ISSN:0975-9646

Visualizing Big Data Mining: Challenges, Problems and Opportunities

Rakesh Ranjan Kumar
CSED, MNNIT Allahabad
Allahabad, India

Binita Kumari
Assistant Professor,GGSESTC, CHAS
Bokaro, India

***Abstract:* — Big Data is fast becoming a big problem since last year. Big data refers to datasets which has large size and complexity. We can't capture, store, manage and analyze with typical database software tools. Data mining is highlighted buzzword that is used to describe the range of Big data analytics, with collection, extraction, analysis and statics. Big Data mining involves to extracting useful information from these huge sets of data and streams of data, due to its volume, velocity and variety. This paper describes an overview of Big Data mining, problems related to mining and the new opportunities. During discussion we include platform and framework for managing and processing large data sets. We also discuss the knowledge discovery process, data mining, and various open source tools with current condition, issues and forecast to the future.**

Keywords: — Data mining, Big data, Big data mining, Big data management, Knowledge discovery, Data mining tools.

1. INTRODUCTION

In 21st century Big Data is the modern kind of electricity power that transforms everything it touches in business, government, and private life. Every day, we generate more than 2.5 quintillion bytes of data and 85% of the data in the world today has been created in the last two years only. In which 80% of data captured today is unstructured such as climate information, data post to social media sites, digital pictures and video, purchase transaction records and getting GPS signals from cell phone. All of this is example of unstructured data is Big data. In 2010, Google estimated that every two days at that time the world generated as much data as the sum it generated up to 2003. In spite of the very recent "Big Data Executive Survey 2013" by New Vantage Partners [15] that states "It's about variety, not volume", lots of people with author would still believe the prime issue with Big Data is scale or volume. Big data has a great variety of data forms: text, images, videos, sounds, and data which have extreme scale. Big data frequently comes in the form of streams of a variety of types. The growth of data will never stop. According to the 2011 IDC Digital Universe Study, in 2005 there were created and stored 130 exabytes of data. The amount grew to 1,227 exabytes in 2010 and is projected to grow at 45.2% to 7,910 exabytes in 2015[14]. In addition to being the hottest new trend in business and government, Big Data is fast becoming a persistent force in modern science. American president Barack Obama administration started a $200M Big Data in Science scheme with the goals of improving economical growth which creates lots of jobs in various sector such as education, health, energy, environmental, public safety, and global development.

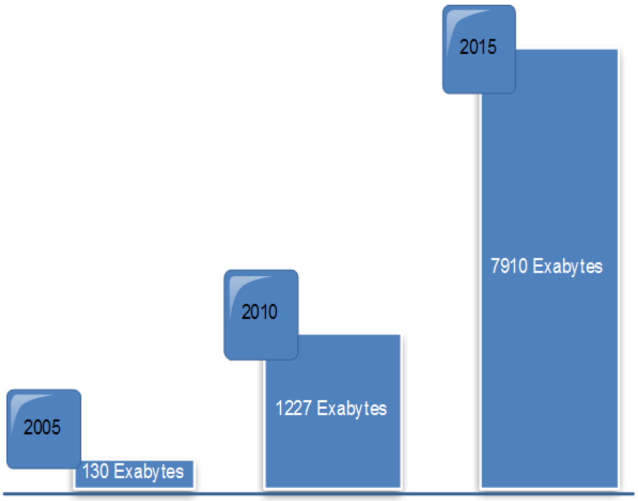

Figure 1:-A decade of digital universe growth: Storage in Exabytes

For all of these applications, we are continuously facing significant challenges in handle the vast amount of data, including challenges such as:-system capabilities, design of appropriate algorithm and business models.

From the perception of data mining, mining Big data has a lot of new challenges and opportunities. Big data allow greater value such as hidden knowledge and more valuable insights. It has great challenges to extract these hidden knowledge and insights from Big data since the established process of know-ledge discovering and data mining from conventional datasets was not designed to and will not work well with Big data.

