RESEARCH ARTICLE OPEN ACCESS

A Review on Data Mining With Big Data

Akanksha, Vinod Maan
Department of Computer Science
Mody University of Science Technology
Rajasthan - India

ABSTRACT
Big Data is a large volume complex datasets with multiple autonomous sources. Big data is rapidly increasing in networking, storage, and the capacity of data collection. In Biomedical Science Big Data is rapidly expanding.in this paper Big Data revolution features characterized by the HACE theorem, and from the data mining proposes a Big Data processing model. With the large quantities of data number of issue is arise especially in analyzing the data in reliable manner. The main objective of Mining Big Data is to discover useful information from the large data which cannot be discovered yet. This paper provides recent Big Data challenges in data driven model.
Keywords:- Big data, structured and unstructured data, Hace Theorem, autonomous sources

I. INTRODUCTION

Big data is a practice used to describe the data generated which is large in volume, complex, and difficult to process the data. Big data refers to the structured data and unstructured data. Basically most of the Big Data is the collection of social media, internet, business organization etc. According to the modern society there are 3.3 billion users in the world and 90 percent of the data is developed in last 2 years. Google has more than 3 million servers which process over 2 trillion searches in 2012 whereas in 2000, it was only 22 million. In a day Google has more than 1 billion queries, twitter process 7 TB data everyday, every second twitter has 7000 tweets and Facebook process 10TB data and have more than 800 billion updates a day. Google, Facebook, Twitter, YouTube and many companies taking this data for generating useful data from these big data. An example, on 4 October 2012, there was more than 10 million tweets within 2 hours on the first presidential debate between President Barack Obama and Governor Mitt Romney [1].Another example, the first strike of Anna Hazzare strike numbers of tweets within 2 hours. These tweets show the public interest and feedbacks on real-time. Another example is Flickr, it's a social pictures sharing site, where 1.8 million photos received per day, from February to March 2012[2].if we assume each picture size is 2 Megabytes (MB), then daily it require 3.6 terabyte (TB) for storing. The Above examples show the rise of Big Data where the data is expanding day by day and it is beyond the ability of software tools which we used commonly to manage, store, capturing within a "tolerable elapsed time".

II.TYPES OF BIG DATA

Basically there are two major types of Big Data: Structured data and unstructured data.

Structured data are the data that is easily analyzed and categorized. Structured data are generated from the smartphones, GPS devices. It includes the things like account balances, transaction data and sales figures [6].

Unstructured data are the data which are received from the pictures uploaded, comments on the social media websites, customer reviews etc. Unstructured data cannot be categorized and analyzed [6].

III. V'S OF BIG DATA

Volume [6] means the large amount of data that are increasing rapidly. The data is in the form of music, videos and large amount of images uploaded on the social media sites.it is very common to have data in petabytes and terabytes. As the volume of data increases it makes more challenging to handle these data.

Velocity is the frequency at which the data is processed. As we think, within a minute we receive SMS messages, credit cards swipes or Facebook updates, these are the good examples of velocity. Amazon web Services Kinesis application is application that handles the high velocity of data and processed it very efficiently.

Variety refer to the types of data which are received and processed in large amount.it is most interesting technology in which more information is digitized. These data includes structured data as well as unstructured data. Structured data is good for relational type data whereas unstructured data are in the format of images, music, video etc. Unstructured have no rule.

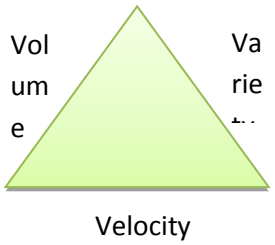

Figure 1:3v's of Big Data

IV. DATA MINING WITH BIG DATA

Collection of large volumes data is big data and these data are received from various social media sites, internet etc. From these large dataset some useful data is generated with the use of Data Mining. Data Mining is a pattern of finding new information from preexisting data.so basically mining work with big data to extract useful information from these large data sets. The figure 1[3] describe the relationship of Data Mining and Big Data. It is observed that Big data provide lots of relationship whereas Data Mining provides lots of information. Data mining is a term of six activities for the specific classes:

