ISSN 2616-5775 Vol. 4, Issue 6: 21-34, DOI: 10.25236/AJCIS.2021.040605

Research on Internet Surfing Behavior of College Students Based on Big Data

Yantao Lan[1], Jinshuai Qu[1,*], Jian Chen[2]

[1]*Key Laboratory of Campus Information and Communication Security Disaster Backup and Recovery, Yunnan Minzu University, Kunming 650500, China*
[2]*College of Electronic Engineering, Shenyang Polytechnic College, Shengyang 110045, China*
[*]*Corresponding author*

Abstract: *With the rapid development of computer technology, the network behavior in the era of big data has become an important activity in students' campus life, which is quietly changing students' study and life. However, the behavior of students' network users directly reflects the purpose and demand of users and the state and performance of the network. The author analyzes the data of student users' online logs, and uses Apriori algorithm to analyze its association rules. Summarize the characteristics of student users' online behavior, including online time analysis, user's visit to websites and other user behavior analysis, which is of great significance for network optimization, personalizedj and differentiated design of services, standardized management, rational allocation of network bandwidth, enhancing information security, improving the efficiency of daily management of college counselors and network administrators, and ensuring the stability and efficiency of campus network environment.*

Keywords: *big data, Internet surfing behavior, association rules*

1. Introduction

With the popularization and development of the Internet, people frequently communicate and communicate through the Internet, presenting diversified online behaviors, such as visiting websites, uploading and downloading, making friends and chatting, video communications, online games, traveling and so on. At present, network transmission speed and data storage technology have been unprecedentedly improved, and mankind has entered the era of big data. Obviously, big data and related technologies are completely changing the way we study, live, work, and socialize, such as online classrooms, online shopping, information login, query, and submission. In the context of the Internet + big data, people's traditional lifestyles are undergoing subversive changes.

In the era of big data, effective analysis of the massive data generated by various behaviors can greatly facilitate the daily activities of the country, society, and individuals. For example, big data analysis has been widely used in military industry, commerce, education and other fields, greatly improving work efficiency. Similarly, with the rapid development of the Internet today, how to sort out, filter, and analyze a large amount of network behavior data generated in daily life so as to provide better information services has become an important research topic.

According to the 46th China Internet Development Statistics Report released by CNNIC in September 2020[1], as of June 2020, the number of Internet users in China reached 940 million, an increase of 36.25 million from March 2020, and the Internet penetration rate reached 67.0%. , An increase of 2.5 percentage points from March 2020. In the professional structure of netizens, students accounted for the largest proportion, reaching 23.7%. Followed by self-employed/freelancers, accounting for 17.4%.

In terms of development trends in recent years, the frequency of network usage by college student users is extremely high, and a large amount of Internet data is generated every moment. These logs contain a lot of valuable information. How to dig out and analyze the characteristics of students’online behaviors from the massive online logs, understand the correlation between student users’online behaviors, obtain the explicit and invisible needs of student users, and then adjust the service management strategy of the campus network to guide students the scientific and reasonable use of campus network resources has become a research field that urgently needs to be expanded.

ISSN 2616-5775 Vol. 4, Issue 6: 21-34, DOI: 10.25236/AJCIS.2021.040605

2. Key technology

2.1 Analysis goal

The basic goal of college students'online behavior analysis based on big data methods is to use the Apriori algorithm of data mining to collect, process and analyze the massive amounts of data generated by campus network users on the Internet every day, and summarize the campus network users' online habits and focus on hot spots. , Interest direction, etc [2].In the process of analyzing the collected student online logs, the investigation is mainly conducted from the following perspectives:

2.1.1 Distribution of online time of student user groups

Nowadays, the campus network is developing rapidly, with a huge number of students on the campus network, with more than hundreds of millions of online information generated every day. Therefore, mastering the distribution of students' online time and exploring the law is particularly important for improving the utilization of campus network resources.

2.1.2 Student users' access to different websites and applications

In the daily Internet logs of student users, through statistical screening of the website and application visits, the applications and websites with high traffic are recorded, and at the same time, the association analysis method is used to explore the types of applications and operating rules that students love, so as to help College counselors and network administrators provide assistance in their daily work.

2.1.3 The type of terminal used by student users to access the network

The daily communication devices used by student users include computers, tablets, and mobile phones. In the face of a large group of student users, the construction of wireless and wired broadband network infrastructure is particularly necessary and urgent [3] .

2.2 Data mining technology

2.2.1 Introduction to Data Mining Technology

Data mining is to mine interesting, useful, implicit, previously unknown and possibly useful patterns or knowledge from a large amount of data [4]. The difference between data mining and traditional data analysis is that traditional data analysis is usually targeted, and information is extracted with clear goals and assumptions. However, current data mining is geared towards ambiguous goals, discovering and summarizing knowledge through in-depth study of information [5].

2.2.2 The general process of data mining

①Data collection: According to the data analysis object, collect all the basic information needed in the data analysis process;

②Data integration: uniformly arrange data with different sources and formats according to certain rules;

③Data screening: sorting out useful data related to analysis work;

④Data transformation: through processing, the data is organized into a form that is convenient for mining and analysis for storage;

⑤Data mining: According to the processed data, select appropriate analysis tools and algorithms to find potentially valuable information [6];

⑥Model evaluation;

⑦Knowledge representation: express the information obtained by analysis in an easy-to-understand way.

2.3 The mathematical basis of association analysis

Suppose $I=\{i1、i2\cdots im\}$ is a collection of all items, the elements of which are called items; the related database D is a collection of transactions T, where each transaction T is a collection of items, and $T \subseteq I$. There is a unique mark corresponding to each transaction, such as the transaction number, which is recorded as TID; let X be a set of items in I, if $X \subseteq T$, then the transaction T is said to contain X. Then, the association rule is an implication of the form $X \Rightarrow Y$, where $X \subseteq I$, $Y \subset I$, and $X \cap Y = \Phi$.

