INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

SURVEY ON BIG DATA MINING USING DIFFERENT TECHNIQUES

MISS.SAYALI SHYAM KARMODE

M.E Student, Department of Computer Science And Engineering, P.R Patil College of Engineering And Technology, Amravati, Maharashtra, India

Accepted Date: 15/03/2016; Published Date: 01/05/2016

Abstract*:* In this paper we find the optimal path from source to destination node in a Duty Cycled Mobile sensor Network with Geographic routing. Geographic routing is one of the most promising techniques due to its simplicity, scalability, and efficiency. The WSN networks employing duty cycle technique in geographic routing has a task to find out the path from sensor to Base Station of the network to send the sensed data. GCKN(Geographic connected k neighbor) sleep scheduling overcome this geographic routing designed for static WSN. We propose a scheme where other node parameters like delays incurred and load on each node are calculated and are taken into consideration to 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. “Big data mining” involves knowledge discovery from these large data sets. “Big data” is gaining huge significance in the current scenario and consequently, big data mining emerged as an innovative and potential research area. This paper gives an overview of big data along with its type, source and characteristics. A review on various big data mining platforms, algorithms and challenges is also discussed in this paper.

Keywords: Big data mining, Big data analytic tools, Big data mining algorithm, Big data mining challenges

Corresponding Author: MISS.SAYALI SHYAM KARMODE

Access Online On:

www.ijpret.com

How to Cite This Article:

Sayali Shyam Karmode, IJPRET, 2016; Volume 4 (9): 1325-1333

PAPER-QR CODE

INTRODUCTION

The Big Data applications where data collection has grown tremendously and is beyond the ability of commonly used software tools to capture, manage, and process within a "tolerable elapsed time." The most fundamental challenge for Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions. In many situations, the knowledge extraction process has to be very efficient and close to real time because storing all observed data is nearly infeasible. For example, the square kilometer array (SKA) in radio astronomy consists of 1,000 to 1,500 15-meter dishes in a central 5-km area.

It provides 100 times more sensitive vision than any existing radio telescopes, answering fundamental questions about the Universe. However, with a 40 gigabytes (GB)/second data volume, the data generated from the SKA are exceptionally large. Although researchers have confirmed that interesting patterns, such as transient radio anomalies can be discovered from the SKA data, existing methods can only work in an offline fashion and are incapable of handling this Big Data scenario in real time. As a result, the unprecedented data volumes require an effective data analysis and prediction platform to achieve fast response and real-time classification for such Big Data [1].

1.1 Need and Importance

Fundamentally, data mining is about processing data and identifying pattern and trends in that information so that you can decide or judge. Data mining principles have been around for many years, but with the advent of big data, it is even more prevalent. Big data caused an explosion in the use of more extensive data mining techniques, partially because the size of information is much larger and because the information tends to more varied and extensive in its very nature in content. You want to know whether those two million are a particular age group and their average earnings so that you can target your customer needs better [2]. Data mining with big data is just about because after you have the information in the format you need, you can apply the different techniques (individually or together) regardless of the required underlying data structure or data set.

2. Related Work

2.1. Literature Review

DR. Yan Mo won the 2012 Nobel Prize in Literature. This is probably the most controversial Nobel Prize of this category. Searching on Google with "Yan Mo Nobel Prize, "resulted in

1,050,000 web pointers on the Internet (as of 3 January 2013). “For all praises as well as criticisms,” said Mo recently, “I am grateful.” What types of praises and criticisms has Mo actually received over his 31-year writing career? As comments keep coming on the Internet and in various news media, can we summarize all types of opinions in different media in a real-time fashion, including updated, cross-referenced discussions by critics? This type of summarization program is an excellent example for Big Data processing, as the information comes from multiple, heterogeneous, autonomous sources with complex and evolving relationships, and keeps growing. Along with the above example, the era of Big Data has arrived [3]. Every day, 2.5 quintillion bytes of data are created and 90 percent of the data in the world today were produced within the past two years.

Our capability for data generation has never been so powerful and enormous ever since the invention of the information technology in the early 19th century. As another example, on 4 October 2012, the first presidential debate between President Barrack Obama and Governor Mitt Romney triggered more than 10 million tweets within 2 hours. Among all these tweets, the specific moments that generated the most discussions actually revealed the public interests, such as the discussions about Medicare and vouchers. Such online discussions provide a new means to sense the public interests and generate feedback in real-time, and are mostly appealing compared to generic media, such as radio or TV broadcasting. Another example is Flicker, a public picture sharing site, which received 1.8 million photos per day, on average, from February to March 2012. Assuming the size of each photo is 2 megabytes (MB), this requires 3.6 terabytes (TB) storage every single day. Indeed, as an old saying states: “a picture is worth a thousand words,” the billions of pictures on Flicker are a treasure tank for us to explore the human society, social events, public affairs, disasters, and so on, only if we have the power to harness the enormous amount of data [4].