1. Classification
2. Estimation
3. Prediction
4. Association rules
5. Clustering
6. Description.

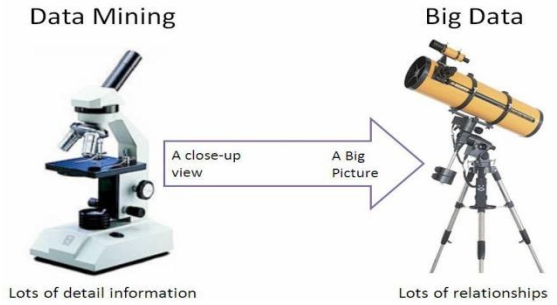

Figure 2: Data Mining with Big Data

Big Data	Data Mining
Now days Big Data is everything	Data Ming is the older version of Big Data
The size of data is large	The size of data is small
Provide lots of relationship	Extract lots of information
All Big Data task can be Data Mining	All Data Mining task cannot be Big Data

Table 1: Difference between Big Data and Data Mining

V.BIG DATA CHARACTERISTICS: HACE THEOREM

HACE stands for heterogeneous, autonomous and complex and evolving relationship. These are the main characteristics of big data which makes it more complex and challenging for discovering useful information from the data. In connection let us imagine a scenario where a blind people are asked to draw a picture of giant elephant. Some people think trunk as " this a nose", tail as a "rope" legs as a "tree "and body as a "wall.to make the scenario more complex let's assume the size of the giant elephant is increasing and the pose is changing and other the blind men can learn from the information exchanging with each other about the elephant. This task is not simple as asking the view of each blind man about the elephant and then an expert to draw the image of elephant from information gathered from all the blind men concerning each individual speaks another language.

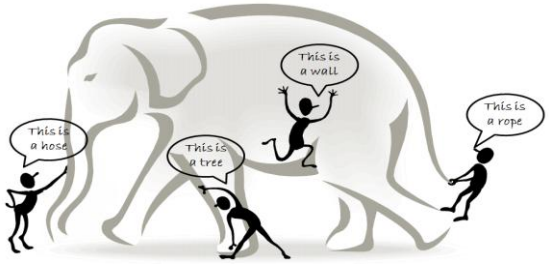

Figure 3 the blind people and the giant elephant

A. Huge Data with Heterogeneous and Diverse Dimensionality

This is the fundamental characteristics of Big Data where data is represented by heterogeneous and diverse dimensionality. Example in biomedical science a human being is represented by information like name, age, sex and so on. Images and videos are used for representing the result of x-ray examination and CT scan of an individual.

B. Autonomous Sources with Distributed and Decentralized Control

This is the main characteristics of Big Data. Being data autonomous, each data sources is able to generate and gathered the information without depending on other sources. The World Wide Web where each web server provides much information and each server is fully functional without depending on other server. For other major Big Data application such as Facebook, Google, Wal-Mart, Flickr, a number of server are deployed all over the world to run nonstop services and quick responses for queries.

C. Complex and Evolving Relationships

As the data size is increasing complexity and relationship is also increasing. When the data size is small the data is easy to handle and the relationship is also small. Nowadays data is increasing through social media which increase complexity in establishing relationship. This complication is now reality of the Big Data, the challenge is to evolve relationship between these complex Big Data.

VI. BIG DATA MINING CHALLENGES

Big Data is a framework which includes three tiers with consideration on accessing the data and computing (Tier I), data security and domain knowledge (Tier II), and last Big Data mining platforms (Tier III).

