The Review of Big Data

Chunhe Shi[1, 2, a *], Chengdong Wu[3, b], Xiaowei Han[4,c],
Zhen Li[5, d], and Yinghong Xie[6, e]

[1, 3,] College of Information Science and Engineering, Northeastern University, Shenyang 110819, China

[2, 4, 6] College of Information and Engineering, Shenyang University, Shenyang 110044, China

[5]Northeast Regional Air Traffic Management Bureau of CAAC, Shenyang 110043, China

[a]schsydx@163.com, [b]wuchengdong@ise.neu.edu.cn, [c]hxw69@163.com, [d]zlalbert@126.com, [e]xieyinghong@163.com

Keywords: Big data; Big data technology; Data mining; Challenge.

Abstract. Big Data is becoming the attentive focus in the current world. With the rapid integration and development of the next generation of information technology, such as Cloud computing, mobile Internet and Internet of Things, data present an exponential growth. This paper illustrates the concept of big data, and present national and international research and application status, particularly analyzes the advantages and disadvantages of the key techniques of big data processing, and summarizes the current challenges that big data is facing. Finally, it views the prospects of the future based on the above research and summaries.

In recent years, big data is rapidly developed into a worldwide hot spot in varied fields of academy, industry, and government agencies. Nature [1] and Science [2] respectively published special issues to discuss the development and research of big data in 2008 and 2011.With the rapid integration and development of the next generation information technology, such as Cloud computing, mobile Internet and Internet of Things, massive interactive and sensor data are successively generated from various industries. Data volume has jumped from TB to PB, EB, and even ZB level. Out of all data obtained by the entire human civilization, 90% have been generated in the past two years [3]. By 2020, the scale of data volume generated worldwide will have soared 44 times to 35 ZB [4].

Concept of Big Data

At present, big data has not a commonly recognized definition, yet what human can directly perceive is its data amount. Gartner Inc. defined big data as information value of big volume, high velocity and large variety [5]; while IDC (International Data Corporation) considered its traits should also include value, [6] which are usually described as 4V features of big data, i.e. Volume, Velocity, Variety and Value. In Wikipedia, Big Data or Mega data is defined as enormous dada, massive data, and large files, referring to the information of so large amount involved that it cannot be interpreted into the form within human understanding through interception, management, processing and organizing in reasonable time by labors. [7] Suggests that generally big data refers to data sets that cannot be perceived, acquired, managed, processed and served by the traditional machine or software and hardware within certain time. Though the concept of big data has not been clearly unified, the pace of exploring it remains proceeding.

The Key Techniques of Big Data Processing

Traditional data processing can no longer satisfy the existing needs, for massive data brings about too many questions, therefore, new technology and methods will necessarily be thriving.

Data Acquisition. The generation of big data mainly has three sources: (1) rich database resources on the Internet; (2) physical information system, such as Internet of Things, Smart City, etc., which is usually acquired by sensors or observing equipment; (3) Scientific experiments and observations, in which the access of Web and sensor data is the key point.

Data acquisition is the first step in the analysis of data processing, and collecting high quality data is vital for subsequent data processing and analysis. This aspect of work focuses much more on data consistency than reducing noisy data and improving accuracy, which can be a direction of further research work.

Storage of Big Data. The traditional mode of storing cannot bear big data. Before the age of big data, the efficiency of relational database can totally meet the fundamental requirements for the volume of GB level, but when data progressively grow, its extensibility disadvantage becomes be more obvious. Therefore, in the era of data explosion, the new way of data storage, i.e. distributed file system is applied to handle this change. Since big data has sparse and high-dimensional traits, and data analysis tasks use relatively fewer fields, column storage possesses larger space compressibility and stands more efficient than row storage [8]. [9] Proposes a row-column hybrid data storage structure to solve problems of fast loading of massive data, shortening the query-response time, and efficiently using the disk space, which currently becomes a de facto standard of distributed storage.

At present there have been many effective methods of big data storage research, which to a certain extent solve the conflicts of increasingly massive data and relatively limited storage space. However, with varied data structures, certain inadequacies remain especially in the concurrent quick storage and management of structured and unstructured data, therefore, follow-up research and further discussion are necessarily needed to accommodate the development of big data storage with complex structure.

Data Mining. Data analysis is dependent on the data mining and big data is no exception. Data mining is a process of acquiring interesting patterns and knowledge from mass data; data sources including the database, data warehouse, Web, and other information repositories or the data dynamically flowing into the system.

