A review of the theories and methods of data analysis and visualization

MA YAO QIANG [a,*], GONG XUE YU[a,1]
[a]School of Electrical Engineering, University of South China, Hengyang, Hunan 421001, China;
[a]School of Electrical Engineering, University of South China, Hengyang, Hunan 421001, China
[1]gongxueyu@126.com
[*] Corresponding author: 007568@gmail.com

ABSTRACT

In this paper, under the background of big data era, the theory and operation methods of data analysis and visualization and their important value in the field of data science are described, and the relationship between them is necessary. Besides, the convenience and shortcomings of using python as a tool for data analysis are analyzed. Finally, the development trend of data analysis and visualization is prospected.

Keywords: Data science, Data analysis, Data visualization, Relationships, Python

1. INTRODUCTION

With the vigorous development of data science, the ability of humans to analyze data has lagged far behind the ability to obtain data, and the amount of data that is more complex and diversified makes the cost of understanding higher. Therefore, how to extract useful information from massive data and how to transform data into information that people can quickly understand has become one of the hottest research directions in the field of data science today. Data analysis plays an important role and role in the whole of data science because it is a critical step in extracting value from data. Data analysis extracts value information through the summary of data, and then achieves the purpose of data science, and data visualization visualizes the analysis results according to different expressions. This paper provides a comprehensive review and summary of data analysis and data visualization.

2. DATA ANALYSIS

Data science includes the whole process of data processing such as data collection, data management, data governance, data analysis, data visualization, data ethics and data application[1], in which data analysis is based on the process of analyzing and summarizing data for a certain purpose. Its significance lies in the extraction and refinement of the information hidden in the data, in order to help people find the internal laws of the object of study, or the laws of the occurrence, development and future change of things, and then help people make judgments and correct decisions.

2.1 Data, Information and Data Analysis

Data refers to the symbols that record and can be identified by objective events, and are physical symbols or combinations of physical symbols that record the nature, state and interrelationship of objective things. It is a recognizable, abstract symbol[2].

Data as a carrier of information, that is, the main information contained in the analysis data, that is, to analyze the main characteristics of the data, that is, to study the numerical characteristics of the data[3], features can exist in symbols, text, numbers, speech, images, videos, etc. Data focuses on data acquisition, cleaning, pre-processing, analysis and mining, graphics focus on solving the problem of receiving optical images, extracting information, processing transformation, pattern recognition and storage display, and visualization focuses on the solution of converting data into graphics and interactive processing.

Information is the connotation of data, and information is loaded on top of data to interpret data with meaning.

Third International Symposium on Computer Engineering and Intelligent Communications (ISCEIC 2022),
edited by Xianye Ben, Proc. of SPIE Vol. 12462, 1246217 · © The Authors. Published under a
Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.2660780

2.2 The difference between data analysis and data mining

From a broad perspective, data analysis covers two parts: data analysis and data mining, see Figure 1. From a narrow point of view, there are differences between data analysis and data mining, which are mainly reflected in the definition description, focus, skill requirements and final output form of the two, see Table 1.

