Analysis and Application of College Students' Network Behavior Based on Data Mining

Wei Zhang[1,2], Ying-ying Gao[2] and Thelma Palaoag[1]
[1]College of Information Technology and Computer Science, University of the Cordilleras, Baguio, Philippines
[2]School of Computer, Xian Yang Normal University, Xianyang, China

Keywords: Data Mining, Machine Learning, Network Behavior, User Portrait, Academic Early.

Abstract: With the continuous development of information construction in universities, campus network has become the main channel for teachers and students to study, live and work. While the Internet brings convenience to teaching, it also has some negative effects. This paper collects one-year data of the university campus network server firstly, and adopts data mining technologies to complete the pre-processing of college students' Internet data through massive data cleaning, label naming, clustering and transformation operations. Then the classification model of college students' network behavior is constructed by user behavior classification algorithm and Pearson Product-Moment Correlation Coefficient is used for correlation analysis. Subsequently K-means algorithm is used to cluster data and machine learning method is used to match data patterns. Finally, the association rule mining algorithm is used to draw conclusions related to research objectives from students' behavior data. The results show, except for the small use of campus network during winter and summer vacations, the average time spent on campus network in other months is more than 300 hours, and the average time spent online accounts for 41.67% of the total school time. The top three of college students' concerned fields are social networking (25.65%), search (17.92%) and video (16.27%). And the top three types of Internet access for excellent students are learning (27.95%), video (18.51%) and social (13.44%). The analysis of results express that the network can improve college students' academic performance, but it also negatively affects their studies. According to this study, university management departments can optimize and guide students to use the network appropriately, and to improve the informatization level of student management gradually.

1 INTRODUCTION

In China, the number of college graduates will reach 11.79 million in 2024, and this growing group has attracted wide attention from all walks of life (Jenkins et al, 2019). For student management workers, the lack of analysis of college students from the level of data science is easy to ignore and miss some hidden problems, such as addiction to games, mental health, consumer borrowing, etc., which may cause major problems. With the continuous popularization and application of computer technology and network technology, network behavior has become one of the main factors affecting college students' learning. The success or failure of college students' studies will affect the stability and development of society, so how to guide and educate college students to complete their studies is very important.

The research on user behavior can be traced back to the 19th century human ethology (Samadi et al, 2017), which mainly studies and analyzes the content of users' visits to websites. Xu Yong et al. (Tollo et al, 2015; Cherniaiev, 2017) studied the principles and data modeling methods of correlation analysis and principal component analysis by analyzing the online behavior data of college students, and analyzed the source data from the aspects of users' online duration, time period and types of websites visited by SAS

This work was supported by the National Natural Science Foundation of China(No:62073218),
Shaanxi Higher Education Teaching Reform Research Project (23BY143),
Academic Leader Project of Xianyang Normal University (XSYXYDT202123),
Shaanxi Provincial Education Reform Project (23BY143), Xianyang Normal University Education Reform Project (2023ZD02)

Zhang, W., Gao, Y. and Palaoag, T.
Analysis and Application of College Students' Network Behavior Based on Data Mining.
DOI: 10.5220/0012876000004536

In *Proceedings of the 1st International Conference on Data Mining, E-Learning, and Information Systems (DMEIS 2024)*, pages 22-30
ISBN: 978-989-758-715-3

software. At present, many colleges and universities use big data to support campus management and decision-making as well as the analysis of students' behavior rules, but the research on college students' online behavior generally stays in the aspect of inquiry and display of online behavior, and there is no correlation analysis between college students' online behavior and academic performance. With the continuous advancement of the smart campus project, the campus network data covers a variety of information, such as students' browsing history, the use of learning platforms, social media interaction, etc. The completion and application of the campus network provides a basis for collecting college students' network behavior data.

This study collects Internet data from a university's campus network server, and uses relevant data mining technologies (Li, 2018) to complete the pre-processing of college students' Internet data through massive data cleaning, label naming, clustering and transformation operations, and uses K-means algorithm (Lakshmi and Krishnamurthy, 2022) to cluster the college students' Internet behaviors. This paper selects the URL based keyword acquisition method to analyze the types, and duration of college students' visits to websites. The Apriori association rule mining algorithm (Lin et al, 2023) is selected to analyze the association between college students' network behavior and academic performance, and the preferences of network behavior of excellent and backward students are obtained, and then a portrait of college students' network behavior is formed.

