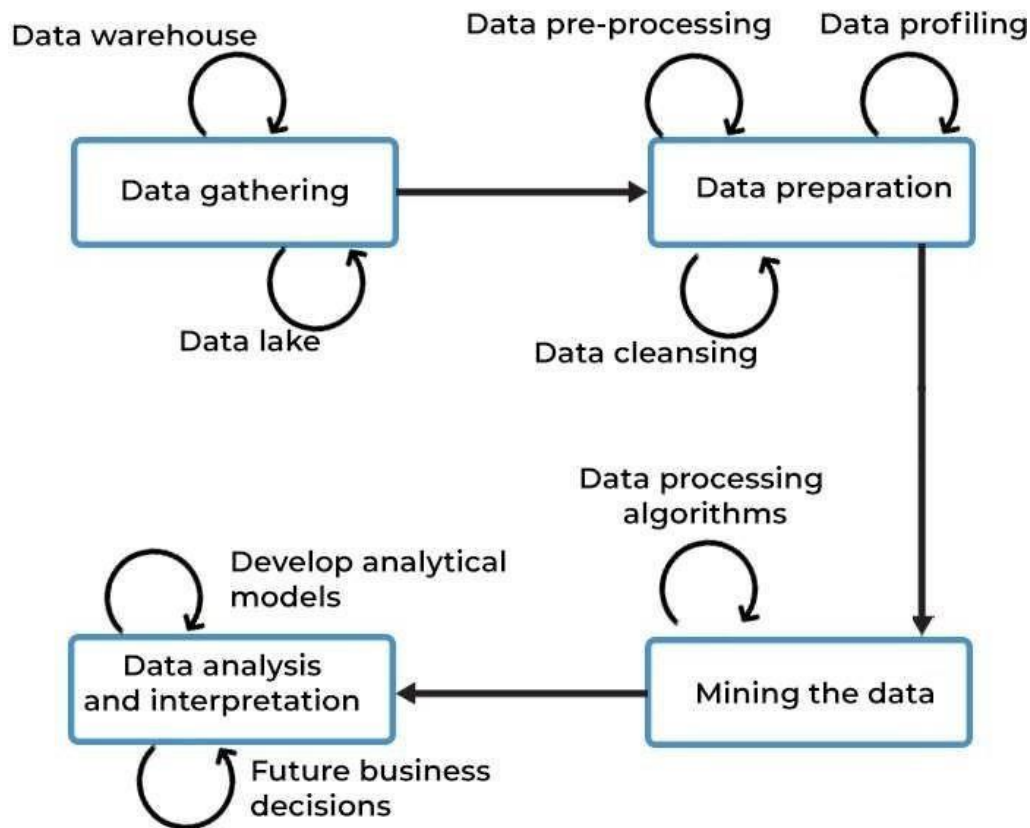


Figure 3: Big Data analytics

DATA MINING PROCESS



DATA MINING TOOLS

Strong software tools are needed for the creation and usage of data mining algorithms. The selection of the best appropriate tool gets harder when there are more and more tools available. Additionally, we suggest criteria for the tool classification that take into account various user groups, platforms, license policies, data structures, data mining tasks and techniques, visualization and interaction styles, import and export options for data and models, and visualization and visualization styles. Each tool has benefits and drawbacks of its own. A set of tools within data mining was created by the data analysis and research communities. They are provided under one of the current open source licenses for free. Businesses are able to make proactive, informed decisions by using data mining techniques that forecast future trends and behaviors. Numerous open-source data mining tools are available.

- Orange. Availability: Open source.
- Weka. Availability: Free software.
- KNIME. Availability: Open Source.
- Sisense. Availability: Licensed.
- Apache Mahout.
- Rapid Miner. Availability: Open source
- Oracle Data Mining

- Data Melt.

USE OF DATA MINING IN VARIOUS SECTORS:-

- Banking
- Marketing
- Health Care
- Manufacturing and Production
- Insurance
- Law
- Government and Defense
- Computer hardware and software
- Airlines
- Brokerage and Securities trading.

CHALLENGES FACED BY DATA MINING: -

- Data quality
- Privacy preservation
- Network Setting
- Data Ownership and distribution
- Complex and Heterogeneous data
- Scalability
- Streaming Data
- Dimensionality.

BIG DATA ANALYSIS ALGORITHMS MINING ALGORITHMS FOR SPECIFIC PROBLEM

Fan and Bifet noted that although the massive information problems have been present for about a decade, the terms "big data" and "big data mining" were initially used in 1998. Finding something from massive amounts of data will be one of the key challenges in this field, as big data and big data mining have almost completely emerged in the interim. In terms of computation cost, memory requirement, and output precision, data mining techniques for data analytics also play a crucial role in big data analysis. Here, we will briefly address this from the perspective of analysis and search computations to make sense of its importance for big data analytics.

CLUSTERING ALGORITHMS

Because traditional bunching calculations typically require that all of the data be in a similar arrangement and be stacked into a similar machine in order to locate some useful things from the entire information, they will prove to be much more limited in the big data age. The characteristics of big data still raised some new challenges for data clustering issues, even though the problem of decomposing large-scale and high-dimensional dataset has drawn in many analysts from different traits in the last century and some arrangements have been demonstrated recently.

