BIG DATA USING DATA MINING CONCEPTS

M.Jancyrani Malli
Department of Information Technology, Dr.Sivanthi Aditanar College of Engineering, Tiruchendur- 628615

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 type of summarization program is an excellent example for Big Data processing, as the information comes from multiple, heterogeneous, autonomous sources with complex and evolving relationships, and keeps growing.

Assuming the size of each photo is 2 megabytes (MB), this requires 3.6 terabytes (TB) storage every single day. Indeed, as an old saying states: “a picture is worth a thousand words,” the billions of pictures on Flicker are a treasure tank for us to explore the human society, social events, public affairs, disasters, and so on, only if we have the power to harness the enormous amount of data.

The above examples demonstrate the rise of Big Data applications where data collection has grown tremendously and is beyond the ability of commonly used software tools to capture, manage, and process within a “tolerable elapsed time.” The most fundamental challenge for Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions. In many situations, the knowledge extraction process has to be very efficient and close to real time because storing all observed data is nearly infeasible. For example, the square kilometer array (SKA) in radio astronomy consists of 1,000 to 1,500 15-meter dishes in a central 5- km area. It provides 100 times more sensitive vision than any existing radio telescopes, answering fundamental questions about the Universe. However, with a 40 gigabytes (GB)/second data volume, the data generated from the SKA are exceptionally large. Although researchers have confirmed that interesting patterns, such as transient radio anomalies can be discovered from the SKA data, existing methods can only work in an offline fashion and are incapable of handling this Big Data scenario in real time. As a result, the unprecedented data volumes require an effective data analysis and prediction platform to achieve fast response and real- time classification- for such Big Data.

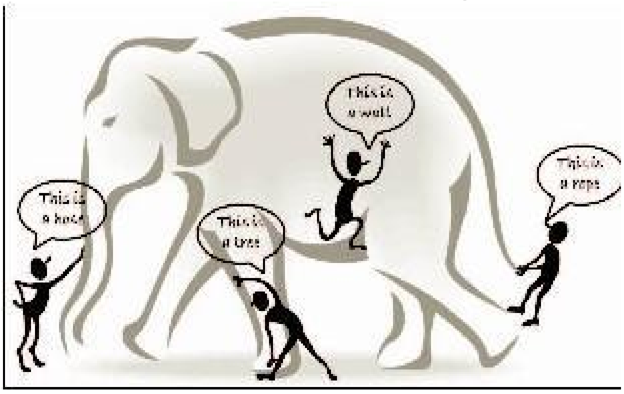


Fig. 1. The blind men and the giant elephant: the localized (limited) view of each blind man leads to a biased conclusion.

2 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 challenge for discovering useful knowledge from the Big Data. In a naı¨ve sense, we can imagine that a number of blind men are trying to size up a giant elephant (see Fig. 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 collects 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 us assume that

1) the elephant is growing rapidly and its pose changes constantly, and 2) each blind man may have his own (possible unreliable and inaccurate) information sources that tell him about biased knowledge about the elephant . 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.

Huge Data with Heterogeneous and Diverse Dimensionality

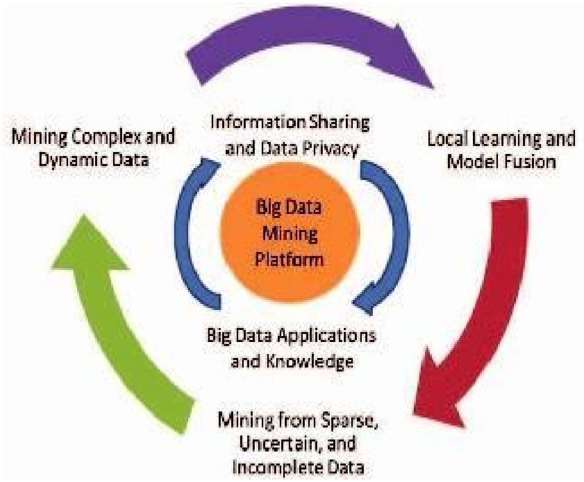
One of the fundamental characteristics of the Big Data is the huge volume of data represented by heterogeneous and diverse dimensionalities. This is because different information collectors prefer their own schemata or protocols for data recording, and the nature of different applications also results in diverse data representations. For example, each single human being in a biomedical world can be represented by using simple demographic information such as gender, age, family disease history, and so on. 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, micro- array 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.

