IJCTA, 9(27), 2016, pp. 455-460 ISSN: 0974-5572
© International Science Press

Visualizing Big Datamining: Issues, Challenges and Opportunities

A. Angelpreethi* and S. Britto Ramesh Kumar**

ABSTRACT

While "big data" has become a tinted catchphrase since last year, "big data mining", i.e., extracting or mining from big data, has nearly straight away followed up as an rising, consistent research area. Big Data refers to large amount of unprocessed complex data. This huge volume of unprocessed data cannot be handled using traditional database management tools. Data mining is extracting the knowledge from the database. Big Data mining engage to extract some functional information from these enormous sets of data and stream of data, due to its volume, velocity and variety. This paper expresses a general idea of Big Data mining, issues and challenges.This paper anticipate our endeavor will help restructure the subject area of today's data mining technology toward solving tomorrow's superior challenges emerging in accordance with big data.

Keywords: Data mining, big data, bigdata mining, knowledge discovery, Data mining tools.

1. INTRODUCTION

Data is distinct pieces of information, which is generally formatted in a special way. Data can be classified into two different types namely structured data and unstructured data. Unstructured data refers to information that does not have a predefined or proper data model. Presentations, slide share videos, images, instant messages, text messages are examples of unstructured data. Structured data is data that can be easily organized. It is clean, analytical and stored in databases. Today's data industry estimates that structured data accounts only 20% of the data available.

In modern years we have seen the spectacular raise in the intensification of information due to collection of data from variousautonomous or associated applications and services. Very popular example is Internet data. Internet is mounting atlightning velocity. The Data stored on the internet is huge. Every second millions of bytes are retrieved and processedeverywhere in the world.This enormous data and information on the internet is increasing at an incredible rate and even estimates are proven wrongevery second.

2. BIG DATA

Huge amount of complex and unprocessed data is called as a Big Data. The 'Big Data' buzz-word is used to describe group of huge,complex or unprocessed data which become complex or unfeasible to process, analyze and store using currentmethodologies, traditional database management tools and analytical solutions. Unlike the immense majority of computer science research, big data has established the considerable public and media interests.

Big data comes from a variety of sources and in different formats. Every day we create 2.5 quintillion bytes of data. 90 % ofthe data in the world today has been created in the last 2 years. According to IBM Report (2014), from 2003 to2010 we have created 5 billion gigabytes data. In 2011 the same amount of data was creating in two days. In 2013 thesame amount of data is created in every 10 minutes. This fast growth is hurried by the striking increase inreceiving of social networking applications like Face book, Twitter, Link In, and flicker, mobile phones, sensor data, you tube, images, mp3audios, educational contents etc.

* Research Scholar, Department of Computer Science, St. Joseph's College, Tiruchirappalli, TN, India, *Email: angelpreethi.mca@gmail.com.*

** Assistant Professor, Department of Computer Science, St. Joseph's College, Tiruchirappalli, TN, India, *Email: brittork@gmail.com.*

Twitter is a social media network which is very popular. In 2012 there are 98000 tweets were generated. In 2013 there are 2.78lakhs, but it 2014 3.42lakhs this shows that the amount of data rate increased every year. Skypecan generate 14Lakhs video calls in 2015. 410000 photos were uploaded in Instagram. These amounts of data are generated in every one minute. These data are called as unprocessed data or big data.

Big data has some characteristics such as volume, velocity, and variety. Volume is quantity of data. It says how machine in generated large amount of data than the traditional data. For example in 2013 in a minute there are79361 comments are posted in face book. Variety refers to various data coming from different data sourceslike, sensor networks, Government data holdings,company market lead databases, bidirectional interactions, E- Health networks, Closed-circuit television, video camerasetc. Streaming of data is called as the Velocity or otherwise velocity refers to data in motion and we need tofind the interesting fact from the streaming of data.

