Data Mining with Big Bang Data

Suma C.[1], Suganya S.[2], Sharmila V.[3]
[1,2]PG Student and [2]Professor
[1,2,3] Computer and Science Engineering, Kathir College Of Engineering, Coimbatore, Tamil Nadu, India

Abstract— Big Data is used to identify the datasets that are due to their large size and complexity. Big Data concerns large-volume, complex, growing data sets with multiple, autonomous sources. Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. Big Data mining is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity, it was not possible before to do it. The Big Data challenge is becoming one of the most exciting opportunities for the next years. This survey paper includes the information about what is big data, a HACE theorem that characterizes the features of the Big Data revolution from the Data mining perspective, Data mining with big data, features and Challenging issues.

Index Terms— Big Data, Data mining, Challenging issues, Datasets, Data Mining Algorithms, HACE Theorem, Map Reduce

1 Introduction

The entire history of humanity has an enormous accumulation of data. Information has been stored and maintained for thousands of years. Data mining is the activity of going through big data sets to look for relevant or pertinent information. This type of activity is really a good example of the old axiom "looking for a needle in a haystack." The idea is that businesses collect massive sets of data that may be homogeneous or automatically collected. Decision-makers need access to smaller, more specific pieces of data from those large sets. They use data mining to uncover the pieces of information that will inform leadership and help chart the course for a business. Big data is the asset and data mining is the "handler" of that is used to provide beneficial results.

2 An Overview

Today large data sources are ubiquitous throughout the world. Anything which is unknown, we Google it and in a fractions of seconds we get the number of links as a result. This is the better example for the processing of Big Data. This Big Data is not any different thing than out regular term data. Data used for processing are obtained from measuring devices, social networks message flows, radio frequency identifiers, meteorological data, remote sensing and location data streams of mobile subscribers, devices, audio and video recordings. So, as Big Data is more and more used all over the world, a new and important research field is being established. The data collection has grown tremendously and is beyond the ability of commonly used software tools to capture, manage, and process within a tolerable time. The mass distribution of the technology and innovative models that utilize these different kinds of devices and services appeared to be a starting point for the penetration of Big Data in almost all areas of human activity, including the commercial sector and public administration. Nowadays, Big Data and the continuing dramatic increase in human and machine generated data associated with it are quite evident.

Big Data is a technology to process high-volume, high-velocity, high-variety data or data-sets to extract intended data value and ensure high veracity of original data and obtained information that demand cost-effective, innovative forms of data and information processing (analytics) for enhanced insight, decision making, and processes control. Big Data Mining is a close up view, with lots of detail information of a Big Data with lots of information.

Data mining involves exploring and analyzing large volume of data to find patterns for big data. The techniques arisen from the fields of statistics and artificial intelligence (AI), with a few database management thrown into the mix.

Generally, the goal of the data mining is either classification or prediction. In classification, the ideology is to sort data into groups. For example, a marketer may be interested in the characteristics of those who responded versus who didn't respond to a promotion.

These are two classes. The idea of prediction is to predict the value of a continuous variable. For example, a marketer may be interested in predicting those who will respond to a promotion.
Typical algorithms used in data mining include the following:

1. Classification trees: A popular data-mining technique that is used to classify a dependent category of variable based on measurements of one or more predictor variables. The result is a tree with nodes and links. It is between the nodes that form if-then rules.

2. Logistic regression: A statistical technique is a variant of standard regression and extends the concept to deal with classification which produces a formula that predicts the

probability of the occurrence as a function of the independent variables.

3. Neural networks: A software algorithm that is modelled after the parallel architecture of animal brains. The network consists of input and output nodes and hidden layers. Each unit is assigned a weight. Data is given in input node, by a system of trial and error; the algorithm adjusts the weight until it meets a certain stopping criteria. Some people have likened this to a black–box approach.

4. Clustering techniques like K-nearest neighbours: A technique that identifies groups of similar records. The K-nearest neighbour technique calculates the distances between the record and also points in the historical training data. Then it assigns this record to the class of its nearest neighbour in a data set.

Here's a classification tree example. Let us consider the situation where a telephone service company wants to determine which residential customers are likely to disconnect their service.