Data mining itself is not a new technique. Though the 10 existing classic algorithms are presented in Fig. 1, unfortunately, these algorithms cannot be directly applied in the field of big data, as big data is usually stored by the distributed file system in the form of key/value pairs, and the way of reading and calculation is distinct from traditional methods.

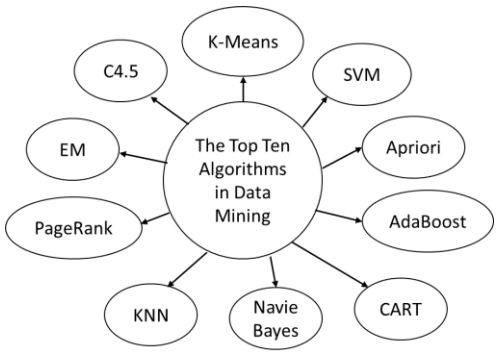

Figure 1. The Top Ten Algorithms in Data Mining

Big data has several key traits, including large heterogeneous structures, different data sources, distributed and dispersed control, complex and constant changing, and knowledge association, etc.. [10] Provides HACE theorem and the big data processing model from the perspective of data mining. The data mining of big data primarily needs a high performance computing platform, which allows fully exploiting the potential of big data; secondly, data mining method is discussed from data, model and system level. For the multi-source and heterogeneous characteristics, [11] describes the Tree-based Association Rules (TARs) to extract useful information from semi-structured documents. [12] Introduces an Analytics-as-a-Service (AaaS) tool for unstructured data, which implements extracting themes and keywords from unstructured data sources, applying the concurrent search and linear search, accomplishing macro tasks with filtering and tagging methods.

Data mining techniques hold an important place in the era of big data. How to combine the original algorithms or interdisciplinary approaches in order to find proper data mining methods should be the

direction of the future development. It is especially need to establish models pertinently in this field, and data mining will have a wide space of development in the future.

Machine Learning. Currently, machine learning problems concerning large-scale data widely exist, but since many of the existing machines learning algorithms are based on memory, while big data cannot be loaded into the computer memory, herein these algorithms cannot be directly process big data. How to put forward new algorithms to meet the requirement of big data processing is one of the hot research directions in the era of big data.

The traditional statistical machine learning method Support Vector Machine (SVM) has two bottlenecks when used in big data classification: (1) compute-intensive, hardly used in large-scale data sets; (2) predictions of fitting model the robust and non-parametric confidence interval are usually unknown. For these problems, [13] provides an online learning algorithm for SVM, dealing with the classification problem of inputting data in sequence. This algorithm is faster and uses fewer support vectors, and has better generalization capacity. How to better apply the classification algorithms to big data environment or improve strategies accordingly becomes the main direction.

Challenges

Data Storage Capacity Problem of Big Data. The development of information technology unceasingly increases the quantity of channels of data production. With the popularity of the Internet and mobile devices and the rise of social networks, data presents a state of blowout. In order to deal with the increasingly huge amounts of data and growingly complex data structures, many companies are working on distributed file systems and distributed parallel database to adapt to the big data era. Designing reasonable hierarchical storage architecture becomes the key of information system.

The Security and Privacy of Big Data. With the development of big data, data analysis is used more widely in various fields. Data contains the records of various behavioral details and even sensitive information, such as spatial and temporal information, consumption habits, etc. For governments, data protection also belongs to the realm of national security, so it is particularly important to ensure big data security and privacy, but traditional data protection methods cannot meet this need.

Information volume and privacy of data is contradictory, and no desirable methods have been found. Meanwhile, a large number of security problems also need settling, such as safety of distributed computing, security of data storage and log management, compulsive access control, etc., which are also the important direction of the future studies of big data.

The Big Data Platform Architecture. Big data platform is to organically integrate the data from different channels, different sources and different structures. Large scale, various types, quick fluidity, dynamic system and great value of the big data is the key consideration of constructing big data platform. Currently the non-relational analysis techniques of non-relational database represented by Map Reduce and Hardtop have been widely applied in the field of big data analysis due to its perfect scale-up capability, which has become the mainstream of big data processing techniques. Presently Spark of Berkeley is open source cluster computing system based on memory computing scalability. Instead of the disadvantages of Map Reduce, i.e. the large amount of network transmission and the low efficiency caused by disk I/O, Spark provides API with much higher layers than Hadoop, which causes the same algorithm’s operating speed 10-100 times faster in Spark than in Hadoop. Spark’s characteristic of efficiently processing distributed data sets enables it to possess a good applying prospect. Even so, Map Reduce and Hardtop still remain unsatisfactory in terms of performance, and distributed parallel algorithms of low complexity and high parallelism still need to be developed and updated according to the practical application, to achieve more efficient and practical data analysis techniques.

