Applied Mechanics and Materials *Online: 2013-09-03*
ISSN: 1662-7482, Vols. 397-400, pp 2326-2329
doi:10.4028/www.scientific.net/AMM.397-400.2326

The Analysis and Data Mining of Students' Online Data Based on Digital Campus

Quan Shi, Yuan Qian, Yan Gong, Xiaomin Zhu, Junqiang Yan,

Nantong University

E-mail: sq@ntu.edu.cn

Keywords: Data Mining, K-means, Microsoft Visual Studio.NET, DMX

Abstract. The research takes the campus network users logging on the Internet as the analysis object, using the data preprocessing technology to clean up the original data, combined with statistical analysis and data mining technology to analyze the users access log records, which will result in the form of dynamic charts for Web display, by using Microsoft SQL Server 2008 and Microsoft Visual Studio 2010. Take use of intelligent .NET platform, combined with K-means algorithm to cluster the students' information. DMX (Data Mining Extensions) will be used to show the mining results on the Web. The realization of the system can not only carry on correct guide to Internet users and regulate the behavior of students, but also have important guiding meaning to managers and policy makers for analysis and making decision.

Introduction

With the development of digital campus, the phenomenon that a large number of students' access data cannot be reasonably utilized has occurred. If we can take the Internet behavior of the students according to the Internet time, Internet traffic, the cost of Internet access for quantitative or qualitative representation, we will find similar trends exist in their use of the network. What's more, if we can make a cluster analysis on the students' data from Internet, their failed courses and their scholarships, we can comprehensively understand the students' information and gain characteristics of the group of students' access to the Internet. Through different management according to the different characteristics of the groups of students, it will be more effective in enhancing the efficiency of school management. At the same time, it can improve the operation and management of campus network.

The Process of Data Mining

Data Mining is a process which picks up the information from the large, incomplete, noise, random data that people do not know in advance, but it's the process containing potentially useful information and knowledge. The first step of Data Mining is to define the task and select the data we need. Secondly, clean up the data, and sort out the data to meet the needs. Then we have to establish a mining structure and a mining model for data analysis according to pre-designed K-means algorithm to find the hidden information. Finally, for the convenience of users, use DMX to select the useful information in the page to display on the page.

K-means Algorithm. The basic idea of K-Means algorithm: Firstly, randomly select k objects from the total n objects as the initial clustering center. Secondly, the rest each object is assigned to its nearest clustering according to the distance from clustering center. Then, calculate the clustering center of each clustering again (the average of the clustering's all objects' value). Repeat the above steps until the clustering center's value will not change.

DMX (Data Mining Extensions). DMX is SQL data mining extension language; the purpose to put it forward is to define the unified concept and query mode for data mining, similar to the role of SQL language in the database. Using the data definition statements in DMX can create and define a

new mining structure and mining model, import and export mining model and mining structure, and delete the existing model from a database. Using DMX can finalize the design of the model for browsing, forecasting and management.

The Analysis and Data Mining of Students' Online Data

The Data Pretreatment. The information about students in the database needs pretreatment when we make analysis on it so that data can meet the need of statistical analysis and data mining analysis. Taking scholarship as an example, all we need is the scholarship of all students but not just the details of the students have got scholarship, so we should make operation about the students without scholarship.

Statistical Analysis of Data. The analysis of students accessing the Internet was statistics according to the schools, departments, classes, students of such category. The jump of the page is made according to the Chart control and the Text document. The analysis of the web log data is analyzed according to different user groups. Using Chart control to display the data analysis, as shown in Figure 1, it shows the undergraduates' online data about connection of month long.