The telephone service company has information consisting of the following attributes: how long the person had the service, how much he spends on the service, whether the service has been problematic, whether he has the best calling plan which he needs and desires, where he resides, how young he is, whether he has other services bundled together in his place, competitive information concerning other carriers plans, and whether he still has the service or not.

Of course, you can find many more attributes than this. The last attribute is the outcome of the variable. This is what the software will use to classify the customers into one of the two groups — perhaps called stayers and flight risks.

The dataset is broken into training data and a test dataset. The training data consists of observations (called attributes) and an outcome variable (binary in the case of a classification model) — in this case, the stayers or the flight risks.

The algorithm is run over the training data and comes up with a tree that can be read like a series of rules. For example, if the customers have been with the company for more than ten years and they are over 55 years old, they are likely to remain as loyal customers.

These rules are then run over the test data set to determine how good this model is on "new data." Accuracy measures are provided for the model. For example, a popular technique here is the confusion matrix. This matrix is a table which provides information about how many cases were correctly versus incorrectly classified.

If the model looks good, it can be deployed on other data when it is available (that is, using it to predict new cases of flight risk). Based on the model the company may decide. For example, to send the special offers to those customers whom it thinks are flight risks.

3 FEATURES

The features of Big Data are:

1. It is huge in size.
2. The data keep on changing time to time.
3. Its data sources are from different phases.
4. It is free from the influence, guidance, or control of anyone.
5. It is too much complex in nature, thus hard to handle.

It's huge in nature because, there is the collection of data from various sources together. If we consider the example of Facebook, lots of numbers of people are uploading their data in various types such as text, images or videos. The people also keep their data changing continuously. This tremendous and instantaneously, time to time changing stock of the data is stored in a warehouse. This large storage of data requires large area for actual implementation. As the size is too large, no one is capable to control it oneself. The Big Data needs to be controlled by dividing it in groups. Due to largeness in size, decentralized control and different data sources with different types the Big Data becomes much complex and harder to handle. We cannot manage them with the local tools those we use for managing the regular data in real time. For major Big Data-related applications, such as Google, Flicker, Facebook, a large number of server farms are deployed all over the world to ensure nonstop services and quick responses for local markets.

4 CHARACTERISTICS

HACE Theorem: Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data.

These characteristics make it an extreme challenge for discovering useful knowledge from the Big Data. Exploring the Big Data in this scenario is equivalent to aggregating heterogeneous information from different sources (blind men) to help draw a best possible picture to reveal the genuine gesture of the elephant in a real-time fashion. Indeed, this task is not as simple as asking each blind man to describe his feelings about the elephant and then getting an expert to draw one single picture with a combined view, concerning that each individual may speak a different language (heterogeneous and diverse information sources) and they may even have privacy concerns about the messages they deliberate in the information exchange process.

4.1 Huge Data with Heterogeneous and Diverse Dimensionality

One of the fundamental characteristics of the Big Data is the huge volume of data represented by heterogeneous and diverse dimensionalities. This is because different information collectors use their own schemata for data recording, and the nature of different applications also results in diverse representations of the data. For example, each single human being

in a bio-medical world can be represented by using simple demographic information such as gender, age, family disease history etc. For X-ray examination and CT scan of each individual, images or videos are used to represent the results because they provide visual information for doctors to carry detailed examinations. For a DNA or genomic related test, microarray expression images and sequences are used to represent the genetic code information because this is the way that our current techniques acquire the data. Under such circumstances, the heterogeneous features refer to the different types of representations for the same individuals, and the diverse features refer to the variety of the features involved to represent each single observation. Imagine that different organizations (or health practitioners) may have their own schemata to represent each patient, the data heterogeneity and diverse dimensionality issues become major challenges if we are trying to enable data aggregation by combining data from all sources.

