Research on the Application of Big Data in the Field of Online Education

Shen Zheng

ABSTRACT

Online education played an important role in the field of education and is becoming more and more attracted people’s attention with the development of the information technology. Because of the growth of the amount of data on online education platform, it is very important to research the big data of online education. This article gives the classification and characteristics of the big data in the field of online education, and briefly introduces the big data technology in online education, such as education data mining, learning analysis and knowledge graph, focusing on the design of a large data model for online education. The goal of this article is to outline the view of the application and to contribute to the development of the online education system with big data.¹

KEYWORDS

Online Education; Big Data; Data Mining; Learning Analysis Technology; Knowledge Graph

INTRODUCTION

Online learning is a way to grow knowledge and skills through information communication technology (ICTs) [1]. With the rapid development of Internet and Web technology, online education has attracted more and more attention. The way of online education has changed significantly with the change of computer

¹Shen Zheng, Quanzhou Normal University, Quanzhou, China
technology and people's educational concept. From the distance education model to the resource platform model, and then to the current popular business online education platform model. From the beginning of the static sharing of resources to the interactive open classroom, the concept and mode of online education are undergoing major changes.

With the rise of online learning tide, online learning platforms have broken into people’s eyes, one of the representative ones is the massive open online course, MOOC. It is a massive open and large-scale online course for the public. The platform borrows the developed network and video technology to alleviate the problem of the lack of education resources and the unbalanced allocation of education resources. Compared with MOOC, SPOC (small private online course) is mainly for classes from dozens to hundreds of students, there are strict limits for registered students, only who meet the corresponding requirements of the course applicants are eligible to study this course. SPOC, this small classroom teaching model makes teachers a real master of the course[2]. In addition to MOOC, SPOC, there are online learning communities, mobile digital libraries, mobile reading, and so on. These online education platforms create a well-resourced, interactive atmosphere for students, strengthening the students’ pre class preview and consolidating practice after class.

The famous MOOC platform, Coursera, has registered more than 4 million 500 thousand in August 2013. By the end of 2012, the Khan College had reached 10 million participants, the daily video frequency ranged from 100,000 to 200,000, and the total number of video views was 221 million. The MOOC from China built by Tsinghua University Platform - School Online, with more than 13,000 total electives as of October 2013, a single "circuit theory" course attracted more than 11,000 students from more than 150 countries.

Facing such huge and complicated data on online education platform, the research and application of big data for online education are particularly important. "Big data" is a massive, high-growth and diversified information asset that needs to undergo a certain processing model to have stronger decision-making power, insight into discovery power and process optimization capabilities [3]. The "big data" mentioned in this article not only refers to massive data, but also a way of thinking that concerns data. That is to say, using objective user data to guide decisions and reduce the interference of personal preferences and other factors. The key point is the conversion process from data to value. The purpose of this study is to assist and guide students' learning, teachers' professors, and the selection of talents by using the data in the field of online education.

The purpose of this study is to enable students to have a personalized and diversified learning path to achieve a global integration of culture and resources. The use of big data on the online education platform can achieve the following purposes: 1) track the students' behavior files and record the students' mastery of the knowledge points and the time they spend on them in real time. According to this information, teachers can get a more true understanding of the students' learning
situation and make personalized tutoring; 2) assist teachers to optimize teaching methods and improve curriculum design so as to make more reasonable syllabus; 3) explore the law of learning and predict the future development of students, so as to provide a more convenient and reliable way for selecting talents in different fields.

DATA IN THE FIELD OF ONLINE EDUCATION

The data in the field of online education are complex and diverse. This section introduces the big data in the online education system from two aspects of data classification and characteristics.

Classification of Data

On the online education platform, students can watch videos and learn course contents anytime, anywhere, and teachers can create courses, upload notices at anytime and anywhere. There are 3 ways to classify the data produced on the online education platform according to the way of obtaining the data, the source of the data and the storage form of the data.

CLASSIFICATION BY WAY OF ACQUISITION

It can be divided into explicit data and implicit data according to the way of data acquisition. Explicit data include user registration information, interest labels, selected courses, course performance exercises, total course results, evaluation of courses, evaluation of discussion contents, etc. Implicit data include search keywords, browsing contents and times, course practice time, time for browsing discussion content, the number of times for dragging the scroll bar, the number of video pause times, collection record of browser, etc.

CLASSIFICATION BY SOURCE

According to source, data can be divided into user information, course information, exercises information, video information, forum information, user behavior information and knowledge graph.

User information refers to the user's personal information, including the user's basic information, academic performance information and the degree of understanding of knowledge points, etc. Course information refers to the content and organizational structure of the course, etc. Exercises information refers to the contents of the exercises and the hints of the exercises, etc. Video information includes video, audio and subtitles, etc. Forum information includes the organizational structure and content of the forum, etc. User behavior information refers to the user's behavior records such as browsing the web, playing video, and so
Knowledge graph is a collection of knowledge points, initially drawn by teachers and then maintained and updated by automatic combined manual methods.

