Advances in Social Science, Education and Humanities Research, volume 555

Proceedings of the 1st International Conference on Education:

Current Issues and Digital Technologies (ICECIDT 2021)

Construction of "Introduction to Big Data" Course Towards Application

Fangxiao Zhou[1,*]

[1] *Panzhihua University, Panzhihua, Sichuan, China*
[*]*Corresponding author. Email: zhoufangxiao2002@163.com*

ABSTRACT

With the rapid increase of society's demand for Big Data talents, cultivating data engineers with qualified professional abilities is the core task of Big Data teaching in applied undergraduate colleges. The course of "Introduction to Big Data" plays an important guiding role in cultivating students' professional ability. On the basis of analysing the ability requirements of data engineers, the paper discusses how to highlight the ability training of data engineers in the aspects of teaching material, experiment organization and case design.

Keywords: *Application-oriented talents, Course construction, Big data.*

1. INTRODUCTION

The purpose of applied undergraduate colleges is to train applied talents for regional economic development. Compared with academic talents, applied talents need more ability of application, which means how to apply the theoretical knowledge they have learned to practice and solve practical problems. Therefore, the opening of new majors in applied colleges should highlight two points: the talents cultivated should focus on "use", and the direction of service should be oriented towards local economic development. Since the Ministry of Education in China established the "Data Science and Big Data Technology" undergraduate program in 2016, more than 600 universities in my country have established this program. Among these universities, applied undergraduate colleges account for the majority. Therefore, for Big Data majors, application-oriented curriculum construction is very necessary. Under the guidance of this idea, the professional curriculum and the content arrangement of core courses should also highlight the characteristics of application first.

"Introduction to Big Data" is an introductory course for Big Data majors and a professional leading course, so it is an important course in the entire professional curriculum system. Qin [1] believes that the goals of the introductory course for Big Data majors are twofold: one is to expand a broad horizon and cultivate a strong interest, and the other is to lay a solid foundation and prepare for subsequent courses. Chao [2] pointed out that the cultivation of data ability is the core task of the introduction course construction. Therefore, the introductory course of the Big Data major should stimulate students' interest in the major, and show students an overview and framework of the entire major. In the learning process, students should pay attention to the cultivation of data ability. For applied colleges, on the basis of the above principles, how to achieve the application-first goal of the introductory course of Big Data majors requires further discussion and analysis.

2. APPLICATION-ORIENTED BIG DATA TALENT DEMAND ANALYSIS

The current domestic and foreign recruitment websites give many different classifications of Big Data professional positions. These positions can be roughly divided into three types of professional roles: data scientist, data analyst, and data engineer. The job responsibilities and required qualities of these three roles are as follows: data scientists are engaged in the most cutting-edge work in the field of Big Data technology, so they need to possess strong mathematical, statistics, and data modeling theoretical literacy. They also need to be proficient in machine learning and data mining algorithms. Data scientists use dynamic technologies such as machine learning to obtain unique and forward-looking insights about data, which will bring a significant impact on the company's proactive decision-making process. Data analysts extract information from a given data pool and perform data analysis. Data analysts use static modeling techniques to analyze data through descriptive statistics and summary statistics, and present the analysis results to decision makers. Data analysts will not directly participate in the decision-making process, but indirectly help the company's decision-making by providing statistical analysis. The data engineer is responsible for the construction and maintenance of the company's Big Data platform and architecture. Using platform data pipelines, data engineers process and store data. This is a highly skilled position that requires experience and skills in Big Data technology and computer technology. After comparing the job responsibilities of these three types of roles, it can be seen that data engineers are mainly responsible for data acquisition, preprocessing and storage. Their work has laid the foundation for various data operations and data analysis for data analysts and data scientists. And the excavation work provides a guarantee. Deep-level analysis and mining of data is not the job responsibilities of data engineers. On the contrary, these tasks are the focus of data analysts and data scientists. From the perspective of competence, data engineers need more technology and skills to solve practical application problems, while data scientists need more research and innovation capabilities, and data analysts are somewhere in between.

