ATLANTIS PRESS
Advances in Computer Science Research, volume 74
2nd International Conference on Computer Engineering, Information Science & Application Technology (ICCIA 2017)

Research Progress on Data Analysis in Big Data Technology

Muqiao Yang
The Hong Kong Polytechnic University, Hong Kong
mqyanghk@gmail.com

Keywords: big data, data mining, machine learning, artificial intelligence.

Abstract. From the concept of "data natural" in DIKW (Data, Information, Knowledge, Wisdom) basic model and data science of information science, human cognition to the world needs to start from the most basic data. Big data technology as the current popular technical field, has been a wide range of attention and research. In this paper, the concept and composition of big data are summarized, including a systematic introduction to data acquisition and perception, data storage and processing, data analysis, data visualization and big data security and privacy protection. Especially, we give a detail analysis in the data analysis aspect, and finally summed up challenges of the big data technology we are facing right now, and technical nature of the big data research. It can give a guide for the future engineering applications.

1. Introduction

Over the past few years, with the rapid increase of the calculation speed, storage capacity, the degree of intelligence and the decline of the prices, as well as the rapid development of mobile Internet, cloud computing and other technologies, the governments are willing to invest in budgeting to build their own information facilities and to collect and analyze more data, the amount of data appeared in explosive growth.

The big data technology is the integrated application of forecasting, data mining, statistical analysis, artificial intelligence, natural language processing, parallel computing, data storage and so on, which makes the most popular new practice of data engineering application technology.

At present, it is believed that the current big data technology mainly originated in Google in the field of IT. Google engineers in 2003 to 2006 has published academic papers on the MapReduce, GFS and BigTable and other core technology, this series of technology quickly caused great repercussions, which attracted Yahoo, Facebook and other Internet companies' attention, it directly led to the current application of the most extensive open source big data framework for the birth of Apache Hadoop.

There are three major characteristics of big data: data volume, data generation velocity and data variety, based on a large number of data, detailed statistical results on the big data technology connotation, extension, status and technical trends were analyzed. On this basis, the industry also summed up other big data characteristics, such as veracity, low value, viability and so on. In the perspective of BI & A (business intelligence and analysis technology), the big data technology is used as the development direction of the next generation BI & A, and its technical connotation is pointed out. The data technology, data analysis, text analysis, Web analysis, Network analysis and other technologies and the e-commerce and market intelligence, e-government and politics, science and technology, intelligent health and medical, public safety and other areas of application model were analyzed. The author [4] argues that big data technology has been extended from four Vs to three dimensions: real-time, interpretable, data accurate / stable. With the advent of era of data engineering, big data needs to be operated, shared, and then tapped and used, which could produce more social value and solve future problems.

Big data technology involves data sensing, acquisition, storage, processing (management), analysis, visual presentation and many other links, the technical means used in all aspects of endless. This paper will conduct a big data research on the concept of composition, development, especially a comprehensive and in-depth analysis in the data analysis.

2. The Concept and Composition of Big Data Technology

McKinsey's report suggests that big data technologies include techniques such as forecasting, data mining, statistical analysis, artificial intelligence, natural language processing, and parallel computing [5]. IBM's Stephen Watt proposed big data ecosystem model, the big data technology is divided into data generation, data storage, data processing, data sharing, data retrieval, data analysis, data visualization etc. [6].

The common technology of big data technology can be divided into: perception, acquisition, storage, analysis and visualization. The technical fields involved include: sensor, computing network, data storage, cluster computing system, cloud computing facilities, artificial intelligence and data visualization.

At present, the range of the sources of big data is increasingly enlarged, various types of sensors, mobile Internet (mobile phones, all kinds of mobile terminals etc.), and Internet of Things (RFID technology, camera etc.) are all sources of big data collection and perception. Data processing and storage of big data is the most basic and most widely used big data technology. The most famous one is the Apache Hadoop series open source platform, which includes: HadoopCommon, HDFS, MapReduce, Zookeeper, Avro, Chukwa, HBase, Hive, Pig and other subprojects [7]. Data analysis is the core of big data technology, which is also the direct value of the part. Data visualization mainly studies on how to make it close between human perception on data and the natural perception of human visualization. It also studies the visual interaction of data expression to enhance human awareness, exhibiting the implicit information of the data. It aims to explore the law of the data, and it is a comprehensive discipline crossing computer graphics, human-computer interaction, statistics, psychology. In the era of big data, the permission, ambiguity and anonymity of privacy information lost efficiency [8], traditional encryption technology, identity authentication and access control and other means also need the assistance of big data technology, and there are needs for the traditional information security and privacy protection legal framework as well.

3. Data Analysis

Through the results of data analysis, the unknown laws and results can be revealed, and it can help people to make more scientific and intelligent decisions. In the big data analysis, besides the traditional BI technology, many technical methods in the field of artificial intelligence technology for a big data analysis provide a variety of analytical methods, including statistical analysis, machine learning, data mining, natural language processing, knowledge and reasoning etc.

