ISSN No: 2348-4845

International Journal & Magazine of Engineering, Technology, Management and Research

A Peer Reviewed Open Access International Journal

Data Mining With Big Data

Y.Prasad
P.G. Scholar (M. Tech),
Department of CSE,
Srinivasa Institute of Technology & Sciences,
Ukkayapalli, Kadapa, Andhra Pradesh.

K.Rajasekhar Reddy
Assistant Professor,
Department of CSE,
Srinivasa Institute of Technology & Sciences,
Ukkayapalli, Kadapa, Andhra Pradesh.

ABSTRACT:

Big Data concern large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. This paper 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 aggregation 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.

Index Terms:

Big Data, data mining, heterogeneity, autonomous sources, complex and evolving associations.

I. INTRODUCTION :

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 mining was very relevant from the beginning, as the first book mentioning 'Big Data' is a data mining book that appeared also in 1998 by Weiss and Indrukya . However,the first academic paper with the words 'Big Data' in the title appeared a bit later in 2000 in a paper by Diebold .The origin of the term 'Big Data' is due to the fact that we are creating a huge amount of data every day. Usama Fayyad in his invited talk at the KDD BigMine‟ 12Workshop presented amazing data numbers about internet usage, among them the following: each day Google has more than 1 billion queries per day, Twitter has more than 250 milion tweets per day, Facebook has more than 800 million updates per day, and YouTube has more than 4 billion views per day.

The data produced nowadays is estimated in the order of zettabytes, and it is growing around 40% every year.A new large source of data is going to be generated from mobile devices and big companies as Google,Apple,Facebook,Yahoo are starting to look carefully to this data to find useful patterns to improve user experience. "Big data" is pervasive, and yet still the notion engenders confusion. Big data has been used to convey all sorts of concepts, including: huge quantities of data,social media analytics, next generation data management capabilities, real-time data, and much more. Whatever the label, organizations are starting to understand and explore how to process and analyze a vast array of information in new ways. In doing so, a small, but growing group of pioneers is achieving breakthrough business outcomes. In industries throughout the world, executives recognize the need to learn more about how to exploit big data. But despite what seems like unrelenting media attention, it can be hard to find in-depth information on what organizations are really doing. So, we sought to better understand how organizations view big data – and to what extent they are currently using it to benefit their businesses.

II.BIG DATA 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. In a naïve sense, we can imagine that a number of blind men are trying to size up a giant elephant (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 person's 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, let's assume that (a) 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 their respective feelings on the elephant. Exploring the Big Data in this 4 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.

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 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.

2.2 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 5 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 behaviors. 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.

2.3 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.

SYSTEM ARCHITECTURE:

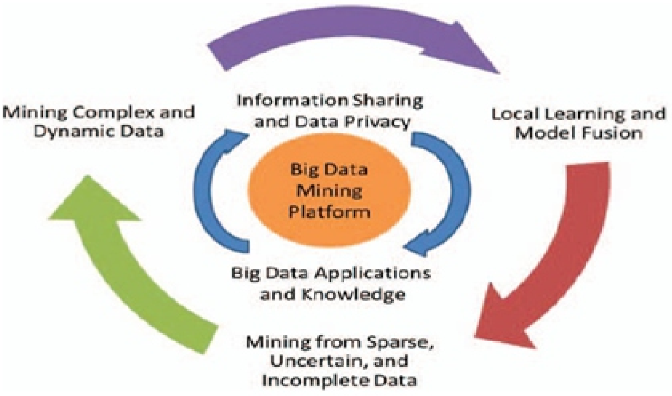
Mining Complex and Dynamic Data
Information Sharing and Data Privacy
Local Learning and Model Fusion
Big Data Mining Platform
Big Data Applications and Knowledge
Mining from Sparse, Uncertain, and Incomplete Data

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.

III. Data Mining Challenges with Big Data :

For an intelligent learning database system (Wu 2000) to handle Big Data, the essential key is to scale up to the exceptionally large volume of data and provide treatments for the characteristics featured by the aforementioned HACE theorem.

