

· 复杂性科学 ·

基于LeaderRank和节点相似度的复杂网络重要节点排序算法

顾亦然, 朱梓嫣

(南京邮电大学自动化学院 南京 210023)

【摘要】复杂网络中重要节点对网络结构和功能的影响引起了广泛关注。本文在现有LeaderRank算法的基础上, 利用节点相似度来衡量节点间的相互作用, 建立了SRank算法进行重要节点排序。利用SIR传播模型和斯皮尔曼等级相关系数在真实社会网络数据上对本文算法与经典的重要节点排序算法进行仿真后, 发现该算法在无向和有向网络中均具有更高的准确性。

关键词 复杂网络; 重要节点; 相似度; SRank算法

中图分类号 TP311, N94 **文献标志码** A **doi**:10.3969/j.issn.1001-0548.2017.02.020

Node Ranking in Complex Networks Based on LeaderRank and Modes Similarity

GU Yi-ran and ZHU Zi-yan

(College of Automation, Nanjing University of Posts and Telecommunications Nanjing 210023)

Abstract The effect of important nodes in complex networks on the structure and function of the networks causes widespread concern. This paper presents a SRank algorithm based on LeaderRank and nodes similarity which is used to measure the interaction between nodes. The simulation of SIR model and Spearman's correlation coefficient on real social networks show that the SRank algorithm preforms better on identifying influential nodes both in directed and undirected networks, compared with the other four classical algorithms.

Key words complex networks; important nodes; nodes similarity; SRank algorithm

随着科学技术的发展, 世界变得越来越小, 也越来越复杂, 涌现了大量难以用经典概念解释的问题, 如微博中谣言通过少量节点快速传播到整个网络, 传染病通过交通网络中的少量节点快速传播扩散^[1]等。借助网络科学的发展与进步, 学者们对这些问题进行了量化的描述和研究, 将这些可以对整个网络的结构和功能产生巨大作用的少量节点, 称之为重要节点。挖掘这些重要节点并研究其内在动力学机理成为当前的研究热点之一。

近年来, 学者在重要节点排序上提出了很多指标和算法, 主要有以下4个方面^[2]: 1) 基于节点邻居节点的排序方法, 如度中心性(degree centrality)^[3], 即节点的邻居节点越多其影响力越大, 其缺点是仅仅考虑了节点的局部信息。2) 基于路径的排序方法, 如接近中心性(closeness centrality)^[4], 通过计算节点与网络中其他所有节点距离平均值来衡量节点的重要性, 缺点是时间复杂度比较高。3) 基于特征向量的排序方法, 如PageRank算法^[5], 该算法认为每一个节点的重要性取决于指向它的其他节点的数

量和质量, 缺点是容易陷入悬挂节点。LeaderRank^[6-7]算法在PageRank算法的基础上, 通过加入一个背景节点, 该节点与网络的所有节点双向连接, 来替代PageRank算法中跳转概率 s , 从而提高了算法的收敛速度和鲁棒性。4) 基于节点的移除和收缩的排序方法^[8-9], 通过移除节点对网络的破坏性来衡量节点的重要性, 但计算复杂度高。

这些算法大多考虑节点在全局中的重要性或者局部重要性, 忽略了节点间相互作用对整个网络的影响。本文考虑在现实生活中, 人们更信任那些和自己关系更为相似的人, 即节点之间的相似度越高则节点间的影响力越大。共同邻居节点相似度^[10]表明, 如果两个节点的共同邻居越多, 这两个节点的相似度越高。因此, 本文引入节点相似度指标作为节点局部重要性的衡量指标, 在LeaderRank的基础上综合考虑节点的局部重要性和全局重要性, 提出SRank算法。将该算法和其他几种经典算法进行比较, 仿真结果表明, 该算法在有向网络和无向网络的重要节点排序结果均优于其他算法。

收稿日期: 2015-12-09; 修回日期: 2016-05-03

基金项目: 教育部人文社会科学研究规划基金(15YJZH016)

