Research on data mining of MOOC platform based on large data environment

Qiuping Wang
Wuhan University of Technology.Wuhan 430000.China
448283915@qq.com

Keywords: big data, MOOC platform, data mining

Abstract. The vigorous development of the Chinese MOOC platform has brought the huge amounts of data. How to tap the value of these data is the current research problems to be solved. This paper puts forward the data mining framework through analyzing the data features of user data in MOOC platform under the large data environment, and also discusses the data mining process and the main data mining methods. Besides, the application of MOOC platform user data mining was explored from the perspective of MOOC platform, learners and universities.

1. Introduction

MOOC (massive open online courses), that is a large open online courses, is presented by the Canadian scholar Dave Cormier and Bryan Alexander in 2008. So far, the world's educators actively engaged in the development of the MOOC.Three big MOOC platform Udacity, Coursera, edX successively was established at the beginning of 2012, attracting a large number of learners from around the world to register to study. For China, the 2013 can be called "Chinese MOOC year", firstly Peking University and Tsinghua University joined the edX in May 2013, secondly a large number of universities and institutions explorated the Chinese style MOOC . “eastern and western university course sharing alliance” was formally established in April 2013 ,which was sponsored by the Chongqing University .Tsinghua University launched the "online school" in October 10,2013,which was based on edix , so far it has launched 166 courses.

The booming development of the Chinese style MOOC platform has brought a huge amount of data. including age, gender, education and other aspects of learner enrollment data; learner behavior data. This paper analyzes the data features of MOOC platform. On this basis, the paper proposes the data mining process, method and application of MOOC platform.

2. MOOC platform data characteristics

(1) body mass: a lot of user data generated by MOOC platform daily has reached TB level, Daily user data plus existing mobile client platform could rise to the PB from TB grade level in the future even to the EB level.

(2) there are many types of data types: MOOC platform. It includes not only the learner's characteristic data, behavioral data, performance data, including data from the data exchange discussion boards, feedback area, the mobile terminal.

(3) the rate is fast: data MOOC platform is accompanied by the generation of learners' learning behavior, these data are high-speed real-time data streams. For example, the study of the elective courses, learning courses on the platform, to participate in the discussion, these will produce real-time data.

(4) the value of the high: MOOC platform data hide huge commercial value. Learners are the core of the MOOC platform operations, through mining these high value data, we can optimize the MOOC platform to enhance the learning efficiency and effectiveness of the learners.

© 2015. The authors - Published by Atlantis Press

3. MOOC platform data mining processes and methods

3.1MOOC platform data mining framework

Because of the large data characteristics of MOOC platform data, the traditional data analysis techniques can not be used to make good use of 4V. There are essential differences between them[2]. (1) data analysis are usually stored in the database or file, data size is generally GB level, and the data size in large data mining is generally PB or even more. (2) data types are different, the traditional data analysis is mainly for static, structured data, and the data mining of large data is not only structured data, but also semi-structured, unstructured data. (3) there are differences between the methods, the main algorithm of the traditional data analysis is based on statistics, classification, large data mining not only need statistical methods, but also machine learning, artificial intelligence, neural networks and other complex algorithms.

When the failure of traditional data analysis, how to gain insight into the value of the data from the large data in MOOC platform will be an urgent problem need to solve. we propose a MOOC platform data mining framework, as shown in Figure 1. MOOC platform data mining framework includes data sources layer, data collection layer, etc.Data collection, data organization, data storage layer belongs to data preprocessing. data analysis layer is applied to data mining model to analyze data. Application layer of data application layer is the application of object-oriented method, which includes the application of the MOOC platform, the application of the University and the application of the learners.

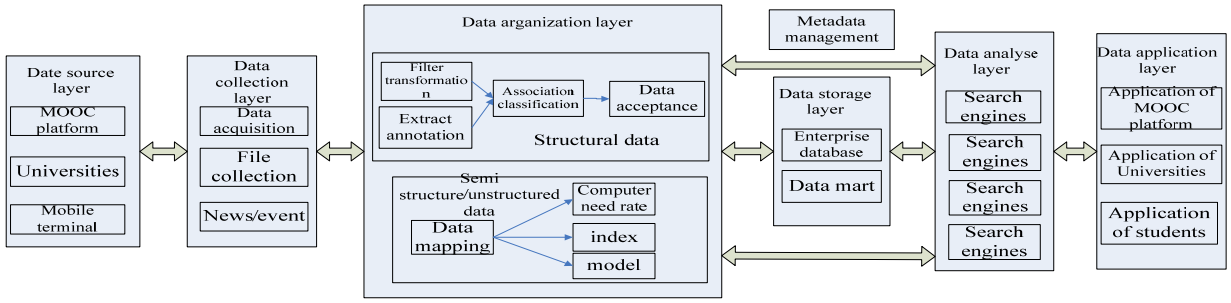

Figure 1 MOOC platform data mining framework

3.2 MOOC platform data mining processes

(1) data collection: the MOOC user data source includes platform learner data, which includes user characteristic data, behavioral data, performance data, course data, and mobile terminal data. These data volume is large, the type is much, the rate is quick ,which bring a lot of difficulty to collect .We can classify the data according to the studying data, course data, observation data. You can use Needlebase, ScraperWiki, and other large data collection tools to obtain data from a variety of data sources.

