FINDING IMPORTANT NODES IN SOCIAL NETWORKS BASED ON MODIFIED PAGERANK

Li-qing Qiu[1], Yong-quan Liang[2], Jing-Chen[3]

[1, 2]College of Information Science and Technology, Shandong University
of Science and Technology, Qingdao, China
[3]Shandong labor vocational and technical college Jinan, China
liqingqiu2005@126.com[1], lyq@sdust.edu.cn[2]
wfchenj@126.com[3]

ABSTRACT

Important nodes are individuals who have huge influence on social network. Finding important nodes in social networks is of great significance for research on the structure of the social networks. Based on the core idea of Pagerank, a new ranking method is proposed by considering the link similarity between the nodes. The key concept of the method is the use of the link vector which records the contact times between nodes. Then the link similarity is computed based on the vectors through the similarity function. The proposed method incorporates the link similarity into original Pagerank. The experiment results show that the proposed method can get better performance.

KEYWORDS

social networks, Pagerank, link similarity

1. INTRODUCTION

In modern society, social networks play an important role in a quickly changing world, and more and more people prefer to obtain information from social networks. This explains the increasing interest in social networks analysis which examines topology of a network in order to find interesting structure within it. Recent works have pointed that some very active nodes have huge impacts on other nodes [1]. Therefore, the problem of how to appropriately find important nodes in social networks among the huge nodes becomes an important issue.

Pagerank[2,3] is a well known method for identifying authoritative pages in a hyperlink network of web pages. Pagerank relies on the democratic nature of the web by using its topology as an indicator of the value to be attached to any page. Pagerank is so useful that we can also apply it to social networks in that the mutual relationship in social networks can be structured as links to the microblogs, and nodes can also be regarded as websites in Pagerank algorithm [4]. In the paper, we present a new importance ranking method based on the modified Pagerank to find the important nodes in social networks. In other words, the proposed method is derives from

David C. Wyld et al. (Eds) : CST, ITCS, JSE, SIP, ARIA, DMS - 2014
pp. 39–44, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4104

Pagerank by considering the link similarity which measures the similarity between the nodes.
The paper is organized as follows. In section 2 we describe our generalization of new measurement, and the experimental results together with the experimental settings are given in section 3. At last we conclude the paper by summarizing our findings in section 4.

2. PROPOSED METHOD

In the section, we propose a new method to find important nodes based on modified Pagerank, by considering the link similarity. We introduce the modified Pagerank model in section 2.1, and then we analyze the importance of nodes using the model in section 2.2.

2.1 Modified Pagerank

The core idea of Pagerank is that of introducing a notion of page authority. In pagerank, the authority reminds the notion of citation in the scientific literature. In particular, the authority of a page p depends on the number of incoming hyperlinks and on the authority of the page q which cites p with a forward link. Moreover, selective citations from q to p are assumed to provide more contribution to the value of p than uniform citations. Therefore, the Pagerank value PR_p of p is computed by taking into account the set of pages $pa[p]$ pointing to p .The Pagerank value PR_p is defined as follows:

$$PR_p = (1-d) + d \sum_{q \in pa[p]} \frac{PR_q}{h_q} \qquad (1)$$

Here $d \in (0,1)$ is a dumping factor which corresponds to the probability with which a surfer jumps to a page picked uniformly at random.

The quality of the links, as measured by Pagerank, is a good choice for ranking nodes but we think there are some other features that can incorporate the activity of the node. We propose to incorporate the features via the link similarity taking into account contact times of the node. The idea behind is that the node has higher similarity must be prized with a higher value. Our main idea consists in constructing the link vector that records the contact times of the nodes, defining a link similarity function to measure the similarity of the nodes according to the link vector, and then reconstructing Pagerank model by considering the link similarity. These ideas are a work in progress.

Definition 1. For node v_i, its link vector is defined as:

$$V_i = \{t_{i1}, t_{i1}, \cdots, t_{i-1}, t_{i+1}, \cdots, t_{in}\} \qquad (2)$$

Where $t_{im} (0 \le m \le n)$ is the contact times of v_i and v_m.

Definition 2. For node v_i and v_j, the link similarity is defined as:

$$Similarity(V_i,V_j)=\overrightarrow{V_i}\cdot\overrightarrow{V_j}=\frac{\sum_{m=1}^{n}V_{im}\cdot V_{jm}}{\sqrt{\sum_{m=1}^{n}V_{im}^{2}}\sqrt{\sum_{m=1}^{n}V_{jm}^{2}}} \qquad (3)$$

Obviously, the link similarity measures are functions that take two link vectors as arguments and compute a real value in the interval $[0..1]$, the value 1 means that the two nodes are closely related while the value 0 means the nodes are quite different.

Definition 3. For node v_i, its modified Pagerank value is defined :

$$PR_{v_i}=(1-d)+d\sum_{v_j\in pa[v_i]}\frac{PR_{v_j}\cdot Similarity(V_i,V_j)}{h_{v_j}} \qquad (4)$$

Here $pa[v_i]$ is the set of nodes which point to v_i, and the dump factor d is set to 0.15 in our experiments.

In the Modified Pagerank, the link similarity of the nodes is taken into account, which is a positive indicator to modify the original Pagerank(see Equation 1).

2.2 Node Analysis based on modified Pagerank

The node analysis based on modified Pagerank is composed of 3 steps as follows:

Step1: read network as graph;
Step2: computer modified Pagerank value according to Equation 4;
Step3: obtain node modified Pagerank value list.

From above steps, we can see that a bit change is made to the original Pagerank by adding the link similarity indicator. The modified Pagerank assumes that the initial Pagerank values of all nodes are the same. Firstly, we calculate the first iterative ranking of each node based on the initial Pagerank values, and then calculate the second rank according to the first iteration. The process continues until the termination condition is satisfied. At last the Pagerank estimator converges to its practical value, which has proven no matter what the initial value is. The whole processing is implemented without manual intervention.

