Volume 5, Issue 4, April 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Big Data and the SP Theory of Intelligence

Nisha Rajasegaran
M.Tech(CSE), Prist University,
Puducherry, India

J. R. Thresphine
A.P, Dept. of CSE, Prist University,
Puducherry, India

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 project 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, autonomous sources, complex and evolving associations

I. INTRODUCTION

Recent years have witnessed a dramatic increase in our ability to collect data from various sensors, devices, in di_erent formats, from independent or connected applications. This data ood has outpaced our capability to process, analyze, store and understand these datasets. Consider the Internet data. The web pages indexed by Google were around one million in 1998, but quickly reached 1 billion in 2000 and have already exceeded 1 trillion in 2008. This rapid expansion is accelerated by the dramatic increase in acceptance of social networking applications, such as Facebook, Twitter, Weibo, etc., that allow users to create contents freely and amplify the already huge Web volume.

Along with the above example, the era of Big Data has arrived [37], [34], [29]. Every day, 2.5 quintillion bytes of data are created and 90 percent of the data in the world today were produced within the past two years [26]. Our capability for data generation has never been so powerful and enormous ever since the invention of the information technology in the early 19th century.

The issue of the measurement of three evaluation dimensions simultaneously has led to another important issue in data stream mining, namely estimating the combined cost of performing the learning and prediction processes in terms of time and memory. As an example, several rental cost options exist:

- Cost per hour of usage: Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable compute capacity in the cloud. Cost depends on the time and on the machine rented (small instance with 1.7 GB, large with 7.5 GB or extra large with 15 GB).
- Cost per hour and memory used: GoGrid is a web service similar to Amazon EC2, but it charges by RAM Hours. Every GB of RAM deployed for 1 hour equals one RAM-Hour.

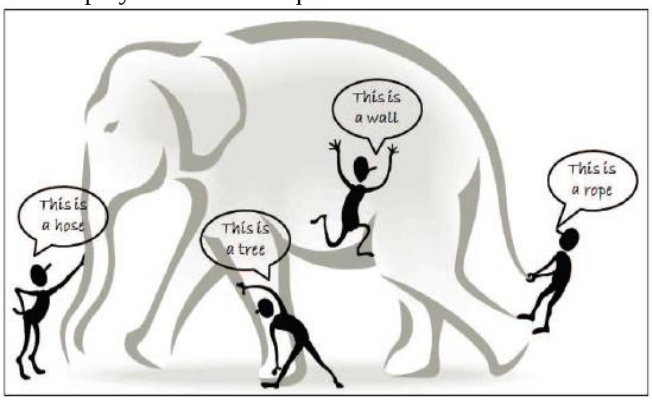

Fig.1. The blind men and the giant elephant: the localized (limited) view of each blind man leads to a biased conclusion.

As an example of the interest that Big Data is having in the data mining community, the grand theme of this year's KDD conference was 'Mining the Big Data'. Also there was a speci_c workshop BigMine'12 in that topic: 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications1. Both events successfully brought together people from both academia and industry to present their most recent work related to these Big Data issues, and exchange ideas and thoughts. These events are important in order to advance this Big Data challenge, which is being considered as one of the most exciting opportunities in the years to come.

II. BIG DATA MINING

Big Data is a new term used to identify the datasets that due to their large size, we can not manage them with the typical data mining software tools. Instead of defining "Big Data" as datasets of a concrete large size, for example in the order of magnitude of petabytes, the definition is related to the fact that the dataset is too big to be managed without using new algorithms or technologies. However, the _rst academic paper with the words 'Big Data' in the title appeared a bit later in 2000 in a paper by Diebold [8] each day Google has more than 1 billion queries per day, Twitter has more than 250 million 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 zetta bytes, and it is growing around 40% every year.