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 source 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, Face book, 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.

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, and so on, 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



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.

3. DATAMINING CHALLENGES WITH BIG DATA

For an intelligent learning database system 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. Fig. 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 arithmetic 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, typical data mining algorithms require all data to be loaded into the main memory, this, however, is becoming a clear technical barrier for Big Data because moving data across different locations is expensive, even if we do have a super large main memory to hold all data for computing.

The challenges at Tier II center on 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. The knowledge is then represented by user communities, leaders in each group, and social influence modeling, and so on. 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. First, sparse, heterogeneous, uncertain, incomplete, and multisource data are preprocessed by data fusion techniques. Second, complex and dynamic data are mined after preprocessing. Third, the global knowledge obtained by local learning and model fusion is tested and relevant information is feedback to the preprocessing 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.

Tier I: Big Data Mining Platform

In typical data mining systems, the mining procedures require computational intensive computing units for data analysis and comparisons. A computing platform is, therefore, needed to have efficient access to, at least, two types of resources: data and computing processors. For small scale data mining tasks, a single desktop computer, which contains hard disk and CPU processors, is sufficient

For Big Data mining, because data scale is far beyond the capacity that a single personal computer (PC) can handle, a typical Big Data processing framework will rely on cluster computers with a high- performance computing platform, with a data mining task being deployed by running some parallel programming tools, such as Map Reduce or Enterprise Control Language (ECL), on a large number of computing nodes (i.e., clusters). The role of the software component is to make sure that a single data mining task, such as finding the best match of a query from a database with billions of records, is split into many small tasks each of which is running on one or multiple computing nodes.

Such a Big Data system, which blends both hardware and software components, is hardly available without key industrial stockholders' support. In fact, for decades, companies have been making business decisions based on transactional data stored in relational databases. Big Data mining offers opportunities to go beyond traditional relational databases to rely on less structured data: weblogs, social media, e-mail, sensors, and photographs that can be mined for useful information. Major business intelligence companies, such IBM, Oracle, Tera data, and so on, have all featured their own products to help customers acquire and organize these diverse data sources and coordinate with customers' existing data to find new insights and capitalize on hidden relationships.

3 3.2Tier II: Big Data Semantics and Application Knowledge

Semantics and application knowledge in Big Data refer to numerous aspects related to the regulations, policies, user knowledge, and domain information. The two most important issues at this tier include 1) data sharing and privacy; and 2) domain and application knowledge. The former provides answers to resolve concerns on how data are maintained, accessed, and shared.

Information Sharing and Data Privacy

Information sharing is an ultimate goal for all systems involving multiple parties. While the motivation for sharing is clear, a real-world concern is that Big Data applications are related to sensitive information, such as banking transactions and medical records. To protect privacy, two common approaches are to 1) restrict access to the data, such as adding certification or access control to the data entries, so sensitive information is accessible by a limited group of users only, and 2) anonymize data fields such that sensitive information cannot be pinpointed to an individual record. For the first approach, common challenges are to design secured certification or access control mechanisms, such that no sensitive information can be misconduct by unauthorized individuals. For data anonymization, the main objective is to inject randomness into the data to ensure a number of privacy goals Common anonymization approaches are to use suppression, generalization, perturbation, and permutation to generate an altered version of the data, which is, in fact, some uncertain data.

One of the major benefits of the data annomization based information sharing approaches is that, once anonymized, data can be freely shared across different parties without involving restrictive access controls. This naturally leads to another research area namely privacy preserving data mining , where multiple parties, each holding some sensitive data, are trying to achieve a common data mining goal without sharing any sensitive information inside the data. This privacy preserving mining goal, in practice, can be solved through two types of approaches including 1) using special communication protocols, such as Yao's protocol, to request the distributions of the whole data set, rather than requesting the actual values of each record, or 2) designing special data mining methods to derive knowledge from anonymized data.

