2020 5th IEEE International Conference on Big Data Analytics

Data Mining and Feature Analysis of College Students' Campus Network Behavior

Liu Kesheng[1,2], Ni Yikun[1,2]*, Li Zihan[2], Duan Bin[2]
[1]School of Economics and Management, Beihang University
[2]Big Data Center of Student Affairs Department, Beihang University,
Beijing, China
*Corresponding author, e-mail: niyikun@buaa.edu.cn

***Abstract*—The rise and promotion of big data methods enables teachers to understand the behavior patterns of students in a timely and accurate manner, especially to find out the groups of students that need to be focused on in time, and to help promote the student affairs management from empirical qualitative knowledge to scientific quantitative analysis. This paper applies the clustering method of data mining to analyze the campus network behavior of 3,245 students in a certain grade of B university, obtains a total of 23.843 million Internet access data in 4 years. The result shows 4 groups of students with different characteristics of Internet access, finds 350 students with large network usage. Achievements and other aspects of performance of these students are affected. This study carried out data mining of student campus network behavior, which can be used as a practical operation case for student affairs management data mining, providing effective data support for the accurate and scientific development of student affairs management.**

Keywords- Data mining, Student Network Behavior; Big Data

I. INTRODUCTION

A major problem for student affairs management is the contradiction between the limited energy of student counselors and the diversity of student behaviors, which results in many potential problem students losing the opportunity for early intervention. Since the beginning of the 21st century, the rapid development of information technology in education and the construction of digital campuses has made it possible for student counselors to conduct quantitative analysis of student school behaviors, especially to provide early warning to students who may have problems, so that the contradiction could be alleviated by applying the analysis and early warning methods.

As contemporary college students who grew up in the Internet era, their daily life, learning and thinking are deeply influenced by the Internet. This provides us with the possibility to understand their campus network behavioral characteristics through big data. How to mine useful information for student counsellors from massive data in the explosive growth of data categories and data scales, is a challenge for current student counsellors, also an important opportunity to conduct work by new means.

This study starting from the actual work problems and was conducted based on the network behavior data of B college students, combining big data thinking and big data mining methods, researching the characteristics of college students' network behavior rules, and detecting the students who need pay close attention because the large amount of campus network usage. This study could also carry out as a practical case of student work data mining for reference.

II. METHODS

A. Data acquisition and data cleaning

The data involved in this study mainly include campus network usage data and other students' campus behavioral data. The campus network usage data mainly includes 8 fields: ID, online date, offline date, online time, offline time, inbound traffic, outbound traffic, and total traffic. Each piece of data records the login information once, and there is a total of 23.843 million pieces of campus network usage data. Other school data include performance data, campus card consumption data, book borrowing data, and physical fitness test data.

The data collected directly is generally incomplete, random, and accompanied by a certain amount of noise. We cleaned the data before using the data mining method. Through data cleaning, some invalid data and private information are eliminated, and useful information for research is retained for further analysis and research. This research focuses on the study of the characteristics of student groups and does not involve the privacy of students' personal Internet content.

B. Index system construction

The initial data in this study only involved 8 fields data of 3245 students for 4 years: ID, online date, offline date, online time, offline time, inbound traffic, outbound traffic, and total traffic. In order to objectively and accurately describe the characteristics of students' network behavior based on the initial fields, this study builds a network behavior index system for the next analysis and research through mathematical statistical methods.

According to the school's existing teaching cycle and school calendar, each academic year is divided into 10 semester months. Since the last semester month is the graduation season, the data quality is poor, so a total of 39 semester months are considered in this study. Taking each semester month as the statistical dimension, 28 periodic analysis indicators (1-39 semester months) related to the length and flow of each student were obtained. The specific indicators are shown in Table 1.