Table 1
Open source tools for big data

S.No	*Type of Tool*	*Tools available in online*
1	Platforms and tools	Hadoop, Map Reduce, Grid Gain, Hpcc systems, Storm
2.	Databases	Cassandra, HBase, Mongo DB, Neo4j, CouchDB, Orient DB, Terrastore, Flock DB, Hibari, Riak, Hypertable, Blazegraph, Hive, Info Bright Community Edition, InfiniSpan, Redis
3.	Business Intelligent Tools	Talend, Jaspersoft, Jedox, Pentaho, SpagoBI, KNIME,BIRT
4.	Mining Tools:	RapidMiner, Mahout, Orange, Weka, DataMelt, KEEL, SPMF, Rattle
5.	File System Programming Languages.	Gluster, Hadoop Distributed File System, Pig, R, ECL
6.	Tools for Transfer and Aggregate	Lucene, Solr, Sqoop, Flume, Chukwa
7.	Miscellaneous Big Data Tools	Terracotta,Avro, Oozie, Zookeeper

Table 1 describes the vast number of open source tools available in free of cost in online. These are the tools are called open source tools. There are immensegap between weight of the Big data and capabilities ofthe current DBMSs for storage, handle, distribution, searchand visualize. To overcome this large gap, Hadoop was introduced which is the heart of Big data. Yahoo and other big companies were created an apache open source version of Google's mapreduce framework, called Hadoop mapreduce. It uses the Hadoop Distributed File System(HDFS).

The map reduce allows two function namely map and reduce. The input is divided into large number of key value pairs then the mapper function is called and divided into many key value pairs [1].After all entries are processed new set of key value pairs are produced and then the reducer will reduce or group the produced values based on common keys. In order to support the map reduce Google developed the Big Table a distributed storage system designed for managing ordered data.

3. DATA MINING

Data mining has concerned a great compact of awareness to the information technology and in the public as a whole in modern years, due to the broad availability of enormous amounts of data and the imminent need for increases such data into functional information and knowledge.Data mining can be used for applications ranging from market analysis, fraud detection, customer retention, to production control and science exploration.

Data mining refers to extracting the knowledge from huge amounts of data. Data mining also known as data dredging, data archaeology, data analysis, pattern analysis, knowledge mining from data, and Knowledge Discovery from Data (KDD) [16]. It consists of several steps.

1. *Data Cleaning*: Used to remove the noise and inconsistent data.
2. *Data Integration*: where several data sources are combined.

3. *Data Selection*: significant data related to analysis are retrieved from the data server.
4. *Data transformation*: data are altered for mining operations.
5. *Datamining*: intellectual methods are applied to the particular data to extract the pattern.
6. *Pattern Evaluation*: To identify the interesting measures some interesting patterns are identified
7. *Knowledge presentation*: visualization techniques are applied to present the mind knowledge to the user.

3.1. Data Mining Functionalities

The most commonly used data mining techniques are

3.1.1. Association Rule

Association Rule Mining Algorithms employ a support- confidence framework. Association rule look for patterns. Support and confidence are two measures of rule interestingness.

$$\text{Support } (A \Rightarrow B) = P(AUB) \quad (1)$$

$$\text{Confidence} = \text{Support } (AUB)/\text{Support } (A) \quad (2)$$

For example

buys (X,"rice") => buys (x,"oil") [support = 1%, confidence = 50%]

A confidence of 50% means that if a customer buys rice, there is a 50% chance that the customer will buy oil as well. A 1% support means that 1% of all of the transactions under analysis show that rice and oil purchased together.

3.1.2. Classification Rule

Classification is the method of finding the model that describes and differentiate the data classes or perceptions for the purpose of being able to use the model to envisage the class of objects whose class label is unknown.

3.1.3. Clustering

To identify with the similarities and differences of each data within the data set clustering is used. It is used to find those types of data sets.

3.1.4. Decision Tree

It is a hierarchical tree based structure in which each internal node specifies the test on attribute and branch represents the result of the test, leaf denotes the class label. Top most single node is the root node of that tree.

