ATLANTIS PRESS
Advances in Social Science, Education and Humanities Research (ASSEHR), volume 182
2018 2nd International Conference on Education, Economics and Management Research (ICEEMR 2018)

Research on Education Big Data Mining based on "Internet +" Environment

Jing-Wei Diao
School of Information Science and Engineering
Hangzhou Normal University
Hangzhou, China

Huan-Song Yang*
School of Information Science and Engineering
Hangzhou Normal University
Hangzhou, China
hzjyhs@163.com

***Abstract*—With the rapid development of the "Internet +" era, the entire data in the world has been growing at a speed that has exceeded double rates every two years. Education Big Data regards as "oil" and "gold mine" of the education industry in 21st century, it can only exert its true value after being meticulously crafted. In order to make Big Data better serve for education, changing the way of cultivating talents, providing personalized services for students, and promoting the evaluation methods and methods of schools, teachers and students to move towards scientific, diversified and intelligent. In this paper, it elaborates the main application techniques, mining goals and the basic process of educational data mining for educational big data mining through the investigation and analysis of relevant literature and works. Through researching, it has been found that the excavation of educational data in China is still in its infancy, and the educators with professional skills are currently an urgent issue that should be paid attention in education data mining. This paper expects to provide reference for workers who engage in educational data mining and lays the foundation for further mining of educational data in the future.**

Keywords—education; big data; data mining; Internet +

I. INTRODUCTION

With the advent of the "Internet +" era, Big Data have brought new soil to every walk of life as the brightest star in the new era. Big data technology provides unlimited development potential for education. In 2013, according to People's Daily online, education data in China were already the largest in the world [1]. On August 31, 2015, the state council in the "Outline of Action for Promoting Big Data Development" noted that the convergence of information technology and economic society has led to the rapid growth of data. Data have become the basic strategic resources of the country. Big Data are increasingly having a significant impact on global production, distribution, consumption activities and economic operation mechanism, social life style and national governance capacity. This outline has proposed to build education culture Big Data [2]. According to IDC's research report, global data are expected to reach 40ZB by 2020, equivalent to 5247GB of data per person. The 40ZB is estimated to be 57 times the amount of sand grains on all the beaches on earth. It is expected that all data will double every two years from now until 2020.

Multimedia Courseware Produce of MOOC Construction Project of Hangzhou Normal University. Program. No. ZX4115F5121632419.

II. RESEARCH SIGNIFICANCE

Education Big Data are a gold mine waiting for deep excavation and mining. It is the most essential goal of big data mining to dig and analyze large-scale data and obtain the most valuable data to solve the real problems in education. Using big data mining technology to analyze massive education data can help countless education workers find patterns, predict trends, assist decision-making, and summarize experience. The application of Big Data in education will realize the personalized training of students, change the traditional teaching management and evaluation methods, and promote the scientific development of education. The effective mining of data directly determines the role and value of data in practice, and data mining are the most important part of big data technology.

III. EDUCATION BIG DATA OVERVIEW

Big Data are also called mass data, it refers to the huge amount of data that exists in databases and networks with different forms.[3] Big Data are different from previous mass data, there are major differences.: (1) With the development of social media and sensor networks, abundant and diverse data are continuously generated.(2) With the development of hardware and software technology, the cost of data processing and storage have also declined.(3) With the gradual development of cloud computing, the storage and processing of big data have no need to build itself.

Big data has four main characteristics: huge volume, various types, low value density and fast digging speed. By investigating the literature, it was found that no scholars have defined the Big Data clearly. In a broad sense, educational Big Data are a subset of Big Data, including all the valuable and meaningful data sets which promote education development.

Education Big Data have four categories:(1) Teaching resources Big Data. (2) Education teaching management Big Data. (3) Teaching and learning behavior Big Data. (4) Education teaching evaluation Big Data. The most difficult data to obtain in education data are the evaluation of teaching behavior.

Copyright © 2018, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

IV. BIG DATA MINING TECHNOLOGY

Big Data mining technology has created opportunities to improve education level. Using big data technology to collect, store, analyze and process education data can effectively predict learners' future learning orientation. Education data mining is a combination of mathematical statistics, machine learning and data mining. The education Big Data were processed and analyzed, and the correlation of various factors in education process was discovered through data modeling. For example, correlation of variables including learners, learning outcomes and learning content, learning resources and teaching behavior can predict the future learning trend of learners in the course of learning. In the study of universities, data mining can be recommended for students, which can help them determine whether they deviated from their chosen major.

A. The Main Content and Technology of Big Data Mining

There are six aspects of data mining: Association, Regression, Classification, Cluster, Prediction and Diagnosis. There are big differences between the six aspects of mining goals, content and the techniques.

