Topological Structure and Interest Spectrum of the Group Interest Network

Ning Zhang

Business School, University of Shanghai for Science and Technology,
Shanghai 200093, P. R. China
zhangning@usst.edu.cn

Abstract. In this paper, the behavior characteristics that the specifical campus group users accessing world wide web has been studied, the dynamic group interest network has been constructed, which was a para-bipartite graph and the topological structure had been discussed. Although the users' visiting time is random and the web pages they visited are different but the interests of a majority of the campus group are accordant. The results indicate that the incoming degree distribution of the group interest network follows power law. And the group interest spectrum was basically steady. The visiting behavior of the campus group had their special disciplinarian.

Keywords: Complex network, para-bipartite graph, group interest spectrum, behavior characteristics, in-degree.

1 Introduction

The rapid development of information technology has brought the great challenges in theoretical study and practical application, and its significant social and economic values attracted significant coverage from all disciplines. The research of information systems' complexity has become one of the important problems for international academic community, especially for the cross-frontier scientific research. Recently, with the research and development of the complex networks, the information system regard as a complex system, has become a cross-research focal point [1-4]. Complex networks are available for studying a great deal of practical system, such as the World Wide Web, the Internet, the electrical power-grid networks, the biological nets and social networks [5-9]. Many empirical evidences indicate that the topological characteristics of practical networks are neither regular nor random [10], they belong to both small world [11] and scale-free [12, 13]. The findings of complex network reflect the basic characteristics of many complex systems, bringing material breakthrough to these systems' research. For instance, scientist collaboration networks [14-16] can make us clear about the relationships among the scientists in different fields, which have short average path length but big clustering coefficient. That is to say, the scientist collaboration networks have the characters of good connectivity and strong clustering. The power law degree distribution of World Wide Web let it has dual-characteristics, robust and frangible [17].

J. Zhou (Ed.): Complex 2009, Part I, LNICST 4, pp. 729 – 736, 2009.

These universal characteristics have significant theoretical meanings and engineering application values. As for the engineering application, the values are obviously. If we can identify all kinds of groups of 137 million Internet users in China and the popular resources of 0.843 million webs [18], then we can pick the most popular resources according to their sort order to store. In this way, the power law distribution can ensure that the mainstream resources are able to meet the needs of the majority.

Information in our life become more and more important. How to conquer 'figure gulf' to let everyone share information resources fairly is always being the global issue and also a full concerned issue to government administrator. Different groups of people need different web resources. In the limited resources, we should to consider which kind of information resources can meet the needs of the majority, and use the most proper way to let the individuals far from cities sharing sunshine information [19,20]. That means we should to study different groups' interests. With the help of clustering analytical method, linearity regressive analytical method, we often dig web users' interests and constitute interest model, according to word, text structure characters, paragraph and sorts expression ability [21]. For user's preference, we can use self-adaptive theory to study user's preference [22] and constitute user's preference model base one data cube [23]. For the data digging method, we can use Markov model to searching for user's behavior characters [24] and find out user interest profile according to the implicit feedback [25], and then combining web content and behavior analysis [26] or base on its searching history [27] to invest interesting model. There are plenty of searching engines and information filtration methods. But none of those researches deal with group users' interest structure characters, or explain group interest spectrum's structure and stability mechanism, or reveal the clustering phenomena and rules of group interests.

Our researches are aimed at studying group users' behavior characters and topological characters of group interesting network, finding out interest spectrum of group interest network and it self's evolvement rules. This essay is only involved in introduction researching results. The further results will be published in other papers as the researches go deeper.

2 Data Analysis

The group users in this paper refer to the faculty and the students in our campus. They visit the Internet by local area network in their offices or dormitories, there are one fixed IP address for one room, but one room may have several computers and more than one users, so we call these users as group users.

There are 3 sets of records were collected during the continuous period (see table 1). For the dataset 1, the time period is during the second semester of our school year. In this semester, the senior students begin their graduation project, they have more free time to visite the Internet. For the dataset 2, the time period is during the first semester of our new school year. In this semester the new senior sdudents had their courses and the freshers just came into the school, began their military Training. For the dataset 3, the time period is during our ordinary school activity and all of the students and faculties

began their normal life. These data exist clearly periods. From fig.1, we can see that the lower traffic volume appeared during 23:00 Pm to 7:00 Am in every day, and the average traffic volume per hour in dataset 1 is more than that in dataset 2 and 3. From fig.2, we can see that the web traffic volume reduced during legal holidays, and the traffic volume in holydays was less than the working days, lower traffic volume appeared during the weekends in every week, the lowest traffic volume appeared at October golden week. From fig.3, we can see common fluctuating cycles and some times of the day the variations are busier than others.

