Empirical Research on the Topological Properties of Internet+ Information Resources Network Nodes

Bin WU[1,a], Ping WU[2,*], Ji-Tao MA[3,a], Ting-Ting GUO[3,a], Jun LI[3,.a], and Wei LIU[3,a]

[1]*No. 542, Beijing Rd. Kunming, Yunnan Province, PRC*
[2]*No.246 East Renming Rd. Kunming, Yunnan Province PRC*
[3]*No. 542, Beijing Rd. Kunming, Yunnan Province, PRC*
[a]*star_amethyst@qq.com*
** Corresponding author: 138087273@qq.com*

Abstract: The "Internet+" is the product of the Internet development, and its network topology isn't the same as the traditional Internet. The relevance of the average daily visiting data and the daily page viewing data are studied empirically, the rich-club coefficient and the node access probability are redefined, and the topological entropy model to measure the degree of nodes information aggregation is built by using the entropy theory. The experimental results showed that the calculation model scaled the degree of information aggregation of the nodes in the "Internet+" topology efficiently. It provides an available computational model for the observation of resource access behaviors in the "Internet+" network.

1 Introduction

Internet network topology can draw the evolution trend of the dynamic characteristics of Internet users' behavior, it is helpful to understand the interaction mechanism among the participants on the network limited resources, and then to design a theoretical tool to avoid network congestion and get a more reasonable method to use the network information resources. At present, the main idea of Internet topology research is to abstract Internet into complex network to deal with, the known research results include small world network[1,2], scale free network[3,4], weighted directed network and its derivative[5,6,7,8].

In recent years, in order to simulate the dynamic changes more realistically in Internet topology, to understand and grasp the characteristics of the dynamic evolution of the Internet, many scholars have conducted a detailed study of Internet topology, by which some of the newer results have been achieved. From the view of the "like attracts like" partial connection phenomenon, someone has studied the network node connection status and internal LAN connection preferences, designed a new network topology model, and reflected the characteristics of small world and aggregation better than the BA model and the GLP model[9]. Garcia Robledo and others have analyzed the linear relationship of multiple complex network test indicators to select a non-redundant set based on the Internet evolution characteristics and the unsupervised machine learning technique, and that the Internet topology structure will be analyzed accurately with these non-redundant indicators[10]. Chai, He and colleague have designed a network cache gain algorithm with the concept of complex network betweenness, it shows that the algorithm not only has the capable of quickly delivering contents, but also reflects the power-law distribution characteristics of Internet topology intuitively[11]. Bo J and Ying Z have studied the spectrum stability of the network nodes interact growth based continuous data snapshots of Internet network topology and found the mapping relations between topology spectrum ranges and values of the network indicators[12]. Holbert B and Tati S have construct a network component of the estimated virtual topology through reasoning the real topology with partial route information, it resolves the problem of network topology detection failure caused by the incomplete route tracking information[13]. Liu Xiao and Zhao Hai have selected topological data (such as ipv4, ipv6 and internet AS) to compare and analyze structure indicators, found the linear relationship among the network structure order, the network scale and the network basic connectivity, and have gotten a result that the internal of the Internet is a dissipative structure[14].

In order to study the real network topology of Internet more deeply, the current research mainly focuses on the aspects of the network connection structure, the measurement index and the characteristic information of the network. However, the "Internet+" is a new form of the development of the Internet, the information content of network information resource stored in the nodes has the characteristics of polymerization and coordination, and the similar

resources is provided by nodes existed in the information resources network[15]. Therefore, the degree distribution of a single node in the information resources network is different from that of in the traditional Internet.

2 Network Node Topological Properties of the "Internet+"

2.1 Analysis of Network Nodes

Because the news content is very similar in the Internet, the same content may appear in multiple news sites in the same day[16]. It means that news nodes are similar in the network topology, so these nodes had been selected as a study object to grasp topological properties. The data, used for an empirical analysis, comes from the China Internet Data Platform(CNIDP), including QQ News, Phoenix News, Netease News and other domestic well-known Chinese news sites, and the time span is from January to August in the year 2015. First, a set of statistical charts of unique visitors(UV) and page views(PV) has been drawing after data-cleaned、data-indexed and other data processing steps (see Figure 1.a and Figure 1.b). As shown in figures 1.a & 1.b, UV and PV of news sites have a near linear relationship and the Phoenix New has the maximum UV and the maximum PV.