978-1-7281-4111-4/20/$31.00 ©2020 IEEE

TABLE I. CAMPUS NETWORK BEHAVIOR RESEARCH INDEX SYSTEM

Index number	Index Name
1	Time
2	Flow
3	AveTime
4	AveFlow
5	DayTime
6	NightTime
7	AveDayTime
8	AveNightTime
9	WeekdayTime
10	WeekendTime
11	AveWeekdayTime
12	AveWeekendTime
13	DayFlow
14	NightFlow
15	AveDayFlow
16	AveNightFlow
17	WeekdayFlow
18	WeekendFlow
19	AveWeekdayFlow
20	AveWeekendFlow
21	Flow/Time
22	DayFlow/DayTime
23	NightFlow/NightTime
24	WeekdayFlow/WeekdayTime
25	WeekendFlow/WeekendTime
26	Times
27	Flow/Times
28	Time/Times

C. Data processing

The campus network usage data involved in this study has the characteristics of large amount of data, high frequency of data recording (data is recorded every time you log in), and continuous changes in semester months. The characteristics of the above data are similar to those of functional data (curve data) [1]. So, consider using the functional data processing method to process the campus network usage data.

The research index data based on the campus network using the initial data statistics is continuous data of an index for each student within 1-39 semester months, which needs to be processed into functional data for subsequent analysis. The principle of processing is to retain the information carried by the original data to the greatest extent, and to perform operations such as smoothing to satisfy some basic assumptions of functional data analysis. Common preprocessing methods include linear interpolation, kernel method smoothing, and spline method approximation [2]. This study used the spline method to process 28 research index data of 3245 students. The processing was completed by writing R language code using Rstudio software.

The smoothed curve data was approximated by a B-spline basis function, the smoothed data was converted into functional data, and further principal component analysis of the functional data was performed. Based on the principle that the variance ratio is greater than 90%, the selected principal component was selected [3]. And extract the selected principal component coefficient, use the selected principal component coefficient vector as the basis for the next data mining.

D. Data mining

Data mining is the process of knowledge discovery which based on a large, incomplete, noisy, fuzzy, random, and original data set, revealing hidden information, previously unknown, but potentially valuable and ultimately understandable information [4]. Conventional data mining deals with traditional data, and mainly treats data as discrete data points. For data mining of functional data, there have been researches that extend traditional methods to functional data processing. These studies have laid a theoretical foundation for the development of this study [5-6]. This study carried out the clustering analysis based on coefficient vectors obtained from principal component analysis of functional data.

Cluster analysis is to classify samples according to their individual characteristics. Through continuous iteration, samples with similar characteristics and rules are in a class, and there are relatively obvious differences between classes. In this study, we use cluster analysis to group student groups with different campus network usage patterns. By analyzing the characteristics of different categories, we can better understand the students' campus network usage patterns and help to discover the group of students who need to pay more attention to the large degree of network usage. It can provide data support for improving the scientific and accurate of students' affairs management.

Currently, the most widely used cluster analysis methods are K-means clustering and hierarchical clustering. K-means clustering algorithm is simple in principle and convenient for processing large amounts of data, but K-values need to be determined through cross-validation and other methods. Hierarchical clustering does not need to specify the number of clusters in advance, and you can find the hierarchical relationship of the classes, but this method is suitable for small data volumes. The calculation speed is slower, and the efficiency is lower when the data volume is large. Therefore, in this study, K-means method was finally selected for clustering.

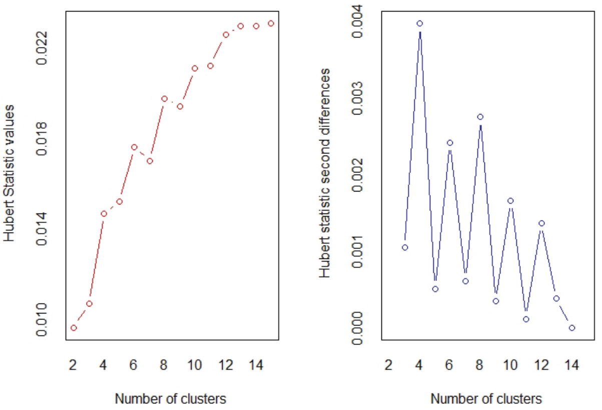

Figure 1. Determine the optimal number of clusters

K-means clustering must first determine the optimal number of clusters. In this study, the optimal number of clusters is determined by Hubert index and cross-validation as 4 categories, as shown in Figure 1. The implementation process is mainly based on the mature algorithms of Nbclust package and Factoextra package in Rstudio.