4.2. Autonomous Sources with Distributed and Decentralized Control

Autonomous data sources with distributed and decentralized controls are a main characteristic of Big Data applications. Being autonomous, each data sources is able to generate and collect information without involving (or relying on) any centralized control. This is similar to the World Wide Web (WWW) setting where each web server provides a certain amount of information and each server is able to fully function without necessarily relying on other servers. On the other hand, the enormous volumes of the data also make an application vulnerable to attacks or malfunctions, if the whole system has to rely on any centralized control unit. For major Big Data related applications, such as Google, Flicker, Facebook, and Walmart, a large number of server farms are deployed all over the world to ensure nonstop services and quick responses for local markets. Such autonomous sources are not only the solutions of the technical designs, but also the results of the legislation and the regulation rules in different countries/regions. For example, Asian markets of Walmart are inherently different from its North American markets in terms of seasonal promotions, top sell items, and customer behaviours. More specifically, the local government regulations also impact on the wholesale management process and eventually result in data representations and data warehouses for local markets.

4.3. Complex and Evolving Relationships

While the volume of the Big Data increases, so do the complexity and the relationships underneath the data. In an early stage of data centralized information systems, the focus is on finding best feature values to represent each observation. This is similar to using a number of data fields, such as age, gender, income, education background etc., to characterize each individual. This type of sample-feature representation inherently treats each individual as an independent entity without considering their social connections which is one of the most important factors of the human society. People form friend circles based on their common hobbies or connections by biological relationships. Such social connections commonly exist in not only our daily activities, but also are very popular in virtual worlds. For example, major social network sites, such as Facebook or Twitter, are mainly characterized by social functions such as friend-connections and followers (in Twitter). The correlations between individuals inherently complicate the whole data representation and any reasoning process. In the sample-feature representation, individuals are regarded similar if they share similar feature values, whereas in the sample-feature-relationship representation, two individuals can be linked together (through their social connections) even though they might share nothing in common in the feature domains at all. In a dynamic world, the features used to represent the individuals and the social ties used to represent our connections may also evolve with respect to temporal, spatial, and other factors. Such a complication is becoming part of the reality for Big Data applications, where the key is to take the complex (non-linear, many-to-many) data relationships, along with the evolving changes, into consideration, to discover useful patterns from Big Data collections.

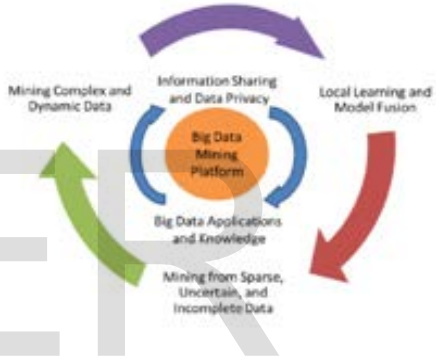

Fig. 1. A Big Data processing framework

The research challenges form a three tier structure and center around the "Big Data mining platform" (Tier I), which focuses on low-level data accessing and computing. Challenges on information sharing and privacy, and Big Data application domains and knowledge form Tier II, which concentrates on high level semantics, application domain knowledge, and user privacy issues. The outmost circle shows Tier III challenges on actual mining algorithms.

5 Big Data Challenges

Big Data has some inherent challenges and problems that can be primarily divided into three groups:

1. Data,
2. Processing and
3. Management challenges

(see Fig. 2). While dealing with large amounts of information we face the following challenges.

The fig. 2. illustrates three main categories of Big Data challenges that are associated with data, its management and processing issues such challenges as volume, variety, velocity and veracity that are also known as 5V of Big Data.

Fig. 2. Big Data Challenges

As those Big Data characteristics are well examined in scientific literature we will only discuss them briefly. Volume refers to the large amount of data, especially, machine-generated. This characteristic defines a size of the data set that makes its storage and analysis problematic utilizing conventional database technology. Variety is related to different types and forms of data sources: structured (e.g. financial data) and unstructured (social media conversations, photos, videos, voice recordings and others).

Multiplicity of the various data results in the issue of its handling. Velocity refers to the speed of new data generation and distribution. This characteristic requires the implementation of real-time processing for the streaming data analysis (e.g. on social media, different types of transactions or trading systems, etc.). Veracity refers to the complexity of data which may lead to a lack of quality and accuracy. This characteristic reveals several challenges: uncertainty, imprecision, missing values, misstatement and data availability. There is also a challenge regarding data discovery that is related to the search of high quality data in data sets.