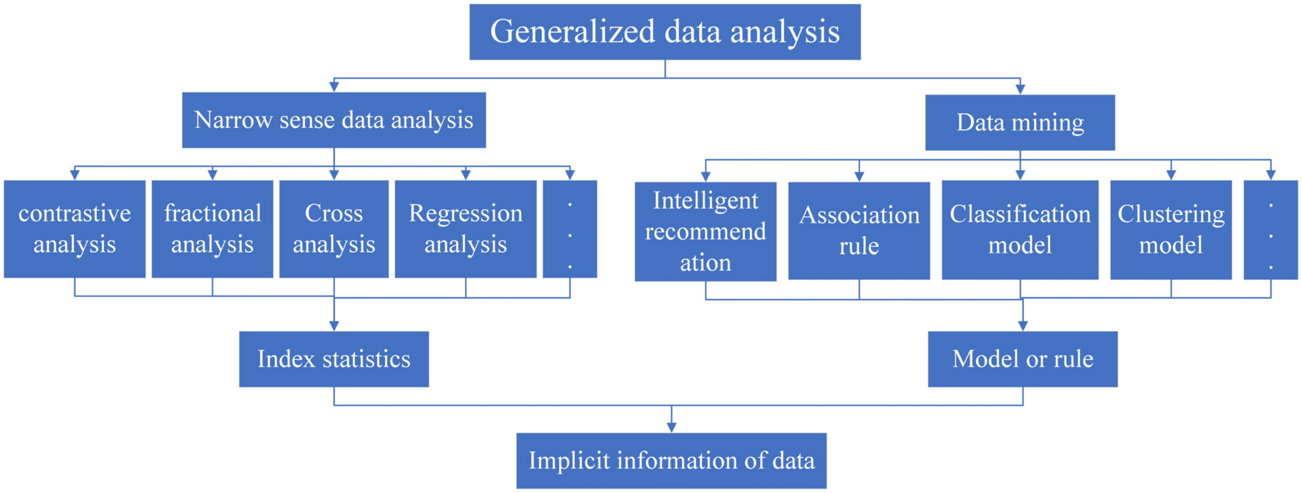

Figure 1. Relationship diagram between data analysis and data mining

Table 1. Data analysis and data mining difference table.

Angle of difference	data analysis	data mining
definition	Describe and explore the analysis to assess the current situation and correct deficiencies in techniques	A technical "mining" process that uncovers unknown patterns and patterns
Emphasis	A technical "mining" process that uncovers unknown patterns and patterns	A technical "mining" process that uncovers unknown patterns and patterns
skill	Statistics, databases, Excel, visualizations, etc.	Excellent mathematical skills and programming skills
outcome	Statistical results need to be interpreted in conjunction with business knowledge	Model or rule

2.3 Data analysis process

The data analysis steps are shown in Figure 2.

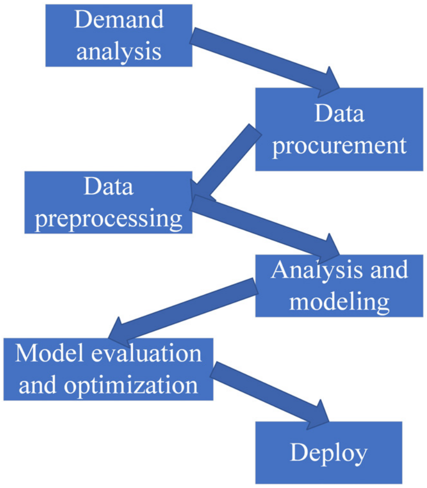

Figure 2. Relationship diagram between data analysis and data mining

3. DATA VISUALIZATION THEORY AND METHODS

Data analytics is exploratory, and it requires curiosity, a desire to find answers, and good resilience, because these answers are not always easy to come by. Data visualization, that is, a visual display of data. Effective visualization can significantly reduce the time it takes for people to process information and gain valuable insights. The terms data analysis and data visualization are inextricably linked. When actually processing data, data analysis precedes visual output, and visual analysis is a good way to present effective analysis results.

From a larger perspective, the technical study of data visualization is actually the study of visual representation. "A piece of information extracted in some form of summary, including the various properties and variables of the corresponding unit of information"[4], which defines the visual representation of the data.

The basic idea of data visualization technology is to represent each data item in the database as a single graph element, a large number of data sets constitute a data image, and at the same time, the various attribute values of the data are represented in the form of multidimensional data[5], and the data can be observed from different dimensions, so that the data can be more deeply observed and analyzed.

The main methods of data visualization are: area or size visualization, color visualization, graphic visualization, spatial visualization, and concept visualization[6].

4. COMMON TOOLS FOR DATA ANALYSIS AND VISUALIZATION

The commonly used tools for data analysis and visualization include MS Excel, R language, python language, SAS language, etc. Here, they are introduced in detail.

Excel is a familiar spreadsheet software that has been widely used for more than two decades, and even a lot of data is now available only in the form of Excel tables. In Excel, it's easy to highlight a few columns and make a few charts, so it's easy to get a rough idea of the data. Excel's limitations are in the amount of data it can handle at once[7], and unless you are familiar with VBA, Excel's built-in programming language, it can be tedious to recreate a chart for different data sets.

R is a language and operating environment for statistical analysis and mapping developed by Ross Ihaka and Robert Gentleman of the University of Auckland, New Zealand, and is a free, free, open source software belonging to the GNU system, and an excellent tool for statistical calculations and statistical mapping[8].The main functions of the R language include data storage and processing systems, stop-and-go computing tools (which are particularly powerful in vector and matrix operations), complete and coherent statistical analysis tools, excellent statistical mapping functions, simple and powerful programming languages, and manipulatable data input and output.

