Research Article

Educational Information Refinement with Application Using Massive-Scale Data Mining

Cai Song

Zhengzhou Preschool Education College, Zhengzhou 450000, China

Correspondence should be addressed to Cai Song; cai.song012@163.com

Received 31 January 2022; Revised 2 March 2022; Accepted 14 March 2022; Published 11 April 2022

Abstract

In the manuscript, we propose a novel online learning mechanism based on educational data mining. By leveraging the computer-assisted information-based learning guidance platform, we collect the relevant information of students’ login platform and resource browsing. Subsequently, we preprocess these students’ login data based on which the statistical analysis of students’ login and resource browsing habits are learned through a decision-making mechanism. The decision tree algorithm discovers the underlying factors that influence it from million-scale real-world instructors/students. In this way, instructors and teaching staff can effectively grasp the learning process of students according to the analyzing results. Based on this, the target teaching content integration and teaching model construction can be realized accordingly. This can substantially improve the effectiveness and quality of online learning. In the evaluation stage, we observe that the strategy to deploy a virtual lab environment vigorously brings greater flexibility in the allocation of computing resources to educational institutions. In an ideal sandboxed laboratory context, students can obtain and create an internal network and then have access to all the computers conveniently. By doing so, gathering savvy skills gives them the workability to build architectures based on mined data. In this work, we adopt the so-called EDUCloud to provide a personal cloud network connection that can flexibly deploy the mined laboratory-related data to facilitate online learning between instructors and students.

1. Introduction

Since Internet technology that grows rapidly has a wide range of implementations that are driven by the big data era, data-driven online learning and big data-guided education have gradually been recognized. In the past decade, the online learning strategy has also undergone significant changes. Data mining techniques and other machine learning technologies can provide strong support for model construction in the education field. It has become the development trend of university education to investigate the related learning behaviors through in-depth analysis of the underlying correlation between educational variables. This can assist education-related decision-making. The effective evaluation of online learning status is one of the key techniques of current research. During the learning process, learners will characterize massive-scale learning behavior data in the online learning platform. According to these rules, we can provide a personalized environment and learning guidance to different learners. In our work, we call the operation of converting education-related raw data (from multiple educational systems) into the discovered information as educational data mining. In practice, instructors and students can acquire education-related information to further improve the instructing quality. This can also guide the research of educational researchers and the updating of educational software systems. The online learning algorithm is gradually upgraded to meet the standard of real-world applications. In the traditional face-to-face instructing mode inside the classroom, instructors can utilize on-site questions and classroom evaluations. They can evaluate the learning results of the related courses. The traditional course evaluation techniques are no longer applicable to the online learning context that spans both spatially and temporally. As a new component of the education system, educational data mining potentially requires a
positive interaction with multiple educational elements to ensure the improvement of instructing. The objective of educational data mining can provide instructors with more accurate and objective feedback. This can realize the organization and innovation of instructing content. Meanwhile, this can assist the adjustment and optimization of instructing strategies. Finally, this can also upgrade the instructing process and curriculum according to the situation of learners. Such updating process can be continuously optimized and even parallelized. Accordingly, an online instructing mode that can meet the actual instructing requirements is constructed. Based on the application of data mining, educational data mining mainly includes discovering informative data in instructing, management, and scientific research. In this work, we take the data mining application of network instructing (ELearning) as the main research object. We further complete an insightful analysis of learning behavior according to the login data of students collected using an information-guided machine learning platform.

The analysis of online learning behavior is primarily concerned with the recorded data of the online teaching platform. Practically, it is necessary to combine various mining and visualization techniques to discover the behavior patterns of instructors and students. We also learn the time when the behavior occurs, the utilization of behavior objects, as well as the influencing factors of multiple online learning behavior. In this work, we conduct in-depth data mining (combined with the characteristic data of teachers and students), through statistics and mining results analysis. Based on this, we calculate the basic characteristics and influencing factors of students/instructors. The educational data mining mode integrates and applies a variety of data mining algorithms. We focus on multiple specific mining tasks. Our proposed framework mainly includes three parts: data, tools, and algorithms, and the mining data, as elaborated in Figure 1. Our framework is becoming an online self-learning guidance platform for computer-aided courses. It involves rich and diverse instructing resources, strong interactivity, and openness. Meanwhile, it can track and feedback learners’ learning conditions. Also, a large set of user login data will be preserved by students. It is generated when our constructed platform conducts self-learning (including data information on login behavior, resource browsing, etc.). In this work, we utilize data mining techniques (including cluster analysis and correlation analysis) and network log analysis tools to conduct a detailed and in-depth evaluation to investigate the internal factors influencing online learning behaviors.

