Research on the Development of Data Scientific Analysis Tools in the Big Data Age

Yuxuan Xu
Department of Economics, New York University, New York, United States of America
email:yx1151@nyu.edu

Keywords: Data science; R language; big data

Abstract. According to the features of big data era, this paper analyzes the main challenges that massive data bring to the analysis tool of data science. The paper introduces the big data analysis tool in response to challenges. Then, the paper carries on the comparative analysis of R language, Rapid Miner and Mahout 3 popular analysis tools of big data in data science, which finds that R language and Rapid Miner have fully functions and the Mahout has more outstanding analysis capability of big data. Finally, the paper points out the development trend of data science analysis tool.

Introduction

The data explosion brought humanity into a big data age. The popularity of big data has led to enthusiasm for big data such as government, academia and industry [1]. "Nature" and "Science", among other top international academic journals, have published special journals to discuss big data. In 2008, "Nature" introduced the issue of "Big Data," which introduced the challenges of vast amounts of Data from Internet technology, environmental science and Internet economics. "Dealing with Data", the publication of the 2011 "Science" publishing Data, explored the challenges of the Data deluge. Government agencies are also paying close attention. In March 2012, the US unveiled "Big Data Development Plans"; The European Union in the past few years for scientific data infrastructure investment were more than 100 million euros, and the data information infrastructure construction was as one of the priorities of Horizon 2020 program.

The big data boom has spurred researchers to consider data science [2-3]. In February 2014, the "scientific data conference" was held in Beijing, the conference took "scientific research data and data science" as the theme, that discussing the big data age scientific research data management, sharing and application of the new trends, and researching the key issues and challenges facing big data, in order to explore the scientific connotation and the development direction of the scientific data.

Data science is a new interdisciplinary field in many fields, such as information science, Internet science and economics, and it is still in the early stages of development [4]. This paper analyzes the development of the subject of data science, and introduces the challenge of data explosion to the data analysis tool in data science and the large data analysis tool which can deal with the challenge. However, not all of the tools are fully functional, they have their own characteristics and advantages and disadvantages, then selected the R language, Rapid Miner, Apache Mahout three major large data analysis tools, an overview of the characteristics of the tool.

Data Science

The boom in big data has led to a new discipline known as data science[5]. Data science is in the early stages of development, a growing discipline. The core of data science involves using automated methods to analyze vast amounts of data and extract knowledge from them. In almost all areas of knowledge discovery, data science provides a powerful new way to discovery, it has a lot of data but don't know how to extract value from the data of company providing a new source of insight. With this approach to automation, data science is helping to create new branches of science and influence the social sciences and the humanities.

Data science is a blend of the multi-discipline science and based on theory and technology of these disciplines, including mathematics, probability model, statistics, machine learning, data warehouse, visualization, etc [6]. In practical applications, the data science includes data collecting, cleaning, analysis, visualization and data applications throughout the iterative process, finally helps to make the correct development decisions. The practitioners of data science are called data scientists [7]. Data scientists are interdisciplinary talents who have widen the vision. They are with solid data science foundation, such as mathematics, statistics, computer science, etc., and have a wide range of business knowledge and experience. Data scientists through intensive technical and professional knowledge in some scientific disciplines solve the problem of complex data, so as to make a plan for the large data of different decision makers and policy. They are considered "sexiest" professional talents in the 21st century [8].

The Main Challenges Facing Data Science Analysis Tools

Big data as a very important aspect in the science of data, for the development of science and education provides a huge opportunity. At the same time, it also brought great challenge to the frontier science project.

The Diversity of Data Formats. In the age of big data, the volume of data is growing rapidly and the format of the data is diverse. For example, the data in a bank or a supermarket is a text format, and the data in YouTube is the image video format, and the data for the digital phone is the voice format. Data form, in addition to the traditional relational data, includes from the web, email, social media BBS, Internet search indexes, log files, such as the original, unstructured and semi-structured data. Thus, in the face of such a large amount of data and various forms of data, data analysis tools are able to deal with the method of structured data and new methods of unstructured data.

The Failure of Traditional Data Algorithms. When analyzing data, a better algorithm for data mining and classification clustering is required. Clustering algorithm is not, however, logarithmic curve type (N log N) or the size of the linear sequence (N), but the typical N cubic size, as N gets large, some methods will fail. As a result, many traditional algorithms lose effectiveness in the face of massive data processing. We need to invent new algorithms and require the algorithm to scale well to address petabytes of data. In addition, much of the big data has the characteristics of real-time application. In this case the primary indicator of the big data applications will no longer be accuracy of the algorithm, and the algorithm needs to achieve a balance between accuracy and real-time performance, such as online machine learning algorithms.