Python is a simple and easy-to-understand programming language that is concise, easy to read, and easy to maintain. Python was originally mainly used in system maintenance and web development, but with the advent of the era of big data, as well as the development of data mining, machine learning, artificial intelligence and other technologies, Python has entered the field of data science. Python also has a variety of third-party modules that users can use to complete tasks in data science[9].

SAS is one of the world's largest software companies and was developed in 1966 by NORTH CAROLINA State University[10]. SAS organically combines data access, management, analysis and presentation with powerful features, comprehensive statistical methods, completeness, novelty and ease of operation.

In addition to the data visualization function module included in the data analysis and mining tools, there are also some dedicated visualization tools that provide more powerful and convenient visual analysis functions. At present, commonly used professional visual analysis tools include Power BI, Tableau, Gehpi and Echarts.

5. PYTHON DATA ANALYSIS AND ADVANTAGES

Python is an interpreted, object-oriented, dynamic data type of high-level programming language, it as the preferred data analysis language for data analysts, but also as the preferred language for intelligent hardware has the following characteristics: simple and easy to learn, set of interpretive and compiled in one, object-oriented programming, extensibility and embeddability, program portability, free and open source.

Python's only drawback is that it doesn't execute fast enough compared to C and C++, because Python doesn't compile the code into the underlying binary code; But Python has the characteristics of embedding, for large programs, it is completely possible to use multi-language mixing strategy, for modules that need to run faster, such as image processing, you can use C language programming[11], where the performance requirements are not very high, you can use Python programming, when you need his image processing, Python program sends the code to the Python interpreter in the internally compiled C code, so that the comprehensive development efficiency and performance are the highest.

Common Python class libraries are: Numpy, SciPy, Pandas, Matplotlib, Seaborn, Scikit-learn.

Python with its own incomparable advantages, is widely used in the field of data science, and gradually derived into the mainstream language, its data analysis has very obvious advantages, including: simple syntax, high readability; Have strong general programming ability; Easy docking of other languages; It has a large and active community of scientific computing[12].

6. CONCLUSIONS

In general, data analysis and data visualization are symbiotic and together form an important part of data science. Good data visualization can make data analysis more effective, and with python's data analysis and visualization common class library, it is more convenient to analyze data and visualize operations. Data analysis and visualization help us strengthen the interpretation and understanding of data information, making it easier for us to obtain information from data, but the categories of data in the future will be more complex and diverse, and data analysis and data visualization will play a greater role in the field of data science in the future.

REFERENCES

[1] YANG Jing, WANG Xiaoyue, BAI Rujiang, etc. Current situation and development trend of data science analysis tools under the background of big data [J]. Intelligence Theory and Practice, Papers 38(3): 134-137 (2015).

[2] ZHENG Zhihong, FANG Haiguang, KONG Xinmei, HONG Xin. An Exploration of Big Data Public Service Model for Regional Education[J].China Education Informatization, Papers 28(01): 18-30 (2022)..

[3] FAN Jincheng, MEI Changlin. Data analysis[M]. Science Press (2002).

[4] LIU Kan, ZHOU Xiaozheng, ZHOU Dongru. Research and development of data visualization[J]. Computer Engineering, Papers 28(8): 1-2 (2002).

[5] ZHENG Chen. Comprehensive data analysis of bridges based on data visualization technology[J]. Shandong Communications Science and Technology, 113-115 (2020).

[6] LONG Chun. The application of teaching evaluation based on visualization technology in high school English listening and speaking classes[J]. Campus English, 139-140 (2021).

[7] XU Yan. Performance optimization and visualization tool development of MapReduce in Hadoop[D]. Beijing Jiaotong University (2016).

[8] ZOU Zongfeng, ZHU Yanke, XIA Yun, ZHANG Ying, LIU Maoling, JIN Juan, JIE Jianwang. ARP analysis program design based on R language[J]. Journal of Mathematical Medicine, Papers 26(04): 468-470 (2013).

[9] Isaac Sacolick. How to choose the right data visualization tools for your apps[J]. InfoWorld.com (2021).

[10] SHI Hailiang, WANG Yuanzheng. Some Thoughts on the Curriculum of "Computer Programming" for Non-Computer Majors[J]. Electronic Design Engineering, Papers 22(08): 15-17+20 (2014).

[11] FENG Guilian. Common Python Data Analysis Method Based on Deep Learning[J]. Journal of Physics: Conference Series, Papers 2037(1) (2021).

[12] PU Yunpeng. Research on Python-based Data Visualization in the Era of Big Data[J]. Information and Computer (Theoretical Edition), Papers 33(23): 179-182 (2021).