CLASSIFICATION BY THE FORM OF DATA STORAGE

It is divided into structured data and unstructured data in the form of data storage. Structured data refers to the data that can be expressed logically by two-dimensional table structure, on the online education platform, including user basic information, course information, etc. Unstructured data refers to data that is not easily represented by a database of two-dimensional logical tables, including video, audio, and knowledge graph.

Characteristics of Data

In the era of explosive growth of data, "big data" is ubiquitous and permeates all walks of life. The data in the field of online education are also different from other areas.

On one hand, data are closely linked. The data sources in the online education system are different, but there are countless connections between them. For example, student information combines courses, forums, grades and so on, knowledge point information combines knowledge graph and exercise database, and each student has contact with every knowledge point.

On the other hand, data are stored in different forms. Video data is stored in an unstructured way. The structure of forums and courses is stored in tree mode, and the basic information of users is stored in a simple two-dimensional table structure. Therefore, the online education system has a variety of storage forms, thus forming a heterogeneous database, which puts forward higher requirements for the management of data.

RELATED WORD

The difficulty of big data technology lies mainly in the complexity of the data, which contains very meaningful information that can help the decision. At present, the research of big data has penetrated into all industries, including Internet, politics, medical treatment and so on. This article focuses on big data in the field of online education.

The research on big data of online education is divided into two main directions: education data mining and learning analysis technology.

Education Data Mining

Data mining (DM) refers to the method and application of extracting or "mining" knowledge from a large number of data. Data mining usually involves
following processes: data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge representation. Educational data mining (EDM) is the application of data mining in the field of education. It refers to the skills, tools and research to extract valuable information automatically from the data generated in the process of human education and learning. This information can be used for educators, learners, managers, educational software developers, and educational researchers. Data of online education are stored in different forms, so the process of data cleaning and integration is more complicated. In addition, due to the diversity and close connection of the data sources, many characteristics and the relationship between them should be considered during the mining process.

**Learning Analysis Technology**

Learning analysis technology in the early days is defined as "use intelligent data, data and analysis model of students to find information and social relations, so as to achieve the aim of predicting and improve people's learning" [4]. In 2012, Greller and Drachsler proposed a more systematic and comprehensive definition of learning analytics from six key dimensions, as shown in Figure 1 [5] These six key dimensions are internal limitations, external constraints, instruments, data, objectives and stakeholders. The stakeholders include institution, teachers, learners etc. Learning analysis technique is based on the data generated by the students, mainly for teachers, institutions, its main goal is to reflect the educational mode and predict students' development prospects, as shown in Figure 2. Using this technology, we can make personalized plans for students, teachers and institutions to achieve self-reflection.
The common goal of educational data mining and learning analysis technology is to build an intelligent learning platform, which is conducive to improving teachers' teaching level, stimulating students' interest in learning and promoting balanced development of learning. In addition to the commonality, there are 3 differences between the two. (i) Educational data mining mainly focuses on specific tools, techniques and algorithms; learning analysis focuses on the whole system to support decisions of different objects. (ii) Educational data mining focuses on details and algorithms; learning analysis technology focuses on the establishment of a
model. (iii) The purpose of educational data mining is to build an intelligent system, which is able to answer automatically. Learning analysis technology is mostly based on the field of education to meet the needs of all kinds of objects.

Knowledge Graph

Knowledge map is a kind of research method to display the core structure, development history, frontier field and overall knowledge architecture of the subject in a visual form, combining the theories and methods of applied mathematics, graphics and information science with the methods of citation analysis and co-occurrence analysis of metrology.

The result of querying a key word in a knowledge graph is a series of related content, not only the content of the key word. Knowledge graph can not only help students establish a complete knowledge system, improve learning efficiency, but also help teachers quickly discover students' learning weaknesses and answer questions pertinent, so as to improve teaching quality.
BIG DATA MODEL IN THE FIED OF ONLINE EDUCATION

The functions of the online education system include teaching video, teaching materials, exercises, tests, experiments, and discussions. Depending on these functions, you can determine the basic modules of the database layer. As shown in figure 3, the big data model based on the online education platform mainly consists of 5 levels: data layer, processing layer, analysis layer, function layer and object layer.

Data Layer

The data layer can be divided into user information, course information, exercise warehouse, video warehouse, forum information, user behavior information and knowledge graph. Among them, the user information contains three parts: the user basic information such as name, age, educational background; the user's understanding of each point of knowledge; the user's performance information in each class. Forum information is divided into chapter forum, course forum and platform forum. The knowledge graph mainly stores the content and the relationship of the knowledge points of each course on the online education platform. The user behavior information records the user's behavior information, including clicking the mouse, playing video, and so on.

Processing Layer

The processing layer is responsible for processing heterogeneous databases, providing effective data for the analysis layer. The layer has functions such as data cleaning, data conversion, data integration, data extraction and data maintenance. Data cleaning refers to deleting invalid or meaningless data based on the implementation function and the use of algorithms. Data conversion refers to providing input to the upper level by cutting, merging, and so on. Data integration integrates data from different databases together. Data extraction means extracting useful data from the database for processing. Data maintenance refers to backup, authority management, etc., so as to ensure the security of the underlying database.

