© 2017 IJSRSET | Volume 3 | Issue 6 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

A Survey of Big Data using Data Mining Algorithms

Megha K. Patel
Information and Technology Department, L.D. College of Engineering, Gujarat Technological University, Ahmedabad, Gujarat, India

ABSTRACT

Big Data concerns large-volume, complex, growing data sets with multiple, autonomous sources. With the fast develop-ment of networking, data storage, and the data collection capacity, Big Data is now rapidly expanding in all science and engi-neering domains, including physical, biolog-ical and bio-medical sciences. 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. This data-driven model involves demand-driven aggre-gation of information sources, mining and analysis, user interest modeling, and secu-rity and privacy considerations. We ana-lyze the challenging issues in the data-driven model and also in the Big Data revolution.

Keywords : Big data, Data Mining Algo-rithms

I. INTRODUCTION

Big Data concerns large-volume, complex, grow-ing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data is now rapidly expanding in all science and engi-neering domains, including physical, biological and bio-medical sciences. This article presents a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data process-ing model, from the data mining perspective. This data-driven model involves demand-driven aggre-gation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. We analyze the challenging issues in the data-driven model and also in the Big Data revolution.

II. METHODS AND MATERIAL

Big Data Characteristics

HACE Theorem

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 chal-lenge for discovering useful knowledge from the Big Data. In a nave sense, we can imagine that a num-ber of blind men are trying to size up a giant ele-phant (see Figure 1), which will be the Big Data in this context. The goal of each blind man is to draw a picture (or conclusion) of the elephant according to the part of information he collected during the process. Because each persons view is limited to his local region, it is not surprising that the blind men will each conclude independently that the elephant feels like a rope, a hose, or a wall, depending on the region each of them is limited to. To make the problem even more complicated, lets assume that the elephant is growing rapidly and its pose also changes constantly, and (b) the blind men also learn from each other while exchanging information on

Figure 1

Figure 1: The blind men and the giant elephant: the localized (limited) view of each blind man leads to a biased conclusion.

Their respective feelings on the elephant.Exploring the Big Data in this scenario is equivalent to ag-gregating heterogeneous information from di erent sources (blind men) to help draw a best possible picture to reveal the genuine gesture of the ele-phant 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 com-bined view, concerning that each individual may speak a di erent language (heterogeneous and di-verse information sources) and they may even have privacy concerns about the messages they deliber-ate in the information exchange process. The size of data which can be considered to be big data is constantly varying factor and newer tools are con-tinuously being developed to handle this big data. Big data can be described by following characteris-tics,

Volume: The quantity of data generated is very important in this context. it is the size of data which determines value and potential of data un-der consideration and whether it is can actually considered as big data or not.

Variety: The next aspect of big data is variety. means that the category to which big data belongs to is also a very essential fact that need to be known by data analyst. this information is very important for people, those will analyze this data. Through this they will understand importance of big data.

Velocity: Velocity refers to speed of generation of data and how fast data is generated and process to meet the demands of business.

Variability: Because of this characteristics , anal-ysis of data is very di cult. this refers to inconsis-tency can be present in data at time. due to this the process required for analysis should manage and handle big data e ectively.

Complexity: The data collected from di erent sources, this data needs to be linked, collected and correlated with each other so through this data we will able to derive information that data wants to represent.

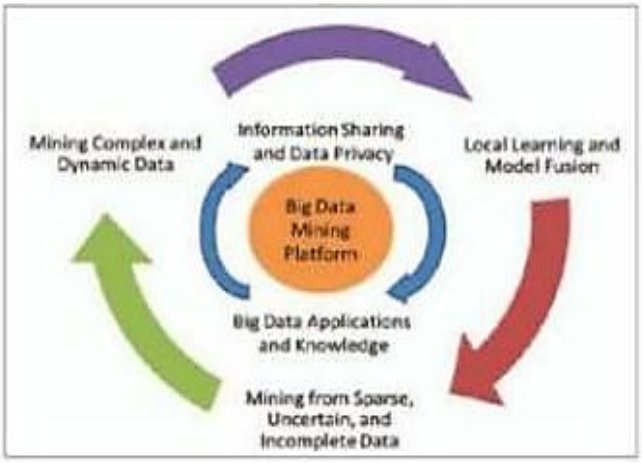

Figure 2: A Big Data processing framework

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.