3.1 Data Mining

Data mining is extracting the implied information in which people do not know in advance but is potentially useful from the large, incomplete, noisy, fuzzy, random practical application of data. It is a combination of statistics, database technology and artificial intelligence technology, and it is the use of statistical and machine learning method to extract a set of patterns of technology through the database management system. Common methods of data mining include association rule learning, cluster analysis, classification analysis, sequence analysis, deviation detection, predictive analysis, pattern similarity mining and regression analysis.

3.2 Statistical Analysis

Statistical analysis is a science of collection, organization and interpretation of the data based on the statistical principles of mathematics. The statistical method is mainly used to analyze the quantitative relationship between the variables. Typical methods are A / B test etc. [5].

In this area, the classic statistical analysis tool is the R language toolkit. R language is a language based on Scheme and S language, developed by Prof. Ross Ihaka and Robert Gentleman of the University of Auckland, New Zealand, for the convenience of statistical courses. R is an open source statistical analysis software that provides a wealth of classical statistical analysis algorithms and mapping techniques, including linear and nonlinear models, statistical tests, time series, classification, clustering and other algorithms, achieving a lot of classic, modern statistical algorithms.

3.3 Natural Language Processing

Natural Language Processing (NLP) is the technology of the use of computer algorithms for human natural language analysis based on computer science and linguistics, belonging to the field of artificial intelligence. The key technologies involve lexical analysis, syntax analysis, semantic analysis, speech recognition, text generation and so on. Many natural language processing algorithms are based on machine learning methods. A typical application in this area of technology is the analysis of language emotions based on social media, electronic detection in the legal field, and other applications including fraud detection, text categorization, information retrieval and filtering, text conversion systems, and machine translation.

3.4 Machine Learning

Under the big data background, the main application areas of machine learning can be summarized into three aspects: search, iterative optimization and graph calculation. Machine learning, as one of the important contents of the field of artificial intelligence, is divided into two categories: supervised learning and unsupervised learning. The supervised learning required the user who used algorithm to know what to predict (ie, the classification information of the target variable). It mainly uses the classification and regression algorithm, if the predicted target value is discrete (if yes / no, A / B / C etc.), it will be suitable for classification algorithm, such as k-nearest neighbor algorithm, decision tree algorithm, naive Bayesian algorithm, support vector machine algorithm, AdaBoost algorithm etc.; if the predicted target value is continuous value (such as 0 ~ 100, 0.1 ~ 150), it will be suitable for regression algorithms, such as Logistic regression, CART algorithm (classification regression tree algorithm) [9].

The current research focus in this field is the use of new machine learning algorithms to achieve deep machine learning. Depth learning is the development of artificial neural network, its essence is through the construction of a lot of hidden layer of machine learning model and massive training data to learn more useful characteristics (relative to the use of traditional machine learning algorithm for shallow machine learning). Thus, it is to improve the accuracy of the final classification or prediction [10]. In the areas of deep learning, Google, Microsoft, IBM, Baidu and other enterprises went at the forefront. The project is based on the Google Brain project, which builds a parallel computing platform with 16,000 CPU cores for training the machine learning model of Deep Neural Network (DNN). The model is based on speech recognition and image recognition and achieve a great success.

4. Conclusion

From the perspective of technological development, the large, diverse, high-speed and complexity of data and the resulting data management and computational storage scalability issues are not new problems in the IT industry. From the goals of data analysis and demand, it is only another new, bigger dataset, breaking the technical conditions of the data processing limit. Therefore, the challenge of big data is to break through the existing data storage, processing, analysis, presentation of technical limitations, which is not a new proposition for the IT sector. Therefore, we should put forward the corresponding strategic research, establish our own data science system, put forward the relevant standards and patents, and finally to form the industrial advantages and technological advantages of this field.

References

[1] Philip Russom. Big Data Analytics. TDWI Best Practices Report [R]. USA: TDWI, 2011.
[2] Paul Zikopoulos, Chris Eaton, Dirk de Roos etc. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data [R]. USA: Mc. Graw-Hill, 2012.
[3] Hsinchun Chen, Roger H. L. Chiang, Veda C. Storey Business Intelligence and Analytics: From Big Data to Big Impact [J]. MIS Quarterly, 2012, 36(04): 1165-1188.
[4] Che P J. The Three Dimensions and Ten Commandments of Big Data [EB/OL]. (2014-03-07) [2014-05-10].

[5] James Manyika, Michael Chui, Brad Brown, etc. Big data: The next frontier for innovation, competition, and productivity [R]. USA: McKinsey Global Institute, 2011.
[6] LI Ming. The Innovators of Big Data Ages [EB/OL]. (2011-11-02) [2014-06-04].
[7] L Jun. Hadoop Big Data Processing [M]. Bei Jing: Posts & Telecom Press, 2013: 45-60.
[8] Victor Mayer-Schonberger, Kenneth Cukier. Big Data: A Revolution That Will Transform How We Live, Work and Think [M]. HangZhou: ZheJiang People Publishing House, 2013: 193-232.
[9] Vinayak Borkar, Yingyi Bu, Michael J. Carey, etc. Declarative Systems for Large-Scale Machine Learning [EB/OL]. (2012-04-25) [2014-05-20].
[10] Y U Kai, J Lei, Y Q Chen. The Yesterday, Today and Tomorrow of Deep Learning [EB/OL]. (2014-06-07) [2014-06-18].