A Study of Data Mining with Big Data

Dr. V.Harsha Shastri[1], V.Sreeprada[2]
[1]Lecturer, Dept. of Computer Science, Loyola Academy Degree and P.G College, Secunderabad, TS, India.
[2]Lecturer, Dept. of Computer Science, St. Mary Centenary Degree College, Secunderabad, TS, India.

Abstract *- Data has become an important part of every economy, industry, organization, business, function and individual. Big Data is a term used to identify large data sets typically whose size is larger than the typical data base. Big data introduces unique computational and statistical challenges. Big Data are at present expanding in most of the domains of engineering and science. Data mining helps to extract useful data from the huge data sets due to its volume, variability and velocity. This article presents a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective.*

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

I. Introduction

Big Data refers to enormous amount of structured data and unstructured data that overflow the organization. If this data is properly used, it can lead to meaningful information. Big data includes a large number of data which requires a lot of processing in real time. It provides a room to discover new values, to understand in-depth knowledge from hidden values and provide a space to manage the data effectively. A database is an organized collection of logically related data which can be easily managed, updated and accessed. Data mining is a process discovering interesting knowledge such as associations, patterns, changes, anomalies and significant structures from large amount of data stored in the databases or other repositories.

Big Data includes 3 V's as its characteristics. They are volume, velocity and variety. Volume means the amount of data generated every second. The data is in state of rest. It is also known for its scale characteristics. Velocity is the speed with which the data is generated. It should have high speed data. The data generated from social media is an example. Variety means different types of data can be taken such as audio, video or documents. It can be numerals, images, time series, arrays etc.

Data Mining analyses the data from different perspectives and summarizing it into useful information that can be used for business solutions and predicting the future trends. Data mining (DM), also called Knowledge Discovery in Databases (KDD) or Knowledge Discovery and Data Mining, is the process of searching large volumes of data automatically for patterns such as association rules [4]. It applies many computational techniques from statistics, information retrieval, machine learning and pattern recognition. Data mining extract only required patterns from the database in a short time span. Based on the type of patterns to be mined, data mining tasks can be classified into summarization, classification, clustering, association and trends analysis [4].

Big Data is expanding in all domains including science and engineering fields including physical, biological and biomedical sciences [1].

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 like internet, social-media, business organization, sensors etc. We can extract some useful information with the help of Data Mining. It is a technique for discovering patterns as well as descriptive, understandable, models from a large scale of data[2].

Volume is the size of the data which is larger than petabytes and terabytes. The scale and rise of size makes it difficult to store and analyse using traditional tools. Big Data should be used to mine large amounts of data within the predefined period of time. Traditional database systems were designed to address small amounts of data which were structured

and consistent, whereas Big Data includes wide variety of data such as geospatial data, audio, video, unstructured text and so on.

Big Data mining refers to the activity of going through big data sets to look for relevant information. To process large volumes of data from different sources quickly, Hadoop is used. Hadoop is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. Its distributed file system supports fast data transfer rates among nodes and allows the system to continue operating uninterrupted at times of node failure. It runs Map Reduce for distributed data processing and is works with structured and unstructured data [6].

III. BIG DATA characteristics- HACE THEOREM.

We have large volume of heterogeneous data. There exists a complex relationship among the data. We need to discover useful information from this voluminous data.

Let us imagine a scenario in which the blind people are asked to draw elephant. The information collected by each blind people may think the trunk as wall, leg as tree, body as wall and tail as rope. The blind men can exchange information with each other.

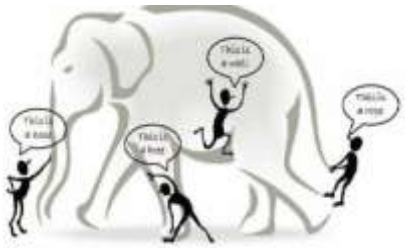

Figure1: Blind men and the giant elephant

Some of the characteristics that include are:

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. 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**: 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 the data becomes infinitely large, the relationship that exists is also large. In early stages, when data is small, there is no complexity in relationships among the data. Data generated from social media and other sources have complex relationships [1].

