ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 4, April 2015

Study on Big data with Data Mining

Shobana.V[1], Maheshwari.S[2], Savithri.M[3]

Assistant Professor, Department of Computer Science, Dr.N.G.P Arts and Science College, Coimbatore, India[1,2,3]

Abstract: We are in a world of increasing data, the day-by-day generation of data becomes very vast. The true of it lies in collecting these data, analyze and perform some computations to obtain some meaningful information. These meaningful information can be derived using some data mining tasks. In short we can call big data as an 'asset' and data mining is a 'handler' that is used to provide beneficial results. To perform these analysis data mining algorithms can be used and also the big data methods.

Keywords: Big data, Data Mining, HACE theorem, structured and unstructured.

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, big data 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.

II. BIG DATA WITH DATA MINING

Generally big data refers to a collection of large volumes of data and these data are generated from various sources such as internet, social media, business organizations etc., With these data some useful information can be extracted with the help of data mining. Data mining is a technique for discovering interesting patterns as well as descriptive, understandable models from large scale data[2].

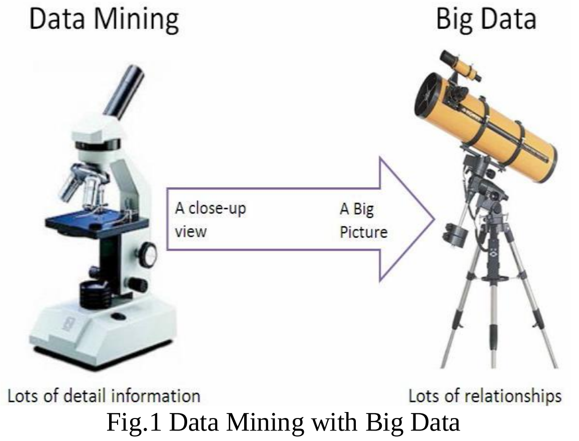

Fig.1 Data Mining with Big Data

The figure 1[3] given above portraits the relation ship of big data with data mining. From the figure it is observed that big data gives lots of relationships and data mining gives lots of information.

III. BIG DATA CHARACTERISTICS -HACE THEOREM

Big Data starts with large volume, **h**eterogeneous **a**utonomous sources with distributed and decentralized control and seeks to explore **c**omplex and **e**volving relationships among data [1].These characteristics makes it an extreme challenge for discovering useful information from big data. In connection with this scenario, let us imagine a scenario where blind people are asked to draw the picture of an elephant. The information collected by each blind people will be such that they may think the trunk as a 'wall', leg as a 'tree', body as a 'wall' and tail as a 'rope'. In this case one blind men can exchange information with other which may be biased.

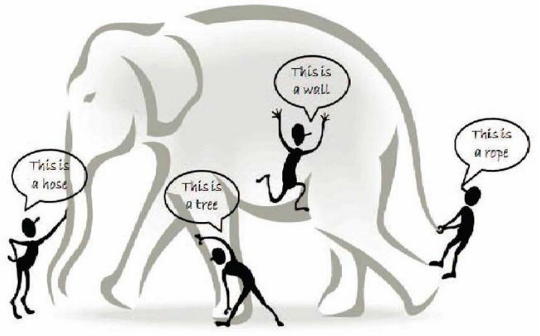

Fig. 2 Blind men and the giant elephant,

i. Vast data with heterogeneous and diverse sources

One of the fundamental characteristics of big data is the large volume of data represented by heterogeneous and diverse dimensions. For example in the biomedical world, a single human being is represented as name, age, gender, family history etc., For X-ray and CT scan images and videos are used. Taking the example heterogeneity refers to the different types of representations of same individual and diverse refers to the variety of features to represent single information [1].

ii. Autonomous with distributed and de-centralized control

These are the main characteristics of big data. Since the sources are autonomous, i.e., automatically generated, it generates information without any centralized control. We can compare it with World Wide Web (WWW) where each server provides a certain amount of information without depending on other servers.

iii. Complex and Evolving relationships
As the size of data becomes infinitely large, the complexity and relationships of data also becomes large. In the early stages when data are so small, there is no difficulty in establishing relationships among data. As the size of data become larger in the current scenario, data are generated from social media and other sources, so there arise complexity in establishing relationships. Such a complication is becoming part of the reality for big data applications, where the key is to take complex data relationships, along with the evolving changes into consideration to discover useful patterns from big data collections [1].

IV. DATA MINING FOR BIG DATA

Generally data mining also known as data or knowledge discovery is the process which analyses data from different perspectives and discover useful information from it. Data mining contains several algorithms which falls into four categories. They are

1) Association Rule.
2) Clustering.
3) Classification.
4) Regression.