Through the user portrait of college students' network behavior, the following three goals can be achieved:

1. College students can better understand their own network behavior and correct their bad network behavior in time.

2. University administrators can give academic early warning and remind students to complete their studies successfully.

3. Help schools better understand the needs of college students and provide personalized education and support services.

Compared with the study of college students' network behavior by questionnaire survey, the data collected in this study is large and more universal. By using K-means algorithm to cluster data and machine learning method (Minxue, 2017) to match data patterns, the analyzed results of college students' network behavior are accurate and efficient, which adapts to the new needs of information technology applied to society and has certain innovation. With this, the study leads to the analysis and application of college students' network behavior based on data mining.

2 METHODOLOGY

2.1 Data Collection and Preprocessing

In this study, logs and session data of a university network management server were used as samples for analysis, including user ID, website access time, access duration, online period, online duration and other information. To protect student privacy, personal identifiers in the data are anonymized.

The data collection and processing process is shown in Fig.1. First, the network behavior data collected on the campus network server is stored in the college students' network behavior database, and then the data is cleaned, classified, feature extraction and other operations by using data mining technology. Finally, the classified data of college students' network behavior is obtained.

This paper collects the historical data of the school's network management system, billing system, Dr.Com logs and NAT logs from November 15,2022 to November 15,2023 for data analysis and processing. The data sources for this article are detailed below.

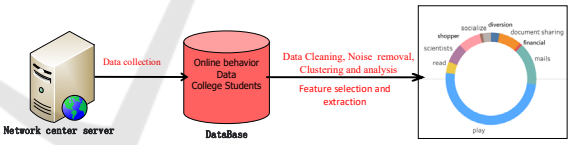

Figure 1: Basic process of data acquisition and processing.

Serial Number	Source Address	Source port	NAT Address	NAT Port	Access Time	Connect Time	Target Address	Target Port
1	192.168.0.27	7871	222.35.184.151	22334	2023/9/25 15:36:30	16	117.146.116.137	80
2	192.168.0.27	7826	222.35.184.151	22209	2023/9/25 15:36:23	16	122.72.6.68	80
3	192.168.0.27	7749	222.35.184.151	22065	2023/9/25 15:36:09	30	122.72.62.222	80
4	192.168.0.27	7684	222.35.184.151	21981	2023/9/25 15:35:54	46	122.72.6.68	80
5	192.168.0.4	51937	222.35.184.151	22449	2023/9/25 15:36:53	3	61.183.12.22	80
6	192.168.0.27	7946	222.35.184.151	22456	2023/9/25 15:36:53	3	122.72.62.222	80
7	192.168.0.27	7411	222.35.184.151	21542	2023/9/25 15:34:35	24	23.44.155.27	80
8	192.168.0.27	7411	222.35.184.151	21542	2023/9/25 15:35:00	101	23.44.155.27	80
9	192.168.0.27	7900	222.35.184.151	22384	2023/9/25 15:36:40	1	122.72.19.147	80
10	192.168.0.27	7921	222.35.184.151	22407	2023/9/25 15:36:47	10	122.72.99.89	80
11	192.168.0.27	7920	222.35.184.151	22406	2023/9/25 15:36:47	2	111.13.110.38	80
12	192.168.0.27	7782	222.35.184.151	22100	2023/9/25 15:36:14	28	122.72.99.89	80
13	192.168.0.27	7895	222.35.184.151	22379	2023/9/25 15:36:38	4	122.72.63.35	80
14	192.168.0.27	7488	222.35.184.151	21656	2023/9/25 15:35:00	42	22.72.6.68	80
15	192.168.0.27	7893	222.35.184.151	22375	2023/9/25 15:36:37	13	103.10.87.142	835
16	192.168.0.27	7695	222.35.184.151	21996	2023/9/25 15:35:57	45	122.72.62.222	80

Figure 2: Panabit application gateway session logs.