Reflecting the rule that when items in X appear, items in Y also appear.

2.3.1 Related parameter description

①Support for association rules :

The support of rule The support of rule $X \Rightarrow Y$ in transaction set D (support) refers to the ratio of the number of transactions containing X and Y to the number of all transactions in transaction set D, denoted as support($X \Rightarrow Y$), abbreviated as s, that is

$$\text{support}\ (X \Rightarrow Y) = \text{support}\ X \bigcup Y = P(XY) \tag{1}$$

The support of the association rules reflects the probability that the items contained in X and Y appear at the same time in the transaction set.

②Confidence of association rules:

The confidence of rule X*Y in transaction set D refers to the ratio of the number of transactions containing X and Y to the number of transactions containing X in transaction D, denoted as confidence(X*Y), that is

$$\text{confidence}\ (X \Rightarrow Y) = \frac{\text{support}(\mathrm{X} \bigcup \mathrm{Y})}{\text{support}(X)} = P(Y \mid X) \tag{2}$$

③Frequent itemsets of association rules:

A collection of items is called an item set, and an item set containing k items is called a k-item set. The frequency of occurrence of an item set is the number of transactions containing the item set in the transaction database, that is, the number of occurrences of the item set in each transaction, referred to as the frequency, support count, or count of the item set. If the support of the itemset is greater than or equal to the product of min_sup (minimum support threshold) and the total number of transactions in D, the item set is said to meet the minimum support min_sup based on the association rule data mining algorithm, or it is called frequent itemsets (Frequent Itemset).

2.3.2 Related methods of association analysis

Apriori algorithm: This algorithm is one of the most influential algorithms for mining frequent itemsets using Boolean association rules. The core is a recursive algorithm based on the idea of two-order frequency sets. Association rules belong to one-dimensional, single-layer and Boolean association rules in classification. Here, all itemsets that support greater than the minimum support (called frequent itemsets) are called frequency sets.

The algorithm first finds out all occurrences of itemsets whose number of occurrences is greater than or equal to the minimum support we set at the beginning, and then generates strong association rules from the frequency set. The rules must meet the minimum support and confidence, and then scan the rules found in the frequency set found in 1 to generate all the rules that only contain the set items. There is only one item on the right side of each rule. After these rules are generated, they are larger than us. The set minimum support will be kept. Then use the recursive method to generate all frequent sets. The Apriori algorithm uses an iterative method of layer-by-layer search. The process is not complicated and relatively easy to implement. But there are some insurmountable shortcomings:

① Too many scans of the database.

② Apriori algorithm will generate a large number of intermediate itemsets.

③Using unique support.

④The adaptability of the algorithm is narrow.

ISSN 2616-5775 Vol. 4, Issue 6: 21-34, DOI: 10.25236/AJCIS.2021.040605

Partition-based algorithm: Savasere and his collaborators designed a partition-based algorithm. The realization of this algorithm is to divide the database at the logical level. It is worth noting that these divided blocks do not intersect and are of appropriate size. Each time a separate block needs to be considered to generate all relevant frequency sets, and then combined together to generate possible itemsets. The last is to calculate the support of the itemset. Each stage only needs to be scanned once. The accuracy of the algorithm is guaranteed by the frequency set of each possible frequency set in at least one block.

The calculation method can be highly parallel. After each processor processes each block, it generates a candidate k-item set through information interaction. The biggest problem with this algorithm is that the information interaction process between the processors is relatively time-consuming, and the time for each processor to generate frequent item sets is not uniform. Some processors are faster, and some are slow. This greatly increases the time-consuming of the algorithm, and becomes the two most difficult problems in the calculation method.

FP-tree frequency set algorithm: For the problems of the classic algorithm Apriori algorithm, Han Jiawei et al. proposed the association analysis algorithm of FP-tree frequency set algorithm in 2000. Using a separate processing method, after the initial scan, the frequent itemsets are compressed into the tree, and their association relationship is preserved, and then the tree is divided into conditional libraries one by one, and these libraries are processed separately. If the amount of data processed is very large, it can also be processed together. Through continuous practice process and processing results, it is found that this algorithm has good universality for rules of any length, and it has obvious advantages compared with Apriori algorithm.

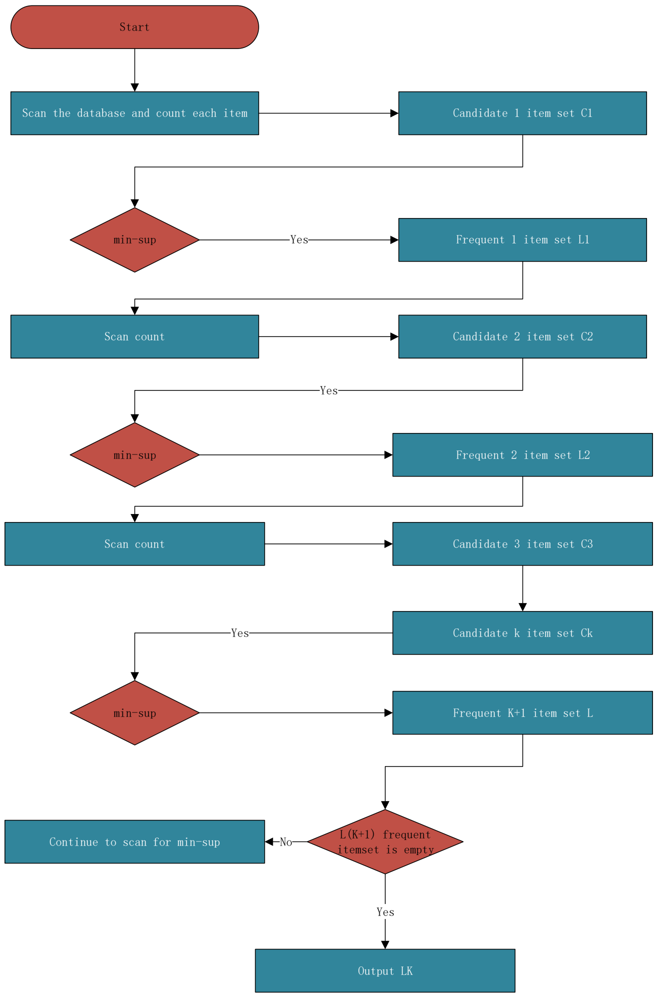

Figure.1 Apriori algorithm flow chart

2.4 Apriori algorithm of association rules

Apriori algorithm is an algorithm for mining frequent itemsets of Boolean association rules. It can

