Homophilies and Communities Detection among a Subset of Blogfa Persian Weblogs: Computer and Internet Category

Adib Rastegarnia[1], Meysam Mohajer[1], and Vahid Solouk[2]

[1] University of Tehran
Dept of Information Technology Engineering
adib.rastegarnia@ieee.org,
meysammohajer@ieee.org
[2] Urmia University of Technology
Dept of Information Technology Engineering
vsolouk@ieee.org

Abstract. The investigation of relationships between various social actors has been the main focus of social network analysis in explaining the structure of social relations, measuring the relationships between the actors and etc. Blogfa is among the popular web service providers for building Persian weblogs in Iran. To the best of our knowledge, there is no social network analysis for the Computer and Internet Category of Blogfa Persian weblogs. The current paper presents a social network analysis for the Computer and Internet category of Blogfa. Each weblog in the target category contains a list of friends to which, it establishes a connection called links or links of friends. These links lead to the formation of a relationship network between weblogs. The current study has particularly focused on the relationship analysis of the network of weblogs based on the friend links. We report on our analyses and measurements of different centrality parameters such as in-degree, out degree, clustering coefficient, modularity for the group of weblogs. Furthermore, the degree of collaborations between these weblogs are analyzed and some homophilies are detected among them. It was found through the analyses that the majority of the bloggers tend to link the weblogs which provide the contents with subjects of common interests among the bloggers, and that the common interests are merely general subjects rather than professional ones.

Keywords: Social network analysis, Blogfa Persian weblogs, Clustering coefficient, Modularity, Relationships Network, Community Detection, Homophily detection.

1 Introduction

The term Social Network (SN) is used for a social structure that provides maps of dyadic ties between individuals and organizations such as friendship, kinship, common interest, financial exchange, etc [1,2]. SNs can operate in different levels

D.-Z. Du and G. Zhang (Eds.): COCOON 2013, LNCS 7936, pp. 811–820, 2013.

from the families up to the nations and thereby, play critical roles in determining the way problems are solved, organizations run, markets evolve and the degree in which individuals succeed in achieving their goals [3]. Hence, the analyses of SN with focus on studying of relationships between various social actors has been used by researchers to explain the structure of social relations, measure the relationships between the actors, etc [4,5]. It is also believed that SN analysis can be used to evaluate the performance of individuals, groups or the entire the social network.

Blogfa is one of the popular service providers for building Persian weblogs in Iran. It contains above 300,000 Persian Weblogs in different categories such as computer, sport, culture, business, entertainment, personal and etc [6].

Several studies on analyzing relationships in SN have been conducted on weblogs as briefly reported in Section 2. To the best of our knowledge, there is no SN analysis for the Computer and Internet Category of the Blogfa Persian weblogs. In this paper, a social network analysis for the Computer and Internet category of Blogfa is presented. Each weblog includes a list of friends to establish connection, called links or links of friends. These links play role in establishing a network of relationships among weblogs. The current study has mainly focused on the network of relationships of weblogs based on their friends' links. We have analyzed different centrality parameters such as in-degree, out degree, clustering coefficient, modularity for the group of the weblogs under investigation. Furthermore, the degree of collaborations between these webglos are analyzed and number of hemophilies are detected among them. The rest of the paper is organized as follows: the following section provides a brief overview of the related works on SN analysis of real world scenarios. Data gathering process is described in section 3. Section 4 presents some of the basic social network analyses based on centrality parameters such as degree distribution, clustering coefficient, connected components, modularity and K-core. The relationships between the number of incoming friends' links and content of the weblogs are explained in Section 5. Section 6 presents an analysis of global relationship network. Visualization and Analysis of the relationships between the Computer and Internet Category Weblogs are presented in Section 7, and Section 8 concludes the work.

