An Overview of Big Data Mining and its Application

Pooja Sharma*, B.D.K. Patro

Department of Computer Science, Maharishi University of Information Technology, Lucknow, India

Review Article

ABSTRACT

The field of big data is currently gaining prominence as a subject of research and is often discussed in contemporary discourse. The use of big data is poised to exert a substantial influence on several domains such as research, business, industry, government and society. These sectors are anticipated to undergo a comprehensive transformation in the forthcoming years. The concept of big data is frequently elucidated by considering its key dimensions, namely volume, velocity, veracity and variety. In order to classify a problem as a Big Data problem, it is generally considered that either all or any one of the specified criteria must be satisfied. The exponential growth in data production is attributed to the remarkable capabilities of computing and communication equipment, as well as our unparalleled ability to collect, capture and distribute extensive amounts of data. In contemporary times, the rapid influx of data surpasses the processing capabilities of conventional algorithms. Consequently, ensuring data quality remains a significant challenge that necessitates addressing either during the data pre-processing phase or through the use of advanced learning algorithms. In addition to the quantitative dimension of data, a significant worry arises from the extensive range of data that is accessible for a particular topic of study. The task of extracting pertinent information from this vast array of data poses a formidable challenge. The objective of the present research paper is "to present an overview of big data, characteristics and its application and discusses challenges and their potential solutions".

Keywords: Big data; Data mining; Data mining application; Characteristics; Big data analytic

INTRODUCTION

Information Technology (IT) has seen tremendous growth in the 21st century. IT, web applications/WWW, social-networking sites and globalisation have all contributed to the exponential growth in data and information production that has been dubbed the "information explosion." Big data is the term used to describe this massive amount of data. Data mining is the process of sifting through and making sense of massive amounts of information stored digitally. Many commercial settings have benefited greatly from data mining's application. The idea is to locate "actionable information," or data that may be put to use immediately to boost earnings.

Big data tries to understand complicated and changing data relationships using massive, heterogeneous, independent sources with distributed and decentralised control. Big data has several main features, making it difficult to uncover useful patterns. Big Data is characterised by its massive dimensionality and heterogeneity. Different users use different schemas and apps, therefore data is represented differently. Big data exceeds the processing power of conventional database systems. This data needs a different processing method to be valuable.

Researchers say big data "exceeds the reach of commonly used hardware environments and software tools to capture, manage and process it within a tolerable elapsed time for its user population". One description is "Big data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze". According to these definitions, big data is expanding as technology advances and its meaning varies by organisation, tools and technologies.

LITERATURE REVIEW

According to Henry G. L. Dillon and Beverley G. Hope, knowledge discovery in databases is a subfield of information science that investigates new methods for analysing data, including data mining [1].

Singh, Dileep Kumar and Swaroop, Vishnu, in this study, we explore problems with mining methodology and user interaction. The scope of data mining should include analysis, finding of new knowledge, classification and prediction. This study provides a crystal-clear focus on finding patterns that might aid in providing and improving data mining requests depending on outcomes. Concepts like high-level data mining query languages and ad hoc data mining are discussed [2-4].

The author Sikha Gupta elaborates on a number of data mining methods in her study. She describes rule-based methods as the most suitable for human judgement and the most straightforward to convert. Keeping the importance of human involvement in the data mining process in mind, this approach is particularly useful for facilitating user participation in databases including both medical and nonmedical information [5].

Data mining approaches were explored by Sharma et al., to process a dataset and determine which data points are useful for a classification test. Classification, clustering, association rule mining and similar technologies are used to tackle massive amounts of problems.

Importance of the research study

It was previously difficult to extract useful information from such large datasets, but with the advent of big data analysis, this became possible. In order to extract meaningful insights from a large dataset, data mining methods are employed. The current study article provides an overview of big data, its applications in numerous industries, the technologies that are involved in generating and handling the big data and various concerns and challenges that distinguish the properties of big data.

Objective of the research study

Objective of the present research paper is "to present an overview of big data, characteristics and its application and discusses challenges and their potential solutions".

Hypothesis of the study

- The fields of computer science and information technology have recently seen the emergence of a potentially fruitful area of research known as "big data analytic."
- Since big data is still changing in mysterious ways and most of it is noisy, highly interconnected, diverse and unreliable, it is likely that data mining techniques will continue to be highly sensitive to its particular qualities in terms of how well they function.

Data collection and research methodology

The present research draws on secondary sources. Secondary information gathered by means of a thorough literature review was also used in the study. Books, research preview studies, magazines, journals, publications, books, websites, etc. were reviewed and compiled by the researcher [6].

