

Statistical analysis of city and the villages Internet users based on user logs

ZHANG Yi-wen, BAI Yan-qi, YANG An-ju

College of Information Engineering, An Hui Xin Hua University, Hefei 230088

Abstract. In recent years, with the rapid increase of users active on the Internet, Internet users access log is also increasing rapidly. According to the user's Internet access log analysis of the characteristics of user behavior on the Internet. In this paper, we classify the statistical analysis of the behavior of Internet users by collecting information and data on urban and rural Internet user behavior. This result may provide a basis for guiding the behavior of Internet software manufacturers or government.

1 INTRODUCTION

When the users visit the Internet, the website will regard log as the carrier and record the interactive information which users use on internet. Statistical analysis [1] is an important method of discovering the law and understanding the essence from the large amount of logs data. The analysis of users' logs [2] and studying in-depth characteristic behaviors of users can explore rules which the users search, reveal the intention which users search, and provide reference for improving the quality of search engine. Thorough thinking and discussions based on Web mining technology show that Web logs file contains a large number of information that users have access. Analysis of these data will help researchers know retrieval law and understand users' search behaviors much better.

The author firstly divides the users into three types of users: urban, suburban and rural. Besides, the author also analyzes differences of three types of users in income, age and education through statistical analysis of users attribute information and internet users behaviors data. According to the SPSS statistical software, analyze the differences of internet behaviors, what is said above is helpful to provide the government, software developers and e-commerce with guidance for understanding of the behaviors' characteristics of each users more accurately.

2 BEHAVIOURS OF INTERNET USER

2.1 web logs of internet users

Users Logs [3], known as Web Logs which generally record the interaction between Web users and websites, are important data source for Web log mining.

A corresponding log file will be formed when a user switches on each time, such as the following log examples "0ab6bbbedff24ec8baac905f45ae314c_2012-05-07_21-22-38.txt", recorded the sample of ID

0ab6bbbedff24ec8baac905f45ae314c during the 2012-05-07 day, since 21:22:38, the operator can further access to the sample of the population attribute information whose ID is "0ab6bbbedff24ec8baac905f45ae314c" by demographic.CSV.

2.2 behaviours of internet user

Extract the fields, the name of programs used by the users. Are represented by N in the behavior logs of the above Internet users [4]. Extract the name of programs used by each users and act it as behavior of Internet users.

3 BEHAVIOIR ANALYSIS OF INTERNET USER

3.1 dada collection and pretreatment

Data of this Paper is derived from a contest held by Didital hall, this study randomly selects 1000 samples of users in log behavior in the following four weeks: from 2012-05-07 to 2012-05-13, 2012-06-04 to 2012-06-04, from 2012-05-13 to 2012-07-08, 2012-08-06 to 2012-08-06, and the corresponding attribute information of sample population.

Demographic information includes the users' birth year, gender, educational background, educational degree, region, category and so on. The logs of behaviors include users' boot time and the corresponding process after startup.

(1) Extract all online processes of each users within 4 weeks.

(2) Calculate the frequency used by each process from all processes of each users within 4 weeks.

(3) The Internet behaviors can be divided into [5] different software about video, chat tools, browser, games, entertainment, music,

office, security, antivirus, download tools, office, input method, graphic images, mobile phones, digital, program ming development, news and information, reading, translation and network applications, systems, tools and other software which hold a total of 17 classes, and calculate the frequency of individual net citizen' behaviors within 4 weeks.

3.2 analysis of internet user' basic information

According to the gender, educational background, occupation, income, region and other attributes of Internet users[6], analysis in sequence is shown:most of the respondents were male (77.6%), the oldest was 75 years old, the youngest was 8 years old, the average was 33.02 years old, and the standard deviation was 9.270. The highest degree was (39.2%), followed by junior college (29.0%), primary school and below(0.3%) is least among those surveies. The kind of people who were surveyed was varoius and dispersive. There are a large number of employees, students and professional technicians, accounting for 21.1%, 17.5% and 14.4% respectively. The income of the respondents is also uneven, with the largest number of people with incomes

of 3001-5000 yuan, accounting for 22.6%, the minimum of 500 yuan and below,accounting for 3.4%, and the non-income accounting for 14.8%. The citizens account for 64.0%.

According to difference of users' location [7], the users are divided into three categories: urban users, suburban users and rural users.

The proportion of these users are shown in table 1:

Table 1 User Distribution table

	num	percentage (%)
city	640	64.0
suburban	73	7.3
rural	127	12.7
unknown	160	16.0
total	100	100.0

Table 2 shows that there are 640 users(64%) in the city, 73 (7.3%) in the suburbs, 127 (12.7%) in rural areas, and 160 (16%) not knowing the geographical location. The users related mainly focus on urban users.

table 2 the diversy of user basics

	Category	City	Suburban	Rural	Chi-Square value	p P Value
Gender	Male	510	61	102	0.621	0.733
	Female	130	12	25		
Income	<=500Yuan	86	14	28	45.611	<0.0001
	501 ~ 1000 Yuan	13	2	7		
	1001 ~ 1500 Yuan	21	4	8		
	1501 ~ 2000 Yuan	34	4	17		
	2001 ~ 3000 Yuan	64	11	11		
	3001 ~ 5000 Yuan	156	14	25		
	5001 ~ 8000 Yuan	147	18	22		
	8001 ~ 12000 Yuan	76	4	5		
	>=12000	30	0	1		
	>=12000	13	2.0	3		
Degree	Primary and below	1	6	2	158.246	<0.0001
	Junior	22	21	18		
	Senior/secondary/technical School	101	14	33		
	College	200	31	35		
	University degree	275	1	35		
	Master and above	41	0	4		