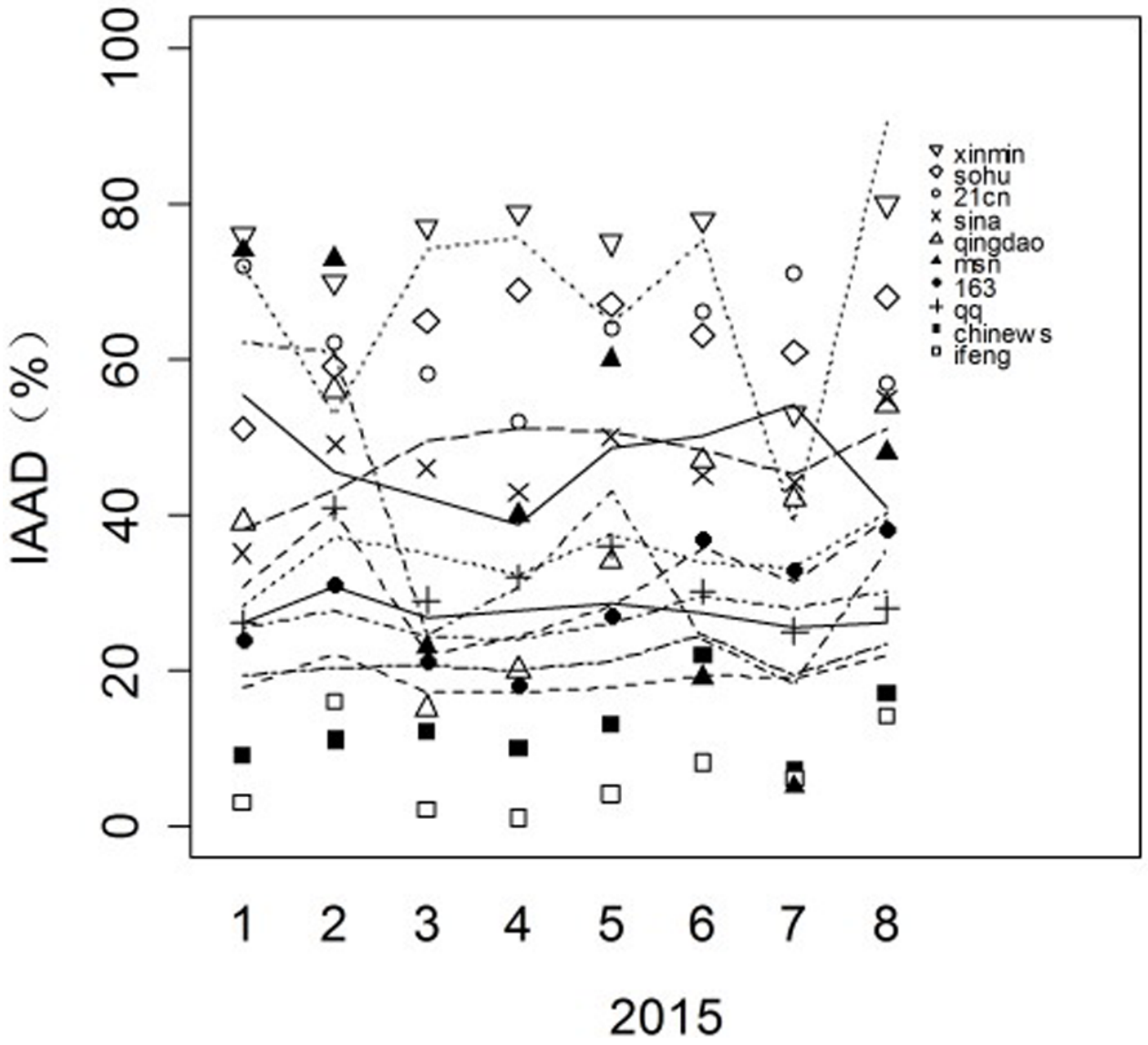


Figure 1. (a) UV statistical data; (b) PV statistical data

For a better observation on the accessing frequency of information resource on a node, the information average access density (IAAD), calculated by PV divide UV, is provided to present the sparse degree of information- accessed. IAAD curves have been drawn to show a reverse trend (see Figure 2) that a site (Xinmin News) with the lowest UV and the lowest PV has the largest IAAD, and a popular site(Phoenix News) has the minimum IAAD value.