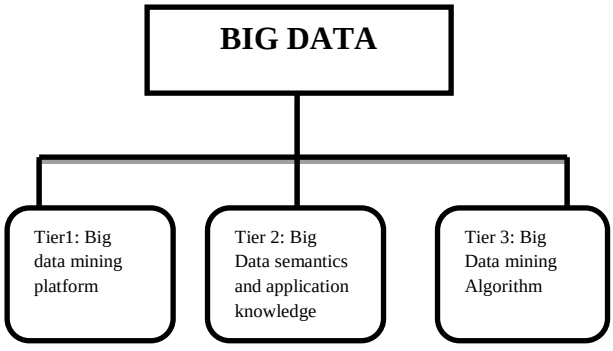

Figure 4 challenges of Big Data

A. Big Data Mining Platform

Big data mining platform basically setup the platform. It basically focuses on arithmetic computing procedure and on data accessing. As data are stored at many location but still data continuously grow, for large-scale data storage an efficacious computing platform distributed into consideration for computing. For small scale data single computer is sufficient which have CPU processors and hard disk. For medium scale data which cannot to be store in main memory, solution is to rely on parallel computing[7] For Big data mining which can't be handle by single computer ,a big data processing framework is deployed which rely on cluster no deny running some programming tools, such as Map Reduce or ECL. This software focuses on to make sure that each task is split into small task which is running on multiple nodes. For example the most powerful super computer Titan which is designed by ORNL in Tennessee, have 16-core CPU each having 18,688 nodes.

B. Big Data Semantics and Application Knowledge

In Big Data semantics and application knowledge refer to large number of aspects related with polices, domain information, regulations, and user knowledge.

In this there are most two important issues

1) Data sharing and privacy

The main goal of systems involving multiples parties is information sharing [8]. It is very clear for data sharing motivation; Big Data are related to the medical records, bills, banking transaction which does not have privacy issue [9], [10], [11]. To protect data privacy two methods are 1) Adding certification and access control to restrict data access so that important information cannot be access by the other user, and 2) anonymize data fields so that for individual records, important information cannot be pinpointed.

For the first method the main challenge is to design access control or certification, so that sensitive information cannot be shared to unauthorized users. For anonym zing data its main focus is to inject randomness in the data.

2) Domain and Application Knowledge

For designing Big Data mining platform, domain and application knowledge provide important essential information. In simple cases, for modeling underlying data domain knowledge help in identifying right features. For achieving business objectives domain and application knowledge helps to design by using Big Data analytics techniques. For example, in every second stock market data generates a large amount of information constantly regarding

buys, puts and, bids. The stock market is continuously compacted by many factors like domestic and international news, natural disasters and so on. A wooing task is to design a system which predicts the stock market movement in the next minutes [12]. It is very challenging to design without correct domain and application knowledge to find useful measure to distinguish the market.

C. Big Data Mining Algorithm

1) Local learning and model fusion for multiple information sources

As big data application has autonomous sources and decentralized control aggregating for a centralized site for mining is strictly prohibitive due to transmission and privacy issues. Although we can always carry out inning activities at each distributed site, biased decision or models are always leads because of the bias view of the data collected just like in case elephant and blind people.

2) Mining from Sparse, Uncertain, and Incomplete Data

The defining features of Big Data applications are scattered, uncertain, and incomplete data. Being sparse, the data points numbers is too less for concluding reliable conclusions. This is normally an issue of the data dimensionality issues, where data is more than 1,000 dimensions; do not show clear trends or distributions. High-dimensional spare data significantly deteriorate the reliability of the models derived from the data in most machine learning and data mining algorithms.

Uncertain data or veracity is the main characteristics of Big Data. In computer science, uncertain data are the type of data reality that makes data deviate from the correct or original values. For example, in an enterprise dataset the address of a customer may be uncertain, or in the sensor the temperature reading is uncertain due to aging of the sensor.

Incomplete data means the missing of the data values for some samples which can be caused by malfunction of sensor nodes, or some policies to skip some values. Data Imputation seeks to impute missing data values to produced efficient models.

3) Mining Complex and Dynamic Data

The data is rapidly increasing in the big data which are complex and changes in volumes and nature. Document posted on social media is also complex. While complex dependency structures underneath the data raise the difficulty for our learning systems. For Big Data application the main challenge is to make use of complex data because in complex data two parties are potentially interested to each other with a social connection. Facebook is large scale networking site where already the number of users reached 1 billion, and the main challenge of Big Data for Data mining task is to analyze these networks. In daily action and interaction if we take into consideration, the scale of difficulty is even more astonishing. Big Data complexity is shown by complex relationship network in data, complex intrinsic semantic association in data, complex heterogeneous data types.