Table 1. The data statistics

Data set	Begin time	End time
1	4:00 on Mar. 14, 2006	3:59 on Mar. 20, 2006
2	4:00 on Aug. 30, 2006	3:59 on Oct. 8, 2006
3	4:00 on Nov. 19, 2006	3:59 on Nov. 22, 2006

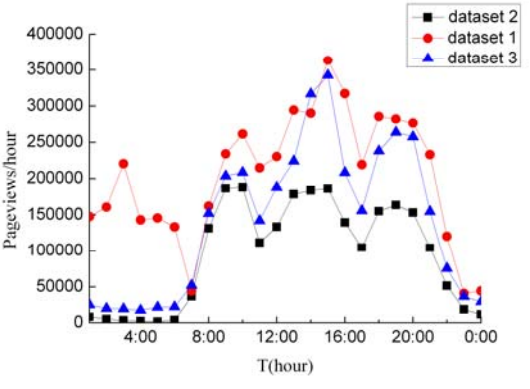

Fig. 1. The average hourly activity per hour

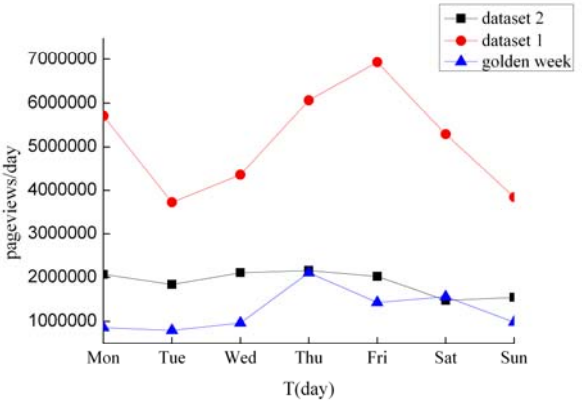

Fig. 2. The average week activity in Data Set 1 and 2

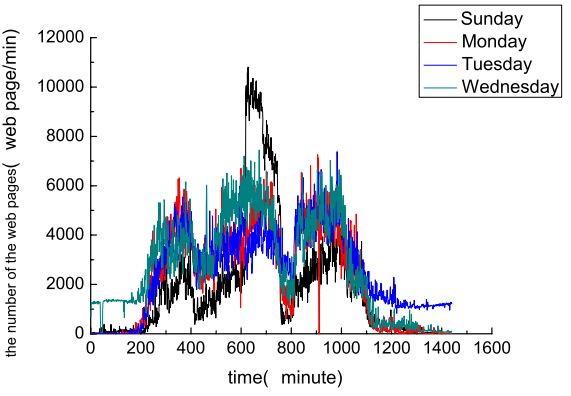

Fig. 3. The total traffic volume per minute in Data Set 3

All these analysis results reflect the characters of the group user' activities, the lowest traffic volume per day is corresponding to the dorms' black out time. During the legal holidays, faculties don't work and most students go home, so the traffic volume is also lower than the work time. That is to say, group users access campus network are mostly related to their work. As a result, although the visiting time of each users visiting the Internet in the campus are random, with the helpp of data analyzing, we still can find some general rules of the group users' behavior.

3 The Group Interest Network

Actually, the activities of group users visiting the Internet should be a dynamic random process, we can use the method of complex network to describe it. The group users' web visiting at each moment can construct a group interest network. The group users' web visiting during a period can constrcte dynamic group interest network. The network have the format of para-bipartite graph and it contain two kinds of vertices, one is user vertex, which refers to group users, the other one is information resource vertex, which refers to the web site resource constructed by many web pages and need

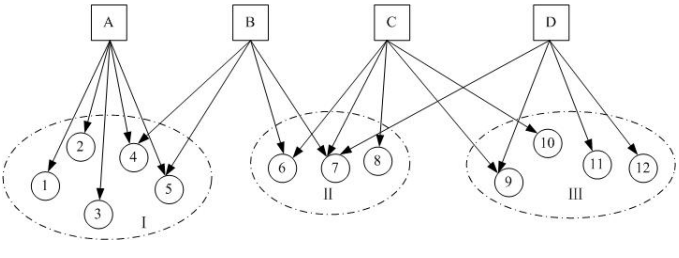

Fig. 4. Schematic illustration of the group interest network. □Nodes A-D represent group users, ○ nodes 1-12 represent web pages, Dash dot circles Ⅰ-Ⅲ represent resources (web sites)

to transfer. Since it's not like the bipartite graph with explicit two different kinds of vertices like the definition in the graph theory, we call it as para-bipartite graph. Group users' visiting activities set up the relationship between group users and information resources, which can be expressed by directed connection. And the complex relationship of many users corresponding to many information resources construct the group interest network (see fig.4). According to the group interest network, we can get users relation network and resources relation network by the projection of users and resources respectively.