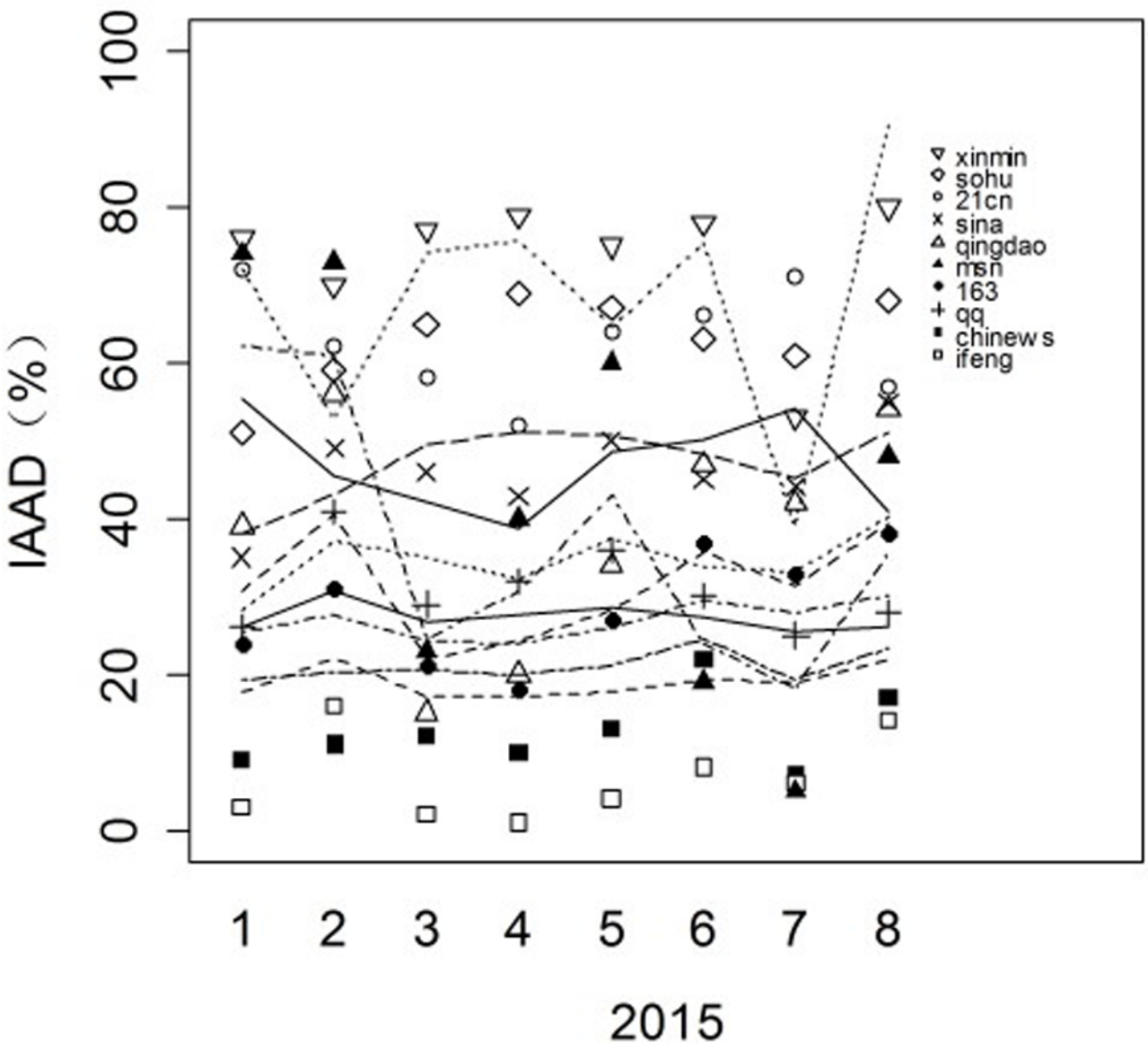


Figure 2. News sites with the information average access density curve in the year of 2015

The traceability of data has been analyzed to find that Shanghai Municipal Committee is in charge of the site of Xinmin News, the information with the site needs to be strictly audited prior to publication, the credibility of information content is high. As a result, the IAAD value and the probability that each message is viewed in the site are higher. In contrast, because the Phoenix News belongs to the Phoenix New Media consulting company, so the audit information has more lenient conditions, the IAAD and the probability is lower correspondingly. However, the Phoenix News average daily traffic is about 40 times the Xinmin News that it is not effective to explain the phenomenon of the "Internet+" network structure based on the topological measure of nodes and edges. In order to describe the "Internet+" network node topology effectively, it is necessary to analyze the rich-club and information aggregation properties of the network nodes.

2.2 The Rich-club Attribute of Network Nodes

Colizza and colleges, who analyzed the rich-club phenomenon of Internet, believe that the rich-club coefficient Φ can more effectively reflect the strength of the network traffic and the interaction process of network nodes. According to the definition of the rich-club coefficient, $\Phi=2l/r(r-1)$, where the l is the actual number of edges of neighborhood service node and the r is the number of nodes in the neighborhood. By definition, the Φ coefficient reflects the intensity of the surrounding nodes connected adjacent points directly. As an independent information node, it is unable to effectively estimate the Φ by comp uting the number of links directly-connected routers. To effectively reflect the rich-club attribute of network nodes, it's assumed that the Φ is the ratio of the number of actual accesses and the number of possible accesses in the unit time.