2.1 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 di erent information collectors use their own schemata for data recording, and the nature of di erent applications also results in diverse rep-resentations of the data. For example, each sin-gle human being in a bio-medical world can be represented by using simple demographic informa-tion such as gender, age, family disease history etc. For X-ray examination and CT scan of each in-dividual, 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 expres-sion 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 fea-tures refer to the di erent types of representations for the same individuals, and the diverse features refer to the variety of the features involved to rep-resent each single observation. Imagine that dif-ferent 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.

2.2 Autonomous Sources with Dis-tributed and Decentralized Con-trol

Autonomous data sources with distributed and de-centralized controls are a main characteristic of Big Data applications. Being autonomous, each data sources is able to generate and collect informa-tion without involving (or relying on) any central-ized control. This is similar to the World Wide Web (WWW) setting where each web server pro-vides a certain amount of information and each server is able to fully function without necessar-ily 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 appli-cations, such as Google, Flicker, Facebook, and Walmart, a large number of server farms are de-ployed all over the world to ensure nonstop services and quick responses for local markets. Such au-tonomous sources are not only the solutions of the technical designs, but also the results of the leg-islation and the regulation rules in di erent coun-tries/regions. For example, Asian markets of Wal-mart are inherently di erent from its North Amer-ican markets in terms of seasonal promotions, top sell items, and customer behaviors. More speci - cally, the local government regulations also impact on the wholesale management process and eventu-ally result in data representations and data ware-houses for local markets.

2.3 Complex and Evolving Relation-ships

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 in-formation systems, the focus is on nding best fea-ture values to represent each observation. This is similar to using a number of data elds, such as age, gender, income, education background etc., to characterize each individual. This type of sample-feature representation inherently treats each indi-vidual as an independent entity without consider-ing 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 so-cial 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 character-ized by social functions such as friend-connections and followers (in Twitter). The correlations be-tween 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 val-ues, whereas in the sample-feature-relationship rep-resentation, two individuals can be linked together (through their social connections) even though they might share nothing in common in the feature do-mains 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 real-ity 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.

III. RESULTS AND DISCUSSION

Big Data Mining Algorithm

3.1 Decision tree induction classi - cation algorithms

In the initial stage di erent Decision Tree Learn-ing was used to analyze the big data. In deci-sion tree induction algorithms, tree structure has been widely used to represent classi cation mod-els. Most of these algorithms follow a greedy top down recursive partition strategy for the growth of the tree. Decision tree classi ers break a complex decision into collection of simpler decision. Hall. et al. [8] proposed learning rules for a large set of training data.

The work proposed by Hall et al generated a sin-gle decision system from a large and independent subset of data. An e cient decision tree algorithm based on rainforest frame work was developed for classifying large data set into number of clusters. Fuzzy- CMeans is a par-tition based clustering algorithm based on Kmeans to divide big data into several clusters[1].

3.2 Hierarchical based clustering al-gorithms

In hierarchical based algorithms large data are or-ganized in a hierarchical manner based on the medium of proximity. The initial or root cluster gradually divides into several clusters. It follows a top down or bottom up strategy to represent the clusters. Birch algorithm is one such algorithm based on hierarchical clustering.To handle stream-ing data in real time, a novel algorithm for extract-ing semantic content were de ned in Hierarchical clustering for concept mining.This algorithm was designed to be implemented in hardware, to han-dle data at very high rates. After that the tech-niques of self-organizing feature map (SOM) net-works and learning vector quantization (LVQ) net-works were discussed in Hierarchical Arti cial Neu-ral Networks for Recognizing High Similar Large Data Sets . SOM consumes input in an unsuper-vised manner whereas LVQ in supervised manner. It subdivides large data sets into smaller ones thus improving the overall computation time needed to process the large data set.