We construct group interest network by classifying the web pages to different web site according to group users accessing habit per day (see table 2), which is a directed dynamic complex network. Basing on dataset 1, the Tuesday's network contains 658 users and 16078 web sites. The Wednesday's network contains 676 users and 17491 web sites. The Thursday's network contains 674 users and 18233 web sites and etc. The network contains 1023 users and 60079 web sites in total during one week. We calculate this group interst networks' in-degree. The in-degree here means the numbers of group users' visiting a web site. Such as in fig.4, the group user A visited 5 pages of web site 1 and B visited 2 pages of web site 1, then the visiting number of web site 1 is 7. In the same way, the visiting number of website 2 is 6 and number of website 3 is 5. Even though the number of user are different and the web sites they visited are different in every day, the in-degree distribution of group interest network have power law character (see fig. 5). Which means that lots of web sites are with a few links (visiting amount), a few web sites are with a medium number of links and a very few noteworthy web sites are with a large number of links in this network. The in-degree frequencies and their per centum of the group interest network can be seen in table 3. The in-degree frequencies refer to the numbers of a in-degree. Such as in fig.4, the in-degree frequencies of in-degree 1 are 7, in-degree frequencies of in-degree 2 are 4, and 3 are 1.

Table 2. The statistics of the group user visiting the Internet

Week	User	Web page	Web site	Degree exponent
3.14 (Tuesday)	658	3727905	16078	1.530477
3.15 (Wednesday)	676	4361211	17491	1.533105
3.16 (Thursday)	674	6068099	18233	1520289
3.17 (Friday)	647	6931486	18390	1.528237
3.18 (Saturday)	381	5291921	12215	1.520854
3.19 (Sunday)	393	3844392	14488	1.532267
3.20 (Monday)	663	5710836	17584	1.521534

Table 3. The in-degree frequencies and their per centum of the group interest network

In-degree	in-degree frequencies	per centum
1-100	51218	85.25%
101-800	6845	11.39%
801-1681974	2016	3.36%

The in-degree distribution of group interest network follows power law, $P_{in}(k) \propto k^{-\gamma}$, $\gamma = 1.52$ (see fig.5), so this network was scale free.

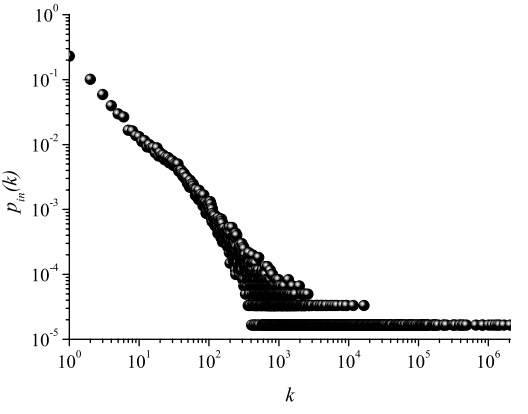

Fig. 5. The in-degree distribution of the group interest network

4 The Interest Spectrum

By ordering the web resources of data set 1, in the web sites with the top 20 visiting volumes, users'number of 7 web sites is 1, users' number, which is no more than 10, web sites are 3. Only 10 popular web sites have more group users. So we got the top ten popular web sites and page numbers of the group users visited (see table 4). These results reveal that which kind of web sites can attract much more our campus' group user, and these users' a part of interest spectrum. We find that the group interest spectrum have a good stability. For instance, although the top 10 web sites of the group users accessed have different orders each day, they are still in each day's top 10 list. We can find out the number of the group users which visited the top ten web sites (see table 5). In the further research, we can get every group users' interst spectrum.

Table 4. The top 10 web sites of the group user surfing during one week

Web site	Web pages
sina.com.cn	1681974
163.com	1451668
sohu.com	834362
usst.edu.cn	662827
online.sh.cn	444066
msn.com	424835
allyes.com	390524
pconline.com.cn	385017
taobao.com	321813
chinaren.com	251568

Table 5. The users' number of visiting the top 10th web sites during one week

Date / Web site	Mar.14	Mar.15	Mar.16	Mar.17	Mar.18	Mar.19	Mar.20
sina.com.cn	341	461	351	339	202	216	351
163.com	338	360	343	331	199	224	349
sohu.com	279	305	293	282	161	179	293
usst.edu.cn	252	268	271	260	119	128	295
online.sh.cn	128	129	130	114	63	61	123
msn.com	283	287	301	283	243	190	318
allyes.com	435	439	432	422	251	260	467
pconline.com.cn	74	79	84	67	53	59	74
taobao.com	111	126	124	128	80	85	147
chinaren.com	112	134	127	109	75	78	123

Researches indicate that the major users have the similar interests, and if we can conform the information resources interested by most individuals and use the abroad storage technique to maintain most peoples request, then we can let people to obtain sunshine information in the most economical way.

