Analysis of Undergraduates' Network Status Based on Data Mining

Qian Yuan, Shi Quan[a]

School of Educational Science, Nantong University, Jiangsu 226000, China;

[a]sq@ntu.edu.cn

Keywords: Data mining, undergraduates, network status.

Abstract. The construction of digital campus gives undergraduates a more convenient way to access the Internet and there have been stored massive network data. In this paper, use clustering algorithm and log file analysis to analyze these data to understand undergraduates' network status and the relationship between network status and academic performance. The result could be helpful to provide reference for network administrator and counselor, and give a new way out for analyze undergraduates' network usage.

Introduction

With the development and popularization of the Internet, education informationization has rapidly developed in the global world. "Development plan of education informatization decade (2011-2020)" issued by the Ministry of Education clearly pointed out that, "promote the modernization of education by educational information and construct educational information system covering urban and rural schools at all levels to promote the popularization and sharing of high quality educational resources; push on the construction of digital campus in colleges and universities to achieve data sharing". Education informationization has promoted the rapid development of the construction of digital campus, and has brought many conveniences for the majority of teachers and students for their work and studies. Internet is everywhere for undergraduates nowadays, but their network usage situation is also hot issues concerned and discussed by the entire social. Domestic scholars have come to the network using having correlation with academic performance and students' worse networks status are related with poor academic results [1, 2]. Foreign scholars consider negative emotions influence network usage, and network usage, especially online time has an inseparable connection with academic results [3, 4]. With the continuous improvement of digital campus, there have been stored much of their network data [5]. Using data mining to analyze the massive network data to understand undergraduates' network status is an effective method.

The Process of Data Mining

Determine the Research Object. This study chose the undergraduates (freshman，sophomore and junior college students) of a local comprehensive university author stayed as the research object to explore the relationship of undergraduates' network status and academic performance. There are totally 23243 students (sample composition as table 1).

Table 1 Sample composition

	Freshman	Sophomore	junior college students
Man	3505	3584	3621
Female	3958	4226	4349
Total	7463	7810	7970

According to the large amount of real data stored in the databases of digital campus, including undergraduates' online data, log file data, personal information, etc, with the usage of undergraduates' online time, academic performance, gender and Internet content, etc, use clustering analysis and other data mining methods to discuss their network status, find out rules and problems, and provide some method for analyzing undergraduates' network status.

Data Acquisition and Preprocessing. Choose undergraduates' online data stored on the databases of campus network system, and select online data of the research object from September 2014 to June 2015. Eventually, there are about 10 million dates. Preprocess with these data by individual, usage time, usage period, etc.

Choose sample undergraduates' result data from the educational administration system, including scholarship acquisition and statistics of their failure subjects. Who got scholarship called 'high-achievers', while who had failure subject called 'low-achievers', and the rest of them called 'the average'. Finally, there are 7140 'high-achievers', 4731 'low-achievers' and 11372 'the average'.

Internet log files include undergraduates' login client, target server, domain name and URL address and other information [6]. With the usage of open directory project (ODP), finding the target sever address and target URL and other methods, divide users' access sites into several types. For the URL hard to divide, give a most reasonable result by artificial judgment.

Clustering Algorithm. Clustering analyze with a collection of data objects. The principle of clustering is that within a cluster the objects have highly similarities, while there are highly differences between clusters. There are many kinds of clustering algorithms, such as K-means and EM. This paper choose K-means algorithm which feature is that every object can only be assigned to one cluster. There is neither connection nor overlap between clusters. The main process of K-means algorithm is as following in fig 1.