We introduce Big Data mining and its applications in Section 2. We summarize the papers presented in this issue in Section 3, and discuss about Big Data controversy in Section 4. We point the importance of open-source software tools in Section 5 and give some challenges and forecast to the future in Section 6. Finally, in Section 7 we provide some conclusion.

2. DATA MINING

Knowledge Discovery (KDD) is a process for extracting useful knowledge from large volume of data in which data mining work as a core step and most interesting step. The constant growth of online data due to the Internet and the widespread use of databases make KDD methodologies very essential. The term data mining first appeared in the 1990s while before that, statisticians used the terms "Data Fishing" or "Data Dredging" to refer to analyzing data without a-priori hypothesis.

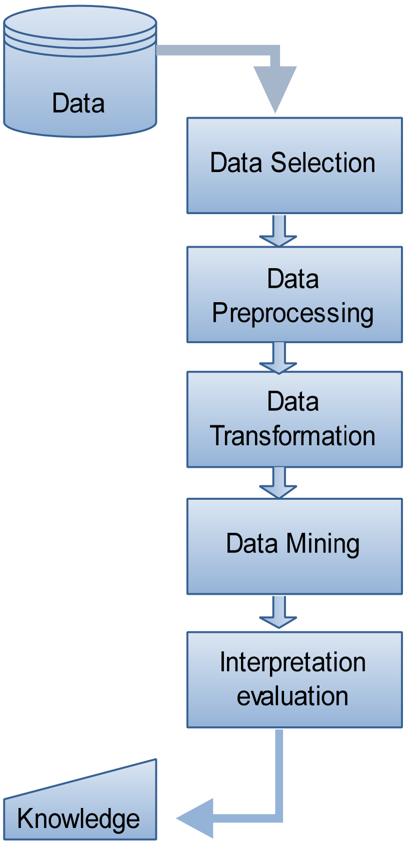

Figure 2:- KDD Process

A widely accepted definition of KDD is given by Fayyad et al. in which KDD is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad-Piatetsky-Smyth 1996). The definition regards KDD as a complicated process comprising a number of steps. Data mining is one step in the KDD process. Typically, data mining discovers interesting patterns and relationships hidden in a large volume of raw data, and its result helps us to make valuable predictions or future observations in the real world. Today data mining has been used by different applications such as business, medicine, science and engineering. It provides lots of beneficial services to real businesses – both the providers and ultimately the consumers of services.

2.1 DATA MINING PARAMETERS

The most commonly used techniques in the data mining are:

Association - Looking for patterns where one event is connected to another event.

Artificial neural networks - Non-linear predictive models that learn through training and resemble biological neural networks in structure

Classification - is a systematic process for obtaining important and relevant information about data, and metadata – data about data.

Clustering - the process of identifying data sets that are similar to each other to understand the differences as well as the similarities within the data.

Decision trees: Tree-shaped structures that represent sets of decisions.

Genetic algorithms: Belong to larger class of evolutionary algorithms that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.

Table 1: Data mining trends comparative statements

Data Mining Trends	Algorithms Techniques Employed	Data Formats	Computing Resources
Past	Statistical, Machine Learning Techniques	Numerical data and structured data stored on traditional databases	Evolution of 4G PL and various related techniques
Current	Statistical, Machine learning, Artificial Intelligence, Pattern Finding Techniques	Heterogeneous data formats includes structured, semi structured and unstructured data	High speed networks, High end storage device and parallel, Distributing computing.
Future	Soft Computing techniques like Fuzzy logic, Neural networks and Genetic Programming	Complex data objects include high dimensional, high speed data streams, sequence, graph, Multi instance objects, temporal data etc....	Multi-agent technologies and Cloud Computing

Current data mining techniques and algorithms are not ready 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 and techniques which cannot stand with the characteristics of Big data. Big data mining demands highly scalable strategies and algorithms, which has preprocessing steps such as data filtering and integration, complex parallel computing situation and effective user interaction. In the next chapter we examine the concept of Big data and related issues, including emerging challenges dealing with Big data.

