International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) ISSN: 2395-5325
Volume 3, Special Issue 3 (September '2017)
Proceeding of - International Conference on Recent Trends in Computational Life Science and Information Technology,
Conference held at Tiruppur Kumaran College for Women,Tirupur,Tamilnadu, India.

A STUDY ON BIG DATA ANALYTICS WITH DATA MINING

R.Narmatha,
M.Phil.Research Scholar (FT),
Department of Research in Computer Science,
Vidyasagar College of Arts and Science,
Udumalpet, Tamilnadu, India.

D.Rajalakshmi ,
Head & Assistant Professor,
Department of Research in Computer Science,
Vidyasagar College of Arts and Science,
Udumalpet, Tamilnadu, India.

Abstract: Big Data is a large-volume, complex, growing data sets with multiple or autonomous sources. This huge amount of data is generated by social media and networks, scientific instruments, mobile devices, sensor technology and networks. These data sets are able to manage, analyze, summarize, visualize and discover knowledge from the collected unstructured data in a timely and scalable manner is very complex task using usual data mining tools. Data Mining is an analytic process with great potential, designed to discover large amounts of data also known as "big data" and explore for consistent patterns and systematic links between variables and then to validate the findings by applying the detected patterns to form new subsets of data. This paper begins with a brief introduction to data mining, followed by the discussions of big data analytics and current status, Controversies and some challenges are also be presented.

Keywords: Data mining, KDD, Big Data Analytics, Challenges of Big Data Mining.

I. INTRODUCTION

Big Data is the term for exceptionally huge large data sets that can be analyzed to find patterns, trends. One technique that can be used for data analysis so that able to help us find abstract patterns in Big Data is Deep Learning. If we apply deep Learning to Big Data, we can find unknown and useful patterns that were impossible so far. With the help of Deep Learning, Artificial Intelligence is getting smart. There is a hypothesis in this regard, the more data, the more abstract knowledge. So a handy survey of Big Data, Deep Learning and its application in Big Data is necessary. Data comes from everywhere, sensors used to gather climate information, posts to social media sites, digital pictures and videos etc. This data is known as big data. Useful data can be extracted from this big data 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. In this paper we overviewed types of big data and challenges in big data for future.

II.DATA MINING

Data Mining is the process of analyzing data from different perspectives and summarizing the results as useful information. It has been defined as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [1].

The process of data mining uses machine learning, statistics, and visualization techniques to discover and present knowledge in a form that is easily comprehensible. The word "Knowledge" in KDD refers to the discovery of patterns which are extracted from the processed data. A pattern is an expression describing facts in a subset of the data.

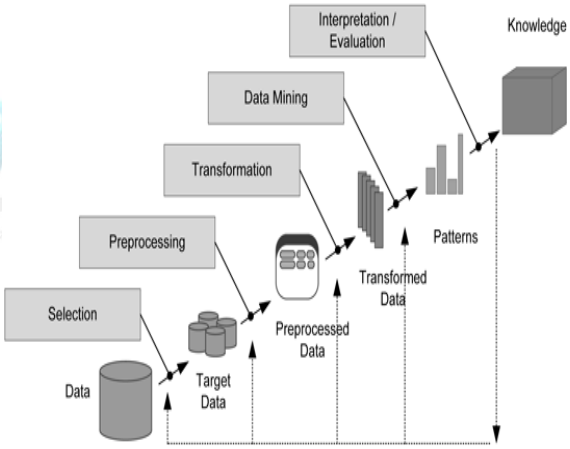

Figure 1: Steps in Data Mining process

Thus, the difference between KDD and data mining is that KDD refers to the overall process of discovering knowledge from data while data mining refers to application of algorithms for extracting patterns from data without the additional steps of the KDD process. However, since Data Mining is a crucial and important part of the KDD process, most researchers use both terms interchangeably. Figure 1 presents the iterative nature of the KDD process [2]. Some of KDD process basic steps as are mentioned as follows

- Providing an understanding of the application domain, the goals of the system and its users, and the relevant prior background and prior knowledge.
- Selecting a data set, or focusing on a subset of variables or data samples, on which discovery is to be performed.
- Preprocessing and data cleaning, removing the noise, collecting the necessary information for modeling,

selecting methods for handling missing data fields, accounting for time sequence information and changes.

- Data reduction and projection, finding appropriate features to represent data, using dimensionality reduction or transformation methods to reduce the number of variables to find invariant representations for data.
- Choosing the data mining task depending on the goal of KDD: clustering, classification, regression, and so forth.
- Selecting methods and algorithms to be used for searching for the patterns in the data mining the knowledge: searching for patterns of interest.
- Evaluating or interpreting the mined patterns, with a possible return to any previous steps.
- Using this knowledge for promoting the performance of the system and resolving any potential conflicts with previously held beliefs or extracted knowledge.

These are the steps that all KDD and data mining tasks development from beginning to end [3].

