International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 10, October 2015
Available at http://internationaljournalofresearch.org

Searching Techniques on Big Data by Using Data Mining

Veeramoni Shiva Kumar
Assistant Professor, KITE College of Professional Engineering Sciences, Hyderabad

Abstract:
The main goal of the data mining process is to extract useful information from Big Data set and transform it into an understandable form for further use. It was not possible to extract useful information from the large datasets or data streams. Now this can be achieved by the capability of Big Data Mining. The overlay-based parallel data mining architecture executes processing by employing the overlay network and fully distributed data management, which can achieve high scalability and service availability. In case of the physical network disruption, an overlay-based parallel mining architecture is not capable of providing data mining services if router/communication line breakdowns, because of this, numerous nodes are removed from the overlay network. An overlay network construction scheme based on physical network structure, including nodes location and a distributed task allocation scheme using overlay network technology is done to overcome with this issue. In this survey paper, a review of Overlay based parallel data mining and its different technologies are studied.

Keywords: Data Mining; Big Data; Overlay-Based; Service Availability; Physical Network Disruption.

I. INTRODUCTION

Big data refers to the enormous amount of structured and unstructured data that overflow the organization. If the overflowed data is used in a proper way it leads to meaningful information. When big data is compared to traditional databases it includes a large number of data which requires more processing in real time. It also provides opportunities to discover new values, to understand an in-depth knowledge from hidden values and also provides space to manage those data effectively. Big Data concern large-volume, complex, growing datasets with multiple data sources. With the fast development of networking, data storage and data collection capacity, bigdata are now expanding in all science and engineering domains, including physical, biological and biomedical sciences.[1]. Data Mining is a task of identifying relevant and significant information from large data set. Data Mining and Knowledge Discovery are usually defined as the extraction of patterns or models from observed data, usually the ability to explore much richer and more expressive models, providing new and interesting domains for the application of learning algorithms.

The wide availability of large-scale data sets from different domains has created a demand to automate the process of extracting valuable information from them. For example, consider Facebook application, we upload various types of information such as text, images and video. The process of effective mining of such data is known as big data mining[2]. Today is the era of Google, the thing which is unknowns searched in Google and within fractions of seconds; we get the number of links as a result. This would be the best example for the processing of Big Data. This Big Data is not any different thing than out regular term data. The Big Data is nothing but a data in an extreme large amount available of heterogeneous, autonomous sources, which get updated in fractions of seconds [2], [3]. Conventional parallel data mining architectures with centralized management schemes lack scalability, which causes bottleneck in the entire system and this leads to decrease in performance of the system as the number of nodes

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 10, October 2015
Available at http://internationaljournalofresearch.org

increases[4]. As a remedy for improving scalability, an overlay based parallel data mining architecture has been proposed. Since all the nodes execute both management and processing functions by using overlay network, this architecture can balance the management load.

II. BIG DATA AND DATA MINING

The Big Data is nothing but a data, available at heterogeneous, autonomous sources, in extreme large amount, which get updated in fractions of seconds. For example, the data stored at the server of Facebook, as most of us, daily use the Facebook; we upload various types of information, upload photos. All the data get stored at the data warehouses at the server of Facebook. This data is nothing but the big data, which is so called due to its complexity. Also another example is storage of photos at Flicker. These are the good 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. Now the term Data Mining, Finding for the exact useful information or knowledge from the collected 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 1 below.

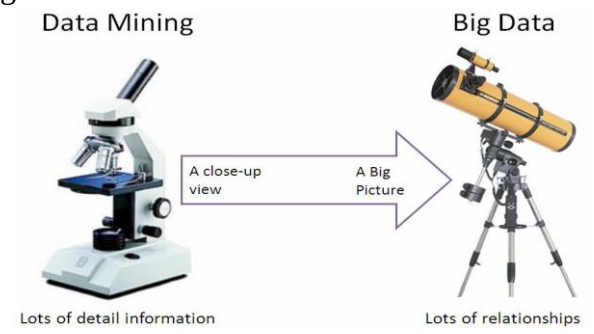

Fig.1 Data Mining with Big Data

The features of Big Data are:

- It is huge in size.
- The data keep on changing time time to time.
- Its data sources are from different phases.
- It is free from the influence, guidance, or control of anyone.
- It is too much complex in nature, thus hard to handle.