Through the above analysis, data engineers belong to the application-oriented talents who are engaged in data implementation and data management in the first line of enterprise data production, and solve the actual problems of the enterprise. From the perspective of the talent training standards of applied colleges, data engineers should be the main goal of talent training for Big Data major in applied colleges [3].

The professional competence of applied talents is a prerequisite for them to engage in social occupations. Although the training of applied talents' professional competence is a systematic project, generally speaking, the curriculum is the main channel of professional competence training [4]. Specific to the Big Data introduction course, because it is a Big Data professional guidance course, it plays an important leading role in the cultivation of students' professional ability.

3. CURRICULUM KEY CLUES DESIGN

3.1 Awareness of the Data Concepts from Multiple Perspectives

Although "data" is the most commonly used term in daily life, it is a very difficult task to make it clear in class. "Data" is the core concept of the introductory course, and it can be understood from multiple perspectives.

3.1.1 The Feature Engineering's Perspective

How to convert the actual objects in the real world into data that can be processed by the computer, or how the machine understands the actual objects, is a key step in processing data. It is also a topic of particular interest to students. By assigning the concept of "feature" to the data, the object can be transformed into a vector with multiple dimensional features. Vector is a common tool in mathematics, and it becomes easy to understand by performing various calculations, storage, and expressions on it. For example, the fruit object can express fruit data through characteristics (attributes) such as color, shape, weight, and texture. Different fruits (such as apples and bananas) will have different eigenvalues, so they will get different eigenvectors. By calculating the difference in feature vectors, the machine can distinguish apples from bananas. The concept of data characteristics is the key to students' understanding of the process of transforming specific things into abstract data. Furthermore, it can also help students understand what the dimension of data is and what the space of data is. The Bag of Words technology, which is widely used in engineering practice, is a concrete manifestation of this concept. Therefore, guiding students to understand the concept of data characteristics in the introductory class is of great benefit to cultivating students' data application ability.

3.1.2 The Simple Statistics Perspective

Given a set of messy data, some basic indicators in statistics can be used to describe this set of data, so as to quickly establish a preliminary understanding of this set of data. The basic statistical indicators are mean, median, mode, variance, quartile, skewness, kurtosis, etc. The central tendency of this set of data can be understood through the mean, median, and mode. The degree of dispersion of this set of data can be understood through variance and quartile. The distribution shape of this set of data can be understood through skewness and kurtosis. These indicators are easy to understand and easy to demonstrate, so that students can get an overall understanding of this set of data. In addition, visualization tools such as histograms, pie charts, two-dimensional curves, and three-dimensional surfaces are also very intuitive to display data distribution. These methods can enable students to have an intuitive understanding of data and cultivate students' intuition about data.

3.1.3 The DIKW Model Perspective

It is required to discriminate and analyze the relationship of Data, Information, Knowledge and Wisdom. Data is the original record of information; information is valuable data after processing; knowledge is the further improvement of information, which is more systematic and theoretical information; and wisdom is the collection of knowledge. Data, information, knowledge and wisdom constitute the four-layer structure of the DIKW pyramid in turn, and each layer is the enrichment and refinement of the bottom layer. From bottom to top of the pyramid, the value manifestation also increases level by level. The wisdom at the top level of the tower is the ultimate goal pursued by the value of data. As one of the 4V characteristics of Big Data, the Value characteristics, students find it difficult to understand. The DIKW model answeres this question in detail. Combined with practical applications in data-driven company management, the intelligence layer corresponds to the company's strategic decisions, the knowledge layer to tactical decisions, and the information layer to the specific implementation process.

