Analysis and Portrait of College Students' Online Behavior Habits

Yupeng Jiang, Yingying Zhao and Hui Xu

School of Computer Sci. & Tech, Nantong University, Nantong, 226019, China
Email: xu.h@ntu.edu.cn
*Corresponding author

Abstract. In the era of information technology, the popularity of college networks has brought great convenience to students. However, some students indulge in the network and thus affect the quality of learning. In order to help college students make use of network resources properly, we analyze students’ surfing behavior from the perspective of data mining. 11 classes of web pages with high degree of attention are considered. The data of more than 100 million log records generated by students in universities within 20 days is collected. We also use Bisecting K-means algorithm to cluster students on different levels. For measuring every student’s behavior, we introduce a stability measure for web surfing behavior. We calculate this parameter for every student and analyze the results. Finally, according to the clustering results, cluster stability analysis results and other attributes, we draw a user portrait for every student. After practice, the method is simple and easy to use, providing reference for universities to organize network management activities and standardizing college students’ online behavior.

1. Introduction
Nowadays, universities have their own intranet. Students will leave a lot of records while surfing the Internet. These log records contain the objective fact attribute, which can reflect the students' online behavior. If rely on traditional manual to find the law of students' online behavior, the workload is large and the efficiency is low. Therefore, it is feasible to analyze network behavior based on data mining. Internal and external researchers choose different data mining methods to study the characteristics of Internet users and their online behaviors. Hu Zuhui [1] correlated students' online time and student achievement, and used decision tree, association rules and logistic regression to analyze the online time and students' learning quality, concluded that the length of the Internet will affect the student's performance. Shuangzhe et al. [2] established a new probabilistic behavior model. The Markov mixed model is used to analyze the relationship between logs, and the probability relationship and feature data are obtained. Xu Changda [3] explore potential habits and patterns of user group from the perspective of user online behavior characteristics. Wei Xuyang [4] proposed a user access pattern mining algorithm based on browsing interest. To solve the sparse nature of web server log data, using browsing interest as weight for data mining. Liu Mengchao [5] proposed a method on constructing online data cubes, Successfully unearthed the online behavior characteristics of network users from different dimensions and granularities such as the duration of the Internet, the time period, and the type of browsed web pages. Most studies are exploring how to define and calculate user online habit parameters. Received a series of research results based on user classification. But just a little study the stability of user habits. Every student has their own online behavior habits, indulging in games or using the Internet to enrich themselves. These problems need to use data mining technology to analyze students' online log data comprehensively. This is the meaning of this study.
2. Related Concepts and Technologies

2.1. Pearson Correlation Coefficient
Correlation analysis is necessary. The primary method can quickly find the connection between data, such as Pearson correlation coefficient. Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. It has a value between (-1, +1), where 1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation.

2.2. Bisecting K-means
Bisecting K-means is a variation of k-means clustering that refines cluster assignments by repeatedly attempting subdivision, and keeping the best resulting splits. It has the following advantages:
- Accelerate the execution speed of the algorithm, because of the reducing of similarity calculation.
- It will not Converge to local minimum.

2.3. User Portrait
User portraits were first proposed by Alancooper, the father of interaction design. The user portrait refers to a tagged and abstracted user model according to the user's attributes, user's preferences, user’s habits, user’s behaviors, etc., and the user needs to be tagged, and the tag is a highly refined feature identifier that is analyzed by analyzing the user information. By tagging, we can use some highly generalized, easy-to-understand features to describe the user, making it easier to understand the user and making it easier for the computer to process. Many scholars construct user portraits of mobile Internet and Internet data, describe the user's group characteristics, and provide data support for accurate decision-making. Select a word cloud to visualize the label.

3. Data Mining Model

3.1. Data Collection
The data used in this paper are collected from the real-time log data of a university's network monitoring platform. Figure 1 plots the structure of this platform. It has Mysql database running on the servers to record the large amount of log data. To ease the analysis we choose last 3 weeks data from monitoring platform in this paper. After deleting some missing records, more than 100 million log data of student’s surfing logs are extracted from the monitoring platform.