5 Conclusions

According to the relationship between group users and information resources, the special group users' web visiting behaviour has been studyed, the time features that the group user visited world-wide-web has been observed, the group interst network has been set up, the topological structure has been discussed in this paper. With the help of complex network's method, the study indicates that the group interest network's in-degree distribution belongs to power law distribution. The given group's interest spectrum is basically stable and the visiting behavior of the campus group had their special disciplinarian, and the interests of a majority of the campus' group users are accordant.

Acknowledgements. This work was supported by Shanghai Leading Academic Discipline Project (No. S30501) and the Natural Science Foundation of Shanghai (06ZR14144).

References

1. Simkin, M.V., Roychowdhury, V.P.: A theory of Web traffic. EuroPhys. Lett. 82, 28006 (2007)
2. Goncalves, B., Ramasco, J.J.: Human dynamics revealed through Web analytics. Phys. Rev. E 78, 26123 (2008)
3. Golder, S., Wilkinson, D., Huberman, B.A.: Rythms of social interaction: messaging within a massive online network, e-print ArXiv cs/0611137 (2006)

4. Meiss, M.R., Menczer, F., Fortunato, S., Flammini, A., Vespignani, A.: A, Ranking Web sites with real user traffic. In: Proc. WSDM (2008)
5. Albert, R., Barabási, A.-L.: Statistical mechanics of complex networks. Rev. Mod. Phys. 74, 47–97 (2002)
6. Dorogovtsev, S.N., Mendes, J.F.F.: Evolution of Networks: From Biological Nets to the Internet and the WWW. Oxford Univ. Press, Oxford (2003)
7. Pastor-Satorras, R., Vespignani, A.: Evolution and Structure of the Internet: a Statistical Physics Approach. Cambridge Univ. Press, Cambridge (2004)
8. Newman, M.E.J.: The structure and function of complex networks. SIAM Rev. 45, 167–256 (2003)
9. Amaral, L.A.N., Ottino, J.M.: Complex networks—augmenting the framework for the study of complex systems. Eur. Phys. J. B 38, 147–162 (2004)
10. Erdös, P., Rényi, A.: On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci. 5, 17–61 (1960)
11. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature 393, 440–442 (1998)
12. Albert, R., Jeong, H., Barabási, A.-L.: Diameter of the World Wide Web. Nature 401, 130–131 (1999)
13. Faloutsos, M., Faloutsos, P., Faloutsos, C.: On power-law relationships of the Internet topology. Comput. Commun. Rev. 29, 251–262 (1999)
14. Newman, M.E.J.: Scientific collaboration networks: I.Network construction and fundamental results. Phys. Rev. E 64, 16131 (2001)
15. Newman, M.E.J.: Scientific collaboration networks: II.Shortest paths, weighted networks, and centrality. Phys. Rev. E 64, 16132 (2001)
16. Newman, M.E.J.: The structure of scientific collaboration networks. Proc. Natl. Acad. Sci. USA 98, 404–409 (2001)
17. Albert, R., Jeong, H., Barabási, A.-L.: Error and attack tolerance of complex networks. Nature (London) 406, 378 (2000)
18. China Internet Network Information Center, Statistical Reports on the Internet Development in China (in Chinese), http://www.cnnic.cn/html/Dir/2007/01/22/4395.htm
19. Li, Y.-P.: Sunshine information——conflict-free share structure. China engineering science 2(1), 24–27 (2000) (in Chinese)
20. Li, Y.-P.: Construct Broad-Storage grid. Computer world 37 (2005) (in Chinese)
21. Lin, H.-F., Yang, Y.-S.: The representation and update mechanism for user profile. Journal of computer research and development 39(7), 843–847 (2002) (in Chinese)
22. Lei, Y.-S., Gan, R.-C., Du, D.: A framework of adaptive vertical website based on user preferences. Computer Engineering 31(24), 18–20 (2005) (in Chinese)
23. Chen, J.-L.: User's interest model based on data cube. Journal of Gulin university of technology 25(1), 84–88 (2005) (in Chinese)
24. Hu, Y.-H., Zhao, H.-J., Lu, H.-R., Wang, H.-J.: Research on extracting patterns from web user behavior. Computer Engineering and Design 27(18), 3416–3418 (2006) (in Chinese)
25. Sun, T.-L., Yang, F.-Q.: An approach of building and updating user interest profile according to the implicit feedback. Journal of northeast normal university 35(3), 99–104 (2003) (in Chinese)
26. Zhao, Y.-C., Fu, G.-Y., Zhu, Z.-Y.: User interest mining of combining web content and behavior analysis. Computer Engineering 31(12), 93–94 (2005) (in Chinese)
27. Xu, K., Cui, Z.-M.: User profile model based on user search histories. Computer technology and development 16(5), 18–20 (2006) (in Chinese)