IV. TOOLS: OPEN SOURCE REVOLUTION

Large companies such as Facebook, Yahoo, Twitter, LinkedIn benefit and contribute work on open source projects. In Big Data Mining, there are many open source initiatives. The most popular of them are:

Apache Mahout [7]: Scalable machine learning and data mining open source software based mainly in Hadoop. It has implementations of a wide range of machine learning and data mining algorithms: clustering, classification, collaborative filtering and frequent patternmining.

R [9]: open source programming language and software environment designed for statistical computing and visualization. R was designed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand beginning in 1993 and is used for statistical analysis of very large data sets.

MOA [8]: Stream data mining open source software to perform data mining in real time. It has implementations of classification, regression; clustering and frequent item set mining and frequent graph mining. It started as a project of the Machine Learning group of University of Waikato, New Zealand, famous for the WEKA software. The streams framework [6] provides an environment for defining and running stream processes using simple

XML based definitions and is able to use MOA, Android and Storm.

SAMOA [11]: It is a new upcoming software project for distributed stream mining that will combine S4 and Storm with MOA.

Vow pal Wabbit [10]: open source project started at Yahoo! Research and continuing at Microsoft Research to design a fast, scalable, useful learning algorithm. VW is able to learn from terafeature datasets. It can exceed the throughput of any single machine networkinterface when doing linear learning, via parallel learning.

V. DATA MINING for BIG DATA

Data mining is the process by which data is analysed coming from different sources discovers useful information. Data Mining contains several algorithms which fall into 4 categories. They are:

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

Association is used to search relationship between variables. It is applied in searching for frequently visited items. In short it establishes relationship among objects. Clustering discovers groups and structures in the data.Classification deals with associating an unknown structure to a known structure. Regression finds a function to model the data

The different data mining algorithms are:

Category	Algorithm
Association	Apriori, FP growth
Clustering	K-Means, Expectation.
Classification	Decision trees,SVM
Regression	Multivariate linear regression

Table 1 Classification of Algorithms

Data Mining algorithms can be converted into big map reduce algorithm based on parallel computing basis.

Differences between Big Data and Data Mining

Big Data	Data Mining
It is everything in the world now.	It is the old Big Data
Size of the data is larger.	Size of the data is smaller
Involves storage and processing of large data sets.	Interesting patterns can be found.
Big Data is the term for large data set.	Data mining refers to the activity of going through big data set to look for relevant information
Big data is the asset	Data mining is the handler which provide beneficial result
Big data" varies depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process andanalyse the data.	Data mining refers to the operation that involve relatively sophisticated search operation

Table 2 Differences between Data Mining and Big Data

VI. Challenges in BIG DATA

Meeting the challenges with BIG Data is difficult. The volume is increasing every day. The velocity is increasing by the internet connected devices. The variety is also expanding and the organizations' capability to capture and process the data is limited.

The following are the challenges in area of Big Data when it is handled:

1. Data capture and storage.
2. Data transmission
3. Data curation
4. Data analysis
5. Data visualization

According to [1], challenges of big data mining are divided into 3 tiers.

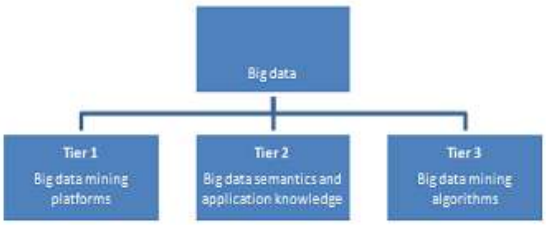

Figure 2: Phases of Big Data Challenges

The first tier is the setup of data mining algorithms. The second tier 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 as size of data is larger. Big data is stored at different places and collecting those data will be a tedious task and applying basic data mining algorithms will be an obstacle for it. Next we need to consider the privacy of data. The third case is miningalgorithms. When we are applyingdata mining algorithms to these subsets of data the resultmay not be that much accurate.