Definition 1 The concurrent access of information network node per unit time is li, the maximum number of concurrent is C, thenΦ= li/C. Considering the number of concurrent users and the service response

time is a near linear relation, the rich-club coefficient is calculated as follow:

$$\varphi_i = \frac{t_i - \alpha}{\beta \cdot C} \quad (1)$$

As the formula (1) shows, the φ_i is equal to the ratio of the network node response time and the maximum response delay time. That is to say, the longer response delay, the greater the load of the network nodes, the higher the number of concurrent. When the network node is full load, the number of concurrent is the highest, and the service response delay is at maximum.

So, the formula (1) can be approximated equivalent to:

$$\varphi_i = \bar{t}_i / t_{\max} \quad (2)$$

In the formula, the t_i is the average response time of nodes, and the t_{max} is the maximum response time among nodes.

2.3 The Information Aggregation Attribute of Network Nodes

With the widespread use of cloud computing technologies, the nodes in the Internet+ network will show the aggregation phenomenon[18]. Namely, with the same time interval t, the larger the arrivals number of the node, the higher the information-viewed probability of the node, which means network structure has an obvious aggregation attribution. In order to evaluate the attribute value of information aggregation quantitatively, it's assumed that the similar news content is evenly distributed on all nodes, the network transmission delay of all nodes is shorter than a timeout length, and the nodes can accept all requests without discard actions.