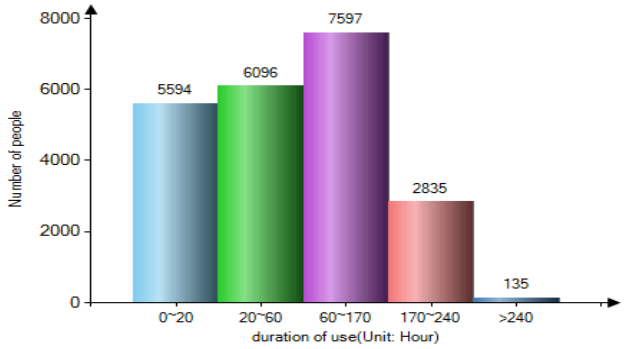

Figure 1 The information of undergraduates' online data

According to the online charging ways of undergraduates, each month every student has 20 hours of free access to the Internet. According to the timing and charging way, each student each month spends 20 Yuan so that they can get totally 60 hours to surf the Internet. If you choose 20 Yuan of domestic monthly package, each student can get 170 hours. It can be seen in Figure 1 that the time on the Internet of the most students is controlled within 170 hours. Only a minority of about 3000 students, whose time is more than 170 hours which accounted for about 10% of the total number of the students. From the graph, we can find that at the time period of 0~20, 0~60 and 60~170, the proportion of users is relatively uniform. The school's network administrators can adjust the charging policy, such as plus a 10 Yuan package of 60 hours. School administrators should also pay attention to the students who surf the Internet for a longer time, especially the freshmen and sophomore who have heavy burden and task on learning but little relationship with the Internet.

Analysis of Data Mining about Students' the number failed subjects and Online Data. According to the above data mining process, BI (Business Intelligence Development Studio) platform can build a data source, a data source view, a mining structure and mining model. Thus, all kinds of students' data mining analysis and the characteristics of each class of students' network behavior can be found. Under the platform of BI model, the model shows in two types: Microsoft classification viewer, Microsoft general content viewer. Because the two viewers under BI platform are not conducive to the query of clustering analysis's results, use DMX (Data Mining Extension) language in the system to show the information of useful results on the Web page such as Table 1.

In Figure 2, the objects of failing mining are the students of grade 2009 who are failed in first 2011-2012 semester. The methods are taken to discrete total number of failed subjects in the mining process, and main statistics are the number of students who are not failed, or the failed subject's number is one, two, three and over three.

Table1 The clustering analysis of the number of failed subjects and the students' online data of one grade

classification	number of students	The total number of failed subjects	time(/H)	Total Internet traffic(/G)）	upload traffic(/G)	download traffic(/G)
class1	4636	none	224.59	19.46	5.78	14.90
class2	490	one	288.17	24.88	6.36	19.37
class3	126	two	342.11	28.93	9.50	21.98
class4	47	three	380.01	23.72	7.47	17.45
class5	34	over three	280.85	24.39	6.41	18.16

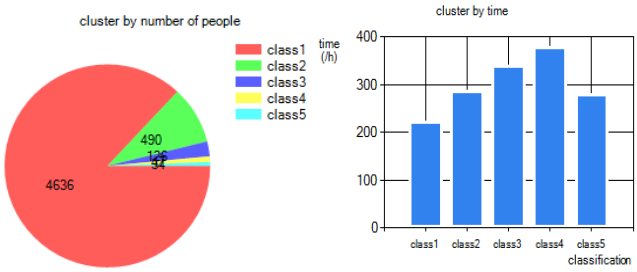

Figure 2 Results by using the DMX

Using K-means algorithm for data mining analysis, the law which is obtained can be shown in Figure 2. The number of students are decreasing with the number of failed subjects, while Internet traffic, Internet costs between class 1 and class 4 are increasing with failed subjects, which is consistent with the general logic. In general, online time is longer, the student is easier to fail.

What's interesting is when the number of failed subjects reached 3 or more, the Internet time and Internet traffic exhibits a clear downward trend. In order to verify the accuracy of mining models, this research also makes the mining analysis of the students of grade 2010 in first 2011-2012 semester who failed the subjects. According to the results, we can see clearly that once the failed subjects are 3 or more, Internet traffic and online time will be less, so there is reason to doubt that these students are often not stay at school, but chose to go to the Internet cafes near schools.