Large Scale Data Visualization. Visualizations are one of the most effective ways to explain vast amounts of data. Visual analysis through interactive interface support not only monitors and validates predictions, but also finds content that is not anticipated. Effective visualization tools are based on mathematical roots and robust algorithms, just like data analysis. However, huge amounts of data brought many challenges to visualization technology. Large data visualization is mainly in the face of the fusion of different scale multiple heterogeneous data, and task complexity scalability and extensibility to interactive challenges. In addition, there are a number of challenges such as in-situ analysis, algorithms, parallelization, data movement, and uncertainty, as well as quantitative, transmission and network architectures. Therefore, we need to promote significant progress in visualizing technologies to support the extraction of meaning from large and complex data sets.

Data Science Analysis Tool

Faced with the explosive growth of data to the scientific project to bring great challenges, it needs to be more good at developing related technologies and tools to support from data collection, data management to data analysis and data visualization of the entire data processing cycle. Through the unremitting efforts of researchers, the development of technology and tools continue to introduce new. In the data storage, the company represented by Google, respectively, developed their own NoSQL database, such as Google's BigTable, VMware's Redis, Microsoft's Azure Tables, etc.,

successfully solved the different format data storage and management issues.

In terms of data analysis, Google created the distributed programming model MapReduce, which is a large data analysis, and implemented parallel computing. With the development of the data from the scientific analysis tool is a variety of forms, and the most popular is undoubtedly Yahoo's open source project Hadoop to distributed file system (HDFS) and MapReduce as the core of Hadoop to provide users with the underlying details of the system transparent distribution type infrastructure. In addition to the Hadoop, many of the large data analysis oriented scientific data analysis tools, such as HPCC, R language, Storm, Apache Drill, Rapid Miner, Mahout, etc., have a plenty of big data analysis specific to a particular application, and have a plenty of full analysis of the platform.

Major Data Analysis Tools. With the development of these scientific data analysis tools, on the one hand, they successfully solved algorithm in data science failure, large scale data visualization and a series of challenges. On the other hand, there are characteristics and advantages and disadvantages. Mahout, for example, has excellent data processing capabilities, not only processing data volumes large and fast, but poor visibility. Then select R language, Rapid Miner, Mahout three mainstream scientific data analysis tool, the overview and in the form of a table have carried on the comparative analysis to the three main features. The basic situations are shown as follows.

(1) R language is a programming language and environment for statistical computing and drawing, using the command line works, and free issued under the GNU agreement. Its source code is available for free download and use. R CRAN website offers a wide range of third-party packages, and the content covers the economics, sociology, statistics, bioinformatics and so on, that is why more and more people from all walks of life love R. The researchers focused on the integration of R language and Hadoop with the poor scalability of traditional analysis software and the weak analysis of Hadoop. As an open source statistical analysis software, R integrates data computing to parallel processing through the deep integration of R and Hadoop, enabling Hadoop to gain a strong analytical capability.

(2) Rapid Miner, formerly known as Yale, is an open source computing environment for data mining, machine learning, and business forecasting analysis. It can be done on a large scale in either a simple scripting language, or in Java, API, or GUI mode. Because it has GUI features, and it is easier for beginners to start with data mining. Rapid Miner 6 has a friendly and powerful toolbox, providing fast and stable analysis, and can design a prototype in a short time, makes the key decisions in the process of data mining as early as possible. It also can help to reduce customer loss, conduct emotional analysis, predictive maintenance, and marketing.

(3) Apache Mahout was originated in 2008, its main goal is to build a scalable repository of machine learning algorithm, and it provides some classic machine learning algorithms, designed to help developers more convenient and quick to create smart applications. Currently, Mahout projects include frequent subitem mining, classification, clustering, and recommendation engines. Mahout currently supports two ways of classifying content based on bayesian statistics. The first is to use the simple Naive Bayes classifier that supports map-reduce. The Naive Bayes classifier is known for its high accuracy and speed, but it assumes that the data is completely independent; The second is Complementary Naive Bayes, which corrects some of the flaws in the Naive Bayes method, while maintaining the simplicity and speed of the former.

Characteristic Analysis. In view of the era of big data challenge for scientific data analysis tool, and the performance of the tool requirements, this paper selects the model algorithm, visualization, big data processing ability and a series of tools characteristic index, and the form of a table for three kinds of tools are analyzed and compared, as shown in Table 1.

In Table 1, several indicators of three tools are compared and analyzed. As shown in table 1, the R language is an open source programming language and software environment with a comprehensive capability. R language supports most of the model algorithms and supports data for different formats, data sources. R language can nnot only analyze the large data set, but also have excellent drawing function, which is a popular data analysis and visualization tools.