Definition 2 The information-viewed number in the node is defined the p_i, the arrivals number of the node is n_i, and the average information-viewed number is calculated by $c_i=n_i/p_i$, which represents the sparse degree that each message is viewed in the node. The information-viewed probability is defined as follow.

$$\rho_c = \frac{c_c}{\sum c_c} \quad (3)$$

Taking into account the rich-club attribute of the news network topology, and each node stores a large number of similar information, so the possibility of a node being accessed has a relationship with the information-viewed probability and the number of instant access concurrency. If the node access probability is high, the concurrency is low, which means that the value of the information stored in the node is not high, then the degree of the information aggregation is low. Thus, the degree of the information aggregation of nodes ξ_i should be related to the information-viewed number ρ_i and the number of instant access concurrency l_i.

$$\xi_i = f(l_i, \rho_i) \quad (4)$$

3 "Internet+" Network Topology Modeling

3.1 The Network Topological entropy Based on the Information Aggregation

"Internet+" is essentially an extension of the application of Internet, so the network structure also has the scale-free characteristic. Meanwhile, a large number of similar information may be spread on a number of nodes, on which there are different access probabilities. As a result, the network topology will have a certain order, and the topological entropy can be used to model the topology.

According to the definition of entropy, the topological entropy of the network node is given in formula (5):

$$H = -\sum_{i=1}^{n} \xi_i \ln \xi_i \quad (5)$$

In order to simplify the process of modeling the "Internet+" network topology, it is assumed that the t_{max} value of all nodes in the "Internet+" network is not obvious, and then the entropy of a single node is derived as follow.

$$p_i = l_i \cdot t = \frac{C \cdot \bar{t}_i}{t_{\max}} \cdot t$$

$$\Rightarrow$$

$$\frac{\bar{t}_i \cdot n_i}{p_i \cdot t_{\max} \sum_i c_i} = \frac{n_i}{t \cdot C \sum_i c_i}$$

$$= \rho_i \cdot \frac{p_i}{t \cdot C} = \rho_i \cdot l_i / C$$

$$\Rightarrow$$

$$H_i = -\frac{\rho_i \cdot l_i}{C} \ln \frac{\rho_i \cdot l_i}{C}$$

$$= -\frac{\rho_i \cdot l_i}{C} (\ln \rho_i + \ln l_i - \ln C)$$

$$= \alpha_i \Delta \rho_i + \beta_i \Delta l_i - \xi_i \Delta C$$

The α and β are coefficients of linear regression equation, $l_i=\alpha+\beta \cdot C$. From the derivation formula of the topological entropy, it is known that the topological entropy of a single node is approximately equal to sum of the topological entropy of the concurrent connection and the topological entropy of the information-viewed probability, and minus the topological entropy of the maximum concurrency in the node.

3.2 Computation of the Minimal Entropy of "Internet+" Network

For measuring the topology of the "Internet+" effectively, first of all, it's assumed that the probability distribution of information-viewed among nodes remains almost unchanged in a unit of time, entropy derivation formula is showed as follow:

$$\frac{dH_i}{d\xi_i} = \alpha_i \frac{d\Delta\rho_i}{d\xi_i} + \frac{d\alpha_i}{d\xi_i}\Delta\rho_i + \beta_i \frac{d\Delta l_i}{d\xi_i} + \frac{d\beta_i}{d\xi_i}\Delta l_i - \Delta C \quad (6)$$

In the formula 6, $\alpha_i=l_i/C$ and $\beta_i=\rho_i/C$.

At second, it's assumed that the concurrent number of all nodes keeps a stable value when the probability distribution of information-viewed is invariable. At then, the topological entropy of a single node has a minimum value with $d\Delta\rho_i/d\zeta_i=0$ and $d\Delta l_i/d\zeta_i=0$, so the formula is simplified as follow.

$$\frac{d\xi_i}{d\rho_i\, dl_i} = \ln C^{-1}\left(\frac{1}{(\rho_i \ln \rho_i)^{-1} d\rho_i} + \frac{1}{(l_i \ln l_i)^{-1} dl_i}\right) \quad (7)$$

The information aggregation degree(IAD) of the node is calculated by integrating on both ends of the formula 7:

$$\xi_i = \frac{l_i \cdot \rho_i \ln(l_i \cdot \rho_i)}{C \ln C} + \lambda \quad (8)$$