2 Related Works

Social network analysis of real world scenarios has been the concern of several studies. In [5] an analysis of top50 political weblogs in people's daily web, based on centrality has been presented. Data mining and centrality analysis have been used to find the network links between the top50 political weblogs. In addition, in order to achieve better understanding of political blogs community, the authors have been tried to find some of the main political blogs groups such as core group members, members with special characteristics, and the important group members. A social network analysis has been presented in [7] for FIT community server (FITCS) which is a popular way for communication between FIT students. Some of the main social network characteristics such as density, closeness, degree and betweenness have been measured for the mentioned social network. In addition, the analysis shows that FITCS can be considered as a small-world scale free network, with several hubs. Furthermore, a large scale study on Persian weblogs has been presented in [8]. Commenting behaviors of Persian bloggers are investigated and a simple model for distribution of comments is introduced by the authors. Social network analysis and data mining methods was used in [9] to investigate the network relationships in on-line forum of university. To solve some of the most important problems such as improving the communication in the university on-line forum, make the students more positive and learn more knowledge, some advices are proposed. In addition, in [10] structural features of the Sina's VIP Blogsphere and the behavior patterns of its members have been investigated by using social network analysis. The authors concluded the behavior patterns of Sinas VIP Blogsphere is consistent with real life behaviors of bloggers.

3 Data Collection Process

In this paper, two kinds of relationships between the weblogs are defined which are described as the follows:

- Local Relationships Network (LRN): denotes the existing links of friends among the Computer and Internet category of Blogfa Persian weblogs.
- Global Relationships Network (GRN): denotes the existing links among the Computer and Internet category and the other categories of Blogfa Persian weblogs such as Entertainment, sports, business, religious, etc.

We have used Win Web Crawler [11] in order to collect the list of weblogs and to extract their in-between relationships. In this study, a number of 16429 webblogs from the Computer and Internet category are crawled and a list of 64000 webblogs from other categories are extracted and categorized.

4 Basic Social Network Analysis

In this section, some basic analysis on the basis of degree, modularity, clustering coefficient and connected components parameters are examined on the LRN of weblogs. The Gephi software is used to visualize and analyze the relationships network of weblogs. Gephi platform is an open-source interactive visualization and exploration tool for all kinds of networks and complex systems [12].

4.1 Degree Analysis

The degree of a node (a node refer to the each of the weblogs) denotes the number of links to a node. In the directional graph like the case in the current study, in-degree and out-degree has been applied to the number of links pointing in and out of a node, respectively [10]. In-degree and out-degree distributions of

LRN of weblogs are illustrated in Figure 1. As shown in Figure 1, only few of nodes have numerous incoming links when compared with the rest with only few incoming links. The nodes with many incoming links act as hubs or connectors in the LRN. On the basis of calculation, the value of in-degrees are between 0 and 3 for more than 95% of the nodes. In addition, only 0.004% of the nodes are linked with more than 10 incoming links. Because of applicability of the 80-20 rule or Pareto law in our case, the LRN of the Computer and Internet category is a scale-free network, which is referred to the network with the degree distribution following a power law. In addition, the out-degree distribution of the local relationships is illustrated in Figure 1, pinpointing that only a small number of the weblogs in the Computer and Internet category are linked to the weblogs of the same group. The calculated results showed that 98% of weblogs in the Computer and Internet category link to the 0 to 3 blogs of the same group. Furthermore, only 0.0029% of the weblogs in this category have been linked to more than 10 weblogs of the same group. Hence, the results evidence weak relationships between the Computer and Internet category weblogs.

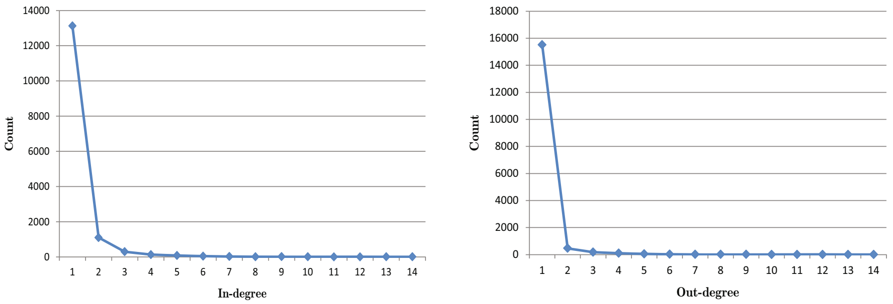


Fig. 1. In-degree and Out-degree Distributions of Computer and Internet Category weblogs

4.2 Clustering Coefficient

The term Clustering Coefficient (CC) is generally referred to as the probability that two randomly selected friends of the set A (in our case a weblog) are friends with each other [2]. In other words, it can be measured by dividing the number of actual links between one's friends by the number of possible links in full friendship case. Full friendship is the case in which everyone is friend with one another. The value of clustering coefficient is assumed to be between 0 and 1. It is believed that the friends of a person (in our case a weblog) are considered good friends with each other once the CC is close to 1. In other words, if the CC is close to 1, strong relationships between the weblogs can be expected. The calculated average CC for the LRN is as small as 0.006, which is the evidence of no strong relationships between these weblogs.