We need new algorithms, and new tools to deal with all of this data. Doug Laney[19] was the _rst one in talking about 3 V's in Big Data management:

- Volume: there is more data than ever before, its size continues increasing, but not the percent of data that our tools can process
- Variety: there are many di_erent types of data, as text, sensor data, audio, video, graph, and more
- Velocity: data is arriving continuously as streams of data, and we are interested in obtaining useful information from it in real time Nowadays, there are two more V's:
- Variability: there are changes in the structure of the data and how users want to interpret that data
- Value: business value that gives organization a compelling advantage, due to the ability of making decisions based in answering questions that were previously considered beyond reach

There are two main strategies for dealing with big data: sampling and using distributed systems. Sampling is based in the fact that if the dataset is too large and we cannot use all the examples, we can obtain an approximate solution using a subset of the examples. A good sampling method will try to select the best instances, to have a good performance using a small quantity of memory and time.

2.1 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.

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.

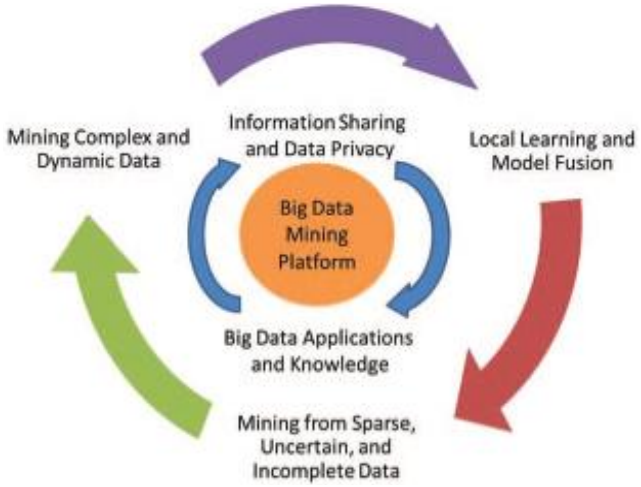

Fig.2. A Big Data processing framework

2.2 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 the human society. Our friend circles may be formed based on the common hobbies or people are connected by biological relationships.

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.

III. DATA MINING CHALLENGES WITH BIG DATA

For an intelligent learning database system [52] 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 (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.

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 fedback 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.

3.1 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 Fig. 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.

The outmost circle shows Tier III challenges on actual mining algorithms. to fulfill the data mining goals. Indeed, many data mining algorithm are designed for this type of problem settings. For medium scale data mining tasks, data are typically large (and possibly distributed) and cannot be fit into the main memory. Common solutions are to rely on parallel computing [43], [33] or collective mining [12] to sample and aggregate data from different sources and then use parallel computing programming (such as the Message Passing Interface) to carry out the mining process.

3.2 Tier 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; whereas the latter focuses on answering questions like "what are the underlying applications ?" and "what are the knowledge or patterns users intend to discover from the data ?"

3.2.1 Information Sharing and Data Privacy

Information sharing is an ultimate goal for all systems involving multiple parties [24]. 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. Simple data exchanges or transmissions do not resolve privacy concerns [19], [25], [42]. 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 [15].

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.

3.2.2 Domain and Application Knowledge

Domain and application knowledge [28] 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 (e.g., blood glucose level is clearly a better feature than body mass in diagnosing Type II diabetes). The domain and application knowledge can also help design achievable business objectives by using Big Data analytical techniques. For example, stock market data are a typical domain that constantly generates a large quantity of information, such as bids, buys, and puts, in every single second. The market continuously evolves and is impacted by different factors, such as domestic and international news, government reports, and natural disasters, and so on.

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 [9]. Without correct domain knowledge, it is a clear challenge to find effective matrices/measures to characterize the market movement, and such knowledge is often beyond the mind of the data miners, although some recent research has shown that using social networks

3.3 Tier III: Big Data Mining Algorithms

3.3.1 Local Learning and Model Fusion for Multiple

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 (or information sources) 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.

3.3.2 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 (such as more than 1,000 dimensions) 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.

Common approaches are to employ dimension reduction or feature selection [48] to reduce the data dimensions or to carefully include additional samples to alleviate the data scarcity, such as generic unsupervised learning methods in data mining.

3.3.3 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 [6]. Documents posted on WWW servers, Internet backbones, social networks, communication networks, and transportation networks, and so on are all featured with complex data. While complex dependency structures underneath the data raise the difficulty for our learning systems, they also offer exciting opportunities that simple data representations are incapable of achieving.