III. Results

A. Clustering results and basic information of various student groups

The K-means clustering algorithm takes 28 research indicators constructed in this study as input, and finally outputs the category label of each student. Through analysis, 3245 students were divided into 4 categories: 230 in the first category; 958 in the second category; 1707 in the third category; and 350 in the fourth category.

TABLE II. Basic Information Of Different Student Categories

Category	Basic Information				
	Number	*Male*	*Percentage of Male*	*Female*	*Percentage of Female*
1	230	58	25.2%	172	74.8%
2	958	795	83.0%	163	17.0%
3	1707	1362	79.8%	345	20.2%
4	350	251	71.7%	99	28.3%

B. Campus Network Usage Characteristics of Various Student Groups

As shown in Figure 2, the total length of Internet access for categories 1, 2 and 4 is similar to the duration of daytime, nighttime, and midweek and weekends. Overall, there were no significant differences in the length of time spent online.

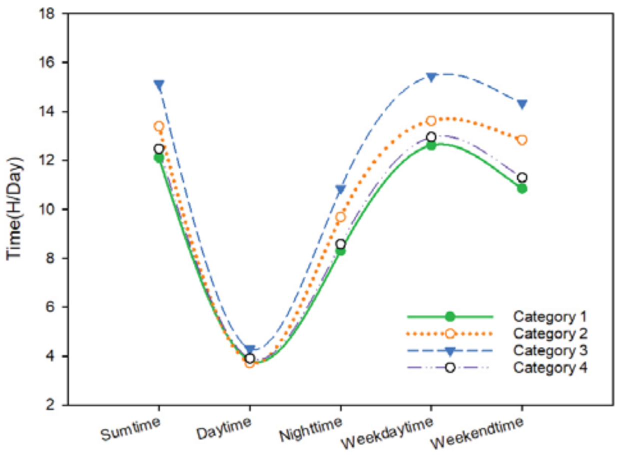

Figure 2. Length of online time of different student categories

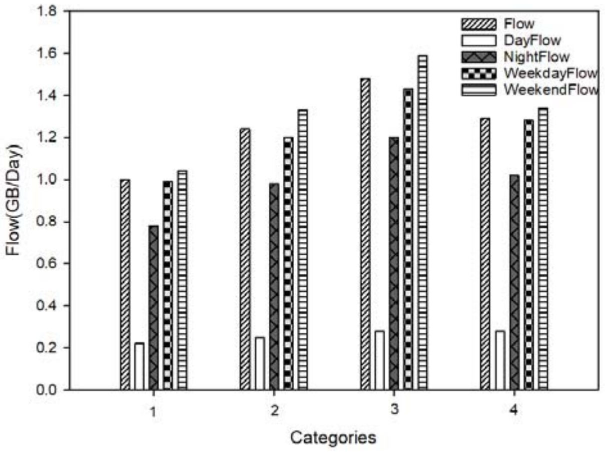

Figure 3. Online traffic of different student categories

It can be seen from Figure 3 that the total traffic of category 3 and the traffic during the day and night are slightly higher than those of categories 1, 2, and 4, and the statistics of Internet traffic of the type 1 student group are the lowest.