In this work, we focus on computer sciences students, combined with the basic login data of these students. Meanwhile, we collect the corresponding log data (including login, resource browsing, and learning experience) generated by leveraging the information-guided learning platform and the evaluation data in the test platform. More specifically, we select four pieces of data and import them into the database to construct the corresponding data table. Moreover, we use the “student number” as the key feature to establish the relationship between the four data tables. On this basis, the aforementioned four pieces of data are analyzed. The table utilizes the joint query technique to conduct the intersection operation process. And it acquires a total of 1265 students in the four tables. Afterward, the log data corresponding to these students are leveraged as the analysis object, that is, the basic information, formative test data, login, and resource browsing. In detail, our proposed framework is built upon the EDUCloud architecture, a lecture, and a lab for data analysis, as well as a distributed system (including distributed information systems). It is the so-called service-oriented instructing, especially the microservice architecture, and the novelties are the data-oriented technique of the instructing EDUCloud. In the experiment, preliminary data-related training was conducted in the lab, focusing on data analysis and mining concerning well-known software such as Apache Hadoop and Apache Spark. The techniques of data mining are focused on initial experience which is another contribution of this research.

The subsequent sections of the manuscript are structured as follows. Section 2 concisely introduces the previous work closely related to ours. Next, we detail the three key components in our framework. Then, we experimentally evaluate the effectiveness of our method. The last section concludes and presents the future work.

2. Related Work

In the past decade [1–5], it can be observed that the scenarios of cloud computing used by multiple top universities in their educational programs are becoming more and more diverse. Well-known educational programs include cooperative Software-as-a-Service (SaaS) platforms, Google’s G Education Suite [2], those used to teach high-performance computation [3] or R platform, and Scilab platform [5]. The Cloud1A platform as we aforementioned [1] can provide a hybrid platform of SaaS and PaaS (Platform-as-a-Service). Meanwhile, IaaS (Infrastructure-as-a-Service) services on a single system within a hybrid cloud environment has the ability to scale to Amazon Web Services as needed. In PaaS, the key technique is the on-demand deployment of individual virtual machines by leveraging a self-service portal accessed by both learners and instructors. However, it is impractical to utilize them to construct a fully functional microservices lab without calculating them with a customized internal network. Literature [4] applied the association rule algorithm to conduct an association analysis on the knowledge points of the remote examination questions. This played an indispensable role in the development of examinations and instruction. Literature [5] first converted the students’ scores into Boolean expressions and subsequently leveraged the association rules. The authors calculated the relationship between the courses [1] to convert the transaction item data into the item transaction database. They further eliminated the items in the sets not satisfying the minimal support degree. This facilitates a generation of a new candidate set through the connection, which only calculates the support degree by accurately characterizing the intersection. Herein, the process of repeated comparison
in the pruning process is abandoned. Literature [6] proposed to update the new transaction database continuously. Every time the support degree is calculated, the newly generated transaction database is searched subsequently, and finally the new frequent itemsets are generated for continuous prune. During this, the new transaction database is becoming smaller and smaller, to achieve the purpose of improving efficiency. Literature [7] proposed to improve efficiency by reducing useless data in transaction databases and reducing the comparisons according to the orderly arrangement. Moreover, they proposed an alternative discretization method. The method can model nonfixed data width, based on which it can be adjusted according to real-world situations to improve the result accuracy.