2.1.1 Application Gateway Flow Control System Session Log Data

In order to comprehensively analyze Internet user

behavior by understanding users' preferences for Internet applications, it is necessary to obtain the types of Internet application data of users. Therefore, we also extracted nearly one year's session logs of Panabit application gateway flow control system as raw data for analysis. The session logs of the panabit application gateway traffic control system record (Biswal, 2011) the behavior characteristics of Internet users. The generated data records are based on each network application access request of the user. The session logs contain the user account, source address, source port, access time, destination address, and NAT address, as shown in Fig.2.

2.1.2 Accounting System Log Data

The Dr.com broadband billing system is widely used in the broadband IP network operation of colleges and universities. It records the network behaviors of users in detail, facilitating campus network administrators to manage the network and users to access the Internet while realizing authentication and billing. The workflow of the authentication server is shown in Fig.3:

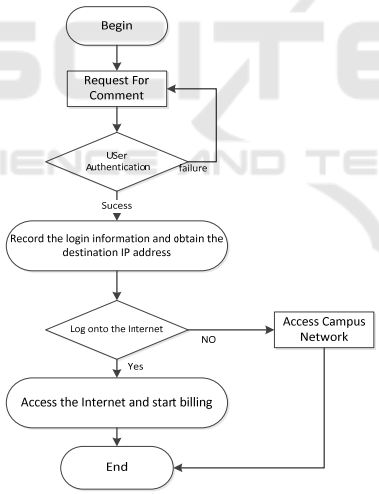

Figure 3: Workflow of the authentication server.

There may be many problems in the collected network data, such as vacant data and data unrelated to the research content, so these data need to be cleaned in the data preprocessing. Cleaning objects include useless attributes, vacant data, and non-research object data of the above system data. For a behavior analysis system, it is not necessary to care about all the fields, so remove the fields that are not needed. In the NMS data history table (Yuan, 2023), you only need to extract the fields of User MAC Address, AP Serial Number, Connection Time, Disconnection Time, Uplink Traffic bytes, and Downlink Traffic bytes.

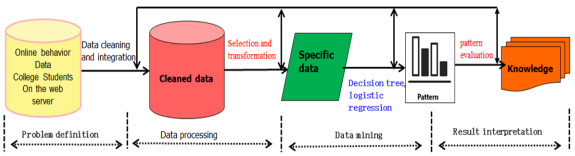

Figure 4: Basic process of data mining.

In addition, campus network users include college students, faculty members, faculty members' families, etc., student users' identification is a 10-digit student number, faculty members' identification is an 8-digit faculty number, and other user names are irregular. Since the content of this study is the analysis of college students' network behavior, the search statement is used to select 10-digit student users with student numbers, and the student users starting with student numbers 19-22 (2019-2022) are restricted. These student numbers represent current students, and other network User records are excluded.

2.2 the Process of Data Mining

Data mining refers to mining hidden, unknown, interesting and potentially valuable knowledge and rules from a large amount of data (Zhao et al, 2020). These can provide some useful information for teaching management, business decision-making, market planning and financial forecasting. Fig.4 shows the basic process of college students' network data mining, which consists of four stages: problem definition, data processing, data mining and result interpretation. In the mining process, the collected data is cleaned and integrated according to the rules formulated by the research needs, and then the operation data such as duplicate records, missing values and outlier values are removed. Then the data format is converted according to the needs to obtain the specific data. Then the decision number and logistic regression are used to predict and analyze the specific data to obtain the prediction structure. Finally, the predicted results are evaluated, and if there is a large deviation from the actual results, the data cleaning and conversion operations are carried out again (Bansal et al, 2020). Data mining is mainly the mining of correlation, trend and characteristics of data. The correlation of data determines a non-deterministic relationship between data, that is, the

correlation relationship, that is, the impact of changes in one attribute of data on other attributes. This study is mainly used to analyze the correlation between college students' Internet behavior and academic performance.

2.3 Classification of Students' Network Behaviors

2.3.1 URL Based Keyword Acquisition

User behavior classification refers to the process of classifying users according to category preferences when browsing web pages. The collected log data contains the URL of the web page visited by the user, extracts the keywords from the URL, and establishes the URL topic list of the web page visited by the user. Based on the user's topic list, you can build a category of web behavior associated with the website category.