Table 1 Comparison analysis of tools

Index	R language	Rapid Miner	Mahout
Support platform	Able to run on different systems of computer systems	Able to run on all major platforms and operating systems	Able to install and configure platforms such as Linux, Windows, and Mac OSX
Model algorithm	Supports most mining algorithms	Support the majority of classification, clustering and association analysis of the model algorithm	Some data mining algorithms are implemented
Data format compatibility	Good	Good	Poor
The running speed	Slow	Fast	Fast
NoSQL database	Support	Support	Support
Secondary development	Not support	Not support	Support
Visualization	Strong	Strong	Weak
Big data processing	Support	Support	Support

Rapid Miner is an easy-to-use visualization predictive analytics software that is easy for regular users. On one hand, Rapid Miner 6 has escaped the shortcomings of older versions in big data processing, and has good data analysis capabilities through integration with Hadoop. On the other hand, the visuals are excellent, with a 3D diagram that is not supported by R language and Mahout.

Because Mahout is an algorithmic framework for Hadoop based data mining and machine learning, the advantages of Hadoop are the advantages of Mahout. Mahout shows obvious advantage in large data set analysis, makes full use of the parallel computing capacity of graphs, the large number of training samples, complete computing tasks efficiently. But the downside is clear: there are fewer algorithms that Mahout currently supports, although the algorithm has been increasing; Second, Mahout visualizes poor performance.

The Development Trend

Through the above introduction and comparison of data science analysis tools and big data era the requirements of the characteristics of the tool, this paper believes that scientific data analysis tools mainly has the following trends:

Large Data Set Analysis. The big data age undoubtedly requires that data science analysis tools be capable of analyzing vast amounts of data. Second, data value is closely related to data capacity and variety. Generally speaking, the larger the data capacity, the more the content, the greater the amount of information it contains, and the greater the potential value of mining. In order to achieve the full data analysis so as to explore new and valuable insight, the scientific data analysis tools are required to comprehensively analyze mass data and format a variety of data.

Excellent Visualizations. Data analysis is the core of data processing steps. However, if the analysis of results is without using the appropriate method to explain correctly, the result will let users is difficult to understand. Visualizes the analysis results effectively and makes it easier for people to receive critical information from the data analysis tool. In the era of big data, the data quantity is big and complicated, not only help people intuitively find the data contained in the information and knowledge, and the visualization is one of the most effective way.

Data Analysis is Distributed Primarily. Big data age, alone in the past alone a single data analysis tool has been unable to master the analysis of massive data, the use of distributed architecture to improve the scalability of the system has become inevitable. There is no doubt that

Hadoop has become the king of today's large data processing technology. Distributed processing technology greatly improves the efficiency and speed of data analysis, such as Mahout and other distributed large data processing tools will replace the traditional tools, occupy the main position.

Conclusion

Big data era, if can be more effectively organized and data used, people will get more opportunities to play a great role to social development of science and technology. Therefore, the continuous development of data analysis tools, can efficiently and accurately unearthed data contains potential value, which is to measure the value of the data analysis tools, also is the key to data science.

The wave of big data has promoted the evolution of "data science" into a separate discipline. At present, the definition of data science has not yet clear, but over time, data science will become a specialized subject, and has perfect theory basis and technology disciplines, cognitive by more and more people. Universities will also set up specialized data science majors to create new jobs related to them. In the next few years, the data scientist must be a shortage of people in all walks of life.

References

[1] Li Jinchang. Statistical Measurement as Foundation: From Statistics to Data Science [J]. Statistical Research, 2015, 83(8): 3-9.

[2] Xu Hao, Qin Yue, Huang Lan. Data science curriculum for education [J]. Computer Education, 2016, 8: 158-162.

[3] Chen Zhenchong, He Tiantian. Data science: the demand and development of talents [J]. Big Data Research, 2016, 5: 95-106.

[4] Li Guojie, Cheng Xueqi. Research Status and Scientific Thinking of Big Data [J]. Bulletin of Chinese Academy of Sciences, 2012, 6: 647-657.

[5] Wang Yuefen, Xie Qingnan, Song Xiaokang. Review and Prospect of Overseas Research on Data Science [J]. Library and Information Service, 2016, 14: 5-14.

[6] Wang Xiaofan. Data Science and Social Network: Big Data, Small World [J]. Science and Society, 2014, 1: 27-35.

[7] Niu Yanfang, Deng Xuemei, Chen Wei. The application of the R language of data science tool in audit data analysis [J]. The Chinese Certified Public Accountant, 2016, 9: 93-97.

[8] Liu Dehuan, Li Xuelian. The Danger Trend of Data Ecology and the Possibility of Data Science [J]. Modern Communication(Journal of Communication University of China), 2016, 1: 21-27.