Figure 2 shows a conceptual view of the Big Data processing framework, which includes three tiers from inside out with considerations on data accessing and computing (Tier I), data privacy and domain knowledge (Tier II), and Big Data mining algorithms (Tier III). The challenges at Tier I focus on data accessing and actual computing procedures. Because Big Data are often stored at different locations and data volumes may continuously grow, an effective computing platform will have to take distributed large-scale data storage into consideration for computing. For example, while typical data mining algorithms require all data to be loaded into the main memory, this is becoming a clear technical barrier for Big Data because moving data across different locations is expensive (e.g., subject to intensive network communication and other IO costs), even if we do have a super large main memory to hold all data for computing.

The challenges at Tier II center around semantics and domain knowledge for different Big Data applications. Such information can provide additional benefits to the mining process, as well as add technical barriers to the Big Data access (Tier I) and mining algorithms (Tier III). For example, depending on different domain applications, the data privacy and information sharing mechanisms between data producers and data consumers can be significantly different. Sharing sensor network data for applications like water quality monitoring may not be discouraged, whereas releasing and sharing mobile users’ location information is clearly not acceptable for majority, if not all, applications. In addition to the above privacy issues, the application domains can also provide additional information to benefit or guide Big Data mining algorithm designs.

For example, in market basket transactions data, each transaction is considered independent and the discovered knowledge is typically represented by finding highly correlated items, possibly with respect to different temporal and/or spatial restrictions. In a social network, on the other hand, users are linked and share dependency structures. The knowledge is then represented by user communities, leaders in each group, and social influence modeling etc. Therefore, understanding semantics and application knowledge is important for both low-level data access and for high level mining algorithm designs. At Tier III, the data mining challenges concentrate on algorithm designs in tackling the difficulties raised by the Big Data volumes, distributed data distributions, and by complex and dynamic data characteristics.

The circle at Tier III contains three stages. Firstly, sparse, heterogeneous, uncertain, incomplete, and multi-source data are preprocessed by data fusion techniques. Secondly, complex and dynamic data are mined after preprocessing. Thirdly, the global knowledge that is obtained by local learning and model fusion is tested and relevant information is fedback to the pre-processing stage. Then the model and parameters are adjusted according to the feedback. In the whole process, information sharing is not only a promise of smooth development of each stage, but also a purpose of Big Data processing.

IV.CONCLUSION:

Driven by real-world applications and key industrial stakeholders and initialized by national funding 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 characteristics suggest that Big Data requires a “big mind” to consolidate data for maximum values. In order to explore Big Data, we have analyzed several challenges at the data, model, and system levels. To support Big Data mining, high performance 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 collection environments, often result in data with complicated conditions, such as missing/uncertain values. In other situations, privacy concerns, noise and errors can be introduced into the data, to produce altered data copies. Developing a safe and sound information sharing protocol is a major challenge.

REFERENCES :

1) Ahmed and Karypis 2012, Rezwan Ahmed, George Karypis, Algorithms for mining the evolution of conserved relational states in dynamic networks, Knowledge and Information Systems, December 2012, Volume 33, Issue 3, pp 603-630 .

2) Alam et al. 2012, Md. Hijbul Alam, JongWoo Ha, Sang-Keun Lee, Novel approaches to crawling important pages early, Knowledge and Information Systems, December 2012, Volume 33, Issue 3, pp 707-734

3) Aral S. and Walker D. 2012, Identifying influential and susceptible members of social networks, Science, vol.337, pp.337-341.

4) Machanavajjhala and Reiter 2012, Ashwin Machanavajjhala, Jerome P. Reiter: Big privacy: protecting confidentiality in big data. ACM Crossroads, 19(1): 20-23, 2012.

5) Banerjee and Agarwal 2012, Soumya Banerjee, Nitin Agarwal, Analyzing collective behavior from blogs using swarm intelligence, Knowledge and Information Systems, December 2012, Volume 33, Issue 3, pp 523-547

6) Birney E. 2012, The making of ENCODE: Lessons for big-data projects, Nature, vol.489, pp.49-51.

7) Bollen et al. 2011, J. Bollen, H. Mao, and X. Zeng, Twitter Mood Predicts the Stock Market, Journal of Computational Science, 2(1):1-8, 2011.