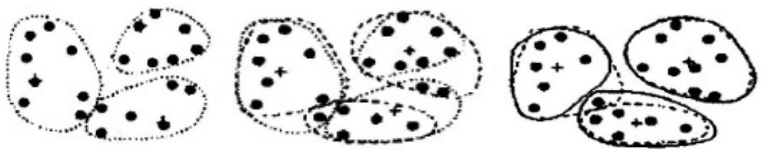

Fig. 1 The main process of K-means algorithm

The main purpose of clustering is to classify the attributes with similar characteristics to one cluster. In this paper, the clustering algorithm is used to cluster by undergraduates' network status and academic performance and investigate the relationship between their network statuses with academic performance, gender, grade and so on. According to the results of clustering, we can provide reference for the college administrators and counselors to help them to strengthen the management of undergraduates' internet usage.

The Result of Data Mining

The Result of Clustering. Using the business intelligence development platform provided by SQL Server 2012, according to undergraduates' online time, academic performance, gender, grade and other information, using K-means clustering to the preprocessed data, finally, get five categories as shown in table 2.

Table 2 The result of clustering by undergraduates' network status

		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Online time (Hour)		19.22±14.3	51.62±29.7	98.61±32.8	149.67±34.9	218.57±90.2
Total number		5012	5214	5170	5520	2369
Divided by academic performance	High-achievers	1809	2111	1618	1418	226
	Low-achievers	2303	2262	2591	2977	1239
	The average	600	641	761	1325	1404
Divided by gender	Man	1658	1725	2245	3083	1999
	Female	3354	3489	2925	2437	370
Divided by grade	Freshman	1512	1627	1456	2015	853
	Sophomore	2746	1404	1707	1316	1150
	Junior college students	754	1883	1707	1789	1366

By the table 2, users of the first 4 clusters whose online time is within 170 hours each month (our school campus network implements the policy of each user has 170 hours online time with 20 RMB.

If online time over 170 hours, per hour counts 0.5 RMB.). Users in cluster 5 whose online time are all over 170 hours and they are undergraduates whose online time are a bit more than others.

Divided by academic performance, 'high-achievers' are mostly in the first 4 clusters and they can better control their online time, while 'low-achievers' are mostly in cluster 4 and cluster 5 and this indicates that 'low-achievers' usually spend a lot of time on Internet. From this we can see that if undergraduates spend more time on Internet, they may be more likely getting lower grades. As a result, when undergraduates' online time are over the rules of school, the network administrators and counselor should pay more attention to avoid students' Internet addiction.

Divided by gender, the proportion of girls is far more than boys in cluster 1 and cluster 2 where users has less online time, while the proportion of boys is far more than girls in cluster 4 and cluster 5 where users has much more online time. This could see that most girls have a stronger control of accessing the Internet, and they can reasonably arrange their online time. While boys' control is weaker, most of them spend much more time on Internet.

Divided by grade, freshmen are mainly distributed in the first 4 clusters, and there is not much difference between each cluster. Freshman has a little bit proportion in cluster 5 where users' online time is over the rule of school. This indicates that the majority of freshmen can strictly control their inline time. The proportion of sophomore in cluster 5 is more than freshmen, which means much more sophomores' online time are over the rule of school. Most junior college students are distributed in cluster 2 to cluster 5 and that means most of them spend more time on Internet. With the growth of grade, undergraduates may more easily spend more time on Internet in a free network environment. College network administrators and counselors should strengthen the network training and guidance of undergraduates, especially the freshmen. Only when they have good Web habits, they can have a reasonable plan to access the Internet and taking into account of study.

The Analysis of Undergraduates' Network Contents. According to the undergraduates' access log file, analyze the network contents of 'high-achievers' and 'low-achievers'. Network contents can help us better understand users' network purposes. Through the analysis of two kinds of users' network contents; we can give specially advice and guidance to each of them to improve their network behavior.

Choose students' Internet log file in April 2015 and only select log file of 'high-achievers' and 'low-achievers' for the analysis of their network contents. Eventually, there are 27752 log file data of 'high-achievers' and 33909 of 'low-achievers'. The result of website type division is as follows in fig 2.