2.2 What is big data mining?

Useful data can be retrieved from this large datasets with the aid of big data mining. Here the data which are handled is big data, hence the term big data mining. Usually, data mining is the technique of analyzing data from different prospects and summarizing these data into interesting, understandable and useful models. For better decision making, the large repositories of data collected from different resources require a proper mechanism for extracting knowledge from the databases. Since big data scales far beyond the capacity of single PC, cluster computers, which have high computing powers and rely on parallel programming paradigms, are used. Thus a large attempt to exploit these huge parallel processing architectures was initiated [5].

2.3 Big data analytics tools

There are varieties of applications and tools developed by various organizations to process and analyze Big Data. The Big Data analysis applications support parallelism with the help of computing clusters. These computing clusters are collection of hardware connected by Ethernet cables. The following are major applications in the area of Big Data analytics [6].

A. Google's Map Reduce, Hadoop and Google Big Table

Google's programming model, Map Reduce, and its distributed file system, Google File System (GFS)

Are the pioneers in the field. Improving the performance of Map Reduce and enhancing the real-time nature of large-scale data processing has received a significant amount of attention, with Map Reduce parallel programming. So with this concept many companies provide big data processing framework that support Map Reduce. After that Yahoo and related companies developed Hadoop uses the Hadoop Distributed File System (HDFS) – an open source version of the Google's GFS. Later in this field to support the Map Reduce computing model strategy, Google developed the Big Table in 2006– a distributed storage system designed for processing structured data with size in the order of petabytes [7].

B. Dynamo

In 2006 Amazon developed Dynamo which uses a key-value pair storage system. Dynamo is a Highly available and scalable distributed data store built for Amazon's platform. Dynamo is used to manage services that need high reliability, availability, consistency, performance and cost effectiveness.

C. HBase

HBase is an open source, non relational, distributed database developed after big table. It works on the top of Hadoop Distributed file system and provides big-table like capabilities for Hadoop [8]

D. Apache Hive

Apache Hive is a data warehouse infrastructure built on top of Hadoop. It provides data summarization, query, and analysis of big data

E. Berkeley Data Analytics Stack (BDAS)

The Berkeley Data Analytics Stack (BDAS) [19] is an open source data analytics stack that integrates Software components being built by the UC Berkeley AMP Lab for computing and analyzing big data. Many systems in the stack provide higher performance over other big data analytics tools, such as Hadoop. Nowadays, BDAS components are being used in various organizations [9].

F. ASTERIX

ASTERIX is an Open Source System for big data management and analysis. With the help of ASTERIX Semi structured data can be easily ingested, stored, managed, indexed, retrieved and analyzed. Many of the drawbacks of Hadoop and similar platforms such as single system performance, difficulties of future maintenance, inefficiency in extracting data and awareness of record boundaries etc are easily overcome by ASTERIX

G. SciDB

SciDB [21] is an open-source data management and analytics software system (DMAS) that uses a multidimensional array data model. SciDB is designed to store petabytes of data distributed over a large number of machines and used in scientific, geospatial, financial, and industrial applications.

H. Hadoop Map Reduce

These algorithms work on top of Hadoop and make use of Map Reduce programming model.

I. NIMBLE

An open source toolkit for the Implementation of Parallel Data Mining and Machine Learning Algorithms on Map Reduce for large datasets. It allows users to compose parallel ML-DM algorithms using reusable (serial and parallel) building blocks that can be efficiently manipulated using almost all parallel programming models such as Map Reduce. It runs on top of Hadoop [10].

J. Big Cloud-Parallel Data Mining (BC-PDM)

Big Cloud Parallel Data Mining mainly relies on cloud computing and works on top of Hadoop and mainly used in intelligence data analysis.

K. Graph Mining tools

Graphs are widely used in data mining application domains for identifying relationship patterns, rules, and anomalies. Certain examples for domains include the web graph, social networks etc. The ever-expanding size of graph-structured data for the above applications needs a scalable system that can process large amounts of data efficiently. Graph, Graph Lab, Bulk Synchronous Parallel Based Graph Mining (BPGM) are the examples for the system to process graph structured data. Many techniques were developed earlier in the analysis of big data. With the advancement in the field of big data, the various analytic techniques such as structural coding, frequencies, co-occurrence, graph theoretic data reduction techniques, hierarchical clustering techniques, multidimensional scaling techniques were developed for large qualitative data sets. It is clearly described that the need for the particular proposal arises with the type of dataset and the way the pattern are to be analyzed [11].

2.4 Big Data Mining Algorithm

A. Decision tree induction classification algorithms

In the initial stage different Decision Tree Learning was used to analyze the big data. In decision tree induction algorithms, tree structure has been widely used to represent classification models. Most of these algorithms follow a greedy top down recursive partition strategy for the growth of the tree. Decision tree classifiers break a complex decision into collection of simpler decision [12].

