

Educational Data Mining and Learning Analysis System Based on Python

Yiwei Wang(Author 1)
School of Information Science and
Engineering
Shandong Normal University
Jinan, China
1479715986@qq.com

Liancheng Xu(Author 2)
School of Information Science and
Engineering
Shandong Normal University
Jinan, China
lxcu@sdnu.edu.cn

Qiang Wang*(Corresponding
Author, Author 3)
Laiwu Vocational and Technical
College
Laiwu, China
wangqiang306@sohu.com

Huiwen Lv(Author 4)
School of Information Science and Engineering
Shandong Normal University
Jinan, China
1079141217@qq.com

Yongsheng Zhang(Author 5)
School of Information Science and Engineering
Shandong Normal University
Jinan, China
zhangys@sdnu.edu.cn

Abstract—Big data in education is of great significance to enhance students' learning efficiency and improve teachers' teaching quality, and is an important application of big data technology in the field of education. The purpose of this paper is to develop a Python-based educational data mining and learning analysis system. The advantages and shortcomings of educational data mining technology in the field of education are illustrated through the comparison of traditional education and online education and the application of related technologies at home and abroad. The basic framework of educational data mining and learning analysis system and the process of educational data mining and analysis are described, hoping to provide a reference for the research of online education system.

Keywords—educational data mining, learning analysis system, Python

I. INTRODUCTION

The purpose of this paper is to develop a learning analytics system based on educational data mining techniques using the Python language. Crawl educational data resources through Python web crawler-related libraries (Requests library, Re library, BeautifulSoup library, etc.), use Python's open source machine learning library Scikit-Learn, [1] open source deep learning library Keras, etc. and methods such as clustering analysis, attribute association analysis, generalized linear regression analysis, etc. to analyze the structural data and unstructured data in SuperStar, Rain Classroom, Wisdom Tree, Xue Tang Online, China University MOOC and other online education platforms to mine and analyze the structured and unstructured data, study the correlation between educational variables, visualize and analyze [2] the changes in students' performance development, provide decision support for teaching resource planning and teaching method implementation, and discover students' learning difficulties, provide help and guidance for improving students' learning status and enhancing students' learning efficiency. It also provides guidance for improving students' learning status and enhancing their learning efficiency, formulates personalized learning programs for

students by using comprehensive data resources, and promotes the reconstruction of the educational evaluation system and the transformation of the scientific research paradigm.

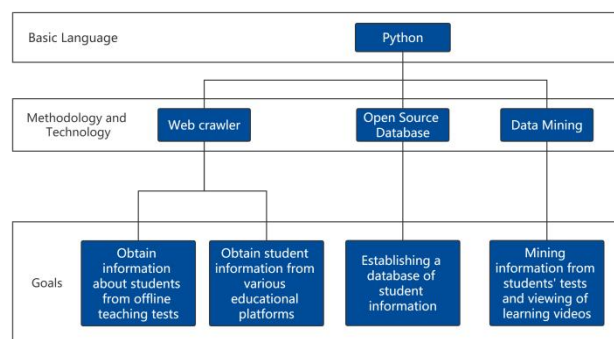


Figure 1. Basic framework of Python-based educational data mining and learning analysis system.

II. PURPOSE OF THE STUDY

Traditional teaching and general online teaching [3] rely on teachers' observation of students and students' performance to judge students' knowledge mastery, which are time-delayed and do not provide real-time information for administrators, teachers and students to take timely and corresponding measures. The purpose of this paper is to mine and process the structured and unstructured data generated by the interaction between students and learning systems, the interaction between students and teachers, the interaction between students and related teaching equipment, and the relationship between students' spatial and temporal states, to explore the degree of correlation between students' learning and each data, to make timely and personalized evaluation of students' learning effectiveness, so we can take corresponding measures to guide students to improve their learning methods, enhance their learning efficiency, facilitate teachers' teaching work, and provide a reasonable basis for

administrators to formulate educational policies and plan educational resources.