VII. Forecast of the future

There are some challenges that researchers and practitioners will have to deal during the next years:

Analytics Architecture: It is not clear yet how an optimal architecture of analytics systems should be to deal with historic data and with real-time data at the same time. An interesting proposal is the Lambda architecture of Nathan Marz [12]. The Lambda Architecture solves the problem of computing arbitrary functions on arbitrary data in real time by decomposing the problem into three layers: the batch layer, the serving layer, and the speed layer. It combines in the same system Hadoop for the batch layer, and Storm for the speed layer. The properties of the system are: robust and fault tolerant, scalable, general, and extensible, allows ad hoc queries, minimal maintenance, and debuggable.

Statistical significance. It is important to achieve significant statistical results, and not be fooled by randomness. As Efron explains in his book about Large Scale Inference [11], it is easy to go wrong with huge data sets and thousands of questions to answer at once.

Distributedmining. Many data mining techniques are not trivial to paralyze. To have distributed versions ofsome methods, a lot of research is needed with practical and theoretical analysis to provide new methods.

Time evolving data. Data may be evolving over time, so it is important that the Big Data mining techniquesshould be able to adapt and in some cases to detect change first. For example, the data stream mining fieldhas very powerful techniques for this task [13].

Compression: Dealing with Big Data, the quantity of space needed to store it is very relevant. There are twomain approaches: compression where we don't loose anything, or sampling where we choose what is thedata that is more representative. Using compression, we may take more time and less space, so we can consider it as a transformation from time to space. Using sampling, we are loosing information, but the gains inspace may be in orders of magnitude. For example Feldman et al. [14] use coresets to reduce the complexity of Big Data problems. Coresets are small setsthat provably approximate the original data for a given problem. Using merge-reduce the small sets can thenbe used for solving hard machine learning problems in parallel.

Visualization. A main task of Big Data analysis is how to visualize the results. As the data is so big, it is very difficult to find user-friendly visualizations. New techniques, and frameworks to tell and show stories will be needed, as for example the photographs, infographics and essays in the beautiful book "The Human Face of Big Data" [15].

Hidden Big Data. Large quantities of useful data are getting lost since new data is largely untagged file based and unstructured data. The 2012 IDC studyon

Big Data [16] explains that in 2012, 23% (643 exabytes) of the digital universe would be useful for Big Data if tagged and analyzed. However, currently only 3% of the potentially useful data is tagged, and even less is analyzed.

VIII. CONCLUSION

The amounts of data is growing exponentially due to social networking sites, search and retrieval engines, media sharing sites, stock trading sites, news sources and so on. Big Data is becoming the new area for scientific data research and for business applications

Data mining techniques can be applied on big data to acquire some useful information from large datasets. They can be used together to acquire some useful picture from the data.

Big Data analysis tools like Map Reduce over Hadoop and HDFS helps organization.

ACKNOWLEDGMENT

The author would like to thank the Principal for his continuous encouragement in publishing the paper. The author would like to thank Black Buck Research and Development for initiating the research in this area.

REFERENCES

[1] Xindong Wu, Xingquan Zhu, Gong Qing Wu, WeiDing, „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.
[6] Puneet Singh Duggal, Sanchita Paul, (2013), "Big Data Analysis:Challenges and Solutions", Int. Conf. on Cloud, Big Data and Trust, RGPV
[7] Apache Mahout, http://mahout.apache.org.
[8] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer. MOA: Massive Online Analysis http://moa.cms.waikato.ac.nz/. *Journal of Machine Learning Research (JMLR)*, 2010.
[9] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.
[10] J. Langford. Vowpal Wabbit, http://hunch.net/˜vw/,2011.
[11] B. Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Institute of Mathematical Statistics Monographs. Cambridge University Press, 2010
[12] L. Neumeyer, B. Robbins, A. Nair, and A. Kesari.S4: Distributed Stream Computing Platform. In *ICDM Workshops*, pages 170–177, 2010.
[13] J. Gama. *Knowledge Discovery from Data Streams*. Chapman & Hall/Crc Data Mining and Knowledge Discovery. Taylor & Francis Group, 2010.
[14] D. Feldman, M. Schmidt, and C. Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In *SODA*, 2013.
[15] R. Smolan and J. Erwitt. *The Human Face of Big Data*.Sterling Publishing Company Incorporated, 2012.
[16] J. Gantz and D. Reinsel. IDC: The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. December 2012.