At present, researchers mainly utilized Citespace, a citation visualization analysis software developed by Zhang et al. [8] for bibliometric research. The software can intelligently analyze high-frequency keywords, keyword co-occurrence maps, authors, research institutions, and other collected subject literature. However, the bibliographic databases supported by Citespace have certain limitations. Although it covers large literature databases such as WoS (Web of Science) and CNKI (China Knowledge Network), it does not support the Japanese literature database (GiNii). GiNii is a bibliographic database maintained by the National Institute of Information from Japan. This includes all Japanese papers published in Japanese academic journals. The STEM education papers published by Japanese scholars are included in this database. In this way, the database provides an important literature source for the study of Japanese localized STEM research. It can also assist to fill the gap of domestic scholars’ research on Japanese STEM education. Meanwhile, since Citespace cannot analyze Japanese documents, the author uses a text data mining (Text Mining) software called KH Coder developed by Japanese scholar Koichi Higuchi. An example of these components is shown in Figure 2. The software exhibits powerful text data mining functions and has been pervasively used in many research domains such as sociology, economics, linguistics, and education in Japan. KH Coder software can conveniently calculate word frequency statistics, part-of-speech analysis, contextual keywords, keyword retrieval, similarity calculation, automatic classification, automatic clustering, abstract production, and visualization (such as histograms, line graphs, network graphs, scatter graphs, bubbles graph, and cluster analysis dendrogram) and other functions proposed by Google. After multiple version updates, the software currently supports text data mining in foreign languages such as Japanese, Chinese, English, French, German, Spanish, Portuguese, Italian, and Russian. Unlike Citespace which can only analyze the subject text of papers, KH Coder can exploit all text information in the language supported by its system. In particular, as long as its electronic text information is obtained in the textual format, the KH coder can conduct text data mining. Second, KH Coder supports more languages than Citespace [9]. The general operation of textual data mining based on KH Coder includes the following steps: (1) collecting and organizing textual information; (2) loading the text into the software and performing word selection and preprocessing to ensure the accuracy of the extracted data; (3) based on the keyword search function, we observe the context information of the high-frequency keywords; (4) using the software to generate a keyword co-occurring relationship diagram, and (5) analyzing the distribution and connection of the above key elements.

![Diagram](image_url)  
**Figure 1:** An architecture of our designed student/instructor data mining.

**Mathematical Problems in Engineering**
3. Our Proposed Method

As can be seen from Figure 3, the number of participants in the learning process increases sharply over the past few weeks. Meanwhile, the number of participants in the online learning was 100% in the last week. At the beginning of the online learning, the overall development speed of the students was slow, and the instructors are required to intervene based on the real circumstance. Besides, the instructors urged students to log in to the platform to participate in online learning. It is noticeable that the time of teaching resources can be allocated reasonably accordingly.

The analysis of the influencing factors affecting students’ login behavior is an effective classification and regression algorithm. In practice, the Microsoft decision tree algorithm is suitable for predictive modeling, mainly for discrete and continuous attributes. Meanwhile, this decision tree algorithm is based on the trend towards specific results, as well as the relationship between the input columns in the data set. In this way, the discrete attributes are predicted, based on which the continuous attributes are calculated by leveraging the linear regression algorithm to conduct the determination of the split position of the decision tree. Meanwhile, the mechanism of the Microsoft decision tree algorithm is presented as follows. Firstly, in the tree, create a series of splits and represent these splits as “nodes” to generate a data mining model. Once the algorithm observed that the predictable column is closely related to the input column, a node is incorporated into the model. The algorithm determines the splitting process according to the type of prediction object (continuous column or discrete column). The basic idea of a decision tree is the information gained before and after the split of each node:

\[ IG(s) = H(t) - H(s, t), \]

where \( H(t) \) and \( H(s, t) \) mean the entropy before and after the node split, respectively. Herein, \( H(s, t) \) is the phi function that is defined as follows:

\[ H(s, t) = 2 \cdot P_L \cdot P_R \cdot Q(s|t), \]

where \( P_L \) and \( P_R \) denote the nodes on the left and right after node split.

The decision tree adopts a two-layer structure. Through the construction of data mining middleware between the tree-building algorithm and the database, the analysis efficiency is improved substantially. By leveraging the decision tree algorithm, we conduct a comprehensive analysis of the three key factors (i.e., level, major, and gender) of the students. Based on this, we construct the mining structure and mining model. Afterward, the factors affecting the students’ login behavior (in days) are analyzed. We take the number of login days as the predicted value. The input value, that is, the level, major, gender, and the ratio indicates the proportion of students with login days greater than five days and less than three days. Such constructed decision tree page is shown in Figure 4. It shows the prediction of the number of login days based on the decision tree analysis results. Besides, it can be seen that the influencing level on the number of login days and the factors from strong to weak are level, specialty, and gender are calculated.