4.3 Connected Components

On the basis of the algorithm proposed in [13], the number of weakly and strongly connected components has been calculated in the LRN as 8215 and 15861 components, respectively. Most of the detected components include less than 1% of the nodes and only one giant component existed, that include 43% of the nodes. This giant component is illustrated in Figure 2 using red color. The most remarkable result is the existence of a connected component as shown in green color in Figure 2. This connected component contains 20 weblogs. The principle of homophily indicates that contact between similar individuals occurs more often than among dissimilar individuals [14]. After investigating the content of these weblogs, a content based homophily is detected in the discussed component. That is, the contents provided by these weblogs are related to the mobile technology and general computer learning issues. It can be concluded that the existence of strong relationships among the weblogs of this community originated due to the existence of this content based homophily.

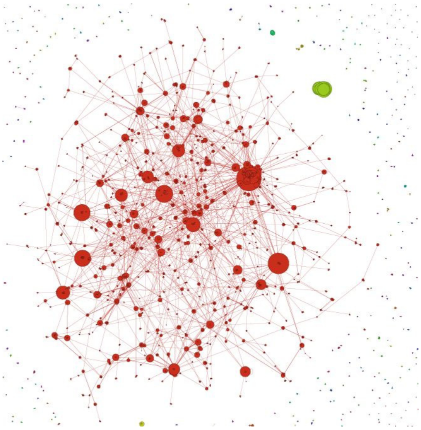


Fig. 2. Connected Components in Local relationships

4.4 Community Detection by Using Modularity

In order to measure the strength of a network separation into clusters or communities, modularity parameter is proposed in [15]. The number of communities calculated by the modularity is 13083, with the largest one including only 1.55% of the nodes as illustrated in Figure 3 using violet color and the second one including only 1.42% the nodes. Less than 1% of the nodes existed in each of the rest of the communities. It can be inferred from the modularity results that the communications between the Computer and Internet category weblogs are limited to a small number of weblogs which proves our observations reported in previous sections.

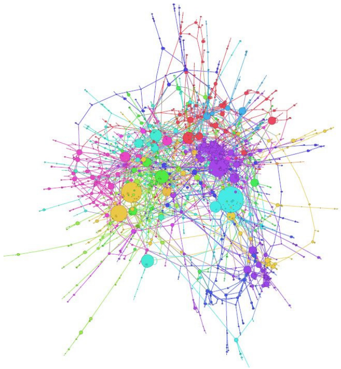


Fig. 3. Detected Communities in Local Relationship Network network by using Modularity Parameter

4.5 Community Detection by Using K-Core Method

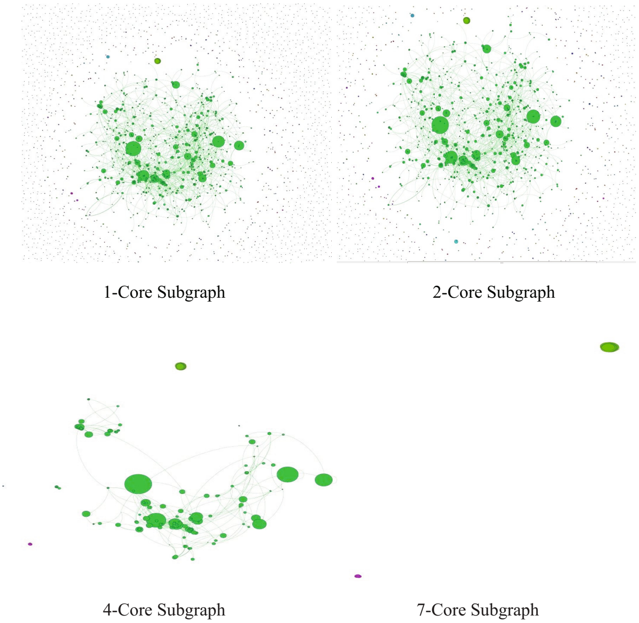
K-core refers to a maximal connected subgraph in which each node is adjacent to at least a minimum number, K, of the other nodes in the subgraph. In social network analysis, K-core method can be used for community detection. LRN of The Computer and Internet category is drawn based on different values of K as shown in Figure 4. By increasing the K value to 4, most of the nodes are discarded from the subgraph and only a small subset of the blogs were found to have relationships with more than 4 nodes. As illustrated in Figure 4, by increasing the value of K to 7, only 2 communities were remained in K-core subgraph, with one community, equal to the 1 as described in Section 4.3. This community were disappeared by increasing the K value to 22.