The webpage classification method based on URL keyword extraction is used to classify the webpage. In order to get the theme of the web page according to the URL string, the most direct method is to get the theme corresponding to the URL by matching the artificially marked website category directory. Website classification directory is the collection of information on the Internet website together, according to different classification topics, placed in the corresponding directory. It is the most direct and accurate method to get the URL topic by matching the website classification directory, but due to the large workload of marking, it needs to consume huge manpower and time, and the amount of marking data is also very limited. In order to solve this problem, the design of webpage classifier is proposed, based on the N-gram language model of webpage classification algorithm, using URL classification directory matching to determine the URL theme, the URL of all web pages are mapped to the corresponding webpage theme one by one. Improve the efficiency and accuracy of web topic determination, so as to obtain more comprehensive user behavior classification information.

2.3.2 User Behavior Classification Information Representation

After accurately obtaining the topic information of the web page visited by the user, the topic list is transformed into the user behavior classification information to provide material for the input part of the user behavior classification model. After the topic list is obtained, by counting the number of occurrences of each topic in the topic list, we get the binary group composed of topic t_i and frequency

Table 1: URL keywords for web page categories.

Topic	Keywords
Game	gamersky, game, 4399, 7k7k, 17173, ali213, yy, douyu,egame,
Social Network	extshort.weixin.qq,weibo, btrace.qq,weibo,tieba.baidu,jiayuan,tianya,zhihu,
Contact	music.163,kugou,y.qq,fm.taihe,xiami,kuwo,yinyuetai,changba,music,
Video	policy.video.iqiyi,video.ptqy,video,ixigua,v.qq,haokan.baidu,youku,v.baidu,mgtv,acfun
Study	wps,cnki ,dict.youdao,wpscdn,flashapp,chinaz,processon,dxzy163,icourse163,mooc
Science	Ludashi,windowsupdate,apple,idianshijia,sandai,duba,ubuntu,zol,ithome
Load	Download，sz-download.weiyun，ardownload.adobe，download.hongbaoshu
Read	xxsy,zongheng,qidian,read,faloo,qidian,novel,jjwxc,lrts.me,zongheng,ximalaya,
Search	Baidu.sohu,news.sina,candian,guancha,mil.ifeng,huanqiu,junshi.china,yahoo,sogou
Shopping	taobao,alibaba,alipay,dangdang,suning,mogu,1688,mi,

c_i $(t_i, c_i)\}$ and form all the resulting binary groups into a binary list {(t1,c1), (t2,c2), …, (t_i, c_i)}, the binary list is the user interest information. It will serve as input to the building part of the college student interest set. When keywords corresponding to the theme appear in a URL, they are mapped to the corresponding theme, and the statistical URL keywords are partially displayed (see Table 1).

2.3.3 The Classification and Construction Process of College Students' Network Behavior

First, the list of urls accessed by users is obtained, and the URL topic is obtained by extracting URL keywords as mentioned above, so as to obtain user behavior information. Then, the classification model of college students' network behavior is constructed by user behavior classification algorithm. The construction process of college students' network behavior classification is divided into four steps:

Step1: Extract the original information. Extract urls accessed by users, count the number of visits to each URL, and generate a binary of urls and visits (URL, counts).

Step2: Obtain keyword information. Use the keyword acquisition method based on URL feature extraction to get the Topic information of URL, that

is, the category of college students' network behavior, and generate the binary group of topic and Counts.

Step3: Calculate the topic weight. Count the number of occurrences of each topic among all visits, and divide the number of occurrences of each topic c_i by the total number of accesses c_{all} to obtain the weight corresponding to each topic:

$$w_i = \frac{c_i}{c_{all}} \quad (1)$$

Step4: Update the topic weights. When a new access theme needs to be calculated, the total number of user visits and the number of visits corresponding to each theme should be updated c_i, and then the number of visits of each theme divided by the total number of user visits to obtain the weight corresponding to each theme, that is:

$$w_{i\_new} = \frac{c_{i\_old}+c_{i\_new}}{c_{all\_old}+c_{all\_new}} \quad (2)$$

First, the list of urls visited by users is obtained, the number of visits to each URL is counted, and the binary group of urls and visits is obtained. Then, the topic corresponding to each URL is obtained for each binary group, and the binary group of topics and visits is obtained. Finally, the user's proportion of each topic is obtained through the above user behavior classification pattern algorithm, and the triplet list of topic, visit times and proportion is obtained. The list is the user behavior classification pattern.