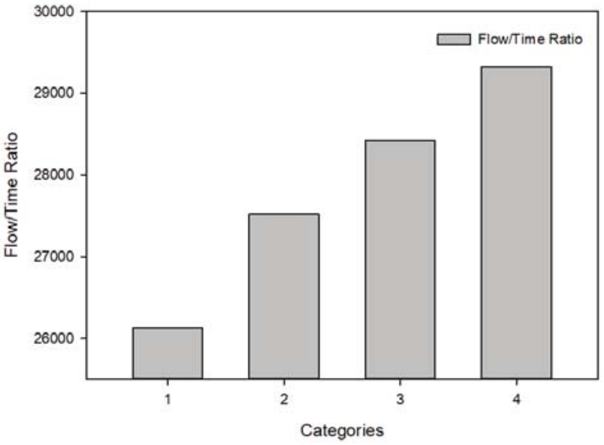

Figure 4. Flow/time ratios of different student categories

It can be seen from Figure 4 that the flow/time ratios of categories 1 to 4 increase sequentially. Further analysis of the flow/time ratio of 39 months for various groups of students (as shown in Figure 5), it can be found that after entering the second year of university (10-20 semester months), the flow/time ratio has significantly increased, and each category have similar rules. After entering the third year of university (after 20 semester months), the traffic time ratio of category 4 and category 3 increased significantly, and category 4 was significantly higher than the other 3 categories.

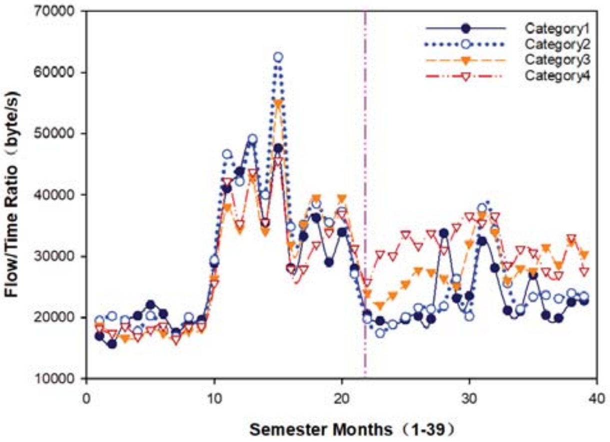

Figure 5. Flow/time ratios of 1-39 semester months

Further analysis of the flow/time ratio of the Internet traffic of the four types of student groups during the day and night. During the 1-39 semester months, the flow/time ratio of various types of students during the day is similar (as shown in Figure 6-a). The flow/time ratio of nights in categories 3 and 4 increased significantly, and category 4 was significantly higher than the other three categories (as shown in Figure 6-b).

