International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 10, October 2015
Available at http://internationaljournalofresearch.org

Data Mining Resolution on High Dimensional Data

Nakka Yeleswara Rao[1]& S.Ranga Swamy [2]
[1]PG Scholar,Dept of CSE, Rao & Naidu Engineering College, Ongole,Prakasam Dist,Andhra Pradesh
[2]Associate Professor,Dept of CSE, Rao & Naidu Engineering College, Ongole,Prakasam Dist,Andhra Pradesh

Abstract—
Nowadays most of the data analysts are working with Big Data. Big Data means large-volume, complex, increasing data sets with various, and independent sources. Currently Big Data are quickly expanding in all science and engineering domains, including physical, biological and biomedical sciences due to the rapid development of networking, data storage, the data collection capability. We present a theorem that illustrates the features of the Big Data, and proposes a Big Data processing model. This data-driven model includes demand-driven aggregation of information bases, mining and exploration, user interest modeling, security, and privacy considerations. We analyze the problematic issues in the data-driven model as well as in the Big Data revolution.

Keywords: Big Data; data sets; heterogeneity; independent sources; difficult and evolving relationships

I. INTRODUCTION

The time of Big Data has arrived, every day, 2.5 quintillion bytes of data are created and 90 percent of the data in the world today were created within the past two years. Our ability for data generation is not so powerful and huge ever since the development of the information technology in the early 19th century. As one more example, on 4 October 2012, the first presidential debate between President Barack Obama and Governor Mitt Romney triggered more than 10 million tweets within 2 hours. Among all these tweets, the specific moments that generated the most discussions actually revealed the public interests, such as the discussions about medicare and vouchers. Such online discussions provide a new means to sense the public interests and generate feedback in realtime, and are mostly appealing compared to generic media, such as radio or TV broadcasting.

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 realtime 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

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 10, October 2015
Available at http://internationaljournalofresearch.org

volumes require an effective data analysis and prediction platform to achieve fast response and real-time classification for such Big Data.

II. CHARACTERISTICS OF BIG DATA

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 naive sense, we can imagine that a number of blind men are trying to size up a giant elephant, 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 arope, 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 (e.g., one blind man may exchange his feeling about the elephant with another blind man, where the exchanged knowledge is inherently biased). 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 areal-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 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, 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.

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 10, October 2015
Available at http://internationaljournalofresearch.org

2.2 Autonomous Sources

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, 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 result in restructured 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, and soon, 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. Our friend circles may be formed based on the common hobbies or people are connected by biological relationships. Such social connections commonly exist not only in our daily activities, but also are very popular in cyber 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 on the data. 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 (nonlinear, many-to-many) data relationships, along with the evolving changes, into consideration, to discover useful patterns from Big Data collections.

III. BIG DATA MINING PLATFORMS

Due to the multisource, massive, heterogeneous, and dynamic characteristics of application data involved in a distributed environment, one of the most important characteristics of Big Data is to carry out computing on the petabyte (PB), even the exabyte (EB)-level data with a complex

computing process. Therefore, utilizing a parallel computing infrastructure, its corresponding programming language support, and software models to efficiently analyze and mine the distributed data are the critical goals for Big Data processing to change from "quantity" to "quality."

Currently, Big Data processing mainly depends on parallel programming models like MapReduce, as well as providing a cloud computing platform of Big Data services for the public. 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 parameters. It calls for intensive computing to access the large-scale data frequently. To improve the efficiency of algorithms, Chu et al. proposed a general-purpose parallel programming method, which is applicable to a large number of machine learning algorithms based on the simple MapReduce programming model on multi-core processors. Ten classical data mining algorithms are realized in the framework, including locally weighted linear regression, k-Means, logistic regression, naive Bayes, linear support vector machines, the independent variable analysis, Gaussian discriminant analysis, expectation maximization, and back-propagation neural networks. With the analysis of these classical machine learning algorithms, we argue that the computational operations in the algorithm learning process could be transformed into a summation operation on a number of training data sets. Summation operations could be performed on different subsets independently and achieve penalization executed easily on the MapReduce programming platform. Therefore, a large-scale data set could be divided into several subsets and assigned to multiple Mapper nodes. Then, various summation operations could be performed on the Mapper nodes to collect intermediate results. Finally, learning algorithms are executed in parallel through merging summation on Reduce nodes. Ranger et al. proposed a MapReduce-based application programming interface Phoenix, which supports parallel programming in the environment of multi-core and multiprocessor systems, and realized three data mining algorithms including k-Means, principal component analysis, and linear regression. Gillick et al. improved the MapReduce's implementation mechanism in Hadoop, evaluated the algorithms' performance of single-pass learning, iterative learning, and query-based learning in the MapReduce framework, studied data sharing between computing nodes involved in parallel learning algorithms, distributed data storage, and then showed that the MapReduce mechanisms suitable for large-scale data mining by testing series of standard data mining tasks on medium-size clusters. Papadimitriou and Sun proposed a distributed collaborative aggregation (DisCo) framework using practical distributed data preprocessing and collaborative aggregation techniques. The implementation on Hadoop in an open source MapReduce project showed that DisCo has perfect scalability and can process and analyze massive data sets (with hundreds of GB).

To improve the weak scalability of traditional analysis software and poor analysis capabilities of Hadoop systems, Das et al. conducted a study of the integration of R (open source statistical analysis software) and Hadoop. The in-depth integration pushes data computation to parallel processing, which enables powerful deep analysis

capabilities for Hadoop. Wegener et al. achieved the integration of Weka (an open-source machine learning and data mining software tool) and MapReduce. Standard Weka tools can only run on a single machine, with a limitation of 1-GB memory. After algorithm parallelization, Weka breaks through the limitations and improves performance by taking the advantage of parallel computing to handle more than 100-GB data on MapReduce clusters. Ghoting et al. proposed Hadoop-ML, on which developers can easily build task-parallel or data-parallel machine learning and data mining algorithms on program blocks under the language runtime environment

IV 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 multi-source data provide essential differences between single-source knowledge discovery and multisource data mining.

Wu et al. 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. 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, 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 mining.

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 of cardiovascular and other chronic epidemiological changes with the passage of time. In the knowledge discovery process, concept drifting aims to analyze the phenomenon of implicit target concept changes or even fundamental changes triggered by dynamics and context in data streams. According to different types of concept drifts, knowledge evolution can take forms of mutation drift, progressive drift, and data distribution drift, based on single features, multiple features, and streaming features.

V. CONCLUSION

Motivated 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 theorem recommends that the main characteristics of the Big Data are 1) huge with heterogeneous data sources, 2)autonomous sources, and 3) complex and evolving in data and knowledge relationships. Such collective 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 realtime. 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

VI. REFERENCES

[1] A. Machanavajjhala and J.P. Reiter, "Big Privacy: Protecting Confidentiality in Big Data," ACM Cross roads, vol. 19, no. 1, pp. 20-23, 2012.

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

[3] 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.

[4] "IBM What Is Big Data: Bring Big Data to the Enterprise," http://www-01.ibm.com/software/data/bigdata/, IBM, 2012.

[5] M. Ye, X. Wu, X. Hu, and D. Hu, "Anonymizing Classification Data Using Rough Set Theory", Knowledge-Based Systems, vol. 43, pp. 82-94, 2013.

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

[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 Mining Large-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.