5 Number of Incoming Links of the Weblogs and Their Contents

Previous studies have shown that the web pages with the highest in-degree are those providing contents in a vast range of subjects. We investigated this phenomena in blogs of the Computer and Internet category. For instance, the weblog with the address http://www.biya2it.blogfa.com has the highest incoming links from the other weblogs. After investigating the contents of the weblog, 37 categories of subjects such as mobile games, web design, movie, health, software, learning computer, etc are detected.

6 Analysis of Global Relationship Network (GRN)

In order to analyze the relationships between the Computer and Internet category and the other categories, number of 64000 Blogfa weblogs are crawled and



1-Core Subgraph

2-Core Subgraph

4-Core Subgraph

7-Core Subgraph

Fig. 4. 1-2-4-7 Core Subgraphs

categorized in 15 categories and 43 subcategories. Then, the number of friends links of the Computer and Internet Category are extracted. Due to the difference in number of weblogs in each category and in order to make fair analysis, the ratio of the Number of Weblogs (NB) to the Number of Friends' links (NF) are investigated. As listed in Table 1, an approximate rate of 88% of the weblogs in the Weblog Themes category together with 18% of the weblogs in the Writing Weblog Issues category are linked by the Computer and Internet category weblogs. Hence, it can be interpreted that there is a common interest between the bloggers to link to the weblogs that provide the contents about installation, maintenance, and configuration of the weblogs. This observation is originated from the fact that the bloggers tends to select a beautiful theme for their weblogs, customize their weblogs and install some applications on them, etc when they create their weblogs for the first time. Consequently, with a high probability, the bloggers visit the weblogs that provide the contents about the mentioned issues and make link to them. In other words, some general subjects are common interests among the most of the bloggers and cause making link to the weblogs that provide contents related to common subjects.

7 Visualization and Analysis of the Relationship

The LRN and GRN networks of relationships are illustrated in Figure 5. The number of connected components is equal to 5609 as determined by the Gephi

Table 1. The ratio of the Number of Weblogs (NB) to the Number of Friends' links (NF)

Category	Sub-Category	NB	NF	NB/NF
News and Media	News	796	36	0.045
	Newspapers and Medias	796	41	0.051
	News writers	796	55	0.069
Ideology	philosophy	1587	44	0.027
	Islam	1899	108	0.056
	Quran	1309	36	0.027
	Christianity	215	3	0.013
	Zarathustra	225	10	0.044
	Jewish	96	5	0.052
Science and Technology	Health	1770	68	0.038
	Medicine	1669	40	0.023
	Nature and Environment	1277	38	0.029
	Foreign Languages	1081	75	0.069
	Agriculture and Biotechnology	1381	36	0.026
	Electrical and Electronic	1880	127	0.067
	Architecture and Civil Engineering	1859	55	0.029
	Stars	960	72	0.075
	librarian-ship	392	24	0.061
	Basic Science	1860	88	0.047
	Nurses	325	7	0.021
Culture and History	Culture and History	1880	90	0.047
Persian Speakers in other Countries	Persian Speakers in other countries	1052	42	0.039
Fotoblog	Fotoblog	1900	163	0.085
Art and Literature	Literature and Poem	1860	95	0.051
	Book	896	88	0.098
	theater and Movie	1940	97	0.05
	Music	1902	87	0.046
	Imagining	639	45	0.070
	Artists	1902	77	0.040
Society and Politics	Politics	1960	39	0.019
	Women	400	17	0.012
	Society	1327	30	0.022
Family and Life	Family and Life	1820	81	0.044
Tourism and Travel	Tourism and Travel	1296	48	0.037
Personal	Personal	1835	83	0.045
Entertainment	Entertainment	1879	192	0.10
Business and Economic	Articles	1371	43	0.031
	Companies and Organizations	1880	34	0.018
	Electronics Commerce	1140	32	0.028
	Earn Money by Internet	1231	41	0.033
Sports	Sports	1880	103	0.054
Weblog Utilites	Politics	1960	39	0.019
	Weblog Themes	377	332	0.88
	Writing Weblogs Issues	237	45	0.18