Conclusion and Expectation

There is no doubt that big data era has come. This paper illustrates the concept of big data, its 4V characteristics and applying status, introduces its key techniques and analyzes the challenges that we are

facing. In general, the development of big data is still in its infancy, and has great space and potential to explore. How to obtain the useful information from massive, sparse and high dimensional data, and how to efficiently deal with big data remains to be explored and discovered.

Acknowledgements

This work was supported by the national natural science foundation of China under Grant（61503274)

References

[1] Nature. Big Data [EB/OL]. http://www.nature.com/news/specials/bigdata/index.html

[2] Science. Special online collection: Dealing with data [EB/OL]. http://www.sciencemag.org/site/special/data/

[3] B. Huberman, "Sociology of Science: Big Data Deserve a Bigger Audience," Nature, Vol. 482, (2012). p. 308.

[4] J. Gantz, D. Reinsel, The Digital Universe Decade-Are You Ready. Hopkinton, MA, USA: EMC, May 2010

[5] C.Q. Ji, L. Yu, and W.M. Qiu, et al. Big Data Processing in Cloud Computing Environments. 2012 International Symposium on Pervasive Systems, Algorithms and Network. 2012,17-23,DOI:10.1109/I-SPAN.2012.9

[6] Barwick H. The "four Vs" of Big Data. Implementing Information Infrastructure Symposium [EB/OL]. http://www.computerworld.com.au/article/396198/iiis_four_vs_big_data/

[7] G.J. Li, X.Q. Cheng. No.6, p.647-657(in Chinese)

[8] Aghav, S. Database compression techniques for performance optimization. Computer Engineering and Technology (ICCET), 2010 2nd International Conference. No.6, p.714-717

[9] Y.Q. He, Lee Rubio, and H Yin, et al. RCFile: A fast and space-efficient data placement structure in Map Reduce-based warehouse systems//Proceedings of the 2011 IEEE 27th International Conference on Data Engineering (ICDE).Hannover, Germany, (2011). p. 1199-1208.

[10] X.D. Wu, X.Q. Zhu, and G.Q. Wu, et al. Data Mining with Big Data. IEEE Transactions on Knowledge and Data Engineering. Vol.26 (2014) No.1, p.97-107

[11] Mirjana M, Elisa Q, and Letizia T. Data Mining for XML query-answering support. IEEE Transactions on Knowledge and Data Engineering, Vol.24 (2011) No.8, p. 1393-1407.

[12] R. K. Lomotey, R. Deters. Analytics-as-a-service (AaaS) tool for unstructured data mining. In Cloud Engineering (IC2E), 2014 IEEE International Conference on, pages 319–324. IEEE, 2014.

[13] Monika J, Chui M, and Brown B. et al. Big data: The next frontier for innovation, competition, and productivity [EB/OL]. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

[14] Hey T, Tinsley S, and Tolle T. The Fourth Paradigm: Data-intensive Scientific Discovery [EB/OL]. http://reserach.microsoft.com/en-us/collaboration/fourthparadigm

[15] J.W. Han, K, Michelin. Data Mining: Concepts and Techniques, vol. 2, Morgan Kaufmann Publisher (2006)

[16] R. K. Lomotey, R. Deters. Real-Time Effective Framework for Unstructured Data Mining. 11th IEEE International Symposium on Parallel and Distributed Processing with Applications (ISPA-13), 16-18 July 2013(1081-1088), Melbourne, Australia.

[17] Q He, N Li, W.J Luo, Z.Z. Shi. A Survey of Machine Learning Algorithms for Big Data. Pattern Recognition and Artificial Intelligence. Vol.27 (2014) No.4, p. 327-336(in Chinese).

[18] Lau K W, Wu Q H. Online Training of Support Vector Classifier. Pattern Recognition, Vol.36 (2003) No.8, p. 1913-1920.

[19] Franco-Arcega A, Carrasco-Ochoa J A, Sánchez-Díaz G, et al. Building Fast Decision Trees from Large Training Sets. Intelligent Data Analysis, Vol.16 (2012) No.4, p. 649-664.

[20] Zaharia M, Chowdhury M, Franklin M, Shenker S, Stoica I. Spark: Cluster computing with working sets. Hot Cloud 2010.