ISSN 2616-5775 Vol. 4, Issue 6: 21-34, DOI: 10.25236/AJCIS.2021.040605

be used in many areas of life. It is often used to study the sales strategy of retail item combinations. It can also be used in university management and uses association rules to Carry out work such as helping students to help the poor [7] .When using this method to generate association rules, it can be divided into two steps: ①find all frequent itemsets in the data list; ②after finding frequent itemsets using Apriori algorithm, use these itemsets to generate strong association rules.

The Apriori algorithm uses an iterative method of searching layer by layer. Scan the list for the first time and get the count of each item, eliminate the items that do not meet the minimum support, and get the set of frequent 1 item sets. Record as L1. Then use the 1-item set set L1 as the set L2 for finding frequent 2-item sets. The 2-item set set L2 is used to find the set L3 of frequent 3-item sets until no frequent set of K items is found. K scans were performed on the database.

The two core steps of the Apriori algorithm are the connection step and the pruning step:

① Connection step: In order to find the set Lk of frequent K itemsets, connect with itself through Lk-1 to obtain candidate K itemsets, denoted as Ck;

② Pruning step: In order to improve the generation efficiency of frequent itemsets, a priori property (if the set has a non-empty subset of frequent itemsets, then the set is not frequent itemsets) is usually used to compress the search space. The empty subset must also be frequent. On the contrary, if the candidate non-empty subset is not frequent, then the candidates will of course not be frequent. So it can be removed from CK.

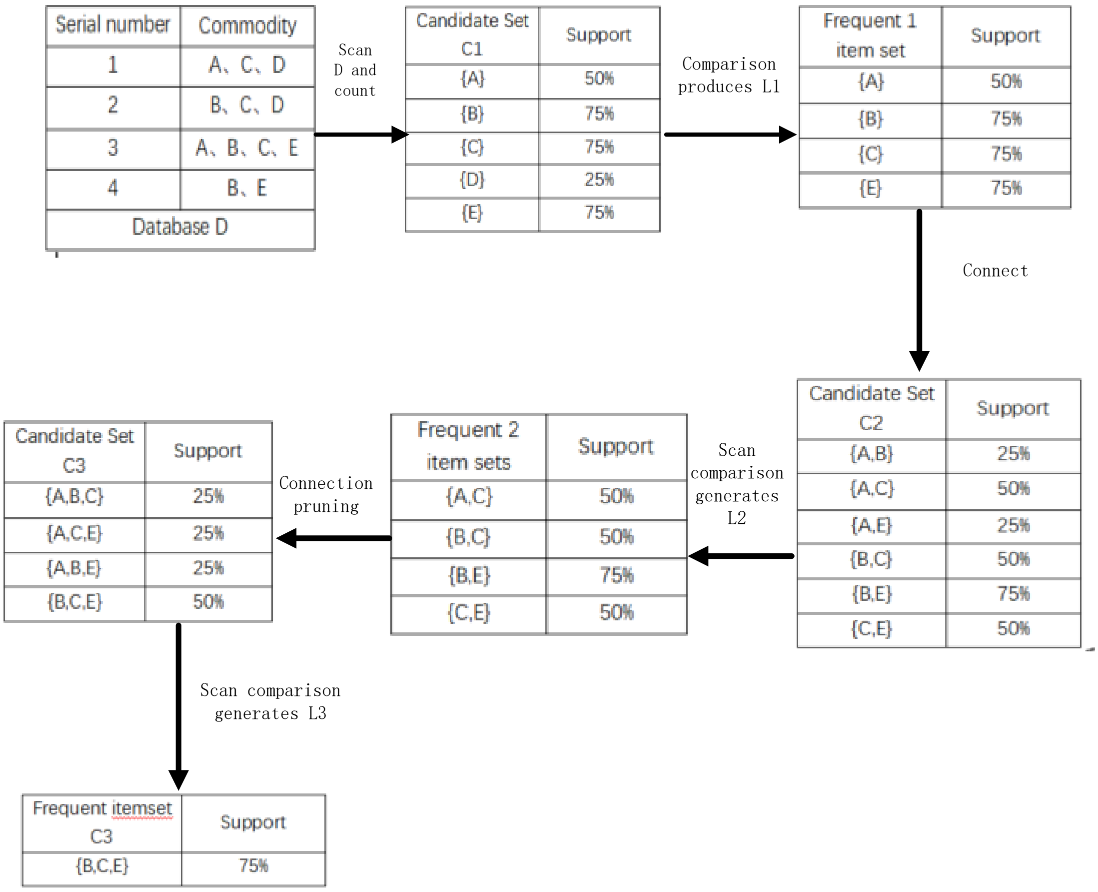

Figure.2 Connection step and pruning step process

3. Realization of big data method in user behavior analysis

3.1 Data collection

A true and reliable data source is the premise of any research. The data used in this article comes from the online behavior management log of the school information center. This article selects