2.4 Analysis of College Students' Network Behavior

After completing the data clustering, this study conducted correlation analysis on different types of websites to understand the characteristics of college students' online behavior. In this study, Pearson Product-Moment Correlation Coefficient (PPMCC) was used for correlation analysis.

2.5 Correlation Analysis

The association rule mining algorithm (Theisen, 2018) can be used to draw conclusions related to research objectives from students' behavior data. This method mainly studies the correlation and influence of college students' online behaviors (such as online duration, online types, etc.) and college students' academic performance, explores college students' online behavior preferences, and provides references for college education management departments and administrators, so as to timely improve or optimize the management mode, improve management efficiency, and promote more college students to become talents as soon as possible.

2.5.1 Support Degree

Support (Bagui, 2018) is an important concept in association rule mining and is used to measure how often a data set contains an item set. Specifically, support represents the proportion of transactions that contain an item set to the total number of transactions. In the Apriori algorithm, support is used to filter out item sets below the threshold, thereby reducing the search space and improving the mining efficiency. The formula for calculating support degree is:

$$Support(X) = \frac{Transactions\ containing\ \mathrm{X}}{Total\ transactions} \quad (3)$$

Where X is the item set.

2.5.2 Credibility

Confidence is another key concept in association rule mining and is used to measure the strength of association rules. Confidence represents the conditional probability that A transaction containing an item set A also contains item set B (Sterner, 2020). Specifically, confidence represents the probability that if A transaction contains item set A, then it also contains item set B. Reliability is calculated as follows:

$$Confidence(A \Rightarrow B) = \frac{Support(A \cup B)}{Support(A)} \quad (4)$$

Where, A⇒B indicates an association rule, Support(A ∪ B) indicates the number of transactions including item set A ∪ B, and Support(A) indicates the number of transactions including item set A.

Confidence values range from 0 to 1, indicating the strength of the association rule. Higher credibility means stronger and more reliable rules.

2.5.3 Pre-Processing and Selection of Students' Grades

The results data of college students are compared and analyzed, and there are data missing (incomplete), duplicate data and wrong data. In order to improve the quality of data, the performance data should be preprocessed before data analysis. In order to simplify the operation, the method of deleting the incomplete data, duplicate data and wrong data in the student

score table is adopted to ensure the consistency of the data. Secondly, we should simplify the student achievement field, only retain the name, student number, grade, department and major of these major fields, and the student number as the key word, delete other fields unrelated to the study, in order to facilitate data processing. The student number is used as a unique identifier, so that the three kinds of data of students' information, students' grades and students' network behavior can be established to facilitate data query.

In our school's student score table, students' test results are represented by three items, namely grades, grade points and grade points (Fig.5). In China, GPA (General Point Average) is an authoritative assessment method (Dhar, 2018). Almost all colleges and universities use GPA to evaluate grades, and all courses they have learned are involved in the calculation. The GPA calculation of students' grades is automatically completed by computers, which converts students' grades into grade points, which is convenient for data analysis. It also has a certain scientific nature.

咸阳师范学院学生成绩

2020-2021学年第二学期

行政班级： 计科2001　　课程/环节：[10000606]C语言程序设计　　学分： 3

类　别： 专业课/必修课　　考核方式： 考试　　学生人数：38

学号	姓名	性别	成绩	绩点	学分绩点	修读性质	辅修标记	备注
1810014109	刘振雄	男	81	3.1	9.3	初修		
2006034121	郭西	女	95	4.5	13.5	初修		
2006034126	张露	女	93	4.3	12.9	初修		
2008054124	韩尚煊	男	76	2.6	7.8	初修		
2010014101	余晨菲	女	74	2.4	7.2	初修		
2010014102	卢欣欣	女	98	4.8	14.4	初修		
2010014103	沈佳乐	男	79	2.9	8.7	初修		
2010014104	徐静涵	女	98	4.8	14.4	初修		
2010014105	薛雯琪	女	88	3.8	11.4	初修		
2010014106	兰凯乐	男	73	2.3	6.9	初修		
2010014107	白宇凌	男	99	4.9	14.7	初修		
2010014108	刘涵玉	女	87	3.7	11.1	初修		
2010014109	杜国栋	男	96	4.6	13.8	初修		
2010014110	李元	男	82	3.2	9.6	初修		
2010014111	沈玉含	女	93	4.3	12.9	初修		

Figure 5: Student score table.