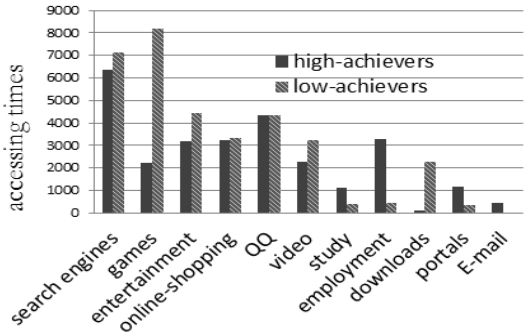

Fig. 2 The analysis of students' network contents

As shown in figure 2, the search engines are the most popular network contents of all the students. 'Low-achievers' have higher accessing times on games, entertainment and download than 'high-achievers'. 'High-achievers' have higher accessing times on studies than 'low-achievers', but the overall level is not high. In addition, 'high-achievers' have higher accessing times on websites of employment, which means 'high-achievers' have a certain sense of urgency of their future jobs. 'High-achievers' have more types of websites than 'low-achievers', and their network contents include portals and E-mail.

Focus on the students whose online hours are over 170 hours in the average month to analyze their network contents, and get the following fig 3. This part of students spends their most online time on

games, and they are also takes a lot of attention on QQ, search engines,online-shopping, entertainment and download. But they only spend a little time on the websites of study or employment. Their network access is a little narrow. So the network management department should take the appropriate measures to prevent this kind of users from Internet addiction.

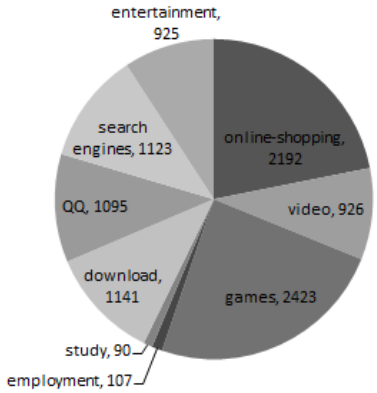

Fig. 3 The network contents analysis of students with much more online time

Summary

In addition, nowadays undergraduates are closer to the Internet. College administrators and counselors should pay more attention to the usage of their network status. And at digital campus, while there are massive network data, we should use data mining and other methods to analyze and provide some useful guidance for help the development of colleges and the undergraduates.

Acknowledgment

This work is supported by the research and innovation program of postgraduates in Jiangsu Province (subject number: YKC14023); National education and information technology research "Twelfth Five Year Plan" key project (subject number:136221504); The 2014 modern educational technology research key issues in Jiangsu province (subject number: 2014-R-30418)；The 2015 modern educational technology research issues in Jiangsu province (subject number: 2015-R-41638); The 2014 natural science research issues in Nantong University (subject number: 14Z016). Shi Quan is the corresponding author of this thesis.

References

[1] He xiangyang, Ma Peifeng. Multi perspective analysis of adolescents' Internet addiction [J]. China Educational Technology, 2009，(6): 53-56.

[2] Fu Hong, Wang Jianzhou, An Yong. Correlation analysis of undergraduates' network behavior and academic performance [J]. Theory and Modernization, 2014, 3: 37-43.

[3] Jun S, Choi E. Academic stress and Internet addiction from general strain theory framework [J]. Computers in Human Behavior, 2015, 49: 282-287.

[4] Tonioni F, D'Alessandris L, Lai C, et al. Internet addiction: hours spent online, behaviors and psychological symptoms [J]. General Hospital Psychiatry, 2012, 34(1): 80-87.

[5] Yan H U, Zhang Y. A design of public-data platform in digital campus based on web service [J]. Journal of China Universities of Posts & Telecommunications, 2014, 21(14):41–45.

[6] Fageeri S O, Ahmad R. An Efficient Log File Analysis Algorithm Using Binary-based Data Structure [J]. Procedia - Social and Behavioral Sciences, 2014, 129: 518-526.