For example, researchers have successfully used Twitter, a well-known social networking site, to detect events such as earthquakes and major social activities, with nearly realtime speed and very high accuracy. In addition, by summarizing the queries users submitted to the search engines, which are all over the world, it is now possible to build an early warning system for detecting fast spreading flu outbreaks [23]. 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 network may be subject to one trillion connections.

IV. RELATED WORKS

4.1 Big Data Mining Platforms (Tier I)

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.

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.

Finally, learning algorithms are executed in parallel through merging summation on Reduce nodes. Ranger et al. [39] proposed a MapReduce-based application programming interface Phoenix, which supports parallel programming in the environment of multicore and multiprocessor systems, and realized three data mining algorithms including k-Means, principal component analysis, and linear regression. Gillick et al. [22] improved the MapReduce's implementation mechanism in Had loop, 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

4.2 Big Data Semantics and Application Knowledge (Tier II)

In privacy protection of massive data, Ye et al. [5] proposed a multilayer rough set model, which can accurately describe the granularity change produced by different levels of generalization and provide a theoretical foundation for measuring the data effectiveness criteria in the anonymization process, and designed a dynamic mechanism for balancing privacy and data utility, to solve the optimal generalization/refinement order for classification. A recent paper on confidentiality protection in Big Data [4] summarizes a number of methods for protecting public release data, including aggregation (such as kanonymity, I-diversity, etc.), suppression (i.e., deleting sensitive values), data swapping (i.e., switching values of sensitive data records to prevent users from matching), adding random noise, or simply replacing the whole original data values at a high risk of disclosure with values synthetically generated from simulated distributions.

4.3 Big Data Mining Algorithms (Tier III)

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 [11], designing a data mining mechanism from a multisource perspective [10], [21], as well as the study of dynamic data mining methods and the analysis of stream data [18], [12]. 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.

V. 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.

REFERENCES

[1] X. Amatriain. Mining large streams of user data for personalized recommendations. SIGKDD Explorations, 14(2), 2012.
[2] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna. Four Degrees of Separation. CoRR, abs/1111.4570, 2011.
[3] A. Bifet and E. Frank. Sentiment knowledge discovery in Twitter streaming data. In Proc 13th International Conference on Discovery Science, Canberra, Australia, pages 1–15. Springer, 2010.
[4] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer. MOA: Massive Online Analysis http://moa. cms.waikato.ac.nz/. Journal of Machine Learning Research (JMLR), 2010.
[5] A. Bifet, G. Holmes, and B. Pfahringer. Moatweetreader: Real-time analysis in twitter streaming data. In Discovery Science, pages 46–60, 2011.
[6] A. Bifet, G. Holmes, B. Pfahringer, and E. Frank. Fast perceptron decision tree learning from evolving data streams. In PAKDD, 2010.
[7] C. Bockermann and H. Blom. The streams Framework. Technical Report 5, TU Dortmund University, 12 2012.
[8] P. Boldi, M. Rosa, and S. Vigna. HyperANF: approximating the neighbourhood function of very large graphs on a budget. In Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011, pages 625–634, 2011.

[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.

[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.

[17] P. Dewdney, P. Hall, R. Schilizzi, and J. Lazio, "The Square Kilometre Array," Proc. IEEE, vol. 97, no. 8, pp. 1482-1496, Aug. 2009.

[18] P. Domingos and G. Hulten, "Mining High-Speed Data Streams," Proc. Sixth ACM IGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '00), pp. 71-80, 2000.

[19] G. Duncan, "Privacy by Design," Science, vol. 317, pp. 1178-1179, 2007.

[20] B. Efron, "Missing Data, Imputation, and the Bootstrap," J. Am. Statistical Assoc., vol. 89, no. 426, pp. 463-475, 1994.

[21] D. J. Leinweber. Stupid Data Miner Tricks: Over_tting the S&P 500. The Journal of Investing, 16:15{22, 2007.

[22] E. Letouz_e. Big Data for Development: Opportunities & Challenges. May 2011.

[23] J. Lin. MapReduce is Good Enough? If All You Have is a Hammer, Throw Away Everything That's Not a Nail! CoRR, abs/1209.2191, 2012.