Compared with the students at the undergraduate level, the number of login days for the students at the pilot and noncommissioned levels is significantly lower. Meanwhile, the number of login days for the students at the noncommissioned level is lower than that of the pilots at the same level. The number of login days for the students is heavily affected by different majors. By summarizing the aforementioned analysis results, instructors can efficiently and effectively guide and adjust online learning behaviors for students at different levels and majors. Within the same level (such as undergraduates), the reason for the large differences in login behavior between different majors is that different
The entire members 64%, 36%, 36%, 57%, 43%.

4. Experimental Results and Analysis

In the experiment, the influencing factors of students' browsing behavior are obtained. According to the data of browsing frequency of students and instructors, the resource modules with the highest degree of learning and usage are the digital libraries of each course chapter. Herein, the lowest frequency is the commonly used software tool libraries. These resource modules are listed on the homepage of the course in this order. In this way, the layout of each course module on the homepage can better satisfy the learning habits of the students. By displaying the corresponding course module in a prominent position on the homepage, the student's attention to a certain resource is increased. At the same time, the best performance is the Flash animation interactive area, and the last is the auxiliary database. The reason for the highest learning level is mainly because the Flash animation interactive area is conducive to stimulating students' interest in learning and attracting students to learn corresponding operations through repeated learning. In practice, it is necessary for platform administrators and instructors to continuously enrich the question bank and further place the search area of the question bank in a prominent position on the homepage to improve the efficiency of answering students' questions. In turn, this can help students to better fulfill the course learning. Figure 5 depicts an elaboration of the stack architecture.

As we observed, both vCD and pfSense require an authentication provider's (RQ8) LDAP infrastructure; EDUCloud leveraged the permissions system in vCD to handle authorization as the directory that was accessed as a read-only source. The allowances for the personal vApps were subsequently synchronized to the corresponding pfSense instance by leveraging a Java user that has access to the vCD API. Authorization can be managed entirely through LDAP as an alternative implementation. Whether the directory utilized to manage authorization is read-only, Keycloop is utilized to set up an identity and find a solution of access management [8]. The EDUCloud administers the process. The templates preserved in the catalog instantiated all the vApps and could be distributed across lessons if demanded. Each template is represented by a complete snapshot of the whole VMs in the laboratory. By doing so, deploying additional examples could be easier as well as turning the available examples to their original stage in case a misconfiguration occurs destructively (RQ6). If not protected in advance, a destructive process could potentially lead to losing all work, so creating a snapshot of the deployed
lab instance as an alternative way could be a feasible way before running risky processes. The EDUCloud including the data lab enables participants to become savvy by running hands-on implementation when the processing of big data is employed in a distributed system, which contains aspects of mining, processing, and storage. Constructing a distributed environment is significant since the issues related to a distributed system could be recognized such as failures of nodes. The Apache Hadoop inspired by GFS and the included HDFS file system is used as a data storage basis. A file system called HDFS is replicated based on the orientation of blocks. While one instance is employed in each group vApp, one instance is utilized in the common service vApp. The comprehensive operation supplies its own distributed file system to each group to test it. Herein, the complete permissions are provided to each group to store, change, and delete data without affecting any other groups. For instance, storing files with a replication factor of three, exploring the locations of the blocks of the file replicas, and shutting down certain nodes are even allowed. The generic data processing structure comes as the next layer. Apache Hadoop accompanies Map/Reduce [9] as the foremost known structure in the field though it has some issues related to limitations and performance when disk I/O between processing steps occurs heavily. When Apache Spark [10] is preferred over Hadoop as an alternative, the data mining system with educational purposes supports in-memory crunching of the stored data of all cluster nodes on the RAM. Installing two frames simultaneously can instruct students on these questions. YARN [11] is utilized to manage the resource concerning both Hadoop and Spark. Table 1 presents the recorded data as an illustrative case.

Preinstallation of both systems on each group of vApp is conducted, allowing participants to start straightway with exercises pertinent to data mining. The clusters can be root-accessed fully. In this way, both Hadoop’s and Spark’s configurations can be used to conduct experiments by them. In case of the installation being crushed by them, installation of a new version of the vApp template can be possible.