TABLE I. COMPARISON OF THE FUNCTIONS OF DIFFERENT EDUCATIONAL METHODS

<i>Educational methods</i>	<i>Teaching formats</i>	<i>Functions available</i>	<i>Real-time evaluation of teaching</i>
Traditional Teaching	1.Offline teacher teaching [4] 2.Examination and assessment	1.Communicate with the teacher in a timely manner according to the content taught by the teacher in class 2.Identify students' learning problems through exams, homework exercises, etc.	Inability to immediately evaluate students' mastery of learning content, poor real-time
General Online Teaching	1.Master teacher live web class 2.Recording class teaching 3.One-on-one online class teaching	1.Homework assessment mode 2.Online classroom teaching 3.Course discussion mode 4.Famous schools teaching video viewing mode	Feedback on students' learning process can be provided by the progress of students watching the video, but it is not possible to provide timely feedback on students' query points, and the real-time nature is generally
This system web teaching	1.Have the teaching mode of general network teaching 2.Obtain student learning information from the Education Division or other online teaching platforms through web crawler technology 3.Visualize and analyze the development of students' performance through data mining analysis 4.Through a large number of student learning data analysis, we get a visual interface for teachers' teaching improvement	1.Functions of general online teaching 2.Pre-developed personalized teaching interface for students 3.Visual interface for faculty teaching improvement	Ability to deeply mine and analyze student learning data through web crawling and other technologies, and to reform student learning and teacher teaching

III. RESEARCH CONTENT

A. Data mining research

- Improving the data in terms of traditional teaching mode, using various techniques such as database, artificial intelligence and mathematical statistics to extract information on students' learning status from their daily learning and examination training data, and linking the data to the online teaching platform and connecting offline learning data to the student information system on the official website of the Academic Affairs Office through a micro service interface to obtain enough initialized information related to students' learning.
- By reasonably using the data related to online teaching, we can analyze the main learning contents of students online, the main types of mistakes in online exams and the analysis of weaknesses in learning contents, and get the analysis conclusion about students' learning status, so as to provide more detailed and accurate data to students and teachers, and facilitate teachers and students to adjust the relevant teaching and learning methods according to the actual situation.

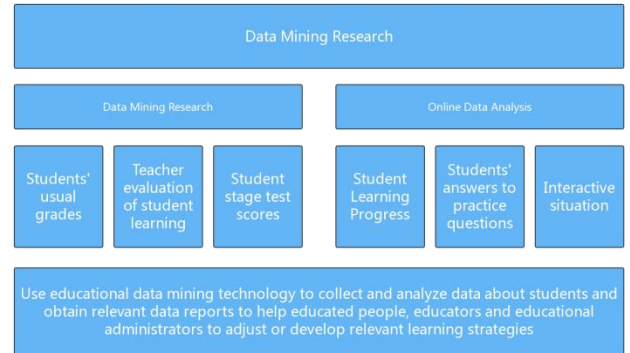


Figure 2. Visual diagram of data mining research.

B. Research on data analysis methods

- We mainly use statistical methods (k-means clustering algorithm and k-modes clustering algorithm, etc.), generalized linear regression predictive analysis, and attribute correlation analysis to [5] count a large amount of online education platform data, establish a most suitable data analysis and extraction model, categorize the factors affecting students' performance, the main weak subjects and error-prone question types of students, and analyze the important factors affecting performance, so as to provide students with study suggestions and predict students' academic performance.
- The extraction model of data analysis integrates the data mining of student information and the main problems existing in the learning process of students, so that teachers can clarify the main problems existing in the learning process of students, and provide the basis for teachers to improve teaching in

a timely manner and develop a teaching plan more suitable for the current situation.