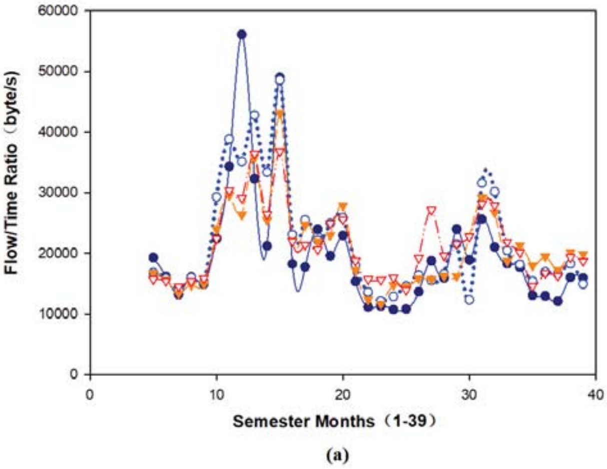

Figure 6-a. Flow/time ratios of during the day

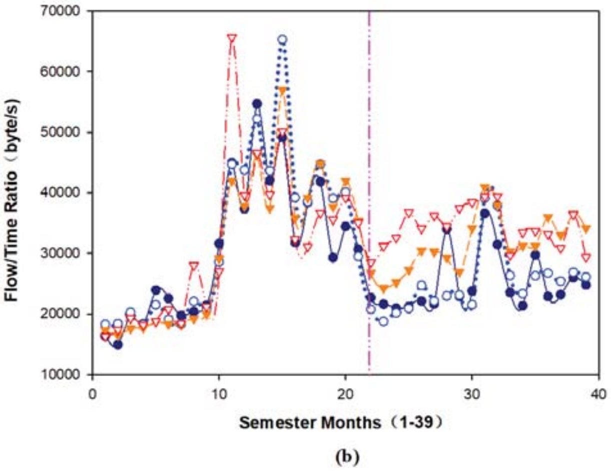

Figure 6-b. Flow/time ratios of during the night

It can be seen from Figure 7-a that the values and rules of the average traffic of single logins of the four types of student groups in freshman and sophomore stages are similar, indicating that the single-time Internet traffic use of various students in freshman and sophomore stages are similar. After entering the junior year, the frequency ratio of category 4 traffic increased significantly, and it continued to be the highest among several categories during the junior and senior year.

In terms of the length of a single Internet connection, the data of the first category students in the first three school years are lower than those in the other categories, and in the fourth school year, they are slightly higher than those in the other categories, as shown in Figure 7-b.

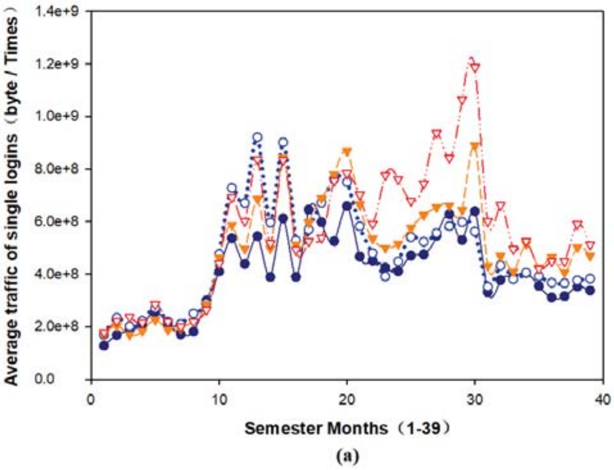

Figure 7-a. Average traffic of single logins

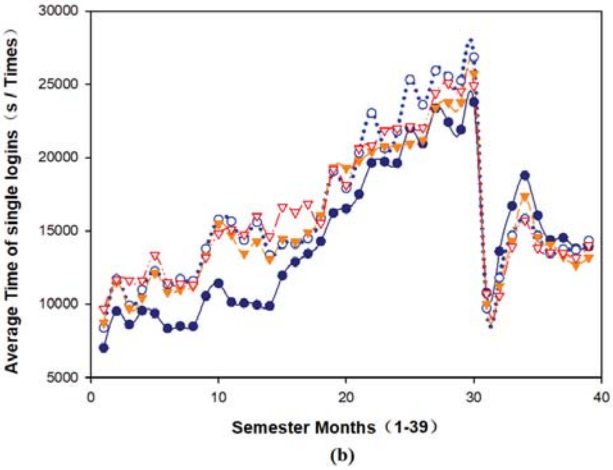

Figure 7-a. Average time of single logins

C. Other school behavior data of different student categories

The network data reflects the characteristics of the use of campus networks of various students. In combination with other behavioral data at school, it helps to comprehensively understand the characteristics of various types of students, and further analyzes the impact of different types of campus network use behaviors on student performance at school. Other school behavior data involved in this study mainly include performance data, campus card consumption data, total book borrowing, and physical fitness test data.

