Analysis and Application of Data Mining by the Implementation of Big Data

Raghav Sethi
Student
Mukesh Patel School of Technology Management and Engineering,
NMIMS University,
Mumbai, India

ABSTRACT

Big Data is a bright expression, which is used to recognize the datasets that are big due to their hefty size and complexity. Big Data is now swiftly on the rise in each and every science, research and engineering domain. Big data can also be included in physical, biological, historical, Geographical and biomedical sciences. Big Data mining is the capability of extracting valuable information from these huge datasets or streams of data, that due to its volume, unpredictability, and velocity was not probable to be done before. The Big Data dare is becoming one of the most exciting opportunities for the next coming years. This study paper includes the full information about the big data, Data mining and Data mining with big data, demanding issues And Its Associated Work.

Keywords

Big Data, Data mining, Challenging issues, Datasets, Data Mining Algorithms

1. INTRODUCTION

Now adays is the period of Google. The information, which is unheard by us, can get just a fraction of a second on the screen with a number of links as a result. This way, we can easily explicable data for the processing of Big Data. An example of big data might be pet bytes (1,024 terabytes) or exabytes (1,024 petabytes) of data consisting of billions to trillions of records of millions of people—all from different sources (e.g. Web, sales, customer contact centre, social media, mobile data and so on). The data is usually loosely prearranged data that is often unfinished and unreachable. This Big Data is not any different thing than out regular term data. Just big is a keyword used with the data to identify the collected datasets due to their large size and complexity? We cannot manage them with our current methodologies or data mining software tools. Another example of big data can visualize on that day when, the speech of Hon'ble Prime Minister of India- Narinder Modi regarding Clean India, then triggered number of tweets within 3 hours. Millions of people have joined this motivational programme and this also generated the most discussions revealed with in the public interests. Feedback of people and their interests generates in a real time to generic media, such as radio or TV broadcasting.This example demonstrates the rise of Big Data applications. The data collection has grown tremendously and is beyond the ability of commonly used software tools to capture, manage, and process within a tolerable time.

2. BIG DATA AND DATA MINING

The Big Data is nothing but a data, which is available at heterogeneous, autonomous sources, in tremendous huge amount, get updated in fractions of seconds. We can observe an example of various social Networking Sites, where the data stored at the server of WhatsApp/Facebook/Instagram, as most of us, daily uses. We upload a variety of types of in sequence, upload photos and videos. All the data get stored at the data warehouses at the server of various social networking sites. This data is nothing but the huge data, which is so called due to its complexity. Also another example is storage of photos at Flicker. These are the good quality real-time examples of the Big Data. Another best example of Big data would be, the readings taken from an electronic microscope of the universe or world. Now the term Data Mining, Finding for the precise useful information or acquaintance from the composed data, for future actions, is nothing but the data mining. So, collectively, the term Big Data Mining is a close up view, with lots of detail information of a Big Data with lots of information. As shown in fig (a) below.

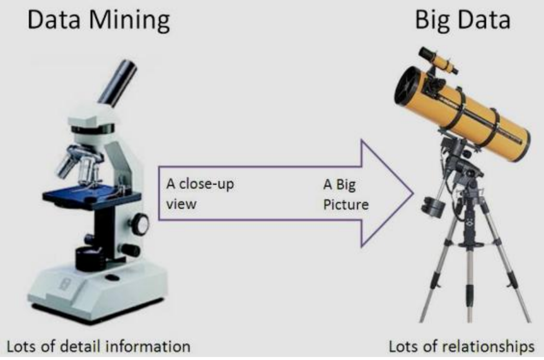

Fig.1 Data Mining with Big Data

3. KEY FEATURES OF BIG DATA

There are various features of Big Data illustrated are as below:

- It is huge in size.
- The data keep on changing time to time.
- Its data sources are from different phases.
- It is free from the influence, leadership, or control of anyone.
- It is too much complex in nature, thus hard to handle.
- Decision-oriented analysis is more similar to traditional business intelligence.
- Action-oriented analysis is used for fast answer, when a pattern emerges or specific kinds of data are detected and action is necessary. Taking advantage of big data through analysis and causing proactive or reactive performance changes offer great potential for early adopters.

It's huge in nature because, there is the collection of data from various sources together. If we consider the example of WhatsApp/Facebook / Instagram, lots of numbers of people are uploading their data in various types such as text, images or videos. The people also keep their data changing continuously. This remarkable and right away, time to time changing stock of the data is stored in a warehouse. This large storage of data requires large area for actual performance. As the size is too large, no one is capable to control it oneself. The Big Data needs to be controlled by dividing it in groups. Due to largeness in size, decentralized control and different data sources with different types the Big Data becomes much complex and harder to handle. We cannot manage them with the local tools those we use for supervision the regular data in real time. For main Big Data-related applications, such as Google, Flicker, Facebook, a big number of server farms are deployed all over the world to ensure nonstop services and quick responses for local markets.

4. CHALLENGING ISSUES IN DATA MINING WITH BIG DATA.

There are three sectors at which the challenges for Big Data arrive. These three sectors are:

- Mining platform.
- Privacy.
- Design of mining algorithms.

Mainly, the Big Data is stored at different places and also the data volumes may get increased as the data keeps on escalating constantly. So, to collect all the data stored at different places is that much costly. Suppose, if we use these classic data mining methods (those methods which are used for mining the small scale data in our personal computer systems) for mining of Big Data, and then it would become an obstacle for it. Because the typical methods are required data to be loaded in main memory, though we have super large main memory. To maintain the privacy is one of the main aims of data mining algorithms. Presently, to mine information from Big data, parallel computing based algorithms such as MapReduce are used. In such algorithms, large data sets are divided into number of subsets and then, mining algorithms are applied to those subsets. Finally, summation algorithms are applied to the results of mining algorithms, to meet the goal of Big Data mining. In this whole procedure, the privacy statements obviously break as we divide the single Big Data into number of slighter datasets.