C. Python based educational data mining and learning analysis system design

- The system adopts browser/server architecture, using PyCharm as the development tool, MySQL and Navicat for MySQL as the database, Python as the programming language, and HBuilder X and JSP as the front-end interface to build the whole system. The system mainly includes student interface, teacher interface and administrator interface, where the student interface includes modular functions such as data collection, information visualization [6], evaluation and suggestions for students; the teacher interface includes modular functions such as basic data of each student, information visualization, evaluation of students' learning and suggestions for teachers; [7] the administrator interface can perform macro operations on all the collected data, thus The administrator interface enables better management and decision making.
- Develop a personalized recommended learning interface using Python-based collaborative filtering recommendation algorithm, build a visualization platform based on student database information, and construct user weak feature values using rating matrix or rating data, non-quantitative data, and other forms. Using collaborative filtering recommendation algorithm, we find similar content collections and use big data to analyze and match them to recommend personalized learning solutions for students.
- The design of pre-processing and topic extraction module, according to the rule of simplifying, can adopt two ways of data extraction: one is to use the custom-developed web crawler to crawl the text data of information pages, after removing irrelevant characters and HTML control characters, to organize and form the standardized text; the other is to read out the text information directly from the database table, and according to the corresponding topic extraction, the basic learning The other is to read the text information directly from the database table, extract it according to the corresponding topics, compress the basic learning status information of students in groups, and enter it into the main learning system.
- The application interface is designed to meet the diversified needs of querying, and the query function is designed for students and teachers to query the related learning status information by single ID number or combination of ID numbers of interactive topics. The flexible calling mode of link and ID number is designed to meet the needs of different online teaching platform applications, and to realize the real-time calling of analysis result page and auxiliary learning page.

- Visualization design for the student interface's data collection interface, information visualization interface, evaluation and suggestion interface for students, and teacher interface for each student's data collection interface, information visualization interface, evaluation of student learning and suggestion interface for teachers.

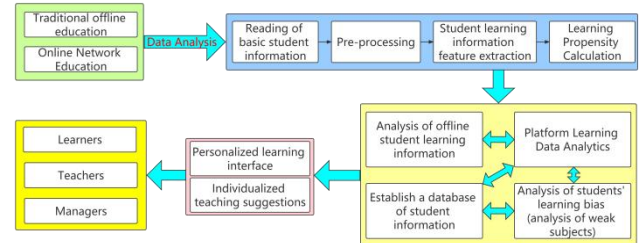


Figure 3. Visualization data mining process.

IV. CURRENT STATUS AND DEVELOPMENT OF DOMESTIC AND INTERNATIONAL RESEARCH

Since the 1980s, data mining technology has been applied to the field of education and teaching, but there are not many relevant achievements. It was not until the 21st century that data mining technology began to be promoted in large numbers in the field of education and teaching under the impetus of education informatization. 2012 saw the release of the blue paper "Promoting Teaching and Learning through Educational Data Mining and Learning Analytics" [8] by the U.S. Department of Education, and in 2015, the State Council of China released the "Action Plan for Promoting the Development of Big Data", which clearly proposed the construction of educational and cultural big data and established the "National Engineering Laboratory of Educational Big Data Application Technology", marking that educational data mining has been raised to the strategic height of education teaching and talent training at the national level in both China and the U.S., greatly promoting the application of data mining in the field of education.

According to the 47th China Internet Development Statistics Report released by China Internet Network Information Center (CNNIC), as of December 2020, the scale of online education users in China reached 342 million, down 81.25 million from March 2020, accounting for 34.6% of the overall Internet users; the scale of cell phone online education users reached 341 million, down 79.5 million from March 2020, accounting for 34.6% of mobile phone users. 79.5 million, accounting for 34.6% of cell phone Internet users. In the second half of the year, with the positive progress in the prevention and control of the epidemic, schools and universities basically resumed normal teaching order, and the scale of online education users fell further, but still grew by 109 million compared with the period before the epidemic (June 2019), with a good development trend of the industry. As can be seen, the scale of online education is gradually expanding, and the development of various online education platforms has facilitated the collection of data.

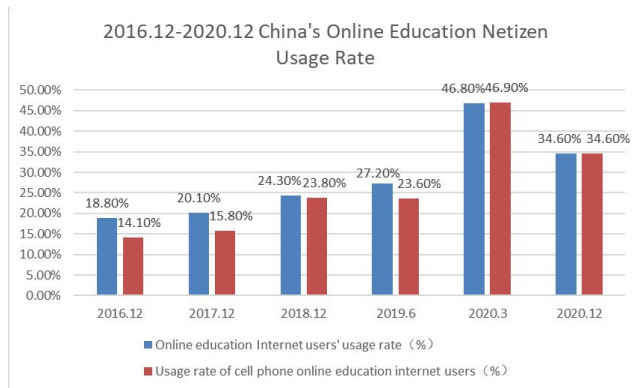


Figure 4. 2016.12-2020.12 China's Online Education Netizen Usage Rate.