III. BIG DATA

The term 'Big Data' appeared for first time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title of "Big Data and the NextWave of InfraStress". Big data is a broad term for so large and complex data collections that traditional data processing applications are inadequate. A new field, Predictive Analytics, is trying to extract value from this big data.

Big data is classically described by the first three properties below [4]. Occasionally referred to as the three but organizations require a fourth value to build big data job.

a) Types of Big Data

There are two types of big data: structured and unstructured. *Structured data* are numbers and words that can be easily categorized and analyzed. These data are generated by things like network sensors embedded in electronic devices, smart phones, and global positioning system (GPS) devices. Structured data also include things like sales figures, account balances, and transaction data. *Unstructured data* include more complex information, such as customer reviews from commercial websites, photos and other multimedia, and comments on social networking sites [5]. These data cannot easily be separated into categories or analyzed numerically.

IV.DATA MINING WITH BIG DATA

Data mining contains extracting and analyzing huge amounts of data to discover models for big data. The methods came out of the basis of artificial intelligence and statistics with a bit of database management. Searching information from data takes two major forms: prediction and description. Data mining is used to summarize and shorten the data in a method that we can distinguish and then allow us to gather things about specific cases based on the patterns [6].

The objective of the data mining as a term used for the specific classes of five activities as follows,

Volume: Massive information sets that are command of size bigger than data managed in habitual storage and analytical results.

Variety: Complex, Variable and Heterogeneous data, which are generated in formats as dissimilar as public media, e-mail, images ,video, blogs, and sensor data as well as shadow data such as access journals and Web explore histories.

Velocity: Data is generated as a stable with real-time queries for significant information to be present up on claim instead of batched.

Value: Consequential insights that transport predictive analytics for upcoming trends and patterns from bottomless, difficult analysis based on graph algorithms, machine learning and statistical modeling.

These analytics overtake the results of usual querying, reporting and business intelligence [4].

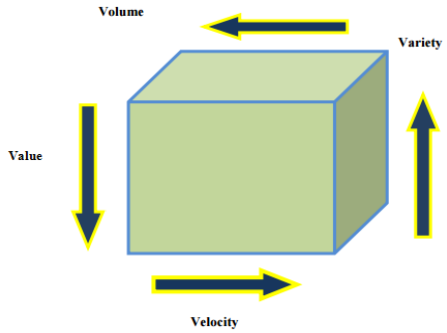

Figure 2:Big Data Management System

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

1. Classification

Classification is a process of generalizing the data according to different instances. Several major kinds of classification algorithms in data mining are Decision tree, k-nearest neighbor classifier, Naive Bayes, Apriori and AdaBoost. Classification consists of examining the features of a newly presented object and assigning to it a predefined class. The classification task is characterized by the well-defined classes, and a training set consisting of reclassified examples [7].

2. Estimation

Estimation deals with continuously valued outcomes. Given some input data, we use estimation to come up with a value for some unknown continuous variables such as income, height or credit card balance.

3. Prediction

It is a statement about the way things will happen in the future , often but not always based on experience or knowledge. Prediction may be a statement in which some outcome is expected [8].

4. Association Rules

An association rule is a rule which implies certain association relationships among a set of objects in a database.

5. Clustering

Clustering can be considered the most important unsupervised learning problem. So as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data [9].

Big data is the term for a collection of complex data sets, Data mining is an analytic process designed to explore data in search of consistent patterns and then to validate the findings by applying the detected patterns to new subsets of data. To support Big data mining, high performance computing platforms are required which imposed by combination of above mentioned data mining techniques and create the systematic designs to unleash the full power of the Big Data. The following table shows the difference between big data and data mining [10].

S.NO	*Big data*	*Data mining*
1	Big data is a term for large data set.	Data mining refers to the activity of going through big data set to look for relevant information
2	Big data is the asset	Data mining is the handler which provide beneficial result.
3	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 and analyze the data.	Data mining refers to the operation that involve relatively sophisticated search operation

Table I. Difference between Big data and Data mining

V.CONTROVERSIES ABOUT BIG DATA MINING

As Big Data is a new advanced topic, there have been a lot of Controversies about it, for example as follows:

- There is no need to distinguish Big Data analytics from data analytics, as data will continue growing, and it will never be small again [11].
- Big Data may be a hype to sell Hadoop based computing systems. Hadoop is not always the best tool [12]. It seems that data management system sellers try to sell systems based in Hadoop, and MapReduce may be not always the best programming platform, for example for medium-size companies.
- In real time analytics, data may be changing. In that case, what it is important is not the size of the data it is its quality.
- Claims to accuracy are misleading. As Taleb explains in his new book [13], when the number of variables grow, the number of fake correlations also grow.
- Bigger data are not always better data. It depends if the data is noisy or not, and if it is representative of what we are looking for. For example, Twitter users are assumed to be a representative of the global population, when this is not always the case.
- Ethical concerns about accessibility. The main issue is if it is ethical that people can be analyzed without knowing it.