VII. LITERATURE SURVEY ON BIG DATA MINING

The major task of Big Data Mining is to extract useful information from large dataset or we can say Big Data. Day by day the data is rapidly increasing which makes it very difficult to handle. Now a day is becoming very important part of our economy, business, professional, orignizational life. Basically Big Data is used to handle the data which is beyond the size of ability of database tools or software.

This paper [4] proposes to extract useful data from Big Data. As the data is increasing mostly in social media as Facebook, Twitter, and many other application storing the data which are very big and difficult to handle. This paper proposes the characteristics of Big Data that's called hace theorem, where H stands for heterogeneous, A stands for autonomous, C stands control (decentralized) control, and E stands for evolving relationship. In this it also summarize the major challenges of Big data mining.the major challenges of Big Data mining is accessing the data, security, domain knowledge and platform for mining the Big data.

In paper [3] it includes the information about the Big Data, what exactly Big Data is, and what are Big Data Mining and the comparison between Big Data and Data Mining, it also includes Big Data with Data Mining and their issues and some of the related works. In this it describe some of the key features of Big Data, as if the data is huge, time to time data is changing, it's very complex, the data have no control. Challenges in Big Data Mining is Mining platform, privacy, and designing of mining algorithm.

Since the data is very huge and from different sources in nature so it's been very difficult to handle. For storing data we required data to be stored in main memory and for Big Data we need a large main memory so that Big Data can be stored. One of the main issue is privacy issue. To mine the data parallel computing based algorithm is used such as map reduce where the large data is divided into small

dataset and after that mining algorithm is performed to them.

In paper [6] it discusses about the types of data and the sources, as describe earlier data is of two types , structured data and unstructured data. It also focus on the 3 V's of Big data i.e. volume, variety, velocity. Rather it will now 5 V's which is value and variability. Some of the data mining techniques are used for Big data as follow

Classification is the process that is used to classify the data according to different instances. Some of classification algorithm that are used for data mining are apriori, decision tree, KNN classifier, naïve bayes.

Estimation is the process which deals with continuously data.it will estimates to the next outcomes such as income, it estimates what will be the estimate income in the coming days and so on.

Predication is used to predict for something or can say to predict for something that will happen in the future. Here always some outcome is expected.

Association *rules* is a process that discovers the probability of the occurrences of the items in the group.

Clustering is process to group or make a class of similar objects. The data in clustering is considered as in 1 group.

VII. CONCLUSIONS

Big Data refer to the large number of complex datasets which are continuously increasing day by day. Data mining is the techniques used for extracting useful information from these large datasets. It becoming more challenging day by day to handle and mining the data. These two terms not different it works together to extract some useful information. In the coming days Big Data becoming an excellent opportunity. There are some challenges and issues in big data which can be required to overcome in the future.

REFERENCES

[1] "Twitter Blog, Dispatch from the Denver Debate,"http://blog.twitter.com/2012/10/dispatch-from-denver-debate.html, Oct. 2012.

[2] F. Michel, "How Many Photos Are Uploaded to Flickr Every Day and Month?" http://www.flickr.com/photos/franckmichel/6855169886/, 2012.

[3]. Rohit Pitre, Vijay Kolekar, "A Survey Paper on Data Mining With Big Data", IJIRAE, Volume 1, Issue 1, April 2014

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

[5] http://www.dummies.com/careers/find-a-job/the-4-vs-of-big-data/

[6]https://www.ijarcsse.com/docs/papers/Volume_4/5_May2014/V4I5-0328.pdf

[7] J. Shafer, R. Agrawal, and M. Mehta, "SPRINT: A Scalable Parallel Classifier for Data Mining," Proc. 22nd VLDB Conf., 1996.

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

[9] G. Duncan, "Privacy by Design," Science, vol. 317, pp. 1178-1179, 2007.

[10] B. Huberman, "Sociology of Science: Big Data Deserve a Bigger Audience," Nature, vol. 482, p. 308, 2012.

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

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

[13] https://en.wikipedia.org/wiki/Uncertain_data.