ISSN 2616-5775 Vol. 4, Issue 6: 21-34, DOI: 10.25236/AJCIS.2021.040605

one-month Internet logs of users of a professional student group from March 2019. The collected logs mainly include: group name, source IP, terminal type, target IP, application type, specific application, control access control, time, details and device name. As shown below:

Table 1 Internet logs of some student users

Source IP	terminal type	Application Type	Concrete application	Time
172.30.241.90	Mobile terminal	Mobile terminal application	Mobile QQ	2019-03-02 23:59：56
172.30.241.90	Mobile terminal	Access network	IT related	2019-03-02 23:59：49
172.30.241.90	Mobile terminal	Access network	IT related	2019-03-02 23:59：49
172.30.241.90	Mobile terminal	Access network	Online video and download	2019-03-02 23:59：49
172.30.241.90	Mobile terminal	Access network	IT related	2019-03-02 23:59：49
172.30.241.90	Mobile terminal	Access network	uncategorized	2019-03-02 23:59：49
172.30.241.90	Mobile terminal	Web streaming	MP4 video	2019-03-02 23:59：49
172.30.246.212	Unknown type	Access network	IT industry	2019-03-02 23:59：48
172.30.243.208	Mobile terminal	Access network	IT related	2019-03-02 23:59：43
172.30.246.212	Unknown type	Access network	IT related	2019-03-02 23:59：27
172.30.241.21	Mobile terminal	Access network	Online Shopping	2019-03-02 23:59：12
172.30.241.21	Mobile terminal	Access network	IT related	2019-03-02 23:59：12
172.30.241.21	Mobile terminal	Access network	IT related	2019-03-02 23:59：12
172.30.241.21	Mobile terminal	Access network	IT related	2019-03-02 23:59：12
172.30.246.212	Unknown type	Access network	News portal	2019-03-02 23:59：09
172.30.246.212	Unknown type	Access network	IT related	2019-03-02 23:59：09
172.30.241.21	Mobile terminal	Access network	IP site	2019-03-02 23:59：08
172.30.246.212	Unknown type	Access network	Life information	2019-03-02 23:59：07
172.30.246.212	Unknown type	Access network	News portal	2019-03-02 23:59：04
172.30.241.21	Mobile terminal	Mobile terminal application	Mobile QQ	2019-03-02 23:58：58
172.30.243.208	Mobile terminal	Access network	IT related	2019-03-02 23:58：45
172.30.241.21	Mobile terminal	Access network	Travel traffic	2019-03-02 23:58：10
172.30.241.21	Mobile terminal	Access network	IT related	2019-03-02 23:58：10
172.30.241.21	Mobile terminal	Access network	IT related	2019-03-02 23:58：10
172.30.241.21	Mobile terminal	Mobile terminal application	Mi App Store	2019-03-02 23:58：10
172.30.240.254	Mobile terminal	Mobile terminal application	Huawei App Store	2019-03-02 23:58：03

3.2 Data processing

3.2.1 Data integration

With the development and innovation of science and technology, the network transmission rate continues to accelerate. Whether it is data upload, file download, website access, instant messaging or audio and video access, it can be completed in a very short time. The scale of data accumulated by users in daily life learning is very large. When faced with such a huge amount of data, our data usually comes from different sources and the source formats are different. Therefore, the process of data integration is to integrate data from various sources. The data used in this analysis comes from the school information center, so the workload on data integration is relatively small. In this article, use the database method to create a table, organize the data into a table, and run the code as follows:

```
CREATE TABLE   ‘’ (zongbiao
  user_name      varchar(64)     DEFAULT     NULL     COMENT ’ user_name’,
user_ip      varchar(64)     DEFAULT     NULL     COMENT ’user ip’,
client_type  varchar(64)     DEFAULT     NULL     COMENT ’ client_type’,
action_types      varchar(64)     DEFAULT     NULL     COMENT ’ action_types’,
Real_type   varchar(64)     DEFAULT     NULL     COMENT ’ Real_type’,
Action_time      varchar(64)     DEFAULT     NULL     COMENT ’ Action_time’,
)ENGINE=InnoDB     DEFAULT     CHARSET=utf8
```

3.2.2 Data filtering

①Since the student online log contains a lot of content, considering the privacy of student users, some irrelevant information is deleted during the data processing process, and data useful to the research process is retained.

ISSN 2616-5775 Vol. 4, Issue 6: 21-34, DOI: 10.25236/AJCIS.2021.040605

②When analyzing the original data, it may be found that not all attributes have analytical value. Excessive data not only consumes huge working hours and reduces the efficiency of analysis, but also does not substantially help data analysis. After comprehensive consideration, several attribute values of group name, target IP, access control, details and device name were deleted. Only the attribute values corresponding to IP, terminal type, application type, specific application and time are left. As shown in the following table:

Table 2 Filtered data

Search result							
Location	Target IP	Application type	Specific application	Access control	Time	Details	Device name
Undefined location	111.30.144.98	Mobile terminal application	Mobile QQ	record	2019-03-11 23:59:58	Terminal details: mobile terminal (And)	AC_192.168.10.3
Undefined location	121.51.8.101	Mobile terminal application	WeChat	record	2019-03-11 23:59:57	Terminal details: mobile terminal (And)	AC_192.168.10.3
Undefined location	184.28.218.8	Visit website	IT related	record	2019-03-11 23:59:55	Access domain name: p16-tiklokcdn	AC_192.168.10.3
Undefined location	183.232.94.44	Mobile terminal application	Mobile QQ	record	2019-03-11 23:59:25	Terminal details: mobile terminal (And)	AC_192.168.10.3
Undefined location	112.34.111.235	Visit website	search engine	record	2019-03-11 23:59:21	Visit domain name: hm.baidu.com	AC_192.168.10.3
Undefined location	183.222.97.213	Visit website	IT industry	record	2019-03-11 23:59:05	Access domain name: assistant-exp	AC_192.168.10.3
Undefined location	111.230.119.240	Mobile terminal application	Mobile QQ	record	2019-03-11 23:58:49	Terminal details: mobile terminal (And)	AC_192.168.10.3
Undefined location	121.51.130.102	Mobile terminal application	WeChat	record	2019-03-11 23:58:01	Terminal details: mobile terminal (And)	AC_192.168.10.3
Undefined location	111.13.42.185	Visit website	IT related	record	2019-03-11 23:58:00	Access domain name: ccc.sys.miui.c	AC_192.168.10.3
Undefined location	47.101.52.119	Visit website	IT related	record	2019-03-11 23:57:55	Access domain name: stats.jpush.cn	AC_192.168.10.3
Undefined location	163.177.89.195	Mobile terminal application	Mobile QQ	record	2019-03-11 23:57:38	Terminal details: mobile terminal (And)	AC_192.168.10.3
Undefined location	14.215.177.38	Visit website	IT industry	record	2019-03-11 23:57:23	Access domain name: servicesuppo	AC_192.168.10.3
Undefined location	111.30.144.98	Mobile terminal application	Mobile QQ	record	2019-03-11 23:57:20	Terminal details: mobile terminal (And)	AC_192.168.10.3
Undefined location	112.23.106.126	P2P	QQ Cyclone P2P	record	2019-03-11 23:57:15	Terminal details: mobile terminal (And)	AC_192.168.10.3
Undefined location	111.19.244.73	Mobile terminal application	WeChat Moments	record	2019-03-11 23:57:11	Terminal details: mobile terminal (And)	AC_192.168.10.3
Undefined location	121.51.13.106	Mobile terminal application	WeChat	record	2019-03-11 23:57:11	Terminal details: mobile terminal (And)	AC_192.168.10.3
Undefined location	9.9.9.9	Mobile terminal application	Google Play Store	record	2019-03-11 23:56:56	Terminal details: mobile terminal (And)	AC_192.168.10.3
Undefined location	114.67.34.43	Visit website	software download	record	2019-03-11 23:56:55	Access domain name: i.theme.oppon	AC_192.168.10.3
Undefined location	120.198.201.160	Mobile terminal application	Mobile QQ	record	2019-03-11 23:56:49	Terminal details: mobile terminal (And)	AC_192.168.10.3
Undefined location	39.130.253.6	Visit website	Online audio and video download	record	2019-03-11 23:56:45	Access domain name: data.bilibili.c	AC_192.168.10.3
Undefined location	111.230.119.240	Mobile terminal application	Mobile QQ	record	2019-03-11 23:56:15	Terminal details: mobile terminal (And)	AC_192.168.10.3
Undefined	118.25.31.186	Visit	Online	record	2019-03-11	Access domain	AC_192.168.10.3

ISSN 2616-5775 Vol. 4, Issue 6: 21-34, DOI: 10.25236/AJCIS.2021.040605

location		website	audio and video download		23:56:10	name: wup.huys.con	
Undefined location	221.179.177.20	Visit website	News portal	record	2019-03-11 23:55:46	Access domain name: m-sohu.comv	AC_192.168.10.3
Undefined location	223.85.58.74	Visit website	Online audio and video download	record	2019-03-11 23:55:46	Access domain name: api.bilibili.com	AC_192.168.10.3
Undefined location	112.34.111.145	Mobile terminal application	Baidu map	record	2019-03-11 23:55:14	Terminal details: mobile terminal (And)	AC_192.168.10.3
Undefined location	111.230.119.240	Mobile terminal application	Mobile QQ	record	2019-03-11 23:55:14	Terminal details: mobile terminal (And)	AC_192.168.10.3

3.3 Data conversion

In the process of data analysis, sometimes it is necessary to transform the form of the data according to our analysis needs to facilitate analysis. Data conversion is to transform the data into a data form that is convenient for us to analyze [8].In the process of processing the data, we found that the types of applications frequently accessed by student users are diverse. About 1.2 million pieces of data are recorded for a month of visits. Each person’s click records are about 40,000. In order to ensure the accuracy of the analysis, we select 10 application types with monthly visit records greater than 20,000 for analysis.

Table 3: Most visited apps

ID	QQ related	IT related	News portal	Online audio and video	software download	uncategorized	IT industry	search engine	IP site	Game Information	total
1	3076	6224	5945	2653	3674	1982	2023	2140	1736	260	29708
2	232575	13767	2995	2544	2042	8385	2248	2678	961	5732	273925
3	16335	1617	4741	1862	7241	1561	896	757	1427	9	36446
4	942	4951	2666	5686	1264	958	2396	787	622	55	20527
5	10674	22425	6935	839	1301	1480	1863	4713	598	3691	54419
6	14290	1540	2794	693	1048	438	311	723	144	392	22373
7	433	634	671	328	230	89	421	474	8	17	3305
8	7974	8371	2751	2014	2183	1589	2654	1937	2918	689	33080
9	2394	3663	2932	2549	1216	1109	1545	1566	1627	43	18744
10	2998	2083	4227	1840	758	855	1645	446	1075	3160	19087
11	6138	2875	3846	7375	6071	2164	302	1188	267	169	30395
12	1661	4684	3284	2398	903	1219	1081	1315	364	103	17022
13	10096	4810	4612	1939	3306	2165	2199	1620	1823	2507	35077
14	715	664	427	343	152	62	51	247	4		2665
15	6739	3673	6315	1875	4601	1696	1483	1274	1306	31	28995
16	33667	14606	9717	5159	4734	3431	2432	6164	3748	1455	85133
17	592	2285	1290	869	317	223	628	583	288	477	7552
18	4169	3123	1820	2131	444	1379	1275	802	173	76	15392
19	12859	3358	3891	5594	1183	2269	2502	1364	3274	16	36310
20	7892	3537	3676	4090	2452	1686	2222	1093	85	1977	28710
21	3880	5878	4881	9674	1049	3679	2627	1354	459	442	33623
22	3780	3435	3109	4230	1231	911	1476	547	6217	371	25607
23	766	3824	3160	2969	1740	878	2150	1400	1425	101	18433
24	9453	3660	3544	2468	1111	1906	1512	1877	311	79	25921
25	469	976	406	1599	349	275	993	330	106	32	5535
26	1901	1021	784	1476	196	473	837	262	171	14	7165
27	11044	3416	7035	2781	2348	2638	3850	1578	3704	121	38515
28	2371	1878	2712	592	935	779	314	673	325	45	10624
29	6961	4745	7858	1833	8335	3349	2750	1844	305	1507	39487
total	416764	137743	109024	80323	62414	49708	46654	42056	356662	23567	1003965