(2) data preprocessing: Data preprocessing is often about the success or failure of data mining. We need to clean, reconstruct, and fill the missing values to improve the quality of the data. Then, unstructured, semi-structured data are transfered into machine language or index, different words are mapped to the standard value; The structured data is filtered to extract meaningful data, and the invalid data is removed to improve the efficiency of the analysis. Finally, data extraction, that is to detect the correlation of data, the related data show more specific user activity characteristics.

(3) data mining: in the process of data mining, according to different application we needs to choose different mining model. The main models are: association rules analysis, clustering analysis, deviation analysis, etc. After the results of data mining, we need to explain the application, general mining applications include ranking and personalized recommendation, anomaly detection, Web mining and search, large data visualization calculation and analysis, etc

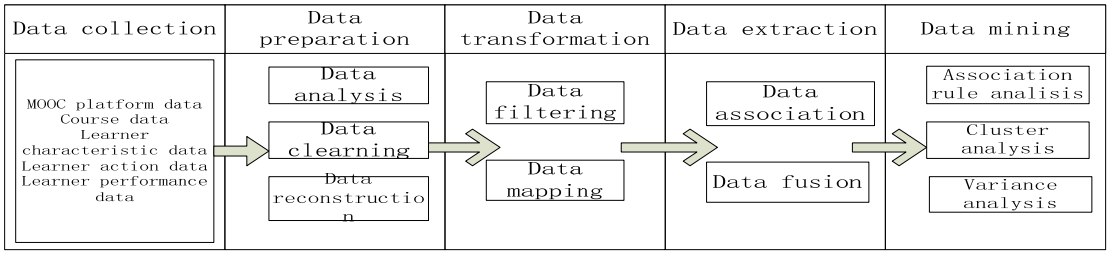

Figure two MOOC platform data mining processes

3.3MOOC platform data mining method

(1) association rule analysis: the most famous example of association rules is WAL-MART's "beer and diapers" case, which is widely used in e-commerce and financial industry, I think the association rules can be applied to data mining in MOOC platform too. In the MOOC platform, it can find out the relationship between the user data and find out the key factors that affect the behavior of learners.

(2) clustering analysis: clustering analysis is one of the main tasks of data mining. Clustering can be used as an independent tool to obtain the distribution of data, Clustering analysis can also be used as a preprocessing step for other algorithms (such as classification and qualitative induction algorithm). Clustering analysis is widely used in business, biology, geography, e-commerce and other industries. Here it can be applied to the classification of learners of the MOOC platform, the learners attribute and characteristic analysis, learner satisfaction analysis, the study of the course selection trend forecast, etc.

(3) deviation analysis: Deviation analysis is a significant change and deviation between the status quo of the survey data, historical records or standards. including a large class of potentially interesting knowledge. Such as the deviation between the observed results and the expected results, the abnormal examples of classification, the exception of the model. Variation and deviation analysis include a lot of potential knowledge, such as abnormal instances of the classification, the special case of the rules, and the variation of the amount of time. It can be applied to the discovery, analysis, identification, evaluation, and loss of the learners of the MOOC platform.

4. Application of data mining in MOOC platform

4.1Application of data mining based on MOOC platform

We can mine the data from the learners in the operating data of the platform to find the user access patterns ,optimize the MOOC platform. We can design and modify the website structure and appearance according to the characteristics of the visitors and the rules of choosing the course. In addition, through the mining of user browsing data of MOOC platform, we can find the relevant page of the user access to the user's desired location. MOOC platform can increase the link between closely related web pages, and the reasonable arrangement of the server page prefetching and caching strategy, reduce the server response delay time, improve customer satisfaction.

Based on the data of the learners in the discussion area or WIKI and other community data We analyze the learning progress of the learner in order to adjust the course of study, make a question and answer, and make the most of the course and the course of the study. In order to attract the learners who have not joined the study and retain the learners who have joined the study already. In addition, community user data from the discussion area or WIKI are very effective in predicting the loss and recommendation of the learner[3].

4.2 the application of data mining based on Colleges and Universities

The traditional teaching mode of college education is facing more and more problems, and The emergence of big data and MOOC technology proposed more possible to our education reform. MOOC education platform is able to face a wide range of students, and enhance the interaction between teachers and students, The teaching effect is analyzed and evaluated in real time to improve the practical teaching level, We should have a full application of this platform, mass storage data analysis study, summed up more physical learning rules and learning mode, so as to build up a personalized learning model, At the same time, it can make the teachers evaluate the teaching effect by the learning process data of each student, and provide the basis for the improvement of the

teaching design, so as to enhance the effect of teaching reform in Colleges and universities, the tuition of students can fully grasp the knowledge and better meet the needs of social development.

4.3 the application of data mining for learners -- a personalized course and service

The results of the data mining of the learner can be provided as the individual course and service to learners. This helps to form a new model in which MOOC platform uses data to communicate with learners, so that learners pay more attention to the MOOC platform, To learners, the MOOC platform provides rich, comprehensive, timely course information, and these can be aimed at similar learners' interest and needs, quickly filter and recommend appropriate courses and services, to provide a strong support for the study of learning.Through data mining, real-time analysis of learners' current scene and historical records, we create a possible learning model to meet the needs of learners and provide personalized service for learners.

Reference

[1] Bin Ye. The application analysis of big data in the MOOC [J]. Micro computer and application. 2015 (11). 1-2

[2] Yun-peng Zhang. Research on Web-based data mining techniques [D]. China Petroleum University 2007.2-3

[3] Chao Huang. MOOC how to attract learners continued participation [J]. China Education Network. 2013 (09) .1-4