software. It can be seen that there is a strong connected component at the center of the graph and all nodes in this component have strong relationships with each other. This giant component includes 75% of the nodes. The relationships in the other components are weak. Moreover, the second giant component as shown in Figure 5 using brown color includes only 8% of the nodes.

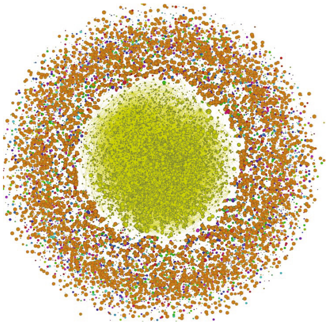


Fig. 5. The LRN and GRN relationships network

8 Conclusion

The current paper, presented an analysis of relationships network of weblogs based on their friends' links for the Computer and Internet category of Blogfa Persian weblogs. Some basic analysis based on in-degree, out degree, clustering coefficient, and modularity centrality parameters has been performed and reported. The results from the basis analysis evidenced weak relationships between the weblogs of this group. In addition, the relationship between the number of incoming links of the weblogs and their contents were investigated. The results proved that the weblogs with the highest in-degree are those providing contents in a vast range of subjects. Finally, an analysis on the relationships of the Computer and Internet category of weblogs with other categories has been carried out. This analysis shown that the most of the bloggers tend to link the weblogs that provide the contents about the subjects of common interests among the most of the blogger.

References

1. Ding, L., Shi, P.: Social network analysis application in bulletin board systems. In: 2011 International Conference on Intelligence Science and Information Engineering (ISIE), pp. 317–320 (August 2011)
2. Easley, D., Kleinberg, J.: Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press (2010)
3. Abbasi, A., Altmann, J., Hossain, L.: Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. J. Informetrics 5(4), 594–607 (2011)

4. Wasserman, S., Faust, K., Iacobucci, D.: Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences). Cambridge University Press (November 1994)
5. Ya-ting, L., Jing-min, C.: The social network analysis of political blogs in people: Based on centrality. In: 2011 International Conference on Consumer Electronics, Communications and Networks (CECNet), pp. 5441–5444 (April 2011)
6. Blogfa: Blogfa weblog service provider (2012), http://www.blogfa.com
7. Hamulic, I., Bijedic, N.: Social network analysis in virtual learning community at faculty of information technologies (fit), mostar. Procedia - Social and Behavioral Sciences 1(1), 2269–2273 (2009)
8. Qazvinian, V., Rassoulian, A., Shafiei, M., Adibi, J.: A large-scale study on persian weblogs. In: The Proceedings of 12th International Joint Conference on Artificial Intelligence, Workshop of TextLink 2007 (2007)
9. Huiqing, N.: Social network analysis of university online forum. In: 2010 International Conference on Computational Aspects of Social Networks (CASoN), pp. 422–429 (2010)
10. Wen-jun, S., Hang-ming, Q.: A social network analysis on blogospheres. In: 15th Annual Conference Proceedings of the International Conference on Management Science and Engineering, ICMSE 2008, pp. 1769–1773 (September 2008)
11. winwebcrawler: Win web crawler (2012), http://www.winwebcrawler.com/
12. Gephi: An open-source graph visualization and maniuplation software (2012), http://www.gephi.org
13. Tarjan, R.E.: Depth-first search and linear graph algorithms. SIAM J. Comput. 1(2), 146–160 (1972)
14. McPherson, M., Lovin, L., Cook, J.: Birds of a feather: Homophily in social networks. Annual Review of Sociology 27(1), 415–444 (2001)
15. Newman, M.E.J.: Modularity and community structure in networks. Proceedings of the National Academy of Sciences 103(23), 8577–8582 (2006)