Domain and Application Knowledge

Domain and application knowledge provides essential information for designing Big Data mining algorithms and systems. In a simple case, domain knowledge can help identify right features for modeling the underlying data. The domain and application knowledge can also help design achievable business objectives by using Big Data analytical techniques. An appealing Big Data mining task is to design a Big Data mining system to predict the movement of the market in the next one or two minutes. Such systems, even if the prediction accuracy is just slightly better than random guess, will bring significant business values to the developers.

Tier III: Big Data Mining Algorithms

Local Learning and Model Fusion for Multiple Information Sources

As Big Data applications are featured with autonomous sources and decentralized controls, aggregating distributed data sources to a centralized site for mining is systematically prohibitive due to the potential transmission cost and privacy concerns. On the other hand, although we can always carry out mining activities at each distributed site, the biased view of the data collected at each site often leads to biased decisions or models, just like the elephant and blind men case. Under such a circumstance, a Big Data mining system has to enable an information exchange and fusion mechanism to ensure that all distributed sites can work together to achieve a global optimization goal. Model mining and correlations are the key steps to ensure that models or patterns discovered from multiple information sources can be consolidated to meet the global mining objective. More specifically, the global mining can be featured with a two- step (local mining and global correlation) process, at data, model, and at knowledge levels. At the data level, each local site can calculate the data statistics based on the local data sources and exchange the statistics between sites to achieve a global data distribution view. At the model or pattern level, each site can carry out local mining activities, with respect to the localized data, to discover local patterns. By exchanging patterns between multiple sources, new global patterns can be synthesized by aggregating patterns across all sites. At the knowledge level, model correlation analysis investigates the

relevance between models generated from different data sources to determine how relevant the data sources are correlated with each other, and how to form accurate decisions based on models built from autonomous sources.

Mining from Sparse, Uncertain, and Incomplete Data

Spare, uncertain, and incomplete data are defining features for Big Data applications. Being sparse, the number of data points is too few for drawing reliable conclusions. This is normally a complication of the data dimensionality issues, where data in a high-dimensional space do not show clear trends or distributions. For most machine learning and data mining algorithms, high-dimensional spare data significantly deteriorate the reliability of the models derived from the data.

Uncertain data are a special type of data reality where each data field is no longer deterministic but is subject to some random/error distributions. This is mainly linked to domain specific applications with inaccurate data readings and collections. For data privacy-related applications , users may intentionally inject randomness/errors into the data to remain anonymous. This is similar to the situation that an individual may not feel comfortable to let you know his/her exact income, but will be fine to provide a rough range like [120k, 160k]. For uncertain data, the major challenge is that each data item is represented as sample distributions but not as a single value, so most existing data mining algorithms cannot be directly applied. Common solutions are to take the data distributions into consideration to estimate model parameters.

Mining Complex and Dynamic Data

The rise of Big Data is driven by the rapid increasing of complex data and their changes in volumes and in nature. Making use of complex data is a major challenge for Big Data applications, because any two parties in a complex network are potentially interested to each other with a social connection. Such a connection is quadratic with respect to the number of nodes in the network, so a million node networks may be subject to one trillion connections.

Complex intrinsic semantic associations in data. News on the web, comments on Twitter, pictures on Flicker, and clips of video on YouTube may discuss about an academic award- winning event at the same time.

Complex relationship networks in data. In the context of Big Data, there exist relationships between individuals. On the Internet, individuals are WebPages and the pages linking to each other via hyperlinks form a complex network.

4. Big Data Mining Algorithms

To adapt to the multisource, massive, dynamic Big Data, researchers have expanded existing data mining methods in many ways, including the efficiency improvement of single-source knowledge discovery methods, designing a data mining mechanism from a multisource perspective, as well as the study of dynamic data mining methods and the analysis of stream data. The main motivation for discovering knowledge from massive data is improving the efficiency of single-source mining methods. On the basis of gradual improvement of computer hardware functions, researchers continue to explore ways to improve the efficiency of knowledge discovery algorithms to make them better for massive data. Because massive data are typically collected from different data sources, the knowledge discovery of the massive data must be performed using a multisource mining mechanism. As real-world data often come as a data stream or a characteristic flow, a well-established mechanism is needed to discover knowledge and master the evolution of knowledge in the dynamic data source. Therefore, the massive, heterogeneous and real-time characteristics of multisource data provide essential differences between single- source knowledge discovery and multisource data mining.

