Comparison of Domestic and Foreign Users’ Learning Characteristics Based on Behavior Logs—Taking Online Resource Logs of Three-dimensional Teaching Materials as an Example

Run-xin YANG and Ji-mei LI*
School of Information Science, Beijing Language and Culture University, Beijing, China
*Corresponding author

Keywords: Data mining, Three-dimensional teaching materials, Behavior logs, Learning characteristics, Comparison between domestic and foreign users.

Abstract. With the extensive advancement of information technology in education, personalized learning has become the most basic needs of learners, and the three-dimensional teaching materials have the characteristics of diverse media forms, flexible learning time and rich online resources. It is one of the main sources of personalized learning resources. Therefore, the three-dimensional teaching materials have become an inevitable trend in the development of teaching materials. This paper makes theoretical research and data mining on users’ learning behavior logs and three-dimensional teaching materials. Some data processing methods based on behavior logs, such as address translation, stay duration statistics and access path tracking, are used to empirically study the logs of three-dimensional teaching materials. The data indicates that users prefer to visit “web page” and “course”, but less access to “file” and “discussion area”, and the average duration of domestic users in each study area is similar, while the average time spent by foreign users on “web page” is much longer than others. In addition, the time data is adjusted by time difference, and it is concluded that domestic users prefer to learn during weekdays, while foreign users prefer to study at night on weekdays. The three-dimensional teaching materials optimization strategy proposed includes enriching the medium modes, strengthening the innovative practice, etc.

Introduction

With the extensive advancement of information technology in education, the combination of educational technology and classroom teaching has become increasingly close, and more and more universities have adopted online teaching platforms for teaching. In the field of education, educational big data is becoming a hot nod. In the continuous development of the three-dimensional teaching materials, online resource website has accumulated a large amount of users’ learning behavior data. The analysis of these data will help the writer to analyze users’ learning characteristics, which is beneficial to the improvement of the three-dimensional teaching materials. However, the logs of the three-dimensional teaching materials have not yet attracted attention. Few people use the users’ behavior logs to conduct in-depth data mining research, especially comparing the learning characteristics of domestic and foreign users.

This paper collects users’ behavior data of “A comprehensive practice course for scenario-based corporate finance business” [1] (referred to as “Comprehensive Practice Course”) in the 2017.1.6-2018.8.7 period from Mingdeng network online teaching platform [2] (referred to as "Mingdeng"), and uses the processing methods such as address translation, stay duration statistics and access path tracking to reveal the usage of domestic and foreign users. In addition, through the analysis of the overall visits, browsing popularity and users’ learning time preferences, the user's behavioral preferences are summarized and the three-dimensional teaching materials characteristics are analyzed. These mining results can be used to analyze the usage and learning characteristics of users of the three-dimensional teaching materials. Finally, based on the experimental data and analysis results, the optimization suggestions for the three-dimensional teaching materials are given.
Related Works

The three-dimensional teaching materials, also known as the integrated teaching materials or diversified teaching materials, refers to relying on modern science and technology, being based on traditional paper teaching materials, focusing on subject courses, integrating audio and video products, electronic products, and online education service platforms [3]. At present, there are many researches on log mining and user behavior analysis.

In China, Wei Shunping conducted data mining on the online resource website in 2011, and took a training course as an example to analyze student visits, training time, student cluster and social network[4]; Wang Ning mainly expounded the effectiveness of distance learning on online teaching platform[5]; Jin Jie took a course as an example to study the correlation between participation, homework completions, teacher participation and learner participation [6]; Zhang Guorong used fuzzy C-means to conduct fuzzy clustering research on the learning data of a certain course[7].

In foreign countries, Romero, C., Ventura, S. & Garcia, E. used data mining to perform statistics, visualization, classification and clustering on various data in online teaching platforms [8]. Jovanovie built a classification model based on homework completions and correct quantities to predict users’ learning [9]; Lavigne clustered students based on student logins, clicks and session time [10].

In summary, most of them are researching the teaching modes and teaching methods of online teaching platforms. There are few analysis of online resource logs of three-dimensional teaching materials. To promote the deep application of the three-dimensional teaching materials and understand the users’ learning characteristics, this paper intends to use the conventional statistical methods, visualization methods, path tracking and other methods to conduct a theoretical research and data mining on logs of the three-dimensional teaching materials on the Mingdeng.

Log Data Preprocessing Method for Comparative Analysis of Domestic and Foreign Users

Log is some operations of the specified system object and a time-ordered set of its operations [11]. Due to the diversity of log forms, log data needs to be pre-processed before the log analysis. This paper uses address translation, stay duration statistics and access path tracking methods.

Address Translation

The log from Mingdeng is in Excel format and the categories are “Time”, “User Name”, “Event Situation”, “Component”, “Event Name”, “Description”, “Source”, and “IP Address”.