3.1.5. Prediction

Prediction values are constant valued functions. It is used to calculate the missing numerical data values slightly than class labels.

3.1.6. Artificial Neural Network

Typically a collection of neuron like processing units with weighted connections between the units.

Present data mining techniques and algorithms are not willing to meet the new challenges of big data.Applying existing data mining algorithms and techniques to real world problems has various challenges due to scalability and adequate of these algorithms which cannot locate with the uniqueness of big data. Big data mining demands highly

scalable strategies and algorithms. In the next chapter we examine the concept of big data mining and related issues, including emerging challenges dealing with big data.

4. BIG DATA MINING

The objective of big data mining techniques go beyond fetching the requested informationor even finding some unseen relationships and patterns between numeral parameters. Evaluating with the outcome resulting from mining the predictable datasets, presenting the enormous volume of consistent assorted bigdata has the latent to exploit our facts and insights in the target field.

The traditional data mining techniques or functionalities explained above are not competent of mining such huge scattered data. Each technique extracts from and analyzes the historical datasets for decision making [10]. The purpose of big data mining is to go away from the historical algorithms like market basket analysis or finding some hidden relationships and patterns between numerical parameters of data. But the purpose is to design and implement very large scale parallel data mining algorithm. This brings a new challenge to the research community. Big data mining has deal with heterogeneity, velocity, privacy, accuracy, trust, and interactivity. There are some open source tools are available for big data mining namely Apache mahout, MOA are some of the available tools. Big data value chain can be divided into three processes

1. *Big Data collection*: It contains huge data sets. Gather and add and provide access to those data sets on the basis of request.
2. *Big Data Aggregation*: It creates the technical infrastructure for data aggregation. Big data pool can provide the software tools to manage and restore the knowledge.
3. *Big Data Integration*:All the aggregated data will be integrated for decision making.

5. ISSUES AND CHALLENGES OF BIG DATA MINING

There are some issues and challenges are arrived wile mining the huge volume of data. Heterogeneity, complexity, privacy, and scalability.

5.1. Hetrogenity

In the past data mining techniques have been used to discover the unknown patterns and relationships of structured, homogeneous data sets. The complexities of big data analysis derive from its large scale as well as the occurrence of mixed data based on diverse patterns or rules in the gathered and stored data. In the case of complicated heterogeneous mixture data, the data has several patterns and rules and the properties of the patterns vary greatly. Deficient data creates uncertainties during data analysis and it must be managed throughout data analysis. Doing this properly is also a challenge[5].

5.2. Complexity

Managing the data and increase the data sources is a challenging one. Conventional software tools are not enough for managing the huge volumes of data. Data analysis, organization, retrieval and modeling are also challenges due to scalability and complexity of data that needs to be analyzed [5].

5.3. Privacy

Privacy is one of the major issues in the area of big data mining. Protecting the data is always a serious issue even from the data mining [5]. This issue has extremely severe with big data mining that frequently requires private information in order to construct accurate results. The huge volume of big data such as face book that contains fabulous amount of highly organized personal information, every piece of information about everybody can be

mined out, and when all pieces of the information about a person are dug out and put together, any privacy about that individual instantaneously vanishes [6].

5.4. Scalability

Managing huge volume of data is a challenging issue for many tasks. The size of big data requires high scalability of its data management and mining tools.

6. TOOLS USED IN BIG DATA MINING

There are many open source tools are available at free of cost in online. Some of them are listed below

6.1. R Language

R is an open source software programming language and software environment designed for statistical computing and visualizing graphics. Statisticians use this software for developing statistical software and data analysis.

6.2. Apache Mahout

Apache Mahout is mainly based on Hadoop. It has implementations of a widerange of machine learning and data mining algorithms clustering, classification, collaborative and frequent pattern mining.

6.3. WEKA

WEKA is a collection of machine learningalgorithms for solving real-world data mining problems. WEKA contains tools for data mining rules such as classification, pre-processing, clustering, regression, association rules, and visualization.