3. BIG DATA

We are awash in a flood of data today. There is variety of application areas, from where data is being collected at unmatched scale. According to McKinsey [5], Big Data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze. There is no exact definition of how Big a dataset is necessary to considered as Big Data. According to O'Reilly "Big data is data that exceeds the processing capacity of conventional database systems. The data is large in size, which moves too fast, and these data does not fit in the structures of existing database architectures. For getting value from these data, definitely there is an alternative way to process it." Big data has 3 V's characteristic which was describe by Doug Laney [6].

- **Volume**: machine-generated data is produced in much larger quantities than traditional data. For example, a single jet engine can generate 13TB of data in 25 minutes.
- **Variety**: In current day's data comes in different types of formats such as text, sensor data, audio, video, graph, and many more.
- **Velocity**: data comes as streams and we need to find interesting facts from it in the real time i.e. social media data stream.

But in current scenarios, there are two more V's:

- **Variability**: defined as the many ways in which the data may be variance in meaning, in lexicon. Differing questions which require different interpretations.
- **Value:** this is the most important feature of Big data. This feature describes for costs a lot of money to implement IT infrastructure systems to store Big data, and businesses are going to require a return on investment.

Gartner [7] in 2012 summarizes the definition of Big data as high volume, velocity and variety information assets which demand cost-effective, information processing tools for enhanced insight and decision making. There are large gap between demands of the Big data and capabilities of the current DBMSs for storage, manage, sharing, search and visualize. To overcome this large gap, Hadoop was introduced which is the core of Big data. Hadoop architecture that has a distributed file system, data storage platforms and an application layer that manages distributed processing, parallel computation, workflow and configuration management for unstructured data. There are many other non-relational databases such as NoSQL databases and MPP system that are also scalable, Network-oriented, semi-structured. With the emergence of Big Data, traditional RDBMS, MPP are transitioning into a new role of supporting Big Data management by processing structured datasets as outputs of Hadoop or MapReduce technologies.

To overcome the scalability of Big Data Google created a programming model named MapReduce [3] Which was facilitated by GFS (Google File System [4]), a distributed file system where the data can be simply partitioned over thousands of nodes in a cluster. Afterward, Yahoo and other Big companies created an Apache open-source version of Google's MapReduce framework, called Hadoop MapReduce. It uses the Hadoop Distributed File System (HDFS) an open source version of the Google's GFS.

The MapReduce framework allows users to define two functions, map and reduce, which process large number data in parallel [8]. Users specify a map function a key/Value pair to generate a set of intermediate key/value pairs, and a reduce function that merges all intermediate value associated with the same intermediate key.

Table 2: Different frameworks on Big Data

	RDBMS	**MPP**	**Hadoop Framework**
Processing	Sequential processing	Some processing parallelism	Massively parallel processing. Grid processing
Scalability	Vertical	Limited horizontal	Massive horizontal
Storage	Relational / SQL database	Propriety data warehouse and data marts	No relational / NoSQL database
Data type	Structure	Structured	All types
Architecture	Shared disk and memory	Shared nothing	Shared nothing
Hardware	Single processor to multi core computing	Data warehouse appliances	Commodity hardware in a distributed grid or clusters
Analysis	Model-based	Model-based	Not model-based

4. BIG DATA MINING

In 1998, 'Big Data' term was appeared for the first time by John Mashey in his slide with title of "Big Data and the Next Wave of InfraStress" [9]. First book was published on the Big data mining in 1998 by Weiss and Indrukya [10]. However, the first academic paper with Big data was present in the 2000 by Diebold [11]. The goals of Big data mining techniques go beyond fetching the requested information or even uncovering some hidden relationships and patterns between numeral parameters. Analyzing fast and massive stream data may lead to new valuable insights and theoretical concepts. Comparing with the results derived from mining the conventional datasets, unveiling the huge volume of interconnected heterogeneous Big data has the potential to maximize our knowledge and in sights in the target domain.