3.2 Learning from Data to Big Data

3.2.1 Big Data Magnitude

From Data to Big Data, the most intuitive understanding for students is that the amount of data has become larger, but the extent is difficult to grasp. so it is necessary to establish the concept of data magnitude to measure the size of the data. In the computer world, binary bits are used as the unit, and the size of 2^{10} is a multiple, resulting in data magnitudes such as KB, MB, GB, TB, PB, EB, and ZB. In order to understand the magnitude of the data intuitively, these symbols should be connected with digital products in real life. For example, photos taken by mobile phones correspond to MB, digital movies and videos to GB, and hard disk storage to TB. The PB level can be used as a critical point, and data at this level can roughly be called Big Data.

3.2.2 Big Data Types

Big Data can be divided into three types: structured, semi-structured and unstructured. Unlike traditional structured data, the era of Big Data faces more data that is unstructured or semi-structured. Unstructured data is currently the key research aims of multi-disciplines. The ever-changing unstructured data is what makes Big Data fascinating. However, establishing an intuitive experience for unstructured data is a difficult point for beginners. Compared with the table data with headers of structured data, unstructured data has various forms, and it is difficult for students to grasp simple examples. We try to use examples of natural language processing techniques such as word segmentation, named entity recognition, and syntactic analysis, and use simple language to describe the process of extracting certain semantic words. Through this process, let students know why they need to deal with unstructured data, and the results obtained after processing, so as to deeply understand the concept of unstructured data.

3.2.3 Understanding Big Data's 4V Characteristics

The 4V characteristics of Big Data: Volume (large amount of data), Variety (various forms of data), Velocity (fast processing speed) and Value. The first two are inherent characteristics of Big Data itself. The Velocity is the requirement of the outside world for Big Data processing technology, and the Value is the ultimate goal of Big Data applications. In a word, Big Data is to quickly find valuable information in the massive and complex data, which is vividly speaking of gold panning in the sand. After understanding the Volume and the Variety, and knowing the goal of data mining, we can forward to study related technologies in depth with the question of how to dig out the value of Big Data in the follow-up content of the Big Data life cycle,

3.2.4 The Changes of Big Data Thinking

The era of Big Data has changed some of people’s traditional ways of thinking. There are three main aspects: the traditional method of sampling statistics has been transformed into an analysis method of full data; the traditional research method of exploring the causality of things has been transformed into a method of mining the correlation between things; traditional pursuit of strict data accuracy has been transformed into both efficiency and accuracy. The knowledge that students learn from high school mathematics basically belongs to the traditional way of thinking about data. However, the thinking mode of Big Data has been verified in the actual application process. Therefore, Big Data majors should think about these three changes and establish a new way of thinking about Big Data.

3.2.5 Big Data Lifecycle

Through the process of data collection, data preprocessing, data storage and processing, data analysis and mining, etc., valuable knowledge is finally obtained. The whole process constitutes Big Data lifecycle. It is very important to be familiar with the whole life cycle process, because each link in the whole life cycle will correspond to one or several professional core courses, and all the links are connected together to basically constitute the course framework of Big Data major. Therefore, after mastering the process of lifecycle, students will establish an overall understanding of the curriculum system of the major, which is one of the important goals to be achieved by the introduction course.

3.2.6 Technologies Related to Big Data Lifecycle

There are many technologies involved in each link of Big Data lifecycle, and the relevant theoretical knowledge is difficult, and it is impossible to explain everything and in-depth. Since there are corresponding courses for in-depth study, the method used in the introduction class is to focus on introducing some application techniques that can arouse students' interest, and avoid esoteric algorithms and mathematics explanations. In addition, these technologies involve many algorithms. From the perspective of application, the algorithm details should be put in the second place, and the output effect and understanding the input data should be put in the first place. Also, the algorithm is regarded as a black box, and the focus is on learning what it can do, not how to do it.

4. CURRICULUM EXPERIMENTS DESIGN

Since introductory courses are generally offered in the early stages of professional learning, students have limited programming skills and tool use capabilities. Therefore, the experiments of the introductory class are mainly arranged for verification and operation, while the comprehensive and design experiments are few. Corresponding to the content of theoretical courses, experiments are divided into three categories.