![Figure 1. Data collection.](image)

3.2. The Websites Classification
In this paper, we use the web surfing frequency of different type of web sites to reveal the user's behavior patterns, therefore website categorization is critical throughout the project. However, for unlabeled data sets, there is no optimal solution for dimension selection. 11 types of online behaviors with high attention and research significance were selected, shown in table 1.
Table 1. Types of Website.

| Type                     | IM (Wechat) | Online shopping | IT related | Entertainment news | Social network | Game |
|-------------------------|-------------|-----------------|------------|-------------------|---------------|------|
| IM (Wechat)             | 1           |                 |            |                   |               |      |
| Mobile application      | 2           |                 |            |                   |               |      |
| Game information        | 3           |                 |            |                   |               |      |
| News                    | 4           |                 |            |                   |               |      |
| Education               | 5           |                 |            |                   |               |      |
| Online video            | 6           |                 |            |                   |               |      |

3.3. Data Preprocessing

The platform transmits the collected data to the database every 60 seconds, therefore we have the original network statistical data on the time granule of 60 seconds. A user's web behavior cannot vary frequently so we choose 24 hour as the time granule of examination. For every user, we count the times he or she visited a certain type sites in 24 hours, and divide it by the total web surfing numbers to get the surfing frequency of this type of sites in 24 hours. So we get a 11-dimensional vector for every user:

\[ V^d_i = (f_1, f_2, \ldots, f_{11}) \]

where \( i \) denotes the ith user, \( d \) denotes the dth day, \( f_1 \) is the surfing frequency of ith website type. Due to the huge amount of data, this paper uses block thinking to process data sets. The server's original logs were sorted, consolidated, then grouped by student number to obtain 86,498 vectors.

4. Linear Correlation Analysis

In real life, people who like to play games often check some game information. People who like to watch variety shows often search for some entertainment information. Before further analysis, we try to figure out whether the surfing frequency of some web site types correlate to each other in a linear way. Using 86498 11-dimensional vectors as the data set, the Pearson coefficient of every two features is calculated to obtain the result.

![Figure 2. The correlation matrix of 11 website types.](image)

It can be seen from the Figure 2 that there is some correlation between some Internet types. Taking 0.3 as the threshold, we can get the following five pairs of connected Website types as followed.
Table 2. The correlation of websites.

|                | Positive correlation | Negative correlation |
|----------------|----------------------|----------------------|
| Game information--Game information | IT related--Game information | IT related--Education |
| IT related--Education | Game--Education | Entertainment news--Online video |

5. User’s Web Surfing Behavior Clustering

For the 86498 vectors, the Bisecting K-means clustering algorithm is used for clustering, and the number of iterations is set to 60 for the center point. After repeated testing, Select 8 cluster centers to get the best effect. The clustering results are shown in Figure 3:

As can be seen from Table 5, the largest cluster center is cluster 7, having 22286 points. Cluster 7 represents a common cluster of users. Among the pages viewed by users, education is the most popular, accounting for 41%. Followed by IT-related web pages, accounting for 0.32%. This user group is not very interested in games and chats, and occasionally look at the news.

Cluster 4, cluster 5, and cluster 6 are also relatively large clusters, each having 15996, 16267, and 17653 points. Cluster 4 users use WeChat or QQ frequently, also like to browse IT related web pages. The IT-related and online audio and video accounts for a large proportion of the webpages browsed by the cluster 5 user group, and occasionally look at news and entertainment information. Cluster 6 user group games and browsing IT related web pages are more, and often chat and watch news.

There are 13,857 users in cluster 3, and the browsing pages are concentrated in IT and news, and there is little interest in other web pages.

Users in clusters 0 and 2 have similarities, and they all prefer to use the mobile phone to browse the web. The cluster 0 user group is a heavy use group of the mobile terminal, and is concentrated on the use of the APP, and also browses a few web pages in a small amount. Cluster 2 users use the app on their mobile phones to chat, watch news, and browse IT-related web pages.

From the results of the analysis, in addition to some of the smaller clustered users, most users like to browse technology-related web pages, browse some news and education websites, in line with the characteristics of contemporary college students advancing with the times. However, there are some games that rely heavily on users, such as cluster 1, which requires network administrators to do related processing.