Research design

The present investigation was conducted using an descriptive approach.

DISCUSSION

The performance of data mining is typically affected by a number of factors, including the presence of missing information, the presence of noise, the presence of outliers and so on. There are a number of aspects of big data that contribute to the difficulty of data mining. Because big data mining is different from simple data mining in the sense that it requires not only the finding of patterns but also the large-scale storage and processing of datasets, extensions need to be introduced into existing data mining techniques. This is because big data mining differs from simple data mining in the way that it requires not only finding patterns but also the large-scale storage and processing of datasets. The use of soft computing as a method to data mining can be viewed as a potential extension of existing data mining approaches that can be applied to the study of massive data. Soft computing approaches are innovative optimisation techniques in artificial intelligence that take advantage of the tolerance for imprecision. To put it another way, these methods take advantage of the capabilities of computers to search through vast amounts of data in a quick and efficient manner. Fuzzy logic, neural networks and evolutionary algorithms are a few examples of the types of soft computing technologies that can be utilised to improve upon conventional data mining practises. Approaches based on the genetic algorithm and swarms have been used (Figure 1).

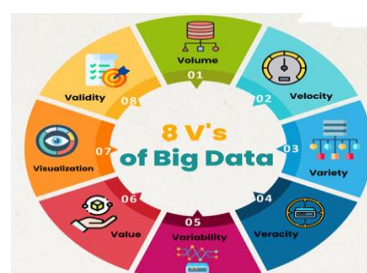**Figure 1.** The characteristics of big data.

Figure 1 illustrates the eight characteristics of big data: Volume; value; veracity; visualisation; variability; velocity; variety; and validity. The term "volume" is used to describe the extent to which data exists. The term "value" describes the insights gained from data, "veracity" describes the varying quality of acquired data, "visualisation" describes how data may be understood quickly and "variety" describes the different kinds of data that can be collected, such as structured, unstructured and semi-structured information. The terms velocity and viscosity describe the ease with which data may be processed, whereas virality refers to the pace with which information spreads from person to person. Big data's validity refers to how accurate the data is for its intended use, while big data's variability refers to how the data can be used and presented.

The term "big data" refers to a collection of data sets that can be categorised as either "structured" (metadata, tabular databases) or "unstructured" (texts, videos, photos, audio, etc.) or "semi-structured" (XML, JSON, Log Files, etc.). The problems with conventional database systems, such as storage, processing, search, analysis, transfer, privacy, etc., make it impossible to analyse big data with such a system [7]. Therefore, new methods for processing this massive amount of data must be discovered and investigated. Big Data Analytics describes this novel approach (Figure 2).

Figure 2. The applications of big data.

Data mining: Real-world implications and applications

Companies with a heavy emphasis on consumers have been the most common users of data mining because such businesses need a deep dive into customer data in order to improve product positioning, customer satisfaction, sales, transactional data, corporate profits, etc. Let's take a look at the most popular data mining uses and current trends:

Uses for data mining in medicine

The medical field is ripe with opportunity for data mining. Best practises can be found and cost-effective solutions provided, with the help of data and analytics. Data mining is an expansive methodology that may be used in many different contexts. It makes use of tools like multi-dimensional databases, statistics, machine learning, data visualisation and soft computing. Insurance fraud and abuse can be uncovered, patient care processes can be optimised to prevent unnecessary delays and the number of patients in each treatment category can be predicted.

Financial institutions and data mining

As a result of the rise of digitalization, the banking industry now processes and stores vast quantities of data and records of financial transactions. When it comes to banking, data mining apps are often the best course of action because of their capacity to reveal hidden patterns, casualties, market dangers and other correlations that management must be aware of. Results can be generated fairly rapidly for the management to make sense of without much effort, despite the amounts of data [8].

Bank management and staff can utilise the data to fine-tune their strategies for acquiring, retaining, and servicing profitable customers. When it comes to issuing credit cards, loans, etc., data mining can also assist banks swiftly identify prospective defaulters.

Uses of data mining for market segmentation

Customers have traditionally been divided into groups using traditional methods of market research, but data mining can be more precise and therefore more useful. It provides a more accurate method of segmentation and facilitates the customization of consumer needs. Decisions based on the insights obtained by data mining can assist increase customer satisfaction by targeting a vulnerable client segment.