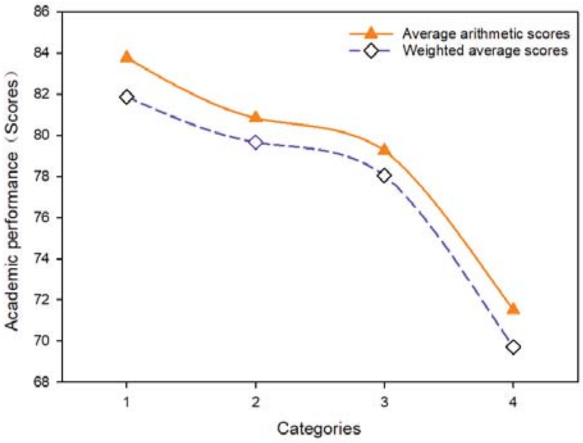

Figure 8. Academic peerformance of different categories

In terms of performance, this study counted the average arithmetic scores and weighted average scores of students in each category, and the results showed that the average scores of the first, second, third, and fourth categories of students decreased sequentially, as shown in Figure 8.

In terms of physical fitness tests, the results of category 1, 2, 3, and 4 long-distance running (1000 meters for boys and 800 meters for girls) decreased by category. With comprehensive physical fitness tests, the performance of category 1 students was higher than the other 3 categories. Grades of category 2, 3, and 4 are close.

In terms of breakfast consumption, this study counts the number of breakfast consumptions of various groups in the canteen, and the results show that the number of breakfast consumptions of the fourth category canteen is significantly lower than the first three. In terms of book borrowing, there is no significant difference between the various student categories.

IV. DISCUSSION

A. New Features of Student Campus Network Use in the Internet Era

The continuous development of networking technology has brought human society into an era of "everywhere on the Internet", making today's college students deeply imprinted on the Internet in their behaviors such as learning, socializing, shopping, and entertainment. In the "2018 Chinese College Students' Online Leisure and Entertainment Behavior Monitoring Analysis Report", a survey on how long college students can accept to leave the mobile phone to access the Internet shows that only 2% of students said that they can accept online without mobile phones for more than 24 hours. 8.1% college students are completely unacceptable to go online without a mobile phone [7]. With the advancement of the times, the Internet has become a basic element of contemporary college students' study and life, and mobile phones have become the most important tool for Internet use.

With the advancement of campus informatization construction, students can use mobile phones and computers to easily connect to campus Wi-Fi in dormitories, classrooms, and even walking on campus. In this scenario, students can stay “online” for a long time or even throughout the day. As a result, the data performance of the “online duration” field has become more and more similar. For example, in the 4 categories of students obtained through cluster analysis in this study, 3 types of data are close to each other (categories 1,2,4), and category 1 is slightly higher than the other 3 categories (category 3). There were no significant differences overall. Therefore, in terms of measuring the degree of student's network usage, "online time" cannot truly represent the amount of network usage. Evaluation of network usage requires other network behavior data as a reference.

Compared with online time, Internet traffic can indirectly reflect the content characteristics and network usage of users. Common high-traffic behaviors include online video watching, online games, downloading, etc. While static content browsing such as text and pictures, and social software applications, etc. are often low-traffic behaviors. Among the four categories of student groups obtained by the cluster analysis in this paper, the total traffic results show that the third category are slightly higher, the first category is the lowest, and the second and fourth types are close. The four types of student categories have small differences in traffic. In terms of flow/time ratio, the first, second, third, and fourth categories increased in order, and the fourth category was significantly higher than the first three. The flow/time ratio indicates the user's network traffic per unit of time. Students who often conducts high-traffic behaviors (such as online games and videos), their flow/time ratio is bound to be higher than other groups. In addition, this study also calculated the average duration and average traffic distribution of the single login campus network for various student groups, as shown in Figure 8. The results show that in terms of the average length of a single login, the fourth category is close to the second and third categories, but the average traffic of a single login is significantly higher than the other three categories (as shown in Figure 8-a). Based on the above analysis, the fourth type of student group uses a large amount of campus network and has higher traffic behavior than the other types.

B. Network behavior characteristics and other school data

Based on student network behavior data, this study uses clustering methods for data mining to obtain 4 types of student categories with different characteristics of campus network usage. By correlating other school behavior data, it is found that course scores, long-distance running test scores (1000 meters for men, 800 meters for women), and the number of breakfasts in the cafeteria of the fourth category are lower than those in other categories, indicating that their learning, physical fitness and living habits have been affected. The degree of campus network usage may be an important influencing factor.