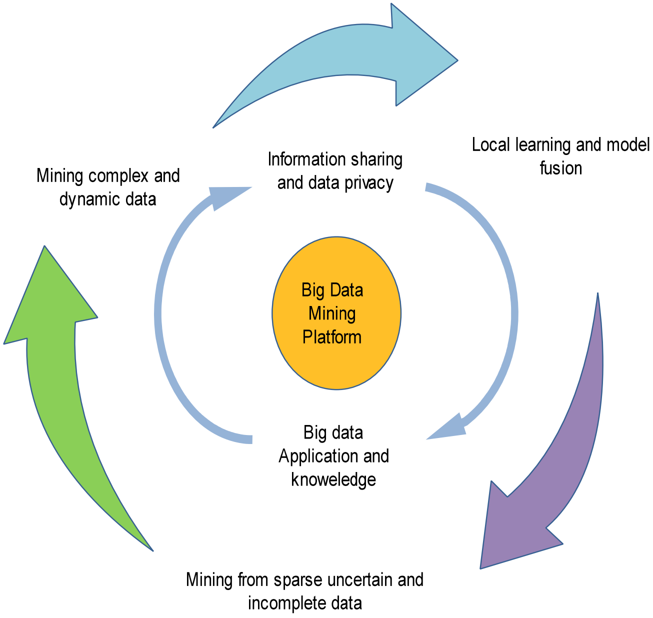
Local learning and model fusion
Mining complex and dynamic data
Information sharing and data privacy
Big Data Mining Platform
Big data Application and knoweledge
Mining from sparse uncertain and incomplete data

Figure3: A Big data mining framework

Big Data mining is necessary in many sectors:

Public sector: enables government departments and developmental organizations to analyze large amount of data across populations and to provide better governance and service.

Financial service: making better trading and risk decisions, improve product by better customer identification and marketing campaign.

Healthcare: mining DNA of each person, to discover, monitor and improve health aspects of every one. **Manufacturing**: finding new opportunities to predict maintenance problems enhance manufacturing quality and reduce costs using Big Data.

Telecommunications: need of real-time data mining of data generated by mobile devices including phone calls, text messages, applications, and web browsing for better customer service and to build on retention and loyalty.

Retails: Big data mining offers numerous opportunities to retailers to improve marketing, merchandising, operations, supply chain and develop new business models

Other industries: mining can also be used in many other industries such as Oil and gas, transportation, GPS system and satellite.

4.1 Issues and Challenges of Big Data Mining

There are lots of issues and challenges facing with Big Data mining such that heterogeneity i.e. variety, scale i.e. volume, timeliness i.e. velocity, garbage mining and privacy

Heterogeneity: in the past, data mining techniques have been used to discover unknown patterns and relationships of interest from structured, homogeneous, and small datasets (from today's perspective). The data from different sources inherently possesses a great many different types and representation forms, and is greatly interconnected, interrelated and inconsistently represented taking out such a dataset, the great challenge is understandable and the complexity is not even imaginable before we deeply get there.

Scale: Of course, the first thing anyone thinks of with Big Data is its size. Managing huge and speedily increasing volumes of data has been a challenging issue for many decades. The exceptional volume/scale of Big data requires commensurately high scalability of its data management and mining tools.

Timeliness: The flick side of size is speed. The larger the data set to be increased, it will take too much of time for analyze. The design of a system that successfully deals with size is likely also to result in a system that can process a given size of data set faster. We must finish a processing/mining task within a specified time; otherwise, the processing/mining results turn into less valuable or even worthless. Ideal applications with real-time requests contain earthquake prediction, stock exchane market prediction and agent-based autonomous exchange (buying/selling) systems. Speed is also relevant to scalability – conquering or partially solving anyone helps the other one. The speed of data mining depends on two major factors: data access time (deter-mined mainly by the underlying data system) and, of course, the efficiency of the mining algorithms themselves.