Local pattern analysis of data processing can avoid putting different data sources together to carry out centralized computing.

Data streams are widely used in financial analysis, online trading, and medical testing, and so on. Static knowledge discovery methods cannot adapt to the characteristics of dynamic data streams, such as continuity, variability, rapidity, and infinity, and can easily lead to the loss of useful information. Therefore, effective theoretical and technical frameworks are needed to support data stream.

Knowledge evolution is a common phenomenon in real- world systems. For example, the clinician's treatment programs will constantly adjust with the conditions of the patient, such as family economic status, health insurance, the course of treatment, treatment effects, and distribution

5. CONCLUSIONS

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 require a "big mind" to consolidate data for maximum values.

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. At the model level, the key challenge is to generate global models by combining locally discovered 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 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. The era of Big Data has arrived.

REFERENCES

[1]R. Ahmed and G. Karypis, “Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks,” Knowledge and Information Systems, vol. 33, no. 3, pp. 603-630, Dec. 2012.

[2]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.

[3]S. Aral and D. Walker, “Identifying Influential and Susceptible Members of Social Networks,” Science, vol. 337, pp. 337-341, 2012.

[4]A. Machanavajjhala and J.P. Reiter, “Big Privacy: Protecting Confidentiality in Big Data,” ACM Crossroads, vol. 19, no. 1, pp.20-23, 2012.

[5]S. Banerjee and N. Agarwal, “Analyzing Collective Behavior from Blogs Using Swarm Intelligence,” Knowledge and Information Systems, vol. 33, no. 3, pp. 523-547, Dec.2012.

[6]E. Birney, “The Making of ENCODE: Lessons for Big-DataProjects,” Nature, vol.489, pp. 49-51, 2012.

[7]J. Bollen, H. Mao, and X. Zeng, “Twitter Mood Predicts the Stock Market,” J. Computational Science, vol. 2, no. 1, pp. 1-8, 2011.

[8]S. Borgatti, A. Mehra, D. Brass, and G. Labianca, “Network Analysis in the Social Sciences,” Science, vol. 323, pp. 892- 895, 2009.

[9]J. Bughin, M. Chui, and J. Manyika, Clouds, Big Data, and Smart Assets: Ten Tech- Enabled Business Trends to Watch.McKinSey Quarterly, 2010.

[10]D. Centola, “The Spread of Behavior in an Online Social Network Experiment,” Science, vol. 329, pp. 1194-1197, 2010.

[11]E.Y. Chang, H. Bai, and K. Zhu, “Parallel Algorithms for MiningLarge-Scale Rich- Media Data,” Proc. 17th ACM Int’l Conf. Multimedia, (MM ’09,) pp. 917-918, 2009.

[12]R. Chen, K. Sivakumar, and H. Kargupta, “Collective Mining of Bayesian Networks from Distributed Heterogeneous Data,” Knowledge and Information Systems, vol. 6, no. 2, pp. 164-187, 2004.

[13]Y.-C. Chen, W.-C. Peng, and S.-Y. Lee, “Efficient Algorithms for Influence Maximization in Social Networks,” Knowledge and Information Systems, vol. 33, no. 3, pp. 577-601, Dec. 2012.

[14]C.T. Chu, S.K. Kim, Y.A. Lin, Y. Yu, G.R. Bradski, A.Y. Ng, and K. Olukotun, “Map-Reduce for Machine Learning on Multicore,” Proc. 20th Ann. Conf. Neural Information Processing Systems (NIPS ’06), pp. 281- 288, 2006.

[15]G. Cormode and D. Srivastava, “Anonymized Data: Generation, Models, Usage,” Proc. ACM SIGMOD Int’l Conf. Management Data, pp.1015-1018, 2009.

[16]S. Das, Y. Sismanis, K.S. Beyer, R. Gemulla, P.J. Haas, and J. McPherson, “Ricardo: Integrating R and Hadoop,” Proc. ACM SIGMOD Int’l Conf. Management Data (SIGMOD ’10), pp.987-998. 2010.