In order to ensure the scientificity and universality of student performance data, we chose the GPA of students' courses to calculate the evaluation of student learning effect. The grade point average for all classes is calculated as follows:

$$AL_{(GPA)} = \frac{\sum_{i=1}^{n} C_i * S_i}{C_1 + C_2 + \cdots + C_n} \quad (5)$$

Where C_i is the credit of class i and S_i is the grade point of class i.

In this way, the average grade point of students ($AL_{(GPA)}$) is calculated, and the data records in the original score table are greatly reduced, and the data records are simplified from tens of thousands to thousands, which not only reduces the total amount of data, but also makes the data more extensive.

In this paper, the website contribution degree is added to the attribute field of the Internet access data table, and the classification algorithm is used to analyze the Internet access data and the academic performance data, and the key attributes of the Internet access data affecting the academic performance of college students are determined. The data set Stu is composed of Internet access data and performance data of students majoring in computer science and technology, software engineering, Internet of Things engineering and intelligent science and technology in Xianyang Normal University. Stu contains 1632 sets of data, and the types of Internet access, online duration, website contribution and online time are taken as the attributes of participation. According to the average grade point, the students are divided into two categories: excellent students and underachiever students. The network behavior of excellent students and underachiever students is analyzed according to the two categories.

3 RESULTS AND DISCUSSION

3.1 Analysis of Online Duration of Campus Network Groups

We have calculated the average online time of students by month, as shown in Fig.6. As can be seen from the Fig.6, except for winter and summer vacations, when students use campus Internet less, the average online time in other months is more than 300 hours, and the average online time accounts for 41.67%, of the total school time of online students. It shows that the dependence of these online students on the campus network is still relatively strong, and this part of students will be the main object of our research on college students' online behavior and academic performance.

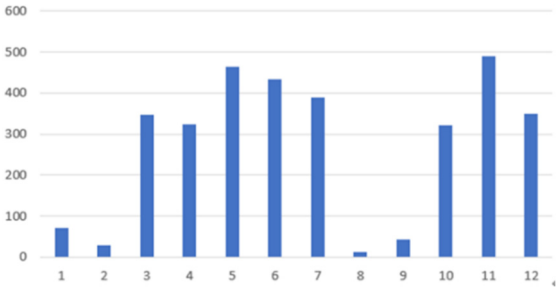

Figure 6: Online duration of students.

3.2 Analysis of Online Duration of Individual Users

In order to obtain the general rule of online data of college students, the data of online users were randomly selected for one week (from November 6 to 12, 2023), and the number of people who spent 1-24 hours online in one day was statistically analyzed. In the statistical process, log in to the campus network less than 1 hour according to 1 hour, more than 1 hour less than 2 hours, according to 2 hours, and so on.

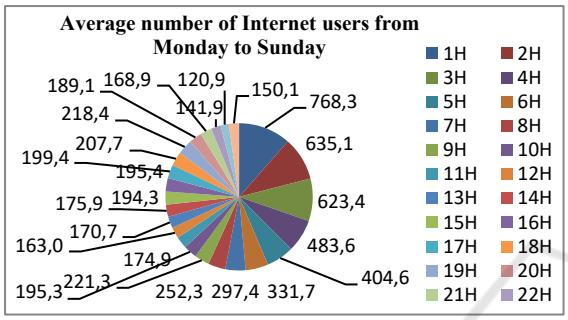

Figure 7: Distribution of online hours of students in a week (7 days).

There are 768 college students who surf the Internet for an average of 1 hour from Monday to Sunday, accounting for 11.50% of the total number of Internet users. The number of college students who surf the Internet for an average of 2 hours and 3 hours from Monday to Sunday gradually decreases, and the number of students who surf the Internet for an average of 24 hours and 23 hours from Monday to Sunday reaches the lowest level. Respectively 2.25% and 1.81%. Fig.7 shows the distribution of the number of student users who spend an average amount of time online in a week.

