A Survey on Data Mining in Big Data

R.Kirubakaran
Assistant Professor, Department of Computer Science
Kumaraguru College of Technology
Coimbatore, India

A.Periya Nayaki
Assistant Professor, Department of Computer Science
Kamaraj College of Engg.& Tech.
Virudhunagar, India

C.Mano Prathibhan
PG Graduate, Department of Computer Science
Thiagarajar College of Engineering
Madurai, India

***Abstract*—Collection of large and complex data is termed as big data. Tons of data are collected in applications such as medical processing, whether reporting, digital libraries, etc. and these data should be managed. Also they contain large amount of varying data such as text, images, video, audio, etc. Data mining is the process extracting useful information or knowledge from a large amount of data. This is also called as Knowledge Discovery in databases (KDD). Various algorithms are used in data mining to apply association rule generation, classification, clustering, etc. These data mining techniques can also be used in big data to extract useful data. In this paper, the various types of big data and the various data mining techniques that can be used in big data are explained based on a literature survey conducted. The various challenges in big data and research for future are also discussed.**

***Keywords*—Big data, data mining, algorithms, knowledge discovery, processing**

I. Introduction

In recent years, the concept of big data has become an important issue in the field of Information Technology [1]. Data is available everywhere since processing has been made automated and in online. The estimates have shown that the amount of data available keeps increasing by more than 35% every year and if this keeps on for some years then it will be a critical problem in many aspects. Some of these issues include the storage and time needed for processing and maintaining the data [2]. The second is the security related issues available in the data that is stored in public networks such as a LAN or public cloud.

Even in a small organization or a company, large amount of data are generated every day. To process, store, handle and extract these data the data mining techniques have been used [3]. But no matter how efficient a data mining algorithm can be, to process really huge data such as a big data is a tough job. The various problems arising due to big data should be handled now and then and the methods used need to be enhanced as the data keeps growing.

The data when undergoes data mining process goes through various changes. Thus any important information should not be lost and any secure information should not be shown outside. But still the mining process should be useful enough to provide the necessary information. By taking all these criteria into consideration, the problem is much harder than it was before since this has to be dealt with big data [4].

In most cases, they use many servers to process large data or use parallel technologies such as Hadoop framework. Even in large technology organizations such as Google, Facebook, etc. they use multiple servers and parallel environments to handle large number of user profiles and searches. But other than these, there are also other methods that can be used such as optimizing the large problem into a smaller problem by breaking it into multiple problems and processing them differently based on the needs. The dependency of the data should be handled in these cases. The next section talks about the concept of big data and the types of data available in big data.

II. Big Data

The concept of big data provides a new perspective in the upcoming field of research and it will be sure to affect the technology field in many aspects. It plays an important role in obtaining the related data or information from a large data collection. It affects the economy as a whole along with the various challenges such as the policy, security, etc. as you move on.

The ability to handle these challenges by analyzing the available data, storing the data and by extracting relative contents is measured by making use of the big data. Without big data, communication is made easy and the cost of processing and storage is relatively low when we consider the latest technologies. So big data even though is challenge, it is still a method to develop more efficient and faster methods of processing data. Fig. 1 shows the various types of data that can form a collection of big data.

The data available in big data can be of two types:

- Structured data
- Unstructured data

The data such a numbers and word that can be easily arranged in a specific order or can be categorized easily are termed as structured data. These kinds of data are available in places such as GPS system where location of various places are displayed in coordinates, account balance, transactions & sales reports in various banking and marketing organizations, various sensors that record reading values, etc.

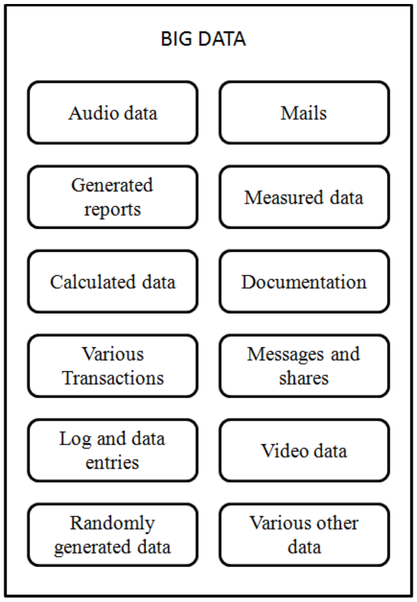


Fig. 1: Types of data available

The second type of data is the unstructured data that includes much complex information such as multimedia, website links and various other formats of text such as symbols and data from online websites. Due to the recent development in internet, social networking and websites the number of unstructured data has increased drastically and it needs to be handled.