C. Implications for student affairs management

This article uses the big data method to study the characteristics of network behavior of an entire grade of undergraduates in four years. It has three implications for student affairs management:

First, in terms of the network behavior of the student group, the use of the campus network by some students after entering the junior year is significantly higher than that of other student groups. For third-year students, it is necessary to pay attention to preventing students from obsessing the Internet. In this study, the flow/time ratio and average traffic per single sign-up of students in categories 3 and 4 increased, of which 1,707 were in category 3 and 350 were in category 4, accounting for 63.4 % in total. Therefore, most of the students will increase their use of the campus network after entering junior year. However, students who have a similar increase in category 4 should pay special attention. These students are likely to be addicted to the Internet and form Internet addiction, which requires timely intervention. In addition, the data shows that the flow/time ratio of various student categories during the day is similar, but for the fourth category of students, the night traffic time ratio has significantly increased after entering the junior year. Teachers and counselors engaged in student work should promptly understand the situation of these students, timely discover and guide students to a healthy schedule and active learning.

The second is the impact of online behavior on learning. The results of this study show that the fourth category of students with a large amount of campus network use is significantly lower on the academic performance than the first and second class of students with a relatively small amount of use, indicating that excessive Internet use may result in a significant decline in academic performance. For those students who are inability to use the network scientifically, timely guidance in this regard is necessary. However, for student counselors, it is still necessary to dialectically view the relationship between the Internet and academic performance. At present, the study and life of college students can hardly be completely separated from the Internet, and it is unrealistic to completely ban students from using the Internet. The front-line counselors should pay more attention to guiding students to use the Internet reasonably, exert the positive effect of the Internet on students 'growth and success, minimize its negative impact, and make the Internet a help for students' learning and growth.

Third, this study completed the entire process of campus network behavior data mining, including raw data cleaning, data processing, data mining, and analysis and application of data mining results. The problems it solves are limited, but its greater value may lie in providing a practical example of student work data mining.

The methods involved in this research and the relevant program code written during the data mining process can be directly applied to the analysis of existing real-time data mining, which can provide data references for the work of front-line counselors and help the counselors understand the campus network of the student categories. It is helpful to timely find out the students who need to focus on the large amount of campus network usage, to take timely educational assistance measures to help them get a healthy schedule and active learning.

Finally, student campus behavior data mining can help teachers and counselors engaged in student affairs management to objectively and timely understand the status and behavior of student groups, and target students with potential problems that need attention. However, the data reveals only "effects". "Cause" also requires face-to-face communication. Student counselor work is essentially a people's work. The exchange of ideas and collisions is an area where data is difficult to describe. Therefore, it is necessary to dialectically apply data mining results to make use of big data technology supports the data of student work, and constantly improves the professional and scientific level of student counselor work.

REFERENCES

[1] Ramsay J. O., Silverman B. W. Functional data analysis[M]. New York: Springer, 1997.

[2] Ramsay J. O., Silverman B. W. Applied functional data analysis: methods and casestudies[M]. Vol. 77. New York: Springer, 2002.

[3] Kesheng Liu, Siyang Wang. Variable selection in regression models including functional data predictors. Journal of Beijing University of aeronautics and astranautics, 2019, 45(10): 1990-1994.

[4] Romero C., Ventura S., Data mining in education[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2013, 3(1):12-27.

[5] Locantore N., Marron J., Simpson D., et al. Robust principal component analysis forfunctional data[J]. Test, 1999, 8(1):1–73.

[6] Yao F., Lee T. Penalized spline models for functional principal component analysis[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2006,68(1):3–25..

[7] iiMedia. 《2018 Chinese College Students' Online Leisure and Entertainment Behavior Monitoring Analysis Report》[EB/OL].2018-11-16.https://www.iimedia.cn/c400/62969.html.