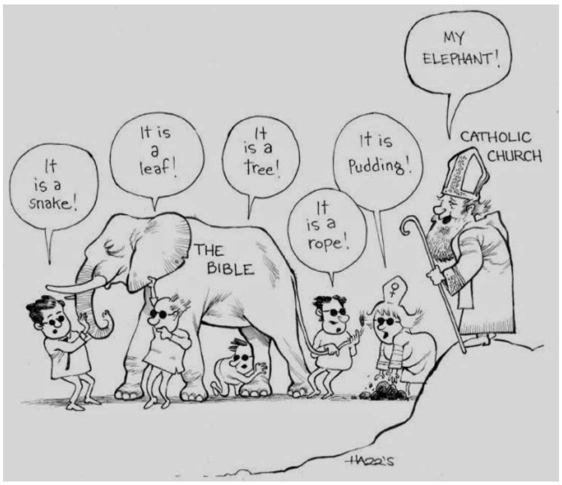

Fig. 2 Blind men and the giant elephant.

While designing such algorithms, we face various challenges. As shown in the figure b above, there are six blind men observing the giant elephant. Everyone is trying to predict their conclusion on what the thing is actually. Somebody is saying that the thing is a leaf; someone says it's a snake; someone says that it's a pudding; someone says that its rope or Bible etc. Actually everyone is just observing some part of that giant elephant and not the whole, so the results of each blind person's prediction is something different than actually what it is. Similarly, when we divide the Big Data in to number of subsets, and apply the mining algorithms on those subsets, the results of those mining algorithms will not always point us to the actual result as we want when we collect the grades together.

5. RELATED WORK

On the level of mining platform sector, at present, parallel programming models like MapReduce are being used for the purpose of analysis and mining of data. MapReduce is a batch-oriented parallel computing model. There is still a certain gap in performance with relational databases. Improving the performance of MapReduce and enhancing the real-time nature of large-scale data processing have received a significant amount of attention, with MapReduce parallel programming being applied to many machine learning and data mining algorithms. Data mining algorithms usually need to scan through the training data for obtaining the statistics to solve or optimize model. For those people, who intend to hire a third party such as auditors to process their data, it is very important to have efficient and effective access to the data. In such cases, the privacy restrictions of user may be faces like no local copies or downloading allowed, etc. So there is privacy-preserving public auditing mechanism proposed for large scale data storage.[1] This public key-based mechanism is used to enable third-party auditing, so users can safely allow a third party to analyze their data without breaching the security settings or compromising the data privacy. In case of design of data mining algorithms, Knowledge evolution is a common phenomenon in real world systems. But as the problem statement differs, accordingly the knowledge will differ. For example, when we go to the doctor for the treatment, that doctor's treatment program continuously adjusts with the conditions of the patient. Similarly the knowledge. For this, Wu [2] [3] [4] proposed and established the theory of local pattern analysis, which has laid a foundation for global knowledge discovery in multisource data mining. This theory provides a solution not only for the problem of full search, but also for finding global models that traditional mining methods cannot find.

6. CONCLUSION

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. This data is going to be more diverse, larger, and faster. We discussed some insights about the topic, and what we consider are the main concerns and the main challenges for the future. Big Data is becoming the new Final Frontier for scientific data research and for business applications. 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. Everybody is warmly invited to participate in this intrepid journey.

7. ACKNOWLEDGEMENT

I would like to express my gratitude towards my respected faculties of my college for there guidance and support. I will forever remain appreciative for there constant support and guidance. Through many discussions, my professors have helped me to form and solidify ideas. The priceless discussions I had with them, the piercing questions they have put to me and the constant motivation, has all led to the development of this project. I would also like to thank my friends and colleagues for providing feedback and suggestions for recuperating my information.

8. REFERENCES

[1] C. Wang, S.S.M. Chow, Q. Wang, K. Ren, and W. Lou, "Privacy- Preserving Public Auditing for Secure Cloud Storage" IEEE Trans. Computers, vol. 62, no. 2, pp. 362-375, Feb. 2013.

[2] X. Wu and S. Zhang, "Synthesizing High-Frequency Rules from Different Data Sources," IEEE Trans. Knowledge and Data Eng., vol. 15, no. 2, pp. 353-367, Mar./Apr. 2003.

[3] X. Wu, C. Zhang, and S. Zhang, "Database Classification for Multi-Database Mining," Information Systems, vol. 30, no. 1, pp. 71- 88, 2005

[4] K. Su, H. Huang, X. Wu, and S. Zhang, "A Logical Framework for Identifying Quality Knowledge from Different Data Sources," Decision Support Systems, vol. 42, no. 3, pp. 1673-1683, 2006.

[5] E.Y. Chang, H. Bai, and K. Zhu, "Parallel Algorithms for Mining Large-Scale Rich-Media Data," Proc. 17th ACM Int'l Conf. Multimedia, (MM '09,) pp. 917-918, 2009.

[6] D. Howe et al., "Big Data: The Future of Biocuration," Nature, vol. 455, pp. 47-50, Sept. 2008.

[7] A. Labrinidis and H. Jagadish, "Challenges and Opportunities with Big Data," Proc. VLDB Endowment, vol. 5,no. 12, 2032-2033, 2012.

[8] Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining," J. Cryptology, vol. 15, no. 3, pp. 177-206, 2002.