Analysis Layer

The analysis layer includes education data mining algorithms library and learning analytics model library. Education data mining algorithms library includes commonly used data mining algorithms, such as SVM, Apriori, FP-tree, etc. Learning analytics model library includes social network analysis model, speech analysis model, content analysis model, etc.
**Function Layer**

The function layer borrows the algorithms and models from the analysis layer to provide specific big data functions for different users, including the service function provided for the students’ individualized learning, the service function provided for the efficient work of teachers and the service function provided for the accurate judgment of decision makers.

**Object Layer**

The object layer mainly consists of students, teachers, and decision-makers. The online education system commonly provides two big subsystems which are LMS (learning management system) and CMS (content management system). LMS is a study system designed for student. It can provide students with video, study materials, discussion, question and many other functions. CMS is designed for teachers. It’s a platform for teachers where they can prepare lessons, give public notice and design teaching program.

Figure 4. The information flow of the online learning platform.
THE GENERAL PROCESS OF USING BIG DATA TECHNOLOGY TO STUDY ONLINE LEARNING BEHAVIOR

Romero and others believe that the data mining technology in the application of online learning system is an iterative process, mining knowledge should be entered into the system of decision support, improve learning system, improve the learners' learning[6]. The information flow of the online learning platform supported by large data technology is shown in Figure 4. The process of applying data mining in online learning behavior research includes four stages, namely data collection, data preprocessing, application of data mining method and interpretation of the results.

Data Collection

Data collection is usually done automatically by online learning platform. Learners' use of learning platforms and interactive information will be recorded in the database or in the form of logs. Data source of the research on online learning behavior, in addition to EdX, Coursera, Udacity and other MOOC platform which is a new source of data, including learning management system (LMS) and intelligent tutoring system (ITS), the learning management system (such as Moodle) is the main source of data for the research on online learning behavior.

Data Prepossessing

Data preprocessing is the process of cleaning data and converting it into a form suitable for data mining, mainly including data cleaning, data conversion, data integration and data reduction. This phase can also filter data from the existing data, integrate the features, and discretize the data.

Application of Data Mining Method

After processing the original data, the data mining method is used to analyze it and find the knowledge that is meaningful to the teachers, students and platform managers. Currently, researchers often use data mining methods such as classification analysis, cluster analysis, association rule mining, sequential pattern mining, text mining and so on. The classification analysis is mainly used to predict learners' achievements, and to extract the characteristics of all kinds of learners according to the classification standards, and thus help to carry out individualized education. Cluster analysis can be used to group students and find students with similar learning characteristics and behavior patterns to improve cooperative learning based on groups. Association rule mining and sequential pattern mining are mostly applied to discover learners' learning and usage habits in online learning process, such as frequent web browsing paths and arranging order of learning activities, so as to implement personalized learning recommendation and learning arrangements.
Interpretation of the Results

Finally, it is necessary to interpret the mining results in combination with the online learning situation, analyze the actual meaning behind it, and make decisions using the acquired knowledge. On the one hand, teachers and administrators can use the results of the analysis, to adjust the process of teaching and learning platform design, understand the learning situation, timely find the poor performance of the learners, so as to timely intervention; on the other hand, learners can know more about the results of the analysis of their own characteristics, learning state and knowledge the degree, to improve the learning efficiency and learning effect.

CONCLUSIONS

This paper expounds the concept of large data on the online education platform from the point of view of the classification and characteristics of the data, and briefly introduces the big data technology in online education, such as education data mining, learning analysis and knowledge graph, focusing on the design of a large data model for online education. The online education platform realizes basic functions such as watching courses and recording courses, the exploration of big data is at the initial stage, this article provides a train of thought for the research of big data in online education.

Big data in the field of online education has both the academic significance and application value of the research direction, with the development of big data technology, people's understanding of the deepening of online education, there will be more and more organizations and individuals to participate in the study of online education related. Believe that through our efforts, we can achieve more convenient and more efficient, better access to acquire knowledge, improve their ability to dream

REFERENCES

1. McGreal R, Anderson T. E-learning in Canada. Inter Journal of Distance Education Technologies (IJDET), 2007, 5(1):1-6
2. Yeqin Kang. An Analysis on SPOC: Post-MOOC Era of Online Education. Tsinghua Journal of Education, 2014, 35(1):85-93
3. Kaisler S, Armour F, Espinosa A. Introduction to big data: Scalable representation and analytics for data science minitrack //Proc of the 46th Hawaii Inter Confon System Sciences (HICSS). Piscataway, NJ:LEEE, 2013:984-984
4. Qing Li, Tao Wang. A review of the research and application of learning analysis technology. China Educational Technology, 2012(8):129-133
5. Greller W, Drachsler H. Translating learning into numbers: A generic framework for learning analytics. Journal of Educational Technology & Society, 2012, 15(3):42-57
6. Romero C, Ventura S. Data mining in education [J]. Wiley Interdisciplinary Reviews Data Mining & Knowledge Discovery, 2013, 3 (1): 12–27.