CLASSIFICATION ALGORITHMS

Similar to the big data mining clustering algorithm, some studies also tried to modify the traditional classification algorithms to make them attempt a parallel computing scenario or to enhance new classification algorithms that function normally in a parallel computing environment. The outline of the classification method is defined in [3] as the information gathered from dispersed data sources and managed by a diverse group of learners.

FREQUENT PATTERN MINING ALGORITHMS

Due to the fact that early frequency pattern mining methodologies attempted to analyze data from large shopping mall transaction data, the majority of frequency pattern mining researchers (association rules and sequential pattern mining) focused on maintaining large-scale datasets at the earliest reference point. Since there are more than "tens of thousands" of transactions, the problems regarding how to handle the large volume of data were studied for a considerable amount of time. For instance, the F P -tree uses the tree structure to incorporate the frequency pattern in order to further reduce the association rule mining calculation time.

COMMUNITY DETECTION ALGORITHMS

Because early approaches to community detection attempted to examine the data, research on community detection first concentrated on managing small group datasets. A multitude of studies have demonstrated the effective time complexity in identifying similar communities over a broad range of data. The integration of different algorithms relies on a top-down or bottom-up strategy.

CONCLUSION AND FUTURE WORK

Big data is going to continue growing during the next years and each data scientist will have to manage much more amount of data every year. The data is going to be larger, diverse and faster. Many technical challenges like implementations and visualizations are to be taken into consideration in future. This is just the survey paper which shows the demand of big data and how big companies are taking interest in it. We are at the beginning of a new era where big data mining will help us to discover knowledge that no one has discovered before. To manage and analyze edge data explore business opportunities deriving from the analytics of edge data. Collaborate with the business to understand existing edge system and the potential use for data. This document provides an overview of the different large data mining algorithms and platforms. High-performance computer platforms are needed to facilitate big data mining. It is acknowledged that the most difficult problems facing big data mining today include finding intriguing patterns, creating an effective global unifying theory, creating effective mining platforms or algorithms, protecting privacy, security, trust, and data integrity. Big Data is quickly emerging as the new frontier in corporate and scientific data applications. The automatic discovery of intelligence entailed in the often occurring patterns and hidden norms is growing dependent on big data analysis. Big data analysis assists businesses in making better decisions, anticipating and recognizing changes, and seeing new opportunities.. It can be concluded from the findings to that amount Enterprise are still looking for the right infrastructure tools so much will enable to them after effectively deal with theirs big-data, among row including theirs business needs. Most companies are already the use of dedicated big-data tools however all still see gaps within capabilities or hold concern related to the suit between these tools or their current and expected needs.

References

- [1] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. In : Proceedings of the ACM SIGMOD International Conference on Management of Data, 2000. pp. 1–12.
- [2] Chiang M-C, Tsai C-W, Yang C-S. A time-efficient pattern reduction algorithm for k-means clustering. Inform Sci. 2011;181(4):356 31. Russom P. Big data analytics. TDWI: Tech. Rep ; 2011.
- [3] Tekin C, van der Schaar M. Distributed online big data classification using context information. In: Proceedings of the Allerton Conference on Communication, Control, and Computing, 2013. pp 14
- [4] IDC, Extracting Value from Chaos: <http://idcdocserv.com/1142>, june 2011
- [5] O'Reilly Radar, What is bigdata? <http://radar.oreilly.com/2012/01/what-is-big-data.html>. January 11,2012. Volume 8, Issue IX, SEPTEMBER/2018 Page No:1623
- [6] Peter Buneman, Semistructured Data <http://homepages.inf.ed.ac.uk/opb/papers/PODS1997a.pdf>, 1997.

- [7] Jaseena K, David J. Issues, challenges, and solutions: Big data mining. *Journal of Computer Science and Information Technology*. 2014;131–40.
- [8] Mouthaan N. Effects of big data analytics on organizations' value creation; 2012.
- [9] Chen C, Zhang C. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*. 2014; 275:314–7.
- [10] Ericsson White Paper. Big data analytics: actionable insights for the communication service provider;2015.
- [11] Buhl H, Röglinger M, Moser F, Heidemann J. Big data. *Wirtschaftsinformatik*. 2013; 55(2):63–8.
- [12] Lawal Z, Zakari R, Shuaibu M, Bala A. A review: Issues and challenges in big data from analytic and storage perspectives. 2016; 5(3):4–6.
- [13] Fan W, Bifet A. Mining big data: current status, and forecast to the future. *ACM SIGKDD ExplorNewslett*. 2013;14(2):1–5.
- [14] Diebold FX. On the origin(s) and development of the term “big data”, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania, Tech. Rep. 2012. [Online]. Available: <http://economics.sas.upenn.edu/sites/economics.sas.upenn.edu/files/12-037.pdf>.
- [15] Weiss SM, Indurkha N. Predictive data mining: a practical guide. San Francisco: Morgan Kaufmann Publishers Inc.; 1998.