- Association: If there is some regularity between the values of two or more variables, it is called association. Association can be divided into simple association, temporal association, and causal association and so on. The purpose of association analysis is to find the hidden network of data. Foreign countries have achieved certain results on education using data association rules. For example, the project of SemTech (Semantic Technologies for Learning and Teaching) funded by JISC (the Joint Information Systems Committee) [4]. The project mainly defines the semantic technology, and differentiates the relationship between the associated data and the traditional metadata. The project of LUCERO (Linking University Content for Education and Research Online) has been launched in the UK. This project used the association data and mining techniques to select, extract and structure the learning resources of the UK Open University. The project of LUCERO integrates publishing associated data to form an associated cloud. [5].
- Regression: Regression is the most basic method in data mining, and it is a statistical analysis method to determine the quantitative relationship between two or more variables. According to the number of dependent variables and the type of regression function in the regression method, the regression method can be divided into unitary linear, unitary nonlinear, multivariate linear and multivariate nonlinear. Multiple linear regression can be used to predict the results of online learning and test scores and how much time an object might take to participate in online learning.
- Classification: Classification is a common problem in data mining. Its typical application is to classify things scientifically according to the characteristics of things in data level. The classical classification methods mainly include: Decision tree algorithm, ID3 algorithm, C4.5 algorithm, KNN algorithm, Neural Network method, Bayesian classification, Discriminant analysis, Support vector machines and other classification methods. Romero C et al. compared the performance of 25 popular classifiers on e-learning data analysis tasks. They selected 438 students from the University of Corcovado to analyze the log data of seven online courses to achieve the classification of students [6].
- Cluster: Cluster is the process of clustering together unique instances of common characteristics. The saying that "birds of a feather flock together" as the best embodiment of cluster. Cluster analysis is a multivariate statistical analysis method for the classification of samples. Cluster algorithm is divided into hierarchical cluster, network based cluster, density cluster and model based cluster according to cluster principle. Cluster mainly groups students in order to implement personalized learning in process of education data mining. The reason why the cluster method is used for cognitive diagnosis is that the performance of traditional cognitive diagnostic methods decreases with the increase of students, knowledge points and courses.
- Prediction: Prediction is to make scientific inference and judgment on the future development trend or state of a particular object according to the development trend and change rule of objective things. It involves collecting historical data and using some mathematical model to predict the future. It can also be a subjective or intuitive expectation of the future, and it can be a combination of the above. The prediction method is divided into quantitative prediction method and qualitative forecasting method. Prediction modeling can be used to predict the academic performance of learners at the end of the study and provide reference for the academic warning and adjustment of teaching strategies.
- Diagnose: Diagnostic method is an important technique in data mining. The goal of diagnostic is to find small amounts of data that are abnormal in the data set and known as outliers or isolated points. The outliers may be noisy or useful information, random deletion of outlier data can cause useful information to be lost. It is significant to find and utilize valuable information in outlier points by outlier diagnosis.

B. Main Research Objective of Education Data Mining

- Education data mining through the integration of learners' knowledge, attitude, motivation, metacognition and other detailed information to construct the learner model and predict the future development trend of learners.
- Exploration and reprocessing mainly include the best teaching content and teaching order model.
- It is one of the main research objectives of education data mining to study the effectiveness of various teaching support programs.

- Construct data computing model which includes learner, domain model and education software model. Data computing model can improve learners' learning efficiency.

C. Education Data Mining Basic Flow

The eight steps of data mining are the data collection, data integration, data reduction, data cleaning, data conversion, data mining process, model evaluation, and knowledge presentation.

- The data collection: Firstly, identify the object of the data analysis and abstract the characteristic information needed in the data analysis, then select the appropriate data collection method and store the collected data to the database. It is important to choose a data warehouse that is suitable for data storage and management for large and complex data. There are many ways to collect data, such as system log collection methods, network data collection methods and manual collection methods. For example, the basic information of students, teachers and schools is mainly collected by hand.
- Data integration: Data integration is centralize data from different features, sources, and formats in logical or physical rules to provide comprehensive data sharing for education. At present, the digital campus system of most colleges and universities in China has been gradually improved, and a large amount of data is continuously generated in these systems. Heterogeneous data integration can solve the problem of information and resource decentralization, weak system decision support and high management cost. School should make full use of these data to serve education.
- Data reduction: Some education data are large, and the data reduction technology can be used to obtain the reduction representation of data sets. This data set is relatively small, but it maximizes the integrity of the original data. The data mining results performed after the statute are identical or almost identical to those before the statute.
- Data cleaning: Some of the data in the database are incomplete. In order to make the results of data mining more perfect, the complete and consistent book information needs to be stored in the data warehouse. Cleaning up the large data collected on the network, and removing any redundant files or blocks can improve storage space utilization and the time cost of analyzing data is saved.
- Data conversion: Data conversion is to transform data into a form suitable for data mining by means of smooth aggregation, data generalization and normalization. It is a very important step in the whole process to transform data through data discretization and concept stratification.
- Data mining process: According to the data in data warehouse, choose the appropriate analytical tools, valuable results are obtained through statistical method, case reasoning, decision tree and fuzzy algorithm to handle data. All pre-data preprocessing procedures are ultimately served for this link.
- Model evaluation: The process of verifying the correctness of data mining results by professionals and experts is called model evaluation. The evaluation data mining model starts with the classification of the error state, and the distribution of all errors is viewed by establishing the classification matrix.
- Knowledge presentation: Visualize the results of data mining and feed back to users, or as a new knowledge in the knowledge base to help other applications. Data visualization is getting more and more attention. There are many tools for data visualization, such as Google Chart, R language, Tableau Software and so on， which can generate graphs, network diagrams, hierarchical diagrams, and other types of charts.