3. EXPERIMENTS

In the following, we use several different networks to study the performance of our generalization of novel algorithm for detecting local community structure.

3.1 First Experiment

We start with the first synthetic dataset, which is shown as Figure 1, to illustrate the process of the proposed method in detail. The network contains 5 nodes, and the contact times are associated with corresponding edges.

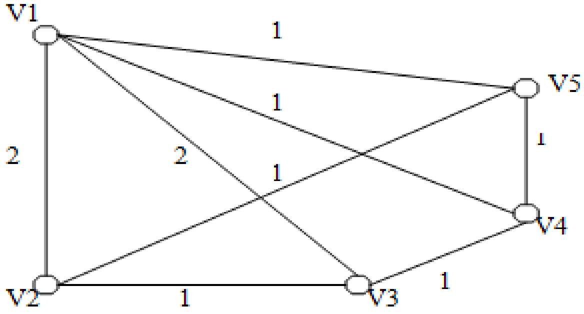

Figure1. A simple network of 5 nodes and 8 edges.

So the mutual relationship matrix can be produced as briefly shown in Table 1.

Table1. Mutual relationship matrix, A.

	V1	V2	V3	V4	V5
V1		2	2	1	1
V2	2		1	0	1
V3	2	1		1	0
V4	1	0	1		1
V5	1	1	0	1	

Parameters in the matrix a_{ij} is defined as the contact times between the nodes. Then the link vectors are: $V1 = \{2,2,1,1\}$, $V2 = \{2,1,0,1\}$, $V3 = \{2,1,1,0\}$, $V4 = \{1,0,1,1\}$, $V5 = \{1,1,0,1\}$ .Thereby, the link similarity can be computed as:

$$Similarity(V1,V2) = \frac{2*2+2*1+1*0+1*1}{\sqrt{2^2+2^2+1^2+1^2} * \sqrt{2^2+1^2+0^2+1^2}} = 0.904$$

The process continues until we get the total link similarity of the network. The link similarity matrix is shown as follows:

Table2. The link similarity matrix,B.

	V1	V2	V3	V4	V5
V1		0.904	0.904	0.730	0.913
V2	0.904		0.833	0.707	0.904
V3	0.904	0.833		0.707	0.707
V4	0.730	0.707	0.707		0.816
V5	0.913	0.904	0.707	0.816	

Then the modified Pagerank can be computed according to Equation 4, and we get the following Pagerank values:

Table3. The modified Pagerank Values.

	Value
V1	0.116
V2	0.099
V3	0.101
V4	0.093
V5	0.109

3.2 Second Experiment

Secondly, we apply our modified Pagerank method to one small network, which is the much-discussed "karate club" network of friendships between 34 members of a karate club at a US university, assembled by Zachary [5] by direct observation of the club's members. This network is of particular interest because the club split in two during the course of Zachary's observations as a result of an internal dispute between the director and the coach. In other words the network can be classified into two communities-one's center is the director, the other's center is the coach.

We select degree centrality and PageRank algorithm as baseline methods, which be identified as "degree" and "PageRank" respectively. And our proposed modified Pagerank method is identified as "M-Pagerank".Table4 shows the result of the experiment. We select the nodes in the top 10 according to different methods.

Table4. The comparison of different methods

degree	Node ID	PageRank	Node ID	M-Pagerank	Node ID
0.515	34	0.101	34	0.093	34
0.485	1	0.097	1	0.089	1
0.364	33	0.072	33	0.065	33
0.303	3	0.057	3	0.050	3
0.273	2	0.053	2	0.036	2
0.182	32	0.037	32	0.028	32
0.182	4	0.036	4	0.017	14
0.152	24	0.032	24	0.016	4
0.152	9	0.030	9	0.014	9
0.152	24	0.030	14	0.014	31

From the above table, we can see that three methods appear to have many common results. However, degree can not distinguish nodes because some nodes have the equal value. For example, node 32 and node 4 have the common value according to degree method, which shows that degree method can not distinguish nodes well. M-Pagerank performs better than Pagerank in that M-Pagerank considers the link similarity of the nodes. For example, node 31 is on the boundary of two communities, which has much connection between the two communities.

According to M-Pagerank, node 31 is important than node 24 obviously. However, Pagerank rank node 24 higher than node 31, which is not reasonable.

4. CONCLUSION

In the paper, we have shown a new method to find important nodes in social works based on modified Pagerank. The method is capable of incorporate the link similarity of the nodes via the link vector. The final goal for this model is to incorporate some features into original Pagerank. The proposed method enables a new way of ranking the nodes in social works. We have analyzed the experiment results using test networks. In our future work, we will further discuss how to achieve better performance to detect importance nodes.

ACKNOWLEDGEMENT

This paper is supported by National Science Foundation for Post-doctoral Scientists of China under grant 2013M541938, and Shandong Province Postdoctoral special funds for innovative projects of China under grant 201302036.

REFERENCES

[1] Khorasgani RR, Chen J, Zaiane OR.Top leaders community detection approach in information networks. KDD 2010, 1-9.

[2] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. WWW7, 1998.

[3] Bianchini M, Gori M, Scarselli F.Inside Pagerank. ACM transactions on internet technology, 2006,92-128.

[4] Abbassi Z, Minokni VS. A recommender system based on local random walks and spectral methods. 9th Web KDD and 1st SNA-KDD 2007 workshop on web mining and social network anaysis,2007,102-108

[5] Zachary WW. An information flow model for conflict and fission in small groups, Journal of anthropological Research, 1977, 33: 452-473.