The various characteristics of big data make it hard for identifying and mining useful information. These characteristics include the following:

- Heterogeneous and diverse data
- Autonomous data
- Data with distributed control
- Complex data with knowledge associations

The diverse amount of big data is a major issue when comes to mining information from the big data. The data available is too large and they are available in many different formats of data as said in Fig. 1. Also the dimensions of the data are different such as in text, audio, video, etc.

In autonomous characteristics, each data source is capable of generating data without the knowledge of the other and so there is no centralized controller in these cases to manage the data that are being generated. All of these different data gets generated in different formats or types in various locations as controlled by the autonomous server or source as in Fig. 2.

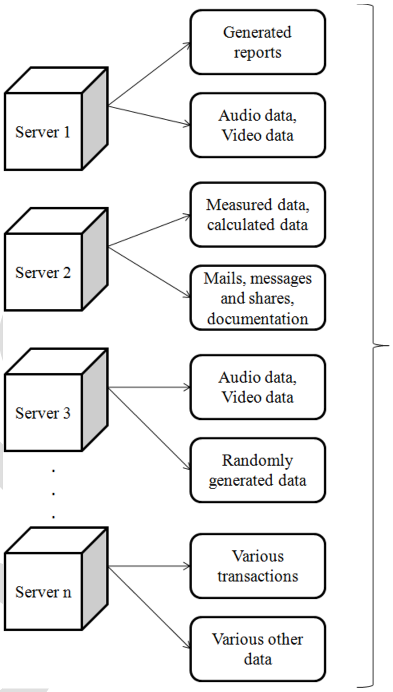
The final characteristic is the availability of large and complex data that are gathered from various sources across the internet. They also have different data formats and types.

Managing the big data with all these characteristics is hard and challenging. The 3 V's that needs to be considered in managing the big data are:

- Volume
- Variety
- Velocity

Apart from these, in recent times there are also other V's that needs to be considered such as:

- Variability
- Value



Variously structured and variously formatted data being generated from different servers placed in different locations online. But to extract some information, data from more than one places need to be used and they should be converted to the same format and structure before processing.

Fig. 2: Big data generated from various servers autonomously

By handling and managing these 5 V's and the above said 3 characteristics as in Fig. 3, the big data analysis will lead to many research topics in future.

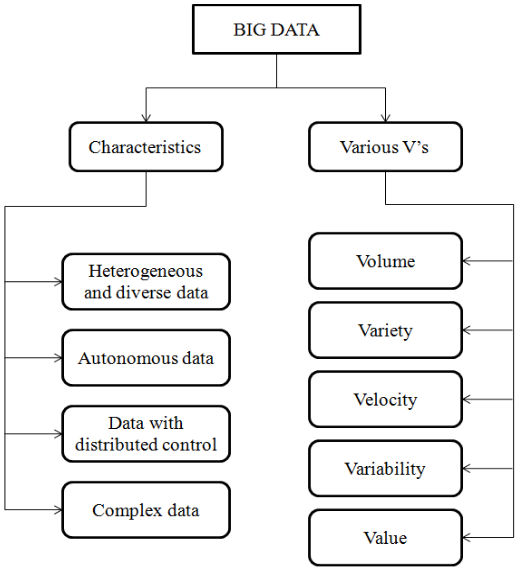


Fig. 3: Characteristics of big data

The volume represents the amount of data that is available on whole. The volume of the data continues to increase rapidly every day in any enterprise or organization that is considered. This increase in data volume should be handled effectively for decision making problems and data mining process.

The variety is the various types of data formats and types that are available. This also includes the different data sources or serves that exist, and provides the different variety of data as discussed before in Fig. 2. This can be managed by handling the various types of structured and unstructured data. Data from both within the organization and from outside the organization should be used for making the mining to generate necessary information for decision making. Any kind of data as mentioned in Fig. 1 can be used for this.

Velocity refers to the constant creation, processing and update of data. Data keeps generating constantly over time and if this is not processed then it keeps accumulating into larger data. Thus the data should be processed as and when it is being generated.

The data keeps updating also as time goes and the obtained mining results also varies. This refers to variability of the big data. Due to this, the decision making process will be hard and especially in case of big data.