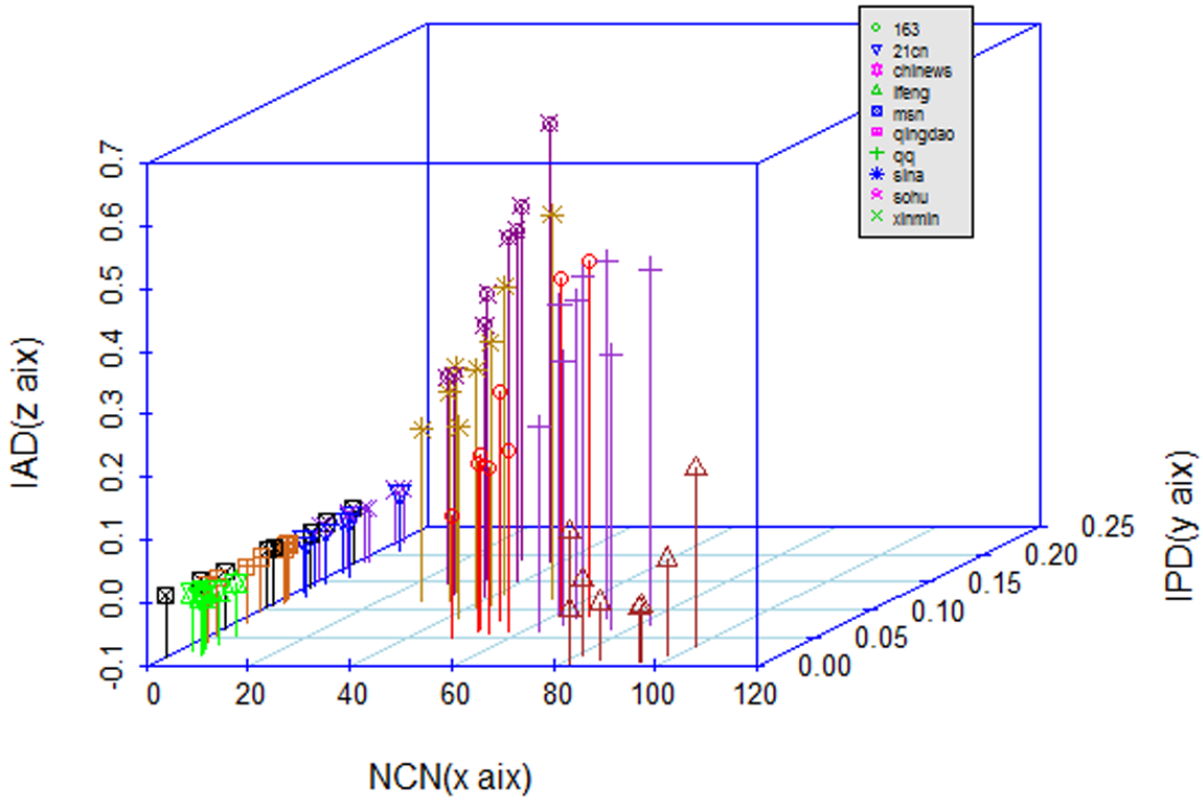
The λ is a integration constant.

4 The Topology Model Test

In general, the clusters of the news site have capable of processing 1 million requests per second at least, so the minimum threshold is taken. The integration constant λ is zero for calculation simplicity. The IAD can be approximately calculated by formula 9.

$$\overline{\xi_i} = \frac{\overline{l_i} \cdot \overline{\rho_i} \ln(\overline{l_i} \cdot \overline{\rho_i})}{C \ln C} \quad (9)$$

Experimental data comes from the daily statistical average number of the ten major news networks, from January to August in the year of 2015. After steps of data cleaning, data sorting and data calculating by using Rstadio tools, the 3-dimension graphical of the IAD is showed as figure 3



(Note: The NCN stands for node concurrent number and the IPD stands for information-viewed probability distribution.)

Figure 3. The AID distribution of news sites

From above the figure, it is concluded that the IAD of the Sohu News is highest(0.388), follow by the Sina News(0.383) and the QQ News(0.296), and the Phoenix News with the highest UV is about 10 times lower than the Sina News and the QQ News. This result shows that although the daily visits of Phoenix News is larger, but the extent of the use of information in the site is not higher than the Sina News and the QQ News. Furthermore, the Xinmin News with the highest IAAD has a negative IAD value(-0.013), it says that every information stored in the site has a higher frequency to be viewed.