3.3 Analysis of Online Time Period of College Students

According to the online information of users, the data of online users for one week (from November 6 to 12, 2023) was randomly selected, and the online analysis was conducted according to 24 periods, as shown in the Fig.8. The number of online users increases from 18:00 to 19:00. The maximum number of online users is from 18:00 to 19:00. According to the analysis, the number of people in the afternoon is relatively more than that in the morning, because there are generally more basic courses in the morning, and those who have no classes are easy to sleep late, and the number of Internet users is the largest from 18:00 to 19:00 in the evening, because the whole course is basically completed during this period, and the evening self-study has not yet begun, and the vast majority of students relax online after dinner. The number of Internet users on Saturday and Sunday is more balanced, and relatively few before 10:00 am, indicating that college students get up late on holidays. There are always more people on Wednesday afternoon, because there are fewer classes arranged by the school after 16:30 on Wednesday, which belongs to the learning time of the staff. The analysis of the students who are still online from 0 to 6 o 'clock on Friday and Saturday shows that they are basically game users. However, on the whole, the number of Internet users during the day on Saturday and Sunday is not large, which indicates that students do not stay in the dormitory to surf the Internet during the two days of rest, but participate in some other outdoor activities, or go out. The number of people surfing the Internet after 17 o'clock on Friday is also much more than on other days, probably because Friday is a holiday, many college students take a break to surf the Internet.

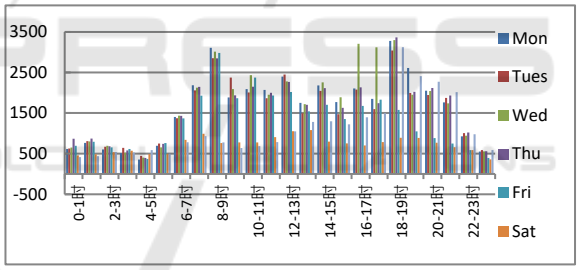

Figure 8: Internet access in different periods of one week.

3.4 Analysis of College Students' Internet Preference

A URL accessed by an Internet user over a period of time, which can be regarded as a unique identification of the user. Through the types of websites visited by users, we can understand the Internet interests of users, so as to classify the Internet behaviors of college students. It is helpful to improve the teaching management of the school through college students' online behavior preference and the proportion of all kinds of staff. According to the urls accessed by college students on the Internet, the system has classified urls according to keywords. In order to accurately analyze the interest tendency of users, statistics on the proportion of the average access types of each college student every day as a feature attribute

can be better used for cluster analysis than statistics on the number of visits. On this basis, k-means clustering algorithm is used to cluster users. Since the clustering method needs to provide the number of categories k in advance, in order to achieve the best effect of the clustering results, the value of k is set to 10, and the clustering results with the highest score are obtained through the contour coefficient.

During the analysis and research on the network behaviors of student users, 80,745,3703 network browsing logs of middle school students in the campus network were collected and analyzed in one semester, and the ranking of topics visited by all student users was calculated, as shown in Table 2.

Table 2: Behavioral Preferences of College Students When They Visit Websites.

Behavior Category Total	Number of visits	Percentage of visits	The most visited website	Number of visits
Social contact	145380687	25.65%	http://www.qq.com/q.cgi	68988187
Search	101553657	17.92%	http://ping.pinyin.sogou.com/ping	4098721
Video	92231901	16.27%	http://policy.video.iqiyi.com/polic	3244202
Study	74521451	13.15%	http://ic.wps.cn/wpsv6internet/inf	3206716
shopping	53776456	9.49%	http://amdc.m.taobao.com/amdc/	5418034
Game	62648577	11.05%	http://stat.game.yy.com/data.do	3614754
Music	19981975	3.53%	http://interface.music.163.com/ea	4036032
Reading	9437921	1.67%	http://api.foxitreader.cn/message/	4351965
Science	5502461	0.97%	http://router15.teamviewer.com/c	2802088
Download	1796294	0.32%	http://ardownload.adobe.com/pub	893871

From the Table2, We found that the most concerned areas of the users are social (25.65%), followed by search (17.92%), video (16.27%), learning (13.15%), shopping (9.49%), games (11.05%), music (3.53%), reading (1.67%), science and technology (0.97%) and downloads(0.32%).