And finally another important aspect to consider is the value of the data that is available and the value of the information that is being extracted from the big data. This information should be satisfying for the organization and should provide necessary information they desire.

III. Data Mining in Big Data and Future Trends

The concept of data mining deals with mining useful information from the available data. The process of data mining involves various algorithms used in different techniques such as:

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

Various algorithms are available in each of these methods as given below in Table I.

TABLE I. DATA MINING ALGORITHMS

Technique	Algorithms used
Classification	• ID3 algorithm • C4.5 algorithm • SLIQ algorithm • Nearest-neighbor algorithm • Naïve-Bayes algorithm • OODG algorithm • Decision Trees • Decision Tables
Estimation	• Instance-based methods • Linear models
Prediction	• The prediction task • Statistical modeling • Bayesian network
Association rules	• Apriori algorithm • Distributed algorithm • Parallel algorithm
Clustering	• k-means algorithm • BFR algorithm • k-medoids algorithm • CURE algorithm • Hierarchical clustering
Description	• Description and report of the entire data mining process

Applying any of these data mining algorithms shown in Table 1, in big data will be difficult and challenging when we consider the amount of data that needs to be processed. Some of the issues in big data mining are shown in Fig. 4.

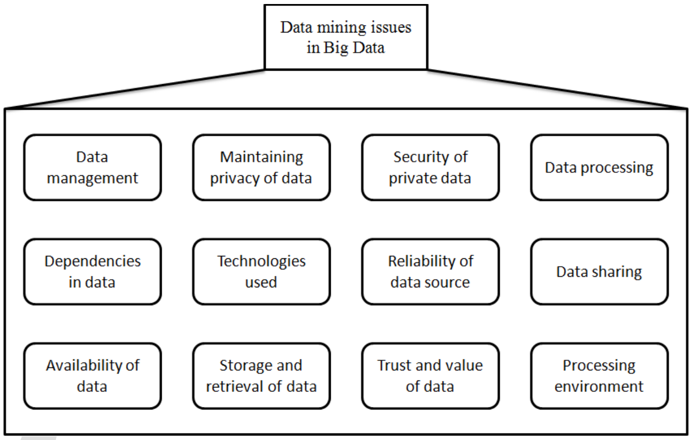


Fig. 4: Issues in data mining of big data

In future, the research areas in big data will include methods and algorithms that will be focused on handling the various issues as mentioned in Fig. 4 and also to provide new methods on how to use this large and evolving data in various fields of data mining and its applications [5][6].

An important field to consider is the use of parallel networks and cloud networks [7] alongside of the big data. Cloud being a hot research topic already can be combined with big data for providing more updates in upcoming future technologies that will be scalable to more areas in technology.

In security, specific and unique algorithms should be designed for managing the security of big data due to the large amount of data that can get leaked out into the public networks. Also private information of the data owner or any private information regarding any organization, person or institute should be made secure.

IV. Conclusion

The collection of large, complex, varying and increasing data is termed as big data. By processing big data and by extracting useful information from big data by using data mining techniques, it is possible to obtain large volumes of necessary information that will benefit an organization or society in different ways. Many research have been conducted in data mining on big data and they all aim to provide novel and efficient methods for implementing the various data mining algorithms. In future many such technologies will be adopted in various data mining applications in real time.

REFERENCES

[1] Elisa Bertino; "Big Data – Opportunities and Challenges", IEEE 37th Annual Computer Software and Applications Conference, 2013, pp. 479-480.

[2] Nikita Jain, Vishal Srivastava; "Data Mining Techniques: A Survey Paper", International Journal of Research and Engineering in Technology, Volume 2, issue 11, 2013, pp. 116-119.

[3] Kalyani M Raval; "Data Mining Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 10, 2012, pp. 439-442.

[4] Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan, Yong Ren; "Information Security in Big Data: Privacy and Data Mining", IEEE Access, Volume 2, 2014, pp. 1149-1176.

[5] Bharati M Ramageri; "Data Mining Techniques and Applications", Indian Journal of Computer Science and Engineering, Volume 1, Number 4, pp. 301-305.

[6] Shu-Hsien Liao, Pei-Hui Chu, Pei-Yuan Hsiao; "Data Mining Techniques and Applications – A Decade Review from 2000 to 2011, ELSEVIER, Expert Systems with Applications, Volume 39, 2012, pp. 11303-11311.

[7] OlukunleA Iyanda; "Big Data and Current Cloud Computing issues and Challenges", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 6, 2014, pp. 1192-1197.