作者简介: 顾亦然(1972-), 女, 博士, 教授, 主要从事复杂网络理论与应用、嵌入式系统及通信网络等方面的研究。

1 重要节点排序方法

1.1 LeaderRank算法

LeaderRank通过在原网络中加入一个与网络中其他节点双向连接的背景节点,增强网络的连通性,从而提高了算法的收敛速度。

定义 1 LR_i 定义为节点 v_i 的全局重要度。

$LR_i(k)$ 为经过 k 步后系统到达稳态,节点 v_i 的初步重要度。稳态后,背景节点将其影响力平均分配给网络中的其他节点,从而获得网络中节点的全局重要度。因此节点 v_i 的最终的全局重要度 LR_i 值为:

$$\begin{cases} LR_i(0) = 1/n \\ LR_i(k) = \sum_{j=1}^{n+1} \bar{a}_{ji} LR_j(k-1) \quad k=1,2,\dots \\ LR_i = LR_i(k) + LR_g(k)/n \end{cases} \quad (1)$$

式中, n 为网络节点总数(不包括背景节点); $LR_g(k)$ 为第 k 步背景节点 v_g 的重要度; \bar{a}_{ji} 为基本Google矩阵中的元素,如下:

$$\bar{a}_{ij} = \begin{cases} 1/k_i^{out} & \text{节点 } i \text{ 指向节点 } j \\ 0 & \text{否则} \end{cases} \quad (2)$$

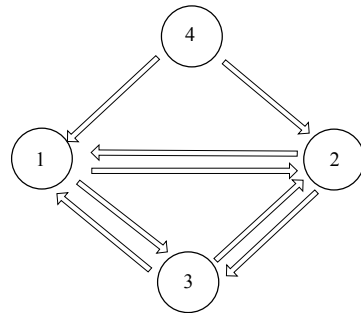
式中, k_i^{out} 为节点 v_i 的出度。

1.2 SRank算法

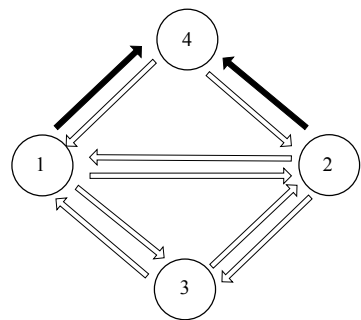
PageRank算法和LeaderRank算法中,每个节点跳转到邻居节点的概率是相等的,但是现实生活中,人们更信任那些和自己关系更为相似的人。在有向社会网络中,节点的出边代表该节点的兴趣所向,入边代表了其他节点对其兴趣的一种认可。可以认为,若两个节点共同指向的节点数越多,其兴趣越相似,相似度值越高;指向两个节点的其他节点数越多,代表其他人认为这两个节点的兴趣越相似,其相似度值越高。因此,两个节点的相似度,不仅和其共同指向的节点数有关,也和共同指向其的节点数有关。

对于一个简单有向网络如图1a所示,考虑节点4不存在的情况,节点1、2、3的影响力是相同的,可得两两节点之间的相似度一致。节点4加入后,直接提高了节点1、2的影响力,对节点3的贡献低,间接削弱了节点3的影响力,即节点对1、2相似度要高于节点对1、3和节点对2、3,符合拓扑结构。因此,节点间相似度和共同指向其的节点数有关,即两个节点受其他节点共同指向的个数越多,其相似度越高。对与图1b来说,由于节点1、2同时指向节点4,提高了节点4的影响力,节点4的影响力提高,也会

提高节点1、2的影响力,因而节点1、2的相似度会高于图1a中的相似度。因此,节点间相似度和其共同指向的节点数有关。两个节点共同指向的节点数越多,其相似度越高。



a. 节点4的单向影响



b. 节点4的双向影响

图1 简单网络

定义 2 节点相似度矩阵 Sim 。节点 v_i 和其邻居节点 v_j 的相似度 Sim_{ij} 为:

$$Sim_{ij} = \gamma k_{i,j}^{out} + (1-\gamma) k_{i,j}^{in} \quad j \in \alpha_i \quad (3)$$

式中, α_i 为节点 v_j 的邻居节点; $0 < \gamma < 1$; $k_{i,j}^{out}$ 表示节点 v_i 和节点 v_j 共同指向的节点个数; $k_{i,j}^{in}$ 表示共同指向节点 v_i 和节点 v_j 的节点个数。

定义 3 SRank矩阵 SR 。节点 v_i 的 SR_i 值如下:

$$SR_i = LR_i \sum_{j \in \alpha_i} \frac{Sim_{ij} + 1}{k_i^{out} + 1} LR_j \quad (4)$$