First, users need to be identified. In the actual access process, the following situations usually occur [12]: multiple users may use the same machine to access the same website; the same user may access the same web server on multiple machines; the same user may use multiple browsers or even use a different operating system to log in to the same website; different users may access the same web server through the same proxy server or firewall. Based on the above cases, there are three rules to identify users: use IP to represent different users; when the IP addresses are the same, different operating systems represent different users; when IP and operating system are same, different browsers represent different users. According to these rules, this paper uses the IP address to represent the user exclusively. Nowadays, it is more and more common for people to use their own computers, so it is feasible to identify users according to this rule. Address translation is to extract the IP information from the log and convert it into the specific geographical location information through the IP address database of the network. Generally, the interfaces for IP address translation include Taobao, Sina, NetEase, Tencent, Sohu, etc. This paper uses the Taobao.

Stay Duration Statistics

To count the user's stay duration, first need to identify the user's session. Session identification refers to a user's effective access to the server [13]. In a log file, the same user may access the same website multiple times, and only a click with a small-time interval can represent a user session. Usually, the process of identifying a user session is to define a time limit $\Delta T$ ($\Delta T = 25$ minutes) [12].
The first step is sorting the tables from large to small by IP address. The second step is extracting data of "time", "event situation" and "IP address" from the log table and store them. Then judge the learning area and calculate the access time difference. The third step is judging the validity of the time difference calculated in the previous step. The judgment depends on two aspects: whether they belong to the same IP or the same session.

Access Path Tracking

Access path tracking is to select "Time", "Description" and "IP Address" from the log table, and to calculate the path of each chapter. First, in the "Description", each event occurs with a numeric ID. Therefore, using the regular expression to extract the operation ID can achieve the purpose of extracting the path. Second, judging the validity of the access path depends mainly on two aspects: the series of visits belong to the same session and two or more paths are counted as valid paths. Third, in the output of results, there are two options: one is to sort according to the frequency; the other is to sort according to the number of operands. Since the maximum operands of the access path is not known in advance, the first sorting scheme is selected, that is, the sorting output is performed according to the frequency size without limiting the maximum path length.

Comparative Analysis of Domestic and Foreign Users Based on Logs of Three-dimensional Teaching Materials

Learning behavior refers to a series of learning-related activities carried out with some motivation [14]. The analysis of users’ learning behavior is based on the log data of a three-dimensional teaching materials, and performs statistical analysis, data mining and data visualization to compare learning behaviors between domestic and foreign users.

This paper applies some basic concepts: PV (page view) refers to the number of page views. An html request is treated as a PV; UV (unique visitor) refers to different natural people who access and browse this webpage; average page views refers to the average number of visitors per page and average duration refers to the average number of the user's duration on the specified webpage.

In this study, the user behavior log of “Comprehensive Practice Course” is collected as a data sample, and all behavior data from 2017.1.6 to 2018.8.7 is extracted. This paper first analyzes the geographical distribution of users. This study extracted 79,095 samples and the UV is 6712. This study uses the address translation method to process the IP address and the result is shown in Figure 1.

As shown in Figure 1, China's Beijing (UV=705), Shanghai (UV=549) and Henan (UV=453) are among the top three. Through statistics on users, the number of domestic users (UV) is 5228, and the number of foreign users (UV) is 1484. It is worth noting that the number of foreign users has also reached 22.12%. Foreign users are distributed in 20 countries including USA, France, Britain, Japan, etc. 1484 foreign users visit the platform, which indicates that the three-dimensional teaching materials has a certain influence abroad. Therefore, this study separates the log data of domestic and foreign users, analyzes the learning characteristics of users separately, and facilitates the better improvement of the teaching materials for different users.
Based on the statistical analysis of the geographical distribution, this paper uses stay duration statistics and access path tracking to compare visits of each learning area, learning time preference, the duration and access path of each learning area.

1. Visits of each learning area

This study counts the visits of the study areas “course”, “discussion area”, “web page”, and “file”. “Course” refers to user clicking on the name of the teaching materials to access the specific webpage. In the “discussion area”, users can view the post and discuss. In the “web page”, users can browse the teaching materials, and in the “file”, users can download document.

As shown in Figure 2, users visits “web page” the most frequently, while “discussion area” and “file” is the least. “Discussion area” belongs to the interaction area, and “file” belongs to the resource download area, which indicates that the behavior of “doing” by users is far less than “watching”.

Three-dimensional teaching materials requires effective interaction between teachers and students, helping to understand the learner's problems, and update the teaching materials timely. So, the feedback should be more effective and communication with students should be more frequent. Teachers should occasionally appear in the discussion area to communicate with users, and quote the contents of “file”.

2. Learning time preference of domestic and foreign users

To deeply analyze the distribution characteristics of users’ learning time, this study firstly adjusts the time difference and counts the user visits.

As shown in Figure 3 and Figure 4, domestic users have the highest PV on Thursday and the lowest PV on Saturday and Sunday, indicating that domestic users prefer to study on weekdays. Foreign users are more inclined to study on Thursdays and Fridays. Domestic users tend to study in the afternoon while foreign users are more inclined to study the teaching materials in the early morning or evening.