It's huge in nature because, there is the collection of data from various sources together. If we consider the example of Facebook, 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 tremendous and instantaneously, time to time changing stock of the data is stored in a warehouse. This large storage of data requires large area for actual implementation. 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 managing the regular data in real time. For major Big Data-related applications, such as Google, Flicker, Facebook, a large number of server farms are deployed all over the world to ensure nonstop services and quick responses for local markets.

III. 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.

IV IMPLEMENTATION

There are three sectors at which the challenges of BigData arrive. They are:

1. Mining platform.
2. Privacy.
3. Security
4. Design of mining algorithms.

Fundamentally, the astronomically immense Data is stored in different places and the data volumes may get incremented as the data keeps on incrementing perpetually. So, to accumulate all the data stored in different places is that much sumptuous. Suppose, if we utilize these typical data mining methods (those methods which are utilized for mining the minuscule scale data in our personal computer systems) for mining of Sizably Voluminous Data, and then it would become an impediment for it. Though we have super sizably voluminous main recollection, the typical methods are required to load the data in main recollection. Variety, Volume, Velocity and Precision are essential characteristics of astronomically immense data. Variety, data from multiple sources inherently possesses many types and different forms like structural, semi structured and unstructured data. Scalability,sizably voluminous volume of sizably voluminous data requires high scalability of its data management and mining implements. Speed of data mining depends on the data access time and efficiency [7]. To maintain the privacy is one of the main aim of data mining algorithms. Presently, to mine information from astronomically immense Data, parallel computing predicated algorithms such as Map Reduce are utilized [8].

The astronomically immense data sets are divided into a number of subsets and then, mining algorithms are applied to those subsets, in such algorithms. Conclusively, summation algorithms are applied to the results of mining algorithms to meet the goal of the Astronomically Immense Data mining. During this whole procedure, the privacy verbalizations conspicuously break as we divide the single Sizably voluminous Data into a number of more diminutive datasets [9]. An emerging topic in data mining is privacy preserving data mining, the fundamental conception of privacy preserving data mining is performing data mining algorithms efficaciously without compromising the security of sensitive information contained in the data. Fuzzy dactyl log ram is one of the data mining techniques that enhance data privacy during data leak detection operation which is predicated on sensitive data. The main goal of privacy preservation is for fending private data while processing or relinquishing sensitive information. S. Moncrieff et.al proposes a solution which is predicated on

environmental context to dynamically alter the privacy levels in the perspicacious house. Fabio Borges proposes a privacy preserving protocol for astute metering systems to ascertain customers' privacy and security in the network data.

The security concerns have become a major barrier to the widespread magnification of cloud computing. Distributed architecture is utilized to eliminate the jeopardizes during data mining predicated attacks. From the above study, we found that, Data mining process is not facile and the algorithm utilized for mining is very intricate. The data needs to be integrated from the sundry heterogeneous data sources as is not available at one place. Data mining derives its name from the kindred attributes between probing for valuable business information in a sizably voluminous database. For example, finding the linked products in gigabytes of store scanner data and mining a mountain for a vein of valuable one [10],[11]. Both processes require either shifting through an immense amount of material, or plausibly probing it to find precisely where the value resides. The databases of sufficient size, quality and the data mining technology can engender incipient business opportunities by providing some capabilities.

V.TECHNIQUES IN DATA MINING PROCESS

A. *Decision trees*

Tree- shaped structures that represent sets of decisions. These decisions generate the rules for classification of a dataset. Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) are included under specific decision tree methods.