In order to facilitate the analysis, the application records with access records higher than 2000 access records are represented by 1, and the application records with access records lower than 2000 times are represented by 0, and they are saved in the database.

Table 4: Converted data

ID	QQ related	IT related	News portal	Online audio and video	software download	uncategorized	IT industry	search engine	IP site	Game Information
1	1	1	1	1	1	0	1	1	0	0
2	1	1	1	1	1	1	1	1	0	1
3	1	0	1	0	1	0	0	0	0	0
4	0	1	1	1	0	0	1	0	0	0
5	1	1	1	0	0	0	0	1	0	1
6	1	0	1	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0
8	1	1	1	1	1	0	1	0	1	0
9	1	1	1	1	0	0	0	0	0	0
10	1	1	1	0	0	0	0	0	0	1
11	1	1	1	0	1	1	0	0	0	0
12	0	1	1	0	0	0	0	0	0	0
13	1	1	1	0	1	0	1	0	0	1
14	0	0	0	0	0	0	0	0	0	0
15	1	1	1	0	1	0	0	0	0	0
16	1	1	1	1	1	1	1	1	1	0
17	0	1	0	0	0	0	0	0	0	0
18	1	1	0	1	0	0	0	0	0	0
19	1	1	1	1	0	1	1	0	1	0
20	1	1	1	1	1	0	1	0	0	0
21	1	1	1	1	0	1	1	0	0	0
22	1	1	1	1	0	0	0	0	1	0
23	0	1	1	1	0	0	1	0	0	0
24	1	1	1	1	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0
27	1	1	1	1	1	1	1	0	1	0
28	1	0	1	0	0	0	0	0	0	0
29	1	1	1	0	1	1	1	0	0	0

4. Experiment analysis

4.1 Analysis of user terminal connection

Through the analysis of one-month Internet records, we sorted out the connection status of the three types of terminals of mobile phones, computers and unknown terminals, and analyzed the connection status of the three types of terminals among the 20 application types most visited by student users.

Judging from the visit records of a known month, in the construction and distribution of school campus network resources, broadband construction and wireless network optimization are still topics that need continuous attention. The results showed that among all the access records, 797,812 records were accessed by PC and 163,140 were accessed by mobile phones. From this point of view, most of the time, student users prefer to use the PC side to access various application websites. For QQ, IT-related, software downloads, IP sites, and game information, the proportion of PC-based access is higher than that of mobile terminals and unknown terminals. Combined with daily applications, PC access to various application software and web page browsing is not only because of the obvious advantages of broadband speed compared with wireless network, but also because the response speed of mobile phone access to games and other web pages is not as fast as that of PC, and the performance is obviously lagging behind. . The emergence of this situation has correspondingly put forward higher requirements for the construction of campus network broadband.

4.2 Application traffic statistics

In the student's online log, by looking up the number of visits to different applications by students, we can see how frequently the applications are accessed, and we can also know the interest of the student user group. Therefore, in the analysis process, the author extracted and counted 20 application types and visit times frequently visited by student user groups. Because there are obvious differences in the software applications that male and female students often use in communication or entertainment, and their hobbies are not the same. For example, male students may prefer to browse related information and information in games and news, while female students may be more inclined to access chat software. Shopping websites, video playback software, etc. At the same time, the author also analyzed the number of clicks and visits of different websites of male and female students and the different degrees of preference of each application, and obtained the following statistical results:

It can be seen from the statistical chart that QQ-related, IT-related, news portals, online audio and video downloads, and software downloads are the most visited users among the user groups. Among them, the most frequently used and visited by student users are QQ-related functions. In daily learning and life, whether it is instant messaging, file transfer, announcements, etc., student groups will operate through the QQ channel. For the 18-24 age group of college students, more needs are reflected in the personalization and convenience of QQ, which is powerful and rich in virtual value. In terms of gender analysis, there are certain differences in the degree of preference between male and female students. In the use of WeChat, the number of visits by girls and boys is not much different, and the attention of boys on game information is significantly higher than that of girls. This also shows that male students are extremely concerned about game information in their daily lives, and are keen on game applications and related game information, which is also reflected in the high frequency of visits to game applications [9] .