Table2 shows that:

(1) There are 510 males and 130 females in the city, 61 males and 12 females in the suburbs, 102 females and 25 males in rural areas, and the largest gender gap lies in urban users. The chi-square test was conducted on the constitute of urban male and female , suburban and rural areas,with the value of 0.621, $p=0.733>0.05$, which indicates that the differences between the male and

female in urban, suburban and rural areas were not statistically significant.

(2) Income mainly concentrates on the range of 2001 yuan to 5000 yuan in urban, suburban and rural users , however ,there is a very big part of rural users has no income,.No matter what kind of income level which people is at, the number of urban users is the largest ,followed by country users and the number of

suburb users .The chi-square test was conducted on the constitute of suburban and rural users with different income levels ,with the value of 45.611, $p = 45.611 < 0.05$, indicates that the differences of subur and rural are statistically significant,the level of urban uses' income is higher than that of suburban and rural users.

(3) Eeducation of urban users is primarily university undergraduate and college, eeducation of suburban users is mainly college and middle school,education of rural users mainly high school,secondary technical

school, college and university undergraduate course. In general, Internet users mainly focus on the completion of nine-year compulsory education and above, few people among the primary school and below use internet . The chi-square test was conducted on the constitute of suburban and rural users with different educational level , with value of 158.246, $p = 158.246 < 0.05$,which indicates that urban, suburban and rural educational level was statistically significant. Urban users have higher educational levels than suburban and rural users.

Table 3 year of birth of urban, suburban and rural users

	NP	Min	MAX	Mean	Sd	Chi-S	p
City	640	1939	2004	1980.21	9.394		
Suburban	73	1963	1998	1982.77	7.455	34.93	<0.0001
Rural	127	1963	2001	1985.1	6.660		

NP: Number of people;MIN: Minimum; MAX: Maximum Value ;Mean: Mean value; Sd: Standard deviation; Chi-S:Chi-Square value.

Table 3 shows that the biggest year of birth of urban users is 1939 , the biggest year of birth of suburban and rural users is 1963 on average, the smallest year of birth of urban users is 2004, the smallest year of birth of suburban users is 1998, the smallest year of birth of rural users is 2001. The chi-square test was conducted on non-parametric test of the birth year of urban, suburban and rural users with value of 34.93, $p=0.0001<0.05$,which indicates the difference of birth year in urban, suburban and rural areas is statistically significant. The average age of urban users is greater than suburban and rural users.

4 CONCLUSION AND FUTURE WORK

The quality of statistical analysis based on users logs [8] mainly depends on the quality of users attribute data and behavior data. The more accurate data of users leads to the more accurate result of statistical analysis.

The author statistically analyzes behavior logs of 1000 users in four weeks .Because sample sizes is small , there are three types- the urban, suburban and rural users who need to be analyzed , and the number of each type users is not very big, three types of result can't accurately representative the attributes and behaviors of urban,suburban and rural Internet users. The results have some bias and still need to continue to expand the sample size to do corresponding research. The next step is to expand the analysis scale, increase the recommendation of other information, expand the data scale, and build a higher level of users statistical analysis. Aiming at the scale of users, it is better to use a more refined method to improve the efficiency of users behavior statistical analysis.

Acknowledgments

This work was Supported by Foundation for The Excellent Youth Scholars of anhui province program under Grant Nos.gxyqZD2018087.

References

1. Duricki D A, Soleman S, Moon L D. Analysis of longitudinal data from animals with missing values using SPSS[J]. Nature Protocols, 2016, 11(6):1112.
2. Starkings S. IBM SPSS statistics 19 made simple by Colin D. Gray and Paul R. Kinnear[J]. International Statistical Review,2012,80(2):333-334.
3. Son H, Friedmann E, Thomas S A. Application of pattern mixture models to address missing data in longitudinal data analysis using SPSS.[J]. Nursing Research, 2012, 61(3):págs. 195-203..
4. Jiliang, TANG, Xufei, et al. Enriching short text representation in microblog for clustering[J]. Frontiers of Computer Science, 2012, 6(1):88-101.
5. Wang Xuefeng W, Jie R, Youguo W. Co-inventor Analysis on China's international technology collaboration in US patent activities: 1976-2010[J]. Procedia Engineering,2012(37):314-322.
6. Secar M. Statistic analysis of international tourism on romanian seaside [J].Annals of the University of Petrosani,Economics,2010,10(1):327-334.
7. Yeung P. SPSS survival manual. A step by step guide to data analysis using IBM SPSS (5th ed) [Book Review][J]. Aotearoa New Zealand Social Work, 2014, 26(4).
8. Dunn P. SPSS survival manual: a step by step guide to data analysis using IBM SPSS[J]. Australian & New Zealand Journal of Public Health, 2013, 37(6):597-598.