3.3 Evolutionary based classi cation algorithms

Evolutionary algorithms use domain independent technique to explore large spaces nding consis-tently good optimization solutions. There are dif-ferent types of evolutionary algorithms such as ge-netic algorithms, genetic, programming, evolution strategies, evolutionary programming and so on. Among these, genetic algorithms were mostly used for mining classi cation rules in large data sets [3].

3.4 Density based clustering algo-rithms

In density based algorithms clusters are formed based on the data objects regions of density, con-nectivity and boundary. A cluster grows in any direction based on the density growth. DENCLUE is one such algorithm based on density based clus-tering.

3.5 Partitioning based clustering al-gorithms

In partitioning based algorithms, the large data sets are divided into a number of partitions, where each partition represents a cluster. K-means is one such partitioning based method to divide large data sets

3.6 Grid based clustering algorithms

In grid base algorithms space of data objects are divided into number of grids for fast processing. OptiGrid algorithm is one such algorithm based on optimal grid partitioning.

3.7 Model based clustering algo-rithms

In model based clustering algorithms clustering is mainly performed by probability distribution. Expectation- Maximization is one such model based algorithm to estimate the maximum likeli-hood parameters of statistical models. In 2013, a new algorithm called scalable Visual Assessment of Tendency. (sVAT) algorithm was developed to pro-vide high scalable clustering in big data sets. Af-terwards a distributed ensemble classi er algorithm was developed in the eld based on the popular Random Forests for big data. This proposed algo-rithm makes use of Map Reduce for improving the e ciency and stochastic aware random forests for reducing randomness. Later in the eld, a mixed vi-sual or numerical clustering algorithm for big data called ClusiVAT was developed to provide fast clus-tering.[3,5].

IV. CONCLUSION

Driven by real-world applications and key indus-trial stakeholders and initialized by national fund-ing agencies, managing and mining Big Data have shown to be a challenging yet very compelling task. While the term Big Data literally concerns about data volumes, our HACE theorem suggests that the key 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 character-istics suggest that Big Data requires a big mind to consolidate data for maximum values (Jacobs 2009). In order to explore Big Data, we have ana-lyzed several challenges at the data, model, and sys-tem levels. To support Big Data mining, high per-formance computing platforms are required which impose systematic designs to unleash the full power of the Big Data. At the data level, the autonomous information sources and the variety of the data col-lection environments, often result in data with com-plicated conditions, such as missing/uncertain val-ues. In other situations, privacy concerns, noise and errors can be

introduced into the data, to pro-duce altered data copies. Developing a safe and sound information sharing protocol is a major challenge. At the model level, the key challenge is to generate global models by combining locally dis-covered patterns to form a unifying view. This requires carefully designed algorithms to analyze model correlations between distributed sites, and fuse decisions from multiple sources to gain a best model out of the Big Data. At the system level, the essential challenge is that a Big Data mining framework needs to consider complex relationships between samples, models, and data sources, along with their evolving changes with time and other possible factors. A system needs to be carefully designed so that unstructured data can be linked through their complex relationships to form useful patterns, and the growth of data volumes and item relationships should help form legitimate patterns to predict the trend and future.

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 technolo-gies, we will hopefully be able to provide most rel-evant and most accurate social sensing feedback to better understand our society at real-time.

V. REFERENCES

[1]. Data Mining with Big Data by Xindong Wu1,2, Xingquan Zhu3, Gong-Qing Wu2, Wei Ding4

[2]. Survey on Big Data and Mining Algorithm by Shweta Verma, Vivek Badhe