4.3 Time period application traffic analysis

Through the click volume analysis of the time period application, it is possible to have a general understanding of the distribution of online time and routines of student users. The following table is the analysis results obtained by analyzing the Internet logs in March, taking 24 hours a day as 24 time periods:

Table 5: Access data in each time period

Time period	Views	Time period	Views
0	13302	12	97342
1	6743	13	76118
2	4738	14	62556
3	4048	15	57575
4	3520	16	62342
5	3638	17	98127
6	6296	18	90801
7	15199	19	82726
8	24306	20	103861
9	27025	21	112614
10	46713	22	113882
11	57228	23	53183

From the time period visits in the above several charts, it is very intuitive to reflect the 24 time periods of the whole day, the activity level of student users and the access situation of the application website: between 0 o'clock and 7 o'clock, the user activity is not High, the visit rate of the application website is not high. Starting from 0:00, as students gradually rest, the application website visits during this period are the lowest in a day. From 7 o’clock onwards, the activity level gradually increases. Affected by the school’s schedule of work and rest, student user activity increases rapidly from 8 am to 12 am, and reaches the highest level at 12 o’clock. At this time, students use the rest time to surf the Internet after class, so the application The number of visits reached the highest value during this period. From 12 o'clock to 15 o'clock, application visits showed a downward trend, and most people were studying courses. From 16:00 to 17:00, the number of user visits increased rapidly in one hour; from 17:00 to 19:00, there was a significant drop. During this period, most people eat in the cafeteria or engage in other extracurricular activities; from 19:00 to At 22:00, application visits and user activity increased significantly, reaching the peak at 22:00. During this period, most of the student users returned to the dormitory after class to go online, so the increase was rapid; after 22:00, it showed a downward trend and declined. The fastest rate.

From the analysis results, the traffic of student users has been increasing since 8 o'clock. During the hour from 11 am to 12 pm, the number of visits to applications and websites increased the fastest.

ISSN 2616-5775 Vol. 4, Issue 6: 21-34, DOI: 10.25236/AJCIS.2021.040605

According to the schedule of students'work and rest in our school, 8 o'clock to 12 o'clock is the class time, but the number of students'access to the Internet has not decreased but increased. This result shows that during class, the number of student users using mobile phones has increased from 8 o'clock. Many, especially the hour from 11 o'clock to the end of class, which is similar to the situation from 16 o'clock to 17 o'clock in the afternoon. It belongs to the time when user visits increase the fastest in the entire analysis period. Since 19:00, although the number of visits has been increasing, the number of students attending evening classes is only part of them. Therefore, this situation is more reasonable. However, attention should be paid to the emergence of these online situations. On the one hand, teachers should strengthen the management and control of the undesirable phenomenon of students playing with mobile phones in the classroom, especially when the hour is close to the end of class; on the other hand, students should learn self-reflection and behavioral restraint, and learn professional knowledge in class. Don't waste time and opportunities for learning.

4.4 Establishment of association rules

In order to facilitate the reading of the data, the following table 4-8 format was changed and saved during the research process.

Table 6: Data conversion

QQ related	IT related	News portal	Online video and download	Software download	Uncategorized	IT industry	Search engine	IP site	Game Information
a	b	c	d	e	f	g	h	i	j

Table 7: Final database file

ID	a	b	c	d	e	f	g	h	i	j
1	a	b	c	d	e	0	g	h	0	0
2	a	b	c	d	e	f	g	h	0	j
3	a	0	c	0	e	0	0	0	0	0
4	0	b	c	d	0	0	g	0	0	0
5	a	b	c	0	0	0	0	h	0	j
6	a	0	c	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0
8	a	b	c	d	e	0	g	0	i	0
9	a	b	c	d	0	0	0	0	0	0
10	a	b	c	0	0	0	0	0	0	j
11	a	b	c	d	e	f	0	0	0	0
12	0	b	c	d	0	0	0	0	0	0
13	a	b	c	0	e	f	g	0	0	j
14	0	0	0	0	0	0	0	0	0	0
15	a	b	c	0	e	0	0	0	0	0
16	a	b	c	d	e	f	g	h	i	0
17	0	b	0	0	0	0	0	0	0	0
18	a	b	0	d	0	0	0	0	0	0
19	a	b	c	d	0	f	g	0	i	0
20	a	b	c	d	e	0	g	0	0	0
21	a	b	c	d	0	f	g	0	0	0
22	a	b	c	d	0	0	0	0	i	0
23	0	b	c	d	0	0	g	0	0	0
24	a	b	c	d	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0
27	a	b	c	d	e	f	g	0	i	0
28	a	0	0	0	0	0	0	0	0	0
29	a	b	c	0	e	f	g	0	0	0

After running the algorithm, the following results are obtained:

ISSN 2616-5775 Vol. 4, Issue 6: 21-34, DOI: 10.25236/AJCIS.2021.040605

Candidate set1			b	c	d	e	f	g	h	i	
Frequent set1			b	c	d						
Candidate set2			c	d	ab	ac	ad	bc	bd	cd	
Frequent set2			c	ab	ac	bc	bd	cd			
Candidate set3	b	c	ad	bc	bd	cd	abc	abd	acd	bcd	
Frequent set3	c	c	abc								
Candidate set4	cd	cd	abcd								
Frequent set4											
Candidate set5											
Frequent set5											
Frequent itemsets		c	b	bc	c	ab	abc	ac	bc	bd	d

Table 8: Run result display

b	-->d	Confidence: 0.71428537	d	-->c	Confidence: 0.9375
	-->a	Confidence: 0.8095238	ab	-->c	Confidence: 0.9411765
ac	-->	Confidence: 0.8	a	-->	Confidence: 0.8095238
a	-->b	Confidence: 0.8095238	b	-->a	Confidence: 0.8095238
c	-->a	Confidence: 0.6956522	b	-->c	Confidence: 0.8888889
	-->c	Confidence: 0.9047619		-->ac	Confidence: 0.7619048
d	-->b	Confidence: 0.9375	b	-->c	Confidence: 0.7619048
c	-->d	Confidence: 0.65217394	c	-->a	Confidence: 0.8695652
b	-->c	Confidence: 0.9047619	ac	-->b	Confidence: 0.8
a	-->c	Confidence: 0.9411765	b	-->ac	Confidence: 0.7619048
a	-->c	Confidence: 0.7619048	bc	-->a	Confidence: 0.84210527
c	-->a	Confidence: 0.84210527	c	-->ab	Confidence: 0.6956522
b	-->	Confidence: 0.85714287		-->b	Confidence: 0.85714287
	-->bc	Confidence: 0.7619048	c	-->b	Confidence: 0.82608694

Association rules:

Through the generation of association rules, we can know the meaning of the first ten association rules:

①b-->d: Most users who are interested in IT also use online applications and download functions, and the confidence between them is 0.71.