6. Stability of User’s Web Behavior

A particular user may belong to the same cluster at different times, or may belong to a different cluster, which represents the invariance and variability of the user's online habits. In this part, we analyze the stability of users' online behavior. Every user belongs to a certain cluster on the dth day, we denote it as \( \hat{C}_d \).
\[ \hat{C}_d = (0,0,0 \ldots,1,0,0,\ldots) \]  
(2)

This vector has 8 dimensions and the ith dimension is set to 1 while others are set to 0 if the user belongs to the ith cluster on the dth day. If we inspect the behavior of the user from day 1 to day m, we will face a vector sequence.

\[ \hat{C}_u = (\hat{C}_1, \hat{C}_2, \ldots, \hat{C}_{20}) \]  
(3)

We define the stability of a user’s behavior as the sigma of the vector sequence as following:

\[ \text{Stab} = \frac{\sum_{i=1}^{m} ||\hat{C}_i - \overline{\hat{C}_u}||^{1/2}}{m} \]  
(4)

\[ \text{stab} \in (0,1), \overline{\hat{C}_u} = \frac{1}{m} \sum_{i=1}^{m} \hat{C}_i, 0 \text{ means stable.} \]

Calculate the stability function for each user and get the stability factor of 15147 users. In order to visualize the results, draw a cluster stability bar chart.

Figure 4. Result of cluster stability.

Some conclusions can be drawn from the figure 4, There is a local maximum of 44%, The corresponding value is 0-0.2. It corresponds to students who have a fixed online habit and a fixed online mode every day. 0.2-0.6 accounts for 8%, which belongs to a group of students with a fixed online mode. Another big value is 28%, which means that there are quite a few users with a wide range of interests and they will not follow a fixed page view mode. For this part of the students, we use user portraits to understand their online behavior.

7. User Portraits

This paper mining different feature dimensions, extracting the user’s objective attributes rather than the user’s self-expressing attributes to give customers different labels, making it easier for administrators to understand the students’ online behavior habits. The customer portraits studied in this paper include the following dimensions:

- Access preference: the preference classification information of the user network access, combined with the log data and the clustering result to characterize the access interest preference;
- App use: portrays students’ behaviors on the Internet, such as Internet communication (QQ, WeChat), social (microblogging, space), news, materials, games, etc.
- Terminal dimension: characterizes the terminal type preference used by the user, and if it is a mobile terminal, includes the model number and parameter configuration of the terminal;

In order to understand the web pages and apps that users often prefer, it is necessary to first count all online log records of users within 20 days. However, because the amount of data is too large, it is impossible to uniformly process all data, and it is necessary to perform block operations in the data processing process. According to the hardware conditions, 1 million records are taken as one chunk, which is divided into a total of 142 chunks. Process and count each record in each chunk. After all the chunk statistics are completed, integrate all the chunks and use the numpy library to convert and count.
the data to improve the statistical speed. Use the classifier to extract the page type and the app application name and count them to get the top 30 keywords in the word frequency sorting. Re-integrate clustering results, access hobby tags, app tags, terminal information tags, etc. The user protrains we get shown in figure 5.

![Word cloud of user portrait](image_url)

**Figure 5.** Word cloud of user portrait.

8. **Conclusions**
This paper collates and counts the online records of all users within 20 days, realizes the binary clustering of all users' Internet access frequency vectors, and also analyzes the stability of users in each cluster. According to these data dimensions, the user's portrait is made by using the word cloud map, which realizes the visualization of the log record and intuitively displays the user's online habits and preferences.

The network administrator can analyze the online behavior of students according to the results of the user's portraits, and improve the efficiency and accuracy of network management. The next research direction is to explore the relationship between students' academic performance and online habits, and gradually improve the accuracy of the clusterer to improve the analysis accuracy.

9. **References**
[1] Hu Z H, Wei S. Research on College Students' Online Behavior Analysis and Data Mining. China Distance Education, 2017(2):26-32.
[2] Shuang Z. Research and application of hidden Markov model in web page information extraction. East China Normal University, 2016.
[3] Xu C D. Multi-dimensional analysis method of network user behavior. Beijing Institute of Technology, 2015.
[4] Wei X Y. Research on user access pattern mining model based on Web log. Southwest University, 2015.
[5] Lu M C. Application and implementation of data mining in user online behavior analysis. Nanhua University, 2013.