The second branch of Big Data challenges is called processing challenges. It includes data collection, resolving similarities found in different sources, modification data to a type acceptable for the analysis, the analysis itself and output representation, i.e. the results visualization in a form most suitable for human perception.

The last type of challenge offered by this classification is related to data management. Management challenges usually refer to secured data storage, its processing and collection. Here the main focuses of study are: data privacy, its security, governance and ethical issues. Most of them are controlled based on policies and rules provided by information security institutes on state or international levels.

Over past generations, the results of analyzed data were represented as visualized plots and graphs. It is evident that collections of complex figures are sometimes hard to perceive, even by well-trained minds. Nowadays, the main factors causing difficulties in data visualization continue to be the limitations of human perception and new issues related to display sizes and resolutions. This question is studied in detail further in the section "Integration with Augmented and Virtual Reality". Preparatory to the visualization, the main interaction problem is in the extraction of the useful portion of information from massive volumes. Extracted data is not always accurate and mostly overloaded with excrescent information. Visualization technique is useful for simplifying information and transforming it into a more accessible form for human perception. In the near future, petascale data may cause analysis failures because of traditional approaches in usage, i.e. when the data is stored on a memory disk continuously waiting for further analysis. Hence, the conservative approach of data compressing may become ineffective in visualization methods. To solve this issue, developers should create a flexible tool for the practice of data collection and analysis. Increases in data size make the multilevel hierarchy approach incapable in data scalability. Hierarchy becomes complex and intensive, making navigation difficult for user perception. In this case, a combination of analytics and Data Visualization may enable more accessible data exploration and interaction, which would allow improving insights, outcomes and decision-making.

Contemporary methods, techniques and tools for data analysis are still not flexible enough to discover valuable information in the most efficient way. The question of data perception and presentation remains open. Scientists face the task of uniting the abstract world of data and the physical world through visual representation. Meanwhile, visualization-based tools should fulfill three requirements: expressiveness (demonstrate exactly the information contained in the data), effectiveness (related to cognitive capabilities of human visual system) and appropriateness (cost-value ratio for visualization benefit assessment). Experience of previously used techniques can be repurposed to achieve more beneficial and novel goals in Big Data perception and representation.

6 Related Work

On the level of mining platform sector, at present, parallel programming models like MapReduce are being used for the purpose of analysis and mining of data. MapReduce is a batch-oriented parallel computing model. There is still a certain gap in performance with relational databases. Improving the performance of MapReduce and enhancing the real-time nature of large-scale data processing have received a significant amount of attention, with MapReduce parallel programming being applied to many machine learning and data mining algorithms. Data mining algorithms usually need to scan through the training data for obtaining the statistics to solve or optimize model.

For those people, who intend to hire a third party such as auditors to process their data, it is very important to have efficient and effective access to the data. In such cases, the privacy restrictions of user may be faces like no local copies or downloading allowed, etc. So there is privacy-preserving public

auditing mechanism proposed for large scale data storage. This public key-based mechanism is used to enable third-party auditing, so users can safely allow a third party to analyze their data without breaching the security settings or compromising the data privacy. In case of design of data mining algorithms, Knowledge evolution is a common phenomenon in real world systems. But as the problem statement differs, accordingly the knowledge will differ. For example, when we go to the doctor for the treatment, that doctor's treatment program continuously adjusts with the conditions of the patient. Similarly the knowledge. For this, Wu proposed and established the theory of local pattern analysis, which has laid a foundation for global knowledge discovery in multisource data mining. This theory provides a solution not only for the problem of full search, but also for finding global models that traditional mining methods cannot find.

7 Future Broadcast

There are many future important challenges in Big Data management and analytics, that arise from the nature of data: large, diverse, and evolving. These are some of the challenges that researchers and practitioners will have to deal during the next years:

1. Analytics Architecture: It is not clear yet how an optimal architecture of an analytics system 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.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.

2. Statistical significance: It is important to achieve significant statistical results, and not be fooled by randomness. AsEfron 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.