② -->a: I have visited several other applications, and the confidence level of visiting QQ-related applications is 0.8.

③ac-->: Users who access QQ related and news portals at the same time have a high probability of accessing several other applications, and the confidence between them is 0.8.

④a-->b: Most users who have visited QQ-related applications also browse IT-related information, and the confidence between them is 0.8.

⑤c-->a: Most users who visit the news portal also use QQ related functions, and the confidence between them is 0.69.

⑥ -->c: Most users visit news portals while visiting other applications, and the confidence between them is 0.90.

⑦d-->b: Users who access online video and download, and also access IT-related information, the confidence between them is 0.93.

⑧c-->d; users who access the news portal also access online video and download, and the confidence between the two is 0.65.

⑨b-->c: Most users who access IT-related information also visited the news portal, and the confidence between them is 0.90.

⑩a-->c: Most users who visit QQ also visit news portals, and the confidence between them is 0.94.

Based on the above analysis results, the following points of network behavior characteristics of campus network student users are summarized;

① When student users access applications and web pages, the number of visits using the PC terminal is significantly higher than the number of visits using the mobile terminal. Based on the daily network status of the campus network, there are two reasons: First, the quality of the campus network' s wireless network is unstable. During the three periods of 11 am to 12 pm, 16 to 18 pm, and 20 pm to 22:00 in the evening, due to With the increase in users and the increase in application and website visits, the campus rate will drop significantly during these periods.

②In terms of application and website visits, on the one hand, QQ, IT-related, IT industry, news portals, online audio and video and downloads are at the forefront of user visits. Student users are both focused and visited on entertainment information and entertainment software. Relatively high. On the other hand, student users have a relatively high number of visits to the IT industry, IT-related applications and websites. This is consistent with the professional situation of the students analyzed in this article.

③The results of the correlation analysis show that QQ-related, IT-related, news portals, online audio and video, and downloads are among the user groups in this study with high correlation confidence. This also shows that these applications and related information are frequently accessed.

5. Summary and outlook

With the rapid development of computer technology, network behavior in the era of big data has become an important activity in student campus life. The leap of information technology is quietly changing students' study and life [10].However, the online user behavior of college students directly reflects the user's preferences, needs, and network status and performance. Therefore, the analysis of campus network user behavior, for network optimization, service personalized and differentiated design, standardized management, reasonable allocation of network bandwidth, enhance information security, improve the efficiency of daily management of college counselors and network administrators, and ensure campus The stability and efficiency of the network environment are of great significance. At the same time, analyzing the user's online behavior data based on the campus network can dig out many valuable hidden features [11].For example, the activity level of student users and the access status of application websites, application access volume and user terminal connection status, etc. Counselors and network administrators can also provide students with convenient, efficient and accurate personalized guidance and services based on these data.

The campus network has brought a lot of convenience to current college students. Through the analysis and data mining of college students’ online behavior, it can provide decision-making support for formulating reasonable and effective network management strategies, and build the network into a good learning aid tool for students. At the same time, the behavioral characteristics and preferences of student users on the Internet also provide a scientific basis for the management and optimization of college campus networks.

Acknowledgements

This work was supported by the Yunnan Provincial Department of Education Science Research Fund Project (NO. 2020J0655).

References

[1] Compiled by the office of the Committee of cybersecurity and informatization of the CPC central Committee. the 45th statistical report on the development of China's internet [M]. Beijing: China internet network information center, 2020: 19-27.
[2] Jiang Yongchao. Research on the algorithm of analyzing students' course selection and learning behavior based on data mining [J]. Modern Electronic Technology, 2016,39(13):145-148.
[3] Nian Mei, Fan Zukui, Huang Xinxin. Analysis and Research on Students' Online Behavior in Campus Network [J]. Computer Age, 2019(09):67-70.
[4] Guo Yubin, Wu Yuhang, Bo Aofeng, Zheng Shumin, Zhang Xiaopeng. Characteristic analysis of students' online time based on authentication data [J]. computer applications and software, 2019, 36(11):101-106+133.
[5] Hu Zuhui, Shi Wei. Research on Internet behavior analysis and data mining of college students [J]. China Distance Education, 2017(02):26-32.
[6] Shi Yingying, Ge Wancheng, Wang Liangyou, Lin Jiayan. Research on the Improvement of K-means Clustering Personalized Recommendation Algorithm [J]. Information and Communication, 2016(01):19-21.

ISSN 2616-5775 Vol. 4, Issue 6: 21-34, DOI: 10.25236/AJCIS.2021.040605

[7] Jesus Maillo,Sergio Ramírez,Isaac Triguero,Francisco Herrera. kNN-IS: An Iterative Spark-based design of the k-Nearest Neighbors classifier for big data[J]. Knowledge-Based Systems,2017,117.
[8] Aobing Sun, Tongkai Ji, Jun Wang, Haitao Liu. Wearable mobile internet devices involved in big data solution for education[J]. Int. J. of Embedded Systems,2016,8(4).
[9] Yin Yu. Research on the corrective measures of college students' classroom mobile Internet surfing behavior [J]. Journal of Jiamusi Vocational College, 2018(03):268+270.
[10] Ren Hua, Zhang Ling, Ye Yu. Analysis and monitoring of big data of users' network behavior in digital campus [J]. Computer and Digital Engineering, 2017,45(09):1814-1818+1823.
[11] Zhou Qianyu, Shi Wei. Analysis of students' consumption and online behavior based on campus one-card data [J]. China Educational Technology and Equipment, 2020(12):8-12+18.