7. CONCLUSION

Big Data is going to continue growing during the next years, and each data scientist will have to manage much huge amount of data every year. We discussed in this paper some approaches about the subject, and what we consider are the main concerns and the main challenges for the future. Big Data is becoming the new ultimate boundary for technical data research and for business applications. We are at the beginning ofa new era where Big Data mining will help us to find out the awareness about data that no one has discovered before. Everybody is genially invited to take part in this courageous journey.

REFERENCES

[1] Dean, J., Ghemawat, S., "MapReduce: Simplified Data Processing on Large Clusters"*Proceedings of the sixth symposium on Operating System Design and Implementation, 137-150, 2004.*

[2] Ghemawat, S., Gobioff, H., Leung, S.T.,"The Google File System",*Proceedings of the19th ACM Sympo-sium on Operating Systems Principles, Bolton Landing, New York,29-33,2003.*

[3] Weiss, S.M.,Indurkhya .N.," Predictive data mining a practical guide", *Morgan Kaufmann Publishers,* Sanfrancisco, USA,1998.

[4] Diebold, F ., "Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting",*Proceedings of theWorld Congress of the Econometric Society, 2000.*

[5] DunrenChe, MejdlSafran, and ZhiyongPeng, "From Big Data to Big Data Mining: Challenges, Issues, and Opportunities", *Proceedings of the DASFAA Workshops , LNCS*, 1–15, 2013.

[6] Wang, Qian Wang, Kui Ren, Wenjing Lou, "Privacy-Preserving Public Auditing for Data Storage Security in Could Computing", *IEEE Transactions on Computers*, **62(2)**, 362-375, 2013.

[7] Ahmed, Karypis, RezwanAhmed, George Karypis, "Algorithms for mining the evolution of conserved relational states in dynamic networks", *Knowledge and Information Systems*, Volume **33(3)**, 603-630, December 2012.

[8] Clifon, C, Marks, D, "Security and privacy implications of data mining," *In Proceedings of SIGMOD'96 Workshop Research Issues on Data Mining and Knowledge Discovery(DMKD'96)*, 15-20, 1996.

[9] Ranger, C., Raghuraman, R., Penmetsa, A., Bradski, G., and Kozyrakis, C., "Evaluating MapReduce for multi-core and multiprocessor systems", *Proceedings of the 13th IEEE International Symposium on High Performance Computer Architecture (HPCA '07)*, 13-24, 2007.

[10] Gopalkrishnan, V., Steier, D., Lewis, H., and Guszcza, J., " Big data, big business: bridging the gap", *In Proceed-ings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*, Big-Mine '12, 7-11, New York, 2012.

[11] Peng Y., Kou G, Shi Y, Chen Z, "A Descriptive Framework for the Field of Data Mining and Knowledge Discovery", *International Journal of Information Technology and Decision Making*, **7(4)**, 639 – 682, 2008.

[12] Nandhakumar A.N, NanditaYambem, "A Survey of DataMining Algorithms on Apache Hadoop Platforms , International Journal of Emerging Technology Advanced Engineering,**4(1)**,2014.

[13] RohitPitre, Vijay Kolekar, A Survey Paper onData Mining With Big Data , *International Journal of Innovative Research and Advanced Engineering,**1(1),*** 2014.

[14] ThirumalaRao B, Reddy S, "Survey on Improved Scheduling in Hadoop MapReduce in Cloud Environments",*International Journal of Computer Applications*, **34(9)**, 2011.

[15] SumanArora, MadhuGoel, "Survey Paper on Scheduling in Hadoop" *International Journal of Advanced Research in Computer Science and Software Engineering*, **4(5)**, 2014.

[16] Fayyad U., Piatetsky Shapiro G., and Smyth P, "From Data Mining to Knowledge Discovery: An Overview", *Advances in Knowledge Discovery and Data Mining*,1-36, Cambridge, 1996.