B. Genetic algorithms

Genetic combination, mutation, and natural selection are the process used in Optimization techniques for design based on the concepts of evolution.

C. Nearest neighbor method

The technique that classifies each record in a dataset based on a combination of the classes of the records(s) most similar to it in a historical Sometimes called the k-nearest neighbor technique, where k is the number of neighbors.

D. Artificial neural networks

Non-linear predictive models, which are based on biological neural systems. At present, on the level of the mining platform sector, parallel programming models like Map Reduce are being used for the purpose of analysis and mining of data. Map Reduce is a batch-oriented parallel computing model [1], [11], [13], [14]. There is still a certain gap in performance with relational databases. Improving the performance of Map Reduce and enhancing the real-time nature of large scale data processing have received a significant amount of attention with Map Reduce parallel programming being applied to many machine learning and data mining algorithms. These data mining algorithms usually need to scan through the training data for obtaining the statistics to solve or optimize the model.

VI. CONCLUSION

The data mining techniques can be applied on big data to acquire some useful information from large datasets. Thus these two terms are not different instead they are coupled together to acquire some useful picture from the data. Thus we conclude that big data will become an excellent opportunity in the forth coming years. We discussed some of the useful information about big data and data mining and have identified the research gaps and open research areas. Due to increasing the size of data day by day, Big Data is going to continue growing during the next years and becoming one of the exciting opportunities in future. This paper

insights about the topic, and controversy, and the main challenges etc. for the future. Hence BigData is becoming the new Final Frontier for scientific data research and for business applications. And unchallenging side, Securely Management of Big Data with today's threat spectrum is a big issue. Because today's we have an overwhelming growth of data interms of volume, velocity and variety. So from a security and privacy standpoint, the threat landscape and security and privacy risks have also seen an unprecedented growth. So as for future research is needed to build a generic architectural framework towards addressing these security and privacy challenges in a holistic manner. Now we are in new era where Big Data mining will help us to discover knowledge that no one has discovered before. So everybody is warmly invited to participate in this intrepid journey to discover the future views.

VII. REFERENCES

[1.] http://big-data-mining.org/

[2.] Ms. Neha A. Kandalkar , Prof. AvinashWadhe," Extracting Large Data using Big Data Mining", International Journal of Engineering Trends and Technology (IJETT) – Volume 9 Number 11 - Mar 2014.

[3.] Wei Fan, Albert Bifet," Mining Big Data: Current Status, and Forecast to the Future"

[4.] James Joshi, BalajiPalanisamy," Towards Riskaware Policy based Framework for Big Data Security and Privacy", 2014.

[5] KatsuyaSuto, Hiroki Nishiyama, XueminShen and Nei Kato1, "Designing P2P Networks Tolerant to Attacks and Faults Based on Bimodal Degree Distribution", Journal of Communications, vol 7, Issue 8, August 2012.

[6] Nikita Jain and Vishal Srivastava "Data Mining Techniques: A Survey Paper", International Journal of Research in Engineering and Technology, vol. 2, Issue 11, November 2013.

[7] Distributed Data Mining in Peer-to-Peer Networks, IEEE Std. 1089-7801, 2006.

[8] S. V. S. Ganga Devi, "A Survey on Distributed Data Mining and it's Trends", IMPACT: IJRET, vol. 2, Issue 3, Mar 2014.

[9] Rekha Sunny and Sabu M. Thampi, "Survey on Distributed Data Mining in P2P Networks".

[10] Wenjun Xiao, Mingxin He and Huomin Liang, "Cayley CCC: A Robust P2P Overlay Network with Simple Routing and SmallWorld Features", Journal of Networks, vol. 6, Issue 9,September 2011

[11]. Xindong Wu, Xingquan Zhu, Gong Qing Wu, Wei Ding, „Data mining with Big data", IEEE, Volume 26, Issue 1, January 2014.