The first type of experiment is aimed at understanding the basic concepts of Big Data. This type of experiment is mainly done by the students themselves looking for original materials on the Internet, and completing the experiment after comparison and synthesis. For example, in the data type comparison experiment, students search for various types of data on the Internet according to the definitions of structured, semi-structured and unstructured data, then classify and compare the data, and finally summarize the differences between the various types of data and respective characteristics.

The second category is to use the purchased Big Data teaching platform and use the installed Hadoop software on the platform to verify the data processing effect of each node in the full life cycle of Big Data collection, extraction, processing, calculation, and storage. The advantage of using the existing teaching platform is that students can avoid complex environment configurations and spend the main time on data processing steps. These experiments are basically verification and operation types. Students only need to operate through interactive commands to gradually complete the experiments. The focus is on learning the entire process of the entire life cycle and how the data has changed on each node. As for the detailed use technology of each software, students can put it in the follow-up related courses for in-depth study.

The third category involves programming calculations. This type of experiment is conducted in many colleges and universities using the Python language, but our introductory course was opened earlier than the Python language course, so we used Excel as the experimental platform. Excel is a kind of software that students are very familiar with. It is easy to get started. At the same time, its functions are very powerful. It can fully implement basic algorithms in statistics, machine learning, and data mining. Using Excel, students can complete experiments such as descriptive statistical calculations, data correlation calculations, linear regression prediction analysis, K-Means clustering, naive Bayes classification, association analysis, and data visualization.

5. CURRICULUM CASES DESIGN

A complete case is designed from begin to end of the curriculum. Thomas Erl [5] provided a complete case about ETI company using Big Data technology to manage the enterprise. In short, the ETI company case combines the real scenarios of the enterprise, each chapter arranges an application scenario, and connects the scenarios according to the Big Data lifecycle to form a complete case that fits the actual situation very well.

Our curriculum includes many classic cases, such as beer and diapers, AlphaGo's victory over the Go champion, wine making prediction formulas, and Google's use of full data to predict influenza.

For purpose of application-oriented, cases with respect to local region and economy are necessary, for example, industrial Big Data in cooperation with Pangang Company, and Health Care Big Data in connection to our city which is a well-known health resort in China.

6. CONCLUSION

With the increasingly widespread application of Big Data technology in many areas, data engineers have become a shortage of jobs in society. The Big Data major in Application-oriented colleges should take the data engineer as the main training target, and construct professional courses standard aimed to improve the professional ability of the data engineer. As a leading course for Big Data majors, the introduction course should pay more attention on shaping and training the professional abilities of data engineers at the beganning stage of entering the professional learning.

AUTHORS' CONTRIBUTIONS

This paper is independently completed by Fangxiao Zhou.

ACKNOWLEDGMENTS

This research was financially supported by the Education Research Funds for Panzhihua University (Grant NO. JJ2027), and the Sichuan FANTAI Industry Development Funds (Grant NO. 2018VTCY-Z-07), and the Foundation Research Funds for Panzhihua University (Grant NO. 2019ZD006)

REFERENCES

[1] Xiongpai QIN, Yueguo CHEN, LI Cuiping et. al. Toward construction of “data science” course group and “introduction to data science” course [J]. Big Data Research, 2018, 4(06): 19-28.

[2] Le-men CHAO. Course Design and Redesign for Introduction to Data Science [J]. Computer Science, 2020, 47(07):1-7.

[3] Yixing Deng, Fang Wang. Study and Design on Big Data Engineer Training [J]. Computer Education, 2018(10): 121-124.

[4] Lixin LIU, Shangqian KOU. Cultivation of Professional Competence of First Class Application-oriented Talents [J]. Journal of Panzhihua University, 2020, 37(06):101-104.

[5] Erl T, Khattak W, Buhler P. Big Data Fundamentals: Concepts, Drivers & Techniques [M]. Prentice Hall Press, 2016.