Privacy: The privacy of data is another important concern in the context of Big Data. Data privacy has been always an issue even from the beginning when data mining was applied to real-world data. This issue has become enormously serious with Big data mining that often requires personal information in order to produce relevant/accurate results such as location-based and personalized services, e.g., targeted and individualized advertisements.

4.2 Related work with Big Data mining

There are many contributions related with Big Data mining in which we are going to introduce some eminent research.

Scaling Big Data Mining Infrastructure: The Twitter Experience by Jimmy Lin and Dmitriy Ryaboy (Twitter,Inc.). Twitter has tremendous growth in term of size, complexity, number of users. In this paper trying to explore ideas about infrastructure and development capability of data mining on Big data over the few past Authors discussed two topics: In first topic trying to discuss a crucial role in understanding how to store petabytes data from many sources, but overall they are unable to providing clear idea about data availability to generate insights. In second, they examine that a most important challenge in building data analytics platforms stems from the heterogeneity of the various components integrated together into production workflows.

Mining Heterogeneous Information Networks: A Structural Analysis Approach by Yizhou Sun and Jiawei Han [20]. The paper presents a new methodology for mining heterogeneous information networks, based on the fact that, in many real-life scenarios, data are available in heterogeneous information networks, which are interconnected multimedia objects which contains titles, descriptions and subtitles. This scenario consists of transform documents into bag-of-words vectors, after that decompose the corresponding heterogeneous network into separate graphs, which compute structural-context feature vectors with Rank of Page. At end, constructing a common feature vector space in which knowledge discovery is performed.

Big Graph Mining: Algorithms and discoveries:

This paper [21] presented by U Kang and Christos Faloutsos and they presents an overview of mining Big graphs, focusing how use Pegasus tool, showing how implement data mining in the Web Graph and Twitter social network. The paper gives inspirational future research directions for Big graph mining. In this paper they explain about Pegasus, which is a Big graph mining system which is erect on top of MapReduce. Also introduce GIM-V, an important primitive which is used by Pegasus as an algorithm for analyzing structure of large graphs.

Mining Large Streams of User Data for Personalized Recommendations:

This paper [22] written by Xavier Amatriain and presents some lessons which discuss the recommender and personalization techniques used in Net. He uses data mining and a machine learning method which is uses for predicting what users has preferences. There are many lessons came out of the competition but Recommender Systems have evolved in these. With help of availability of dissimilar kinds of user data in industry and the interest, revolution has carried among the research community. The purpose of this paper provides, what is the exact figure of uses of the data mining approaches for personalization and recommendation techniques.

4.3 Tools used in Big data mining

In Big Data Mining, there are many open source tools. Some of the most popular tools are the following:

Apache Mahout [12]: This is provided by Apache Software Foundation which offers free implementations of distributed algorithms and data mining techniques, which is mainly based on Hadoop. It has implementations of a wide range of machine learning and data mining algorithms clustering, classification, collaborative and frequent pattern mining.

R [13]: It is an open source software programming language and software environment designed for statistical computing and visualizing graphics. Data miners and statisticians use this software for developing statistical software and data analysis. R was designed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand. Source code for R software environment is written in C, FORTRAN and R and available under the GNU General Public License.

WEKA [16]: WEKA is a collection of machine learning algorithms for solving real-world data mining problems. These algorithms can either be applied directly to a dataset or you can use this algorithm from your own Java code. Weka includes tools for data mining rules such as classification, pre-processing, clustering, regression, association rules, and visualization. This tool is well-suited for developing new machine learning schemes. It is written in Java and runs on almost any platform.

RapidMiner [18]: RapidMiner is a software platform developed by the company of the same name that provides software, solution and service in the field of machine learning, predictive analytics, data mining, text mining, and business analytics. It automatically and intelligently analyzes data on a large scale. We can use Rapid Miner for business and industrial applications. This tools can also useful in research, education, training, rapid prototyping, and application development. With the help of RapidMiner tool we support all steps of the data mining process such as results visualization, validation of results and optimization of results.