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

8 Conclusion

Big Data is going to continue growing during the next years, and each data scientist will have to manage much more amount of data every year. This data is going to be more diverse, larger, and faster. We discussed some insights about the topic, and what we consider are the main concerns and the main challenges for the future. Big Data is becoming the new Final Frontier for scientific data research and for business applications. We discussed HACE theorem suggests that the characteristics of the Big Data are

1. Huge with heterogeneous and diverse data sources,
2. Autonomous with distributed and decentralized control, and
3. Complex and evolving in data and knowledge associations.

Such combined characteristics suggest that Big Data requires a "big mind" to consolidate data for maximum values.

Hence we regard Big Data as an emerging trend and the need for Big Data mining is arising in all science and engineering domains. With Big Data technologies, we will hopefully be able to provide most relevant and most accurate social sensing feedback to better understand our society at real-time. We can further stimulate the participation of the public audiences in the data production circle for societal and economical events. We are at the beginning of a new era where Big Data mining will help us to discover knowledge that no one has discovered before. The era of Big Data has arrived.

Acknowledgment

First and foremost, We would like to thank our Teacher, Asst. Prof. Ms. Sharmila V. M.E., for her guidance and support. We will forever remain grateful for the constant support and guidance by her. Through our many discussions, she helped us to form and solidify ideas. The invaluable discussions we had with her, the penetrating questions she has put to us and the constant motivation, has all led to the development of this Survey. I would also like to thank to my friends for providing feedback and suggestions for improving my ideas.

References

[1] C. Wang, S.S.M. Chow, Q. Wang, K. Ren, and W. Lou, "Privacy- Preserving Public Auditing for Secure Cloud Storage" IEEE Trans. Computers, vol. 62, no. 2, pp. 362-375, Feb. 2013.

[2] X. Wu and S. Zhang, "Synthesizing High-Frequency Rules from Different Data Sources," IEEE Trans. Knowledge and Data Eng., vol. 15, no. 2, pp. 353-367, Mar./Apr. 2003.

[3] X. Wu, C. Zhang, and S. Zhang, "Database Classification for Multi-Database Mining," Information Systems, vol. 30, no. 1, pp. 71- 88, 2005

[4] K. Su, H. Huang, X. Wu, and S. Zhang, "A Logical Framework for Identifying Quality Knowledge from Different Data Sources," Decision Support Systems, vol. 42, no. 3, pp. 1673-1683, 2006.

[5] E.Y. Chang, H. Bai, and K. Zhu, "Parallel Algorithms for Mining Large-Scale Rich-Media Data," Proc. 17th ACM Int'l Conf. Multimedia, (MM '09,) pp. 917-918, 2009.

[6] D. Howe et al., "Big Data: The Future of Biocuration," Nature, vol. 455, pp. 47-50, Sept. 2008.
[7] A. Labrinidis and H. Jagadish, "Challenges and Opportunities with Big Data," Proc. VLDB Endowment, vol. 5, no. 12, 2032-2033, 2012.
[8] Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining," J. Cryptology, vol. 15, no. 3, pp. 177-206, 2002.
[9] Alex Berson and Stephen J.Smith Data Warehousing,Data Mining and OLAP edition 2010.
[10] Department of Finance and Deregulation Australian Government Big Data Strategy-Issue Paper March 2013
[11] NASSCOM Big Data Report 2012
[12] Wei Fan and Albert Bifet " Mining Big Data:Current Status and Forecast to the Future",Vol 14,Issue 2,2013
[13] Algorithm and approaches to handle large Data-A Survey,IJCSN Vol 2,Issue 3,2013
[14] 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
[15] Xu Y etal, balancing reducer workload for skewed data using sampling based partioning 2013.
[16] X. Niuniu and L. Yuxun, "Review of Decision Trees," IEEE, 2010 .
[17] Decision Trees for Business Intelligence and Data Mining: Using SAS Enterprise Miner "Decision Trees-What Are They?"
[18] Weiss, S.H. and Indurkhya, N. (1998), Predictive Data Mining: A Practical Guide, Morgan Kaufmann Publishers, San Francisco, CA.