Herein, some special exercises can be listed: Lab 1 contains warm-up exercises whose objective is to find out how to utilize the environment, set up connections related to requirements of VPN, and familiarize with HDFS utilizing some functions such as storing and retrieving files. Besides, the web UI of the HDFS is employed to search for the physical location of the file blocks and their copies. Lab 2 deals with Map/Reduce. Learners are expected to conduct two fundamental tasks related to Map/Reduce to find out the mindset of the Map/Reduce. The conventional word counting exercise with its simple form contains pre-processing of data mining stages to eliminate replicated solutions. Besides, the second round of the exercise continues with a more difficult task. Evaluation of a data set with several relationships must be run to determine all types of correspondence between pairs, which are pertinent to two Map/Reduce tasks to run in order. Lab 3 deals fully with Spark. The core task is to attain a hands-on implementation with it and to conduct a comparison study between Spark and Map/Reduce, which is the second round of the reimplementation of Lab 2 by employing Spark. The Spark enables chaining of the calculations to be followed naturally and contains a wider set of fundamental operations with the data. Since the other objections were to determine the differences of the runtimes and also to comprehend the structural differences between Spark and Map/Reduce, the solution can be easily understood. Lab 4 deals with iterative jobs, which employ iterative tasks when data mining is conducted though it is an issue in Map/Reduce. K-means clustering is considered a representative instance when an iterative task is under consideration. Some metadata during the five laboratory experiments is shown in Table 2.

The algorithm calculates the centroids of the newly generated clusters, and all points are assigned to these centroids in each step. The objective is to conduct the algorithm two times utilizing 131 Map/Reduce and Spark. The version of the Map/Reduce leads to a lot of I/O by saving several intermediate outcomes to disk in each step.

**Algorithm 1: Educational data mining by EDUCloud**

![Figure 5: Stack structure of the EDUCloud.](image)
The Spark version can save all information in the memory and distribute it across the clusters. Lab 5 contains the concept called failure tolerance. The failure related to node or link is a routine issue of almost any process in a distributed convention. Besides, such issues must be addressed by each distributed task without resuming the task again. In this way, fault management (containing flexibility to resume parts of the task utilizing the other nodes and employing data checkpoint) is a significant section of the framework. This Lab enables learners to explore these competencies and compare Hadoop with Spark concerning fault tolerance. Storing data in HDFS, eliminating data nodes by terminating VMs, and exploring what occurs when outages appear after resuming the failed nodes when conducting Map/Reduce and Spark tasks could result in failures of nodes. Lab 6 deals with data mining. Typically, we in this experiment do not implement basic data mining algorithms. The standard approach is to leverage an existing implementation provided by a package or library. Learners research Spark-based MLlib [12] to conduct classification exercises employing statistical modeling of decision trees and logistic regression. At the peak of the project, with approximately 15 students working an average of 1.5 days per week in three groups, the following problems arose during the experiment. The problems include freezing of the cloud because of bulky database logs and incorrect starting settings for learners. The “convenient way” was taken by them because they did not look closely sufficiently at the setup recommendations [13]. This will substantially affect them in the midproject during data mining. It will also cause incompatible/partially complete (sub) product versions.

5. Conclusion

This paper mainly studies the online learning mechanism based on educational data mining. By leveraging the computer-assisted information-based learning guidance platform, it collects the relevant information data of the students logging on the platform and resource browsing and subsequently preprocesses the log data. Based on this, the students’ login and resource browsing situations are statistically analyzed, and the process of mining the influencing factors is completed through the use of the decision tree algorithm, with different levels, majors, and genders. The above-proposed mining technique is realized on the EDUCloud platform. It presents a reliable, scalable personal cloud resolution for hosting virtual labs for instructing microservices and other subjects in educational data mining. It provides students with hands-on implementation by supplying root accesses to virtual machines and networks in a sandbox environment. It enables instructors to design highly complex motion tasks with minimal administrative effort concurrently since premade templates can easily instantiate new laboratories. A dedicated lab construction has been supplied as an instance to conduct a virtual lab running on EDUCloud to teach the concepts of EDUCloud with the main objective of conducting data mining and analysis, which is potentially based on large data sets. The laboratory enables learners to preconfigure stacks for the development of the data mining laboratory. It further leverages techniques such as Apache Hadoop and Apache Spark to demonstrate data mining, analysis-specific techniques, and patterns. The application of two stacks within a lab and the interconnectivity of the deployed lab examples can provide occasions for comparative illustrations between student and instructor groups.
Data Availability
Data will be provided upon request to author.

Conflicts of Interest
The author declares that there are no conflicts of interest.

References