式(4)表明: 1) 与节点连接的邻居节点个数越多,节点的重要度越高,这与直观判断相符。2) 节点的邻居节点重要度越高,节点的重要性越高。3) 节点对之间的相互影响力不同,邻居节点数多的一方对另一方的影响力更大。符合现实网络中,当两个人的共同好友数一致时,好友数少的一方对认为另一方和其的关系会比好友数多的一方认为对方和其的关系更为亲密。4) 节点对的相似度为0时,节点之间仍然有相互作用。

SRank算法步骤:

1) 使用LeaderRank算法,即使用式(1)得到每个

节点的LR 值; 2) 使用式(3)计算节点对之间的相似
度; 3) 使用式(4)得到节点的SR 值。

2 算法仿真与分析

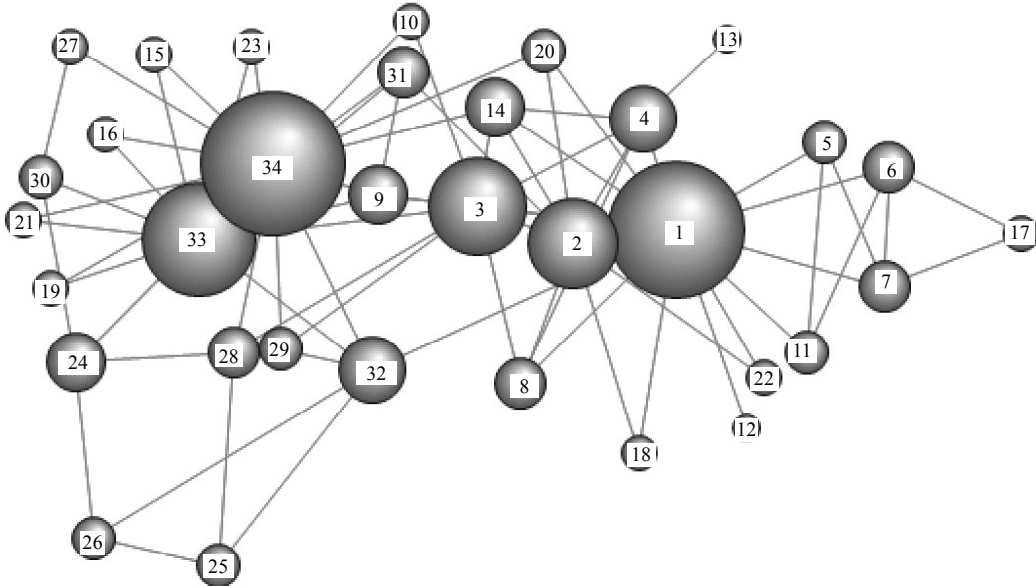
2.1 仿真数据

本文选取了4种真实网络, Zachary空手道网络、
海豚社会网络(Dolphins)、线性虫神经网络(Neural)

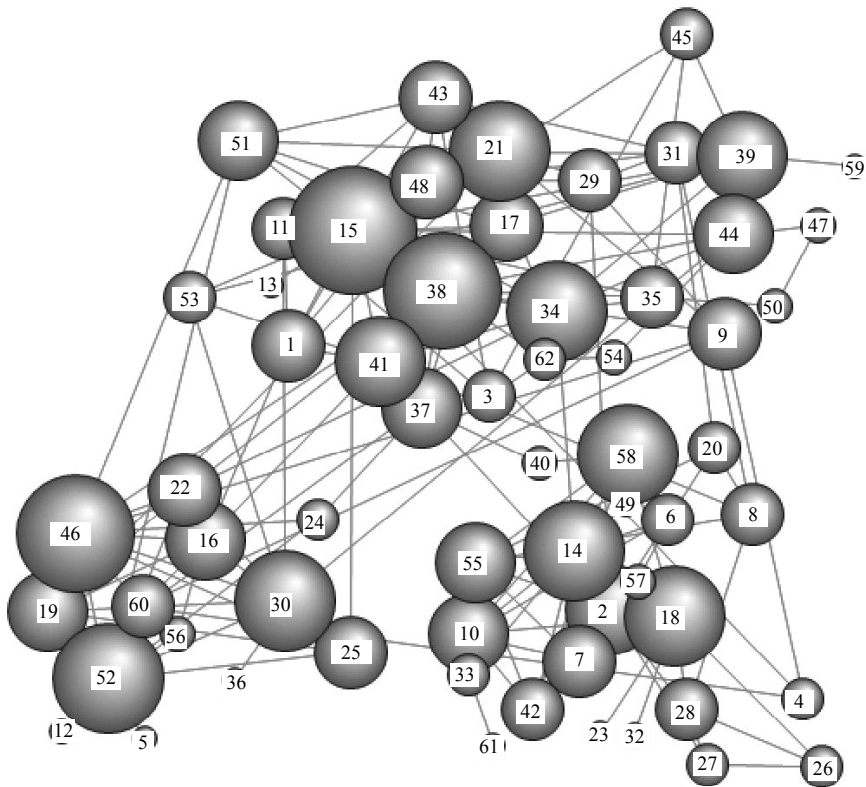
及美国政治家博客网络(Polblogs), 其网络参数如表1
所示。本文节点相似度计算时选取参数 $\gamma = 0.5$ 。