KNIME [19]: KNIME is a user-friendly graphical workbench for the entire analysis process such as data access, data transformation, and initial investigation as well as for great predictive analytics, visualization and reporting the reults. This open source platform provides over 1000 modules (nodes), including those of the KNIME community and its extensive partner network.

Vowpal Wabbit [17]: it is an open source learning system library and program which is developed at Yahoo! This software is very popular online machine learning implementation for solving linear models like LASSO, sparse logistic regression, etc. It was started by John Langford.

5. Conclusions And Future Work

We have entered an era of Big Data. Mining Big Data is currently big challenging task. While the term Big Data has characteristics (1) huge with heterogeneous and diverse data sources, (2) independent with distributed and decentralized control and (3) complex and evolving in data and knowledge association. Big Data is going to be more diverse, larger, and faster. We discussed in this paper some insights about the topic and the main challenges for the future. Big data mining shows potential research area, still it in growing stage. Because limited work has done on Big Data mining, we believe that much work is required to overcome their challenge which is related to heterogeneity, speed, accuracy, scalability, trust, provenance, privacy, and instructiveness. This paper also provides an overview of different Big Data mining tools. We must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data.

6. References

[1] Wei Fan, Albert Bifet “Mining Big Data: Current Status, and Forecast to the Future SIGKDD Explorations

[2] Dunren Che, Mejdl Safaran, and Zhiyong Peng “Big Data to Big Data Mining”, Springer-Verlag Berlin Heidelberg 2013

[3] Dean, J., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters. In: 6th Symposium on Operating System Design and Implementation (OSDI), pp. 137–150 (2004)

[4] Ghemawat, S., Gobioff, H., Leung, S.T.: The Google File System. In: 19th ACM Sympo-sium on Operating Systems Principles, Bolton Landing, New York, pp. 29–43 (2003)

[5] James Manyika, et al. Big data: The next frontier for innovation, competition, and productivity.

[6] D. Laney. 3-D Data Management: Controlling Data volume, Velocity and Variety. META Group Research Note, February 6, 2001

[7] Gartner, http://www.gartner.com/it-glossary/big-data.

[8] Dean, J., Ghemawat, S.: MapReduce: a Flexible Data Processing Tool. Communication of the ACM 53(1), 72–77 (2010)

[9] F. Diebold. On the Origin(s) and Development of the Term "Big Data". Pier working paper archive, Penn Institute for Economic Research, Department of Eco-nomics, University of Pennsylvania, 2012.

[10] S.M.Weiss and N. Indurkhya. Predictive data mining: a practical guide. Morgan Kaufmann Publishers Inc.,San Francisco, CA, USA, 1998.

[11] “F. Diebold. "Big Data" Dynamic Factor Models for Macroeconomic Measurement and Forecasting. Discus-sion Read to the Eighth World Congress of the Econo-metric Society, 2000.

[12] Apache Mahout, http://Mahout.apache.org.

[13] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Com-puting, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[14] IDC. The 2011 Digital Universe Study: Extracting Value from Chaos.

[15] NewVantage Partners: Big Data Executive Survey (2013)

[16] A.Bifet, G. Holmes, R. Kirkby, and B. Pfahringer. Massive online analysis Journal of machine learning (JMLR),2010

[17] J. Langford. Vowpal Wabbit, http://hunch.net/~vw/, 2011

[18] From Wikipedia: en.wikipedia.org/wiki/RapidMiner

[19] The KNIME Text Processing An introduction by Dr. Killan Thiel and Dr micheal Berthold

[20] Mining Heterogeneous Information Networks: A Structural Analysis Approach by Yizhou Sun and Jiawei Han

[21] Big Graph Mining: Algorithms and discoveries: U Kang and Christos Faloutsos

[22] Mining Large streams of user Data for personalized by Xavier Amatriain