V. CURRENT EDUCATION BIG DATA MINING CHALLENGES

The essence of Big Data is the use of machines to extract information from data and anticipate future possibilities, but education data are more complex branch of data. The use of Big Data mining technology requires specialized data talents, and the general manager on education also needs to improve the data processing level. It is a new requirement for teachers' teaching ability in the era of big data that business skills or teacher teaching ability enhancement are required to collect, analyze and interpret different types of data, and translate into the ability to improve the knowledge and practice of teaching process and behavior[8]. Starting from September 2017, the high school information technology textbooks in Zhejiang province will focus on the course of Python, and increase the knowledge points of programming. The programming language will be replaced by Python. Using the combination of "Python+Matplotlib+Pdandas" to draw icons, manipulate data and perform data visualization analysis. This reform is sufficient to show that our country has begun to pay attention to the cultivation of data mining talents.

At present, higher education universities in China should focus on cultivating and forming a team of professional and technical personnel with Big Data backgrounds to strengthening data analysis and value mining capabilities. While mastering the professional knowledge and technology also need to acquire the ability of knowledge quickly, high expression ability and high management ability to adapt to the development of the Big Data era.

VI. CONCLUSION

In the education field of developed countries, relevant research and application of Big Data have been achieved. The research of education Big Data in China is still in the exploratory stage. There are still many problems to be solved. Therefore, we urgently need to integrate the existing data resources, analyze mass data, and use the valuable data to realize education Big Data research and application leap development. Through data mining and analysis, the most

valuable data are selected and the evaluation model is established. We will be better able to read different students. Education Big Data mining gives us a chance to understand the real learning situation of each student, and can better provide excellent and personalized education resources for each student, and continuously realize the fairness of education. Education big data are not only a technology, but also an art. It is worth learning deeply from education workers.

ACKNOWLEDGMENT

I would like to express my gratitude to all those who have helped me during the writing of this thesis. I gratefully acknowledge the help of my supervisor Professor. I do appreciate his patience, encouragement, and professional instructions during my thesis writing. The corresponding author of this paper is Huan-Song Yang.

REFERENCES

[1] People's Daily online (2013). Ministry of education: education is the largest in the world [EB/OL].[2015-12-10], http://politics.people.com.cn/n/2013/1015/c1001-23206346.html. (In Chinese)

[2] Ministry of industry and information technology of the People's Republic of China (2015), Notice of the state council on the issuance of a platform for the development of big data [EB/OL], http://www.miit.gov.cn/n1146290/n1146392/c3882451/content.html. (In Chinese)

[3] Jia-Ping Wu, Yi-Dong Gu, Huan-Song Yang, "Big data and education big data exploration, " Journal of Jiamusi vocational college, pp.143-144, December 2016. (In Chinese)

[4] Semtech | ECS | University of Southampton[EB/OL]. [2012-10-12]. http://www.semtech.ecs.soton.ac.uk/.

[5] The LUCERO Project [EB/OL].[2012-10-12], http://lucero-project.info/lb/.

[6] Romero C, Ventura S, "Data mining algorithms to classify students" , EDM, 2008, pp.8－17.

[7] Zhou Su, Wang Wen, Big data introduction, Beijing: Tsinghua university, 2016, pp.2-40. (In Chinese)

[8] Ying Zhou, Jin-Wu Zhuo, Yue-Qing Bian, Method and case analysis of big data mining system, Beijing: Mechanical industry press, 2017, pp.20-50. (In Chinese)

[9] Macfadyen L P, Dawson S, "Mining LMS data to develop an early warning system for educators:A proof of concept",Computers & Education,vol.54, pp.588-599,2010.