In summary, the model can be simple to evaluate the information aggregation degree of all nodes from the information-viewed number and the number of the instant access concurrency.

5 Conclusion

Through the relationship analysis of the News site's PVs and UVs in the "Internet+" structure, a phenomenon has been found that the site with a lower value of PV and UV has the larger IAAD value, and the phenomenon can't be explained effectively by the conventional complex network theories. In order to explain this phenomenon reasonably, the rich-club attribute of the "Internet+" node is analyzed at first, then the information-viewed probability is defined, and the model for calculating the IAD is established finally based on the topological entropy theory.

The experimental result shows that the proposed model can be used to evaluate the information aggregation degree in the "internet+" network, which provides a new evaluation for the analysis of the "Internet+" network structure.

Acknowledgement

This paper was supported by the *Applied Basic Research Key Program of Yunnan Province* (2012CC031), with the project named as The *Cloud Key Technology and its Applicable Research for Scientific and Technical Resources Public Service*.

References

1. Watts D J, Steven H S. Collective dynamics of 'small-world' networks. Nature, 1998,393(6684):440-442.
2. Strogatz, Steven H. Exploring complex networks[J]. Nature, 2001, 410(6825):268-276.
3. Barabási A L, Réka A. Emergence of scaling in random networks. Science, 1999, 286(5439):509-512.
4. Réka A, Barabási A L. Statistical mechanics of complex networks. Reviews of modern physics, 2002, 74(1):47.
5. Boccaletti S, Latora V, Moreno Y, et al. Complex networks: Structure and dynamics. Physics reports, 2006, 424(4):175-308.
6. Barrat A, Barthélemy M, Pastor S R, at el. The architecture of complex weighted networks: Proceedings of the National Academy of Sciences of the United States of America,2004. Boston:pnas, 2014, 101(11):3747-3752.
7. Antal T, Krapivsky P L. Weight-driven growing networks. Physical Review E, 2005, 71(2): 026103.
8. Dorogovtsev S N, Mendes J F. Minimal models of weighted scale-free networks. 2004, arXiv:cond-mat/0408343. http://arxiv.org/abs/cond-mat/0311416.
9. LIANG S, LI M, LEUNG K, et al. An Experimental Study of Response Times of Web Applications. Journal of Computer Research and Development, 2003, 40(7): 1076-1080.
10. Garcia-Robledo A, Diaz-Perez A, Morales-Luna G. Correlation analysis of complex network metrics on the topology of the Internet. Emerging Technologies for a Smarter World (CEWIT), 2013 10th International Conference and Expo on. IEEE, 2013: 1-6.
11. Chai W K, He D, Psaras I, et al. Cache "less for more" in information-centric networks (extended version). Computer Communications, 2013, 36(7): 758-770.
12. Bo J, Ying Z, Jing D, et al. Study on the stability of the topology interactive growth mechanism using graph spectra. Communications, IET, 2014, 8(16): 2845-2857.
13. Holbert B, Tati S, Silvestri S, et al. Network Topology Inference With Partial Information. Network and Service Management, IEEE Transactions on, 2015, 12(3): 406-419.
14. LIU X, ZHAO H, WANG J, et al. Dissipation Analysis of Internet Topology Structure. Journal of Northeastern University(Natural Science),2015, 36(9): 1237-1241.
15. LU X, WANG H, WANG J. Virtual Computing Environment iVCE:Concept and system structure. Science in China(Series E:Information Sciences), 2006, 10:1081-1099.
16. LIANG H. A Similarity Detection System of Network News. JILIN: JILIN University, 2011.
17. Calvert K L, Doar M B, Zegura E W. Modeling internet topology. Communications Magazine, IEEE, 1997, 35(6): 160-163.
18. LIANG S, LI M, LEUNG K, et al. An Experimental Study of Response Times of Web Applications. Journal of Computer Research and Development, 2003, 40(7): 1076-1080.