![Figure 3. Visits by week.](image)

![Figure 4. Visits by period.](image)

Through the above-mentioned results, teachers can conduct online tutoring during the working day for domestic learners and the early morning for foreign learners.

3. Duration and access path

1) The statistical table of duration in each study area is shown in Table 1.

<table>
<thead>
<tr>
<th>Study area</th>
<th>Average duration [min]</th>
<th>Total duration [min]</th>
<th>UV [per person]</th>
<th>PV [per person]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Domestic</td>
<td>Foreign</td>
<td>Domestic</td>
<td>Foreign</td>
</tr>
<tr>
<td>Discussion</td>
<td>5.65</td>
<td>0.62</td>
<td>322</td>
<td>96</td>
</tr>
<tr>
<td>File</td>
<td>4.68</td>
<td>3.16</td>
<td>332</td>
<td>155</td>
</tr>
<tr>
<td>Web page</td>
<td>6.83</td>
<td>7.55</td>
<td>121517</td>
<td>42088</td>
</tr>
<tr>
<td>Course</td>
<td>5.66</td>
<td>4.4</td>
<td>38223</td>
<td>2965</td>
</tr>
</tbody>
</table>

As shown in Table 1, domestic users’ learning behavior has the following characteristics: First, although “web page” is far greater than the others in total duration, UV and PV, the gap in the average duration is not so large. Second, although “discussion” is far lower than the “web page” and “course” in the last three indicators, it is similar in average duration, which indicates although the number of
users visiting the “discussion” is much lower than “web page” and “course”, the average duration is relatively high and it has a certain user viscosity. If “discussion area” can be popularized among users, the average duration will be significantly improved. While the total duration, UV, PV and average duration of the “web page” of foreign users are far greater than others, teachers should upload some English files in the “file” to help foreign learners learn further.

2) The statistical table of access path is shown in Table2:

As shown in Table2, users visit the most is “Course: comprehensive practice course” (referred to as “course”). In addition, users prefer to browse a page of a relevant section. There are two reasons: First, if users need to access the detailed chapters of teaching materials, they need to access through this course. Second, people tend to focus only on certain knowledge points that they need to learn or are interested in, without having to complete the entire teaching materials. Compared with domestic users, foreign users do not pay special attention to "courses", but prefer to browse a small section of the chapter interests them.

<table>
<thead>
<tr>
<th>ID (frequency)</th>
<th>Path</th>
<th>ID (frequency)</th>
<th>Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>(35-35', 258)</td>
<td>Course</td>
<td>(291-291..., 18)</td>
<td>Discussion area: News and notification</td>
</tr>
<tr>
<td>(35-35-35', 98)</td>
<td>Course</td>
<td>(35-35-35..., 8)</td>
<td>Course</td>
</tr>
<tr>
<td>(35-410', 13)</td>
<td>Course -- Web page: Fill in the requisition</td>
<td>(35-35-35', 4)</td>
<td>Course</td>
</tr>
</tbody>
</table>

Therefore, in the three-dimensional teaching materials, the links between the chapters should be strengthened instead of fragmenting the chapters, in order to help the users to learn more efficiently. Besides, the innovative practice can be strengthened as well. Combined with the existing three-dimensional teaching materials, universities can set up specific classrooms corresponding to the needs of talents, encourage the participation in relevant innovation and entrepreneurship contests, and carry out targeted training for learners to lay the foundation for future work.

Summary

This paper mainly studies the logs on the online teaching platform, and makes an empirical study on the learning characteristics of domestic and foreign users based on the logs of the three-dimensional teaching material "comprehensive practice course" in the Mingdeng. The main achievements are as follows.

Firstly, the research content of users’ logs based on online resource website is defined. Based on the log of the three-dimensional teaching material, the data processing methods such as address translation, stay duration statistics and access path tracing are used to preprocess user data. Secondly, use these methods to make an empirical study of behavior logs. The geographic location is analyzed first. Based on the geographic location, this paper divides users into domestic users and foreign users and makes a comparative analysis in four aspects: visits of each learning area, learning time preference of users, duration and access path. Finally, according to the experimental data, it can be analyzed that the different learners have different access to each learning area as well as the distinction in the learning time preference, duration and access path. The teaching materials optimization strategy proposed includes optimizing the structure of three-dimensional teaching
material, differentiating three-dimensional teaching material, strengthening effective feedback, enriching the mediums, and strengthening the innovative practice.

There are still some imperfections in the work. User identification are defective. Due to the small number of registered users, most of them are visitors. Therefore, this paper does not consider the situation that domestic users may use external network or foreign users are all Chinese students.

The next stage will be based on more online logs, analyze users, especially registered users, and optimize the three-dimensional teaching material; through the analysis of student feedback information, dig deeper into the existing problems and optimize the online resources.

Acknowledgment

This research project is supported by project of Beijing Language and Culture University (supported by “the Fundamental Research Funds for the Central Universities”, No.16YJ030001, 17ZDJ02,17PT01, OTP201607), the National Computer Basic Education Research Association Project (No. 2018-AFCEC-177) and the Teacher Support Program (No. OTP201607).

Reference