4 CONCLUSION

Through analyzing and mining the data of college students' network behavior on the campus network server of a university, this paper obtained the preference of college students' network behavior. The results shows that the time and frequency of visiting gaming websites and social networking websites are negatively correlated with students' test scores, while the time and frequency of visiting learning websites and reading websites were positively correlated with students' test scores. Based on the average online time and types of excellent students and students who have poor performance, the network behavior portraits of excellent students and backward students are drawn. Then the bad behaviors of college students can be reminded or corrected in time. Through the study of college students' network behavior, it can be seen that the network can not only promote the improvement of college students' academic performance, but also affect their studies negatively. University management departments must understand the network behavior of college students accurately and guide them to use the network properly. This study provides a reference for optimizing and managing the college students' rational use of the network, and also improve the informatization level of student management.

REFERENCES

https://baijiahao.baidu.com/s?id=1784509414358338975 & wf= spider &for =pc [EB/OL].[2023-12-06/2024-1-4]

Jenkins, E.L., Ilicic, J., Barklamb, A.M., et al, 2019. Assessing the credibility and authenticity of social media content. *Lessons and applications for health communication: a scoping review of the literature.*

Samadi, S.Y., Billard, L., Meshkani, M.R., et al, 2017.Canonical Correlation for principal components of time series.*Computational Statistics*.

Tollo, G.D., Tanev, S., Liotta, G., et al, 2015.Using online textual data, principal component analysis and artificial neural networks to study business and innovation practices in technology-driven firms.*Computers in Industry*.

Cherniaiev, O.V., 2017. Systematization of the hard rock non-metallic mineral deposits for improvement of their mining technologies. *Naukovyi Visnyk Natsionalnoho Hirnychoho Universytetu*.

Xiaoyu,L.I, 2018. Optimal Neighbor Parameter of K-Nearest Neighbor Algorithm for Collaborative Filtering Recommendation. *Computer & Digital Engineering*.

Lakshmi, N., Krishnamurthy, M., 2022. Association rule mining based fuzzy manta ray foraging optimization algorithm for frequent itemset generation from socialmedia.*Concurrency and computation: practice and experience*.

Lin, K., Zhao, Y., Wang, L., et al, 2023. MSWNet: A visual deep machine learning method adopting transfer

learning based upon ResNet 50 for municipal solid waste sorting.*Frontiers of Environmental Science & Engineering*.

Zhiling, C., Minxue, H.E., 2017. Research on the perspective and development path of the behavior characteristics of sports Internet users in China——Based on the investigation of the behavior of sports users in Sohu. *Journal of Liaoning Normal University(Natural Science Edition)*.

Biswal, D.K., 2011. Redundant version information in history table that enables efficient snapshot querying.*US*.

Yuan, L., Cao, J., 2023. Application of data mining in female sports behavior prediction based on FCM algorithm.*Soft Computing*.

Zhao, J., Yang, X., Qiao, Q., et al, 2020. Sentiment Analysis of Course Evaluation Data Based on SVM Model. *IEEE International Conference on Progress in Informatics and Computing (PIC)*.

Bansal, A., Khare, A., Moriwal, R., 2020. ECLAT Algorithm for Frequent Item Set Generation with Association Rule Mining Algorithm. *International Journal of Scientific Research in Computer Science Engineering and Information Technology*.

Theisen, M.R., Mcgeorge, C.R., Walsdorf A A, 2018. Graduate Student Parents' Perceptions of Resources to Support Degree Completion: Implications for Family Therapy Programs. *Routledge*.

Bagui, S., Dhar, P.C., 2018. Mining positive and negative association rules in Hadoop's MapReduce environment. *Proceedings of the ACMSE 2018 Conference.*

Sterner, E.A., 2020. Impact of academic libraries on grade point average (GPA) :a review. *Performance Measurement and Metrics, ahead-of-print(ahead-of-print)*.