Limited access to Big Data creates new digital divides. There may be a digital divide between people or organizations being able to analyze Big Data or not. Also organizations with access to Big Data will be able to extract knowledge that without this Big Data is not possible to get. We may create a division between Big Data rich and poor organizations [14].

VI.CHALLENGES OF BIG DATA MINING

There are many future significant challenges in Big Data mining and analytics, that occur from the nature of data. These are some of the challenges that researchers will have to resolve during the next years.

a) ***Analytics Architecture***

It is not clear yet how an optimal architecture of an 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 [15]. 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 [16]. 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 [17].

b) ***Time evolving data***

Data may be evolving over time, so it is important that the Big Data mining techniques should be able to adapt and in some cases to detect change first for e.g. the data stream mining field has very powerful techniques for this task [16] .

c) ***Distributed mining***

Many data mining techniques are not trivial to paralyze [16]. To have distributed versions of some methods, a lot of research is needed with practical and theoretical analysis to provide new methods to predict future.

d) ***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 it is easy to go wrong with huge data sets and thousands of questions to answer at once [18].

e) ***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 study on Big Data explains that in

2012, 23% (643 hexabytes) 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 [19],[20].

VII.CONCLUSION

Big Data is data whose scale, diversity, and complexity require new architecture, Techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. To hold Big Data mining, high performance computing platforms are indispensable. Data Mining is an analytic process with great potential, designed to explore large amounts of data also known as "big data" and search for consistent patterns and systematic relationships between variables, and then to validate the findings by applying the detected patterns to form new subsets of data. This paper focuses on a keen overview of several insights about the big data mining and the major concern , Controversies and the core challenges for the future.

VII.REFERENCES

[1]. Bharti Thakur, Manish Mann "Data Mining for Big Data: A Review", 2014, IJARCSSE,Page-469, Volume 4, Issue 5, May 2014 ISSN: 2277 128X.
[2]. Frawley, W. J., Piatetsky-Shapiro, G. and Matheus, C. J.: 1992, Knowledge discovery in databases: An overview, AI Magazine 13(3), 57–70.
[3]. Brachman, R. and Anand, T. 1996. The Process of Knowledge Discovery in Databases: A Human Centered Approach. In AKDDM, Cambridge, MA: AAAI/MIT Press.
[4]. Department of Finance and Deregulation Australian Government Big Data Strategy-Issue Paper March 2013.
[5]. Wei Fan and Albert Bifet "Mining Big Data: Current Status and Forecast to the Future",Vol 14,Issue 2,2013.
[6]. Xindong Wu , Gong-Quing Wu and Wei Ding " Data Mining with Big data ", IEEE Transactions on Knoweledge and Data Enginnering Vol 26 No1 Jan 2014.
[7]. M.H. Alam, J.W. Ha, and S.K. Lee, ―Novel Approaches to Crawling Important Pages Early,‖ Knowledge and Information Systems, vol. 33, no. 3, pp 707-734, Dec. 2012.
[8]. S. Aral and D. Walker, ―Identifying Influential and Susceptible Members of Social Networks,‖ Science, vol. 337,pp. 337-341, 2012.
[9]. FUJIMAKI Ryohei, MORINAGA Satoshi :The Most Advanced Data Mining of the Big Data Era
[10]. E. Birney, ―The Making of ENCODE: Lessons for Big-Data Projects,‖ Nature, vol. 489, pp. 49-51, 2012.
[11]. D. Boyd and K. Crawford. Critical Questions for Big Data. Information, Communication and Society,15(5):662–679, 2012.
[12]. D. J. Leinweber. Stupid Data Miner Tricks: Overfitting the S&P 500. The Journal of Investing, 16:15–22, 2007.
[13]. J. Lin. MapReduce is Good Enough? If All You Have is a Hammer, Throw Away Everything That's Not a Nail! CoRR, abs/1209.2191, 2012.
[14]. N. Taleb. Antifragile: How to Live in a World We Don't Understand. Penguin Books, Limited, 2012.
[15]. Mrs. Deepali Kishor Jadhav," Big Data: The New Challenges in Data Mining", IJIRCST, ISSN: 2347-5552, Volume-1, Issue-2, September, 2013.
[16]. http://albertbifet.com/big-data-mining-future-challenges
[17]. http://www.emc.com/about/news/press/2012/20121211-01.htm
[18]. Wei Fan, Albert Bifet," Mining Big Data: Current Status, and Forecast to the Future"
[19]. Bharti Thakur, Manish Mann," Data Mining for Big Data: A Review", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014, ISSN: 2277 128X.
[20]. A. Ghoting and E. Pednault, ―Hadoop-ML: An Infrastructure for the Rapid Implementation of Parallel Reusable Analytics,‖ Proc. Large-Scale Machine Learning: Parallelism and Massive Data Sets Workshop (NIPS '09), 2009.