Analysis of Data Mining about Students' Scholarship and Online Data. As shown in Table 2 and Figure 3, it can be found that in the second category and the third category the two indicators are similar, and there are obvious gap with the forth category, which is in accordance with the standard of school scholarship. The number of the students of the first-prize scholarship account for only 5%, and the second-prize and the third-prize for the total number of the ratio of 30% and 50%, that is to say, the students with first-prize scholarship must be the most outstanding, and this also make these students must study hard. The rules of the second-prize and the third-prize for students are not so strict, thus the difference of the ratio is relatively close to the Internet, so the online situation of the two categories are similar. In addition, through observation, it can also find that length of time on the Internet and Internet traffic of the students who get access to individual scholarship is higher than other types of students got scholarships, and even higher than those who do not get the scholarship. In order to verify the accuracy of the model, this paper selects the students of grade 2010 in the mining model in first 2011-2012 semester. According to the analysis of individual scholarship, we can find that this phenomenon is true. Because the school scholarship award mainly refers to activities award, research and innovation award, the social work prize, sportsmanship award. These awards are primarily concerned with the students' organization

activities, social work ability, so that students need to cultivate the ability of their own through various channels, which may also make these students' length of time on the Internet and Internet traffic will be higher than those of the other students'.

Table2 The clustering analysis of the number of scholarships and the students' online data of one grade

class	the number of student	Scholarship category	Time (/H)	upload traffic(/G)	total traffic (/G)	download traffic(/G)
class 1	3170	none	565.3	11.15	42.78	32.77
class 2	1438	the third-prize scholarship	429.41	9.12	34.94	26.90
class 3	587	the second-prize scholarship	414.93	8.92	34.18	26.81
class 4	288	the first-prize scholarship	353.44	7.50	28.15	21.51
class 5	35	Individual scholarship	554.69	11.60	46.57	36.49

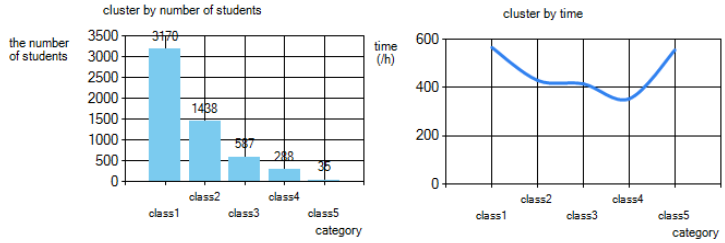

Figure 3 Results by using the DMX

Conclusions

The data mining analysis of student's online data from digital campus is based on data mining technology, knowledge discovery as the guiding ideology, finding the users' online behavior characteristics according to the records the student's access to the Internet, and combined with additional information about the students, learning students' information, which is taken for the campus network through the stable operation and construction concept. To apply the data mining techniques can not only be able to generalize the characteristics of each class of students, but also make use of the model which has been derived to predict the behavior of students. It has great guiding significance for practical work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61171132); Natural Science Foundation of Jiangsu Province (No. BK2010280); Applying Study Foundation of Nantong(No.BK2011003, BK2012001, BK2012034).

References

[1] Xie, J.Y. , Jiang, S., Xie, W.X., et al, An efficient global K-means clustering algorithm. Journal of Computers, 2011. (6):271-279.

[2] Jatmiko S., Refianti R., Mutiara A.B., et al. Analysis data of student's GPA and travelling time to campus using clustering algorithm Affinity propagation and K-means. Journal of Theoretical and Applied Information Technology, 2012.40(2):213-217.

[3] Jamie M., ZhaoHui T., Bogdan C. The principle and application of data mining：SQL Server2008Data base[M]. Beijing: Tsinghua University press，2010，482-485.

[4] Sridhar A., Sowndarya S., Efficiency of K-Means Clustering Algorithm in Mining Outliers from Large Data Sets. International Journal on Computer Science and Engineering, 2010.3043-3045.