Benefits of data mining for the classroom and education

The goal of Educational Data Mining (EDM), a newly developing field, is to develop methods for mining information contained in databases collected from educational institutions. EDM seeks to foresee how students can learn, investigate how various forms of academic support might affect students' performance and expand our understanding of the science of learning [9]. Educational institutions can use data mining to make educated decisions and forecast student accomplishment levels, allowing them to devote more resources to lesson plans and classroom instruction. Developing these methods of instruction is possible through observation of students' studying and behavioural habits.

Mining data for basket-of-goods performance

For merchants, one of the most useful modelling approaches is market basket analysis, which allows them to see how various products in stock relate to one another. Basically, it is looking for common groupings of products to order.

Consumers' shopping habits can be deduced through market basket analysis. Retailers can use this data to better accommodate their customers by catering to their needs and preferences. Differential analysis allows for straightforward comparisons to be made across various outlets and various demographics of shoppers.

Data mining's potential in stopping fraud

Traditional methods for detecting fraud are both labor-intensive and error-prone. Data mining is used in this context because it can produce reliable data and useful conclusions [10]. In a perfect world, a fraud detection system would take every precaution to safeguard user data. A model is developed to recognise and categorise this data as fraudulent or non-fraudulent through the use of a sample set of records.

Data mining's use in customer relationship management

In order to get new clients, keep existing ones happy and grow repeat business, businesses often employ CRM systems. Maintaining positive connections with clients is facilitated by data mining. Data mining technology, which gathers pertinent

data and information for analysis, is chiefly responsible for this. The information gleaned allows for the development of more efficient solutions. Use this manual as a resource to learn more about classification in data mining.

The use of data mining in industrial production engineering

The manufacturing process is often highly complex, so having access to relevant and trustworthy information is crucial. In this case, data mining applications can be quite helpful. In system-level design, they aid in spotting trends and patterns and extracting connections between the product portfolio, product architecture and user needs. Costs associated with product development, asset depreciation, manufacturing cycle times, dependencies, etc. can all be estimated with the use of data mining. Manufacturers may plan ahead for maintenance, which helps to cut down on unplanned downtime [11].

Uses of data mining in analysing research

Researchers will find data mining particularly useful for data pre-processing, database integration and general data cleansing. In order to effect change in the research, data mining can be used to determine the relationship between two events or sequences of events. Clarity in data and research can be gained through the use of data mining in conjunction with data visualisation and visual data mining.

Criminal investigation uses for data mining

Exploring and defining crime characteristics and the links between criminals and those factors are central to crime analysis. Due to the abundance of data, criminology can be a challenging field to study. Data mining is clearly a useful technology with wide-ranging applications in this area. All textual crime reports can be exported as Word documents for use in crime-matching applications.

There are many more industries that make heavy use of data mining beyond the 10 listed here. Various industries, including but not limited to: Advertising; intrusion detection; lie detection; corporate surveillance; bioinformatics; e-commerce; retail; service providers; insurance; communications; and many more.

CONCLUSION

The process of data mining provides insight into the past as well as projections for the future. This opens the door for a huge variety of applications and uses across a wide range of sectors and businesses. Data mining, despite the fact that it is not an independent entity, is the essential process that collaborates with pre-processing techniques, such as data preparation, data exploration and so on and post-processing techniques, such as model validation, model performance monitoring, scoring and so on, to produce the most beneficial insights and solutions.

REFERENCES

1. Adhikari A, et al. Mining multiple large data sources. Int Arab J Inf Technol. 2010;7:241-249.
2. Yuri D, et al. Architecture framework and components for the big data ecosystem. J Syst Netw Eng. 2013:1-31.
3. Singh DK, et al. Data security and privacy in data mining: research issues and preparation. Int J Comput Trends Technol. 2013;4:194-200.
4. Fan W, et al. Mining big data: Current status and forecast to the future. ACM SIGKDD Explorations Newsletter. 2013;14:1-5.

5. Hashmi S, et al. Big data mining techniques. Indian J Sci Technol. 2016;9:37.

6. Jadhav DK, et al. Big data: The new challenges in data mining. IJIRCST. 2013;1:39-42.

7. Kalra S, et al. Application of big data: Current status and future scope. IJACTE. 2014;3.

8. Shilpa MK, et al. BIG data and methodology-A review. Int J Adv Res Comput Sci Softw Eng. 2013;3:991-995.

9. Gupta S, et al. Analysis of data mining in health care management. Nat J Comput Sci Technol. 2012;4:28-36.

10. Singh D, et al. A survey on platforms for big data analytics. J Big Data. 2015;2:1-20.

11. Sharma TC, et al. WEKA approach for comparative study of classification algorithm. Int J Adv Res Comput Commun Eng. 2013;2:1925-1931.