B. Partitioning based clustering algorithms

In partitioning based algorithms, the large data sets are divided into a number of partitions, where each partition represents a cluster. K-means is one such partitioning based method to divide large data sets into number of clusters [13].

C. Hierarchical based clustering algorithms

In hierarchical based algorithms large data are organized in a hierarchical manner based on the medium of proximity. The initial or root cluster gradually divides into several clusters. It follows a top down or bottom up strategy to represent the clusters.

D. Density based clustering algorithms

In density based algorithms clusters are formed based on the data objects regions of density, connectivity and boundary. A cluster grows in any direction based on the density growth. DENCLUE is one such algorithm based on density based clustering

E. Grid based clustering algorithms

In grid base algorithms space of data objects are divided into number of grids for fast processing. OptiGrid algorithm is one such algorithm based on optimal grid partitioning.

F. Density based clustering algorithms

In density based algorithms clusters are formed based on the data objects regions of density, connectivity and boundary. A cluster grows in any direction based on the density growth. DENCLUE is one such algorithm based on density based clustering [14].

2.5 Challenges of big data mining

A. Evaluating the interestingness of mined patterns

B. Building a global unifying theory of big data mining

C. Scaling up to meet the growing needs of large data sets

D. Building efficient big data mining platform

E. Building efficient mining algorithms/models for big data

F. Maintaining security, trust and data integrity

G. Data privacy issues [15]

3. Conclusion

We are living in a digital world of big data where massive amounts of heterogeneous, autonomous, Complex and evolving data sets are constantly generated at unprecedented scale.

In this paper, an overview of big data along with it types, sources, characteristics and challenges are discussed. This paper reviews about the various big data mining platforms and algorithms. To support big data mining, high-performance computing platforms are required. It is understood that interestingness of discovered patterns, developing a global unifying theory, building efficient mining platforms or algorithms, privacy preserving, security, trust and data integrity are the major challenging issues in the current big data mining scenario. It is known that big data mining is an emerging trend in all science and engineering domains and also a promising research area. In spite of the limited work done on big data mining so far, it is believed that much work is required to overcome its challenges related to the above mentioned issues.

4. REFERENCES

1. R. Ahmed and G. Karypis, "Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks," Knowledge and Information Systems, vol. 33, no. 3, pp. 603-630, Dec. 2012.

2. M.H. Alam, J.W. Ha, and S.K. Lee, "Novel Approaches to Crawling Important Pages Early," Knowledge and Information Systems, vol. 33, no. 3, pp 707-734, Dec. 2012

3. S. Aral and D. Walker, "Identifying Influential and Susceptible Members of Social Networks," Science, vol. 337, pp. 337-341, 2012

4. A. Machanavajjhala and J.P. Reiter, "Big Privacy: Protecting Confidentiality in Big Data," ACM Crossroads, vol. 19, no. 1, pp. 20-23, 2012.

5. S. Bannered and N. Agarwal, "Analyzing Collective Behavior from Blogs Using Swarm Intelligence," Knowledge and Information Systems, vol. 33, no. 3, pp. 523-547, Dec. 2012.

6. E. Birney, "The Making of ENCODE: Lessons for Big-Data Projects," Nature, vol. 489, pp. 49-51, 2012.

7. J. Bollen, H. Mao, and X. Zeng, "Twitter Mood Predicts the Stock Market," J. Computational Science, vol. 2, no. 1, pp. 1-8, 2011.

8. S. Borgatti, A. Mehra, D. Brass, and G. Labianca, "Network Analysis in the Social Sciences," Science, vol. 323, pp. 892-895, 2009.

9. J. Bughin, M. Chui, and J. Manyika, Clouds, Big Data, and Smart Assets: Ten Tech-Enabled Business Trends to Watch. McKinney Quarterly, 2010.

10. D. Centola, "The Spread of Behavior in an Online Social Network Experiment," Science, vol. 329, pp. 1194-1197, 2010.

11. M. Helft, "Google Uses Searches to Track Flu's Spread," TheNewYorkTimes, 2008.

12. D. Gillick, A. Faria, and J. DeNero, MapReduce: Distributed Computing for Machine Learning, Berkley, Dec. 2006.

13. A. Jacobs, "The Pathologies of Big Data," Comm. ACM, vol. 52,no. 8, pp. 36-44, 2009.

14. A. Yao, "How to Generate and Exchange Secretes," Proc. 27th Ann. Symp. Foundations Computer Science (FOCS) Conf., pp. 162-167,1986.

15. E. Schadt, "The Changing Privacy Landscape in the Era of Big Data," Molecular Systems, vol. 8, article 612, 2012.