(1) Association is used to search relationship between variables. It is applied in searching for frequently visited items. In short it establishes relationship among objects.
(2) Clustering discovers groups and structures in the data. i.e., it classifies the data belongs to which group. Classification deals with associating an unknown structure to a known structure.
(3) Regression finds a function to model the data.

The different data mining algorithms are shown below.

Category	Name of the Algorithm
Association	Apriori, Partition, FP growth, ECLAT
Clustering	K-Means, Expectation Maximization, DB SCAN, fuzzy C Means.
Classification	Decision Trees, C4.5, KNN, Naïve Bayes, SVM.
Regression	Multivariate Linear regression

Table I Classification of Algorithms

These data mining algorithms can be converted into big data map reduce algorithm which is based on parallel computing basis. As data clustering has attracted a significant amount of research attention in past decades, many clustering algorithms has been proposed. However the enlarging data in applications makes clustering of very large scale of data a challenging task. A fast parallel K-means clustering algorithm has been proposed based on Map reduce which has embraced both academia and industry[4].

Differences between big data and data mining.
We cannot say data mining as 'big data' and big data as 'data mining'. There are some differences between these two and they are shown below.

Data Mining	Big Data
Data mining is the old big data	Big data is everything in the world now
Data size is smaller	Data size is Larger
Finding interesting patterns	Involves large scale storage and processing of large data sets.
All data mining tasks are not big data	All big data tasks are data mining

Table II Differences between Big data and Data mining

V. CHALLENGING ISSUES WITH BIG DATA

Big data has been one of the current and future research problem. In the year 2014, Gartner listed 'Top ten Strategic Technologies trends for 2013' and 'Top ten critical Technology Trends for the next five years' and big data is listed in both two.[5].

Challenges in big data are very large. On one hand big data had many opportunities and on the other hand it is facing lot of challenges too. When handling big data challenges occurs in the following areas.
i. Data Capture and Storage.
ii. Data Transmission.
iii. Data Curation.
iv. Data Analysis.
i. Data Visualization.
According to [1] challenges of big data mining is generally divided into three tiers.

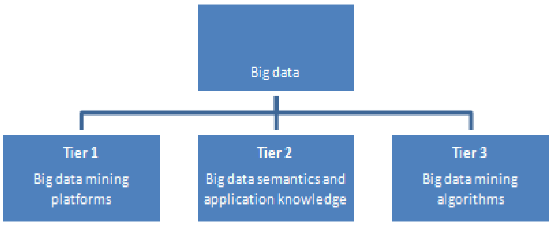
Big data
Tier 1
Big data mining platforms
Tier 2
Big data semantics and application knowledge
Tier 3
Big data mining algorithms

Fig .3. Phases of Big data challenges

The first tier includes the setup of data mining platforms. The second one includes
1. Information sharing and Data privacy.
2. Domain and Application Knowledge. The third one includes Local Learning and model fusion for multiple information sources.
3. Mining from sparse, uncertain and Incomplete data.
4. Mining complex and dynamic data.

Generally mining of data from different data sources is tedious one as the data size is larger. And also big data is stored at different places collecting those data will be a tedious task and applying basic data mining algorithms will be an obstacle for it. The second case is the privacy of data. Since in big data platform the data is processed using parallel computing algorithms such as map reduce framework is applied on those data. After that the data are combined using summation algorithms. In these steps the privacy of data is very much broken and privacy is a question mark in this case. The third case is mining algorithms. Consider the drawing of elephant example here each blind person will predict one result and it does not mean actually what it is. Also when we are applying data mining algorithms to these subsets of data the result may not be that much accurate.

VI. OPEN AREAS FOR FURTHER RESEARCH

- According to [4], we find out the following problems.
- They are
- Communication Overhead increases, So a method to decrease this should be devised.
- Synchronization problems cannot be solved.
- Representation of image processing in an optimal manner.
- Conversion of serial algorithms to Hadoop map reduce algorithms.
- Optimized data partitioning methods.
- Support for block key value update which improves the performance

VII. 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.

REFERENCES

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

[2]. Bharti Thakur, Manish Mann ‘Data Mining for Big Data- A Review’, IJARCSSE, Volume 4, Issue 5, May 2014.

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

[4]. Dr. A.N. Nandhakumar, Nandita Yambem ‘A Survey of Data Mining Algorithms on Apache Hadoop Platforms’, IJETAC, Volume 4, Issue 1, January 2014.

[5]. C.L. Philip Chen, C.-Y. Zhang, “Data-intensive applications, challenges, techniques and technologies: A survey on Big Data”, Inform. Sci. (2014), http://dx.doi.org/10.1016/j.ins.2014.01.015.

DOI 10.17148/IJARCCE.2015.4485