The development of online education platforms and their scale and data mining techniques provide the technology and data to support the development of Python-based educational data mining and learning analytics systems.

V. STRENGTHS, WEAKNESSES AND OUTLOOK

This system is a learning analysis system specially designed for college students, which is more targeted and can provide effective suggestions for students' learning and teachers' teaching through dynamic monitoring of students' learning, help students improve their learning performance and promote more efficient learning. The system is designed and developed by combining the current popular data collection methods and the mainstream compiled language Python, which collects structured and unstructured data and provides timely feedback to teachers and students on relevant learning data. The system can collect and process real-time information from popular online learning platforms (Super Star, Rain Classroom, Wisdom Tree, Xue Tang Online, China University MOOC, etc.), and perform clustering and visualization analysis on the collected data, and the system can be flexibly extended and the application platform can be developed twice.

A. Advantages

The template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization). This template was designed for two affiliations.

1) *For students, real-time learning feedback and comprehensive learning support are realized.* [10] Firstly, through real-time data analysis and feedback, each student can know their own learning situation at any time; secondly, students' regular learning and daily quizzes are stored in the database, so that they can check and fill in the gaps in time after class; finally, with the classroom interactive system, students can participate in activities such as questioning, answering, extended reading and interacting with the teacher.

2) *For teachers, the system enables statistics and analysis of lecture process information.* The classroom

interactive system can make statistics and analysis of the information pushed before class, classroom interactive data, and teacher-student communication and interaction data after class. Firstly, through the statistical analysis of learning resources before class and answering questions before class, we can initially grasp the students' pre-study situation. Secondly, the classroom interactive system can provide real-time statistics of students' learning progress and answer situation in the course of the lesson, and teachers can adjust the lesson progress timely according to the analysis of students' learning progress and answer situation. Finally, after the class, teachers can analyze the data of discussion, quiz and extended reading to further understand the students' mastery of knowledge, so as to adjust the teaching plan and method according to the actual situation.

3) *For Managers, it enables the management and analysis of teaching data.* First of all, the classroom interactive system can accurately record students' learning process, including attendance, question answering, material reading, classroom performance, extracurricular interaction and other activities. If these information are weighted accordingly, a more scientific process evaluation can be obtained. Secondly, the classroom interactive system can be used to have a general understanding of teachers' teaching situation, such as the initial assessment of teachers' workload through the materials, test questions and discussions issued by teachers, and the characteristics of each teacher through different teachers' lectures, so as to give some constructive suggestions to the teachers concerned. Finally, through the statistical analysis function of the classroom interactive system, we can grasp the teaching situation of the whole school in real time; through the statistics and analysis of the teaching resources, we can realize the quality assessment of the educational resources.

B. Main problems and shortcomings

1) *The problem of information silo is prominent.* In the current information-based teaching environment, the curriculum center, special teaching resource library, intelligent teaching environment and classroom interactive system are independently operated systems, and the data among the systems are not connected, which leads to the problem of data gap and the information silo. Due to the existence of information silos, coupled with the low degree of intelligence of existing teaching equipment, it is impossible to collect detailed data of teaching process, which brings many difficulties to the mining and analysis of education big data.

2) *Big data mining is not deep and comprehensive.* The core features of big data are often summarized as "4V", namely, large data volume, fast input and processing speed, diverse data and huge value. [11] At present, most of the data that can be mined through the classroom interactive system is the data of the teaching process, and much dynamic information data cannot be obtained, resulting in the problem that the data information is not deep enough and comprehensive.

3) *Privacy and security issues of educational data.* [12] School systems collect and integrate information such as

names, family information, physical conditions, daily attendance, interactions and interests of students and teachers, etc. If these recorded data are leaked, it can cause irreparable harm to students, teachers and educational managers.

C. Future outlook and suggestions

The template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization). This template was designed for two affiliations.

- Promote the internal circulation and integration of information data and eliminate information silos. Although we have built some information platforms such as smart education platforms, there are problems such as data disconnection between them. How to strengthen the effective cooperation between platforms and departments is the key to effectively solve the problem of information silos.
- Use information technology to build multi-faceted big data analysis models. Strengthen the introduction and training of education big data talents, and then build a professional team to improve the field of education big data. Use big data analysis to build a multi-faceted model for students, teachers and administrators, collect comprehensive data in static and dynamic, process and evaluative aspects, and continuously improve the model and build a visualization platform so that modern technology can be better applied in the field of education.
- Increase investment in education funding and strengthen cooperation between schools and enterprises. [13] At present, there is a lack of relevant technical talents in schools, and the connection between schools and enterprises should be strengthened to build a spiraling education big data development ecosystem. At the same time, relevant policies should be formulated to strengthen the security and confidentiality of campus data, accelerate the formation of a special big data storage system in education, monitor and manage education big data in real time, so that education big data can better serve education and avoid problems such as data privacy leaks.

ACKNOWLEDGMENT

This research was supported by the provincial innovation and entrepreneurship training program for college students (S2020100445134) and the innovation and entrepreneurship training program for college students of Shandong Normal University (202010445620). In addition, the author would like to thank the reviewers for their valuable comments and suggestions.

REFERENCES

- [1] G. Varoquaux, L. Buitinck, G. Louppe, O. Grisel, F. Pedregosa, and A. Mueller, "Scikit-learn", GetMobile: Mobile Computing and Communications, vol. 19(1), 2015, pp. 29-33.
- [2] Jing Leng, Rifa Guo, Yanru Hou, and Xiaoqing Gu, "Research on online activity design and visual analysis to promote college students' critical thinking", Research on Audio Visual Education, vol. 39 (10), 2018, pp. 75-82, DOI: 10.13811/j.cnki.eer.2018.10.011
- [3] Wenjuan Wu, "An Analysis of Traditional Education and Online Education", Knowledge Library, vol. 7, 2020, pp. 12-14.
- [4] Yun Xiang and Xianjin Zhou, "Comparative Study of Online Education and Traditional Education," China Education Technology Equipment, vol. 16, 2017, pp. 9-13.
- [5] Yufan Li, Huifu Zhang, Shangli Liu, and Bing Tang, "Research Progress in Education Data Mining", Computer Engineering and Application, vol. 55 (14), 2019, pp. 15-23.
- [6] Ruoming Wang, "Adaptive system design for educational big data based on data mining and learning analytics", Wireless Connected Technology, vol. 16(24), 2019, pp. 45-46.
- [7] Shuhui Sun, Bangqi Liu, and Xin Li, "Data Mining and Learning Analysis Framework and Application for Smart Classroom", China Audio Visual Education, vol. 2, 2018, pp. 59-66.
- [8] Peng Xu, Yining Wang, Yanhua Liu, and Hai Zhang, "Big Data Perspectives on Analyzing Learning Change--Interpretation and Insights from the U.S. Report "Promoting Teaching and Learning through Educational Data Mining and Learning Analytics"", Journal of Distance Education, vol. 31(6), 2013, pp. 11-17, DOI:10.15881/j.cnki.cn33-1304/g4.2013.06.008.
- [9] Tianyi Liu, "Research status and future development trend of big data in education", China Information Technology Education, vol. 16, 2021, pp. 89-93.
- [10] Sheng Li and Niaodong Feng, "Research on educational big data mining and learning analysis based on classroom interaction system", Educational Information Technology, no. Z2, 2018, pp. 57-60.
- [11] Yingna Wang and Chenglin Shen, "Reform of secondary education in the "big data era"", Education and Career, vol. 17, 2014, pp. 18-20, DOI:10.13615/j.cnki.1004-3985.2014.17.005.
- [12] Yannan Zhang, "Research on the application of big data in education", East China Normal University, 2016.
- [13] Xianmin Yang, Sisi Tang, and Jihong Li, "Big Data of Education Development: Connotation, Value and Challenge", Modern Distance Education Research, vol. 1, 2016, pp. 50-61.