2.2 排序结果与分析

Zachary空手道网络和海豚网络如图2所示, 分
别用度中心性、接近中心性、PageRank算法、
LeaderRank算法、SRank算法进行节点重要性排序,
所得排序结果如表2、表3所示。



a. Zachary空手道网络



b. Dolphins海豚网络

图2 两个无权无向网络

表1 4种网络的参数

网络	节点数	边数	有向	平均度	平均路径	直径	聚类系数
Zachary	34	78	无向	4.59	2.41	5	0.570 6
Dolphins	62	159	无向	5	3.37	8	0.265 2
Neural	239	1912	有向	8	3.99	5	0.196 5
Polblogs	793	15 783	有向	19.90	3.19	8	0.288 1

表2 Zachary网络节点重要性排序

排名	度中心性		接近中心性		PageRank		LeaderRank		SRank	
	节点	DC	节点	CC	节点	PR	节点	LR	节点	SR
1	34	17	34	23.25	34	0.098 3	34	0.080 4	1	0.018 7
2	1	16	1	23.17	1	0.094 6	1	0.075 9	34	0.017 8
3	33	12	3	21	33	0.070 1	33	0.058 0	33	0.010 6
4	3	10	33	20.92	3	0.055 1	3	0.049 1	3	0.007 65
5	2	9	32	19.33	2	0.051 5	2	0.044 6	2	0.007 23
6	32	6	2	19.17	32	0.036 7	32	0.031 3	4	0.004 15
7	4	6	14	18.5	4	0.035 2	4	0.031 3	14	0.002 373
8	24	5	9	18.5	24	0.031 4	14	0.026 8	9	0.002 09
9	14	5	4	17.67	7	0.029 7	9	0.026 8	8	0.001 91
10	9	5	20	17.5	6	0.029 7	24	0.026 8	32	0.001 91

由表2可知,相较于其他算法排名,SRank算法认为节点1比节点34更为重要。节点1和节点34的共同邻居节点为节点9、14、20、32,其重要性不同是因为两节点特有的邻居节点。对于节点1而言,其邻居节点中节点2、3、4的重要性在其他算法中排名靠前,而节点34的邻居节点中只有节点33的重要性在其他算法中比较高,因此,本文算法认为节点34重要性低于节点1更为合理。本文算法认为节点8的重要性比节点32高,因为节点8的邻居节点为节点1、2、3、4,这些节点重要性在其他算法中也比较高,而节点32的邻居节点为节点1、25、26、29、33、34,虽然节点32的部分邻居节点在网络中的重要性比较高,但由于节点32的邻居节点个数比较多,弱化了重要性较强的邻居节点如节点33、34对其的作用。因此,本文算法认为节点32重要性低于节点8更接近事实。

由表3可知,SRank算法选出的10个节点和其他算法结果相似,区别在于节点的重要性排序不同,说明本文算法具有一定合理性。本文算法和度中心性都认为节点46比节点38重要。从网络结构中可以看出两个节点相互连接,其影响力不同在于其存在不同的邻居节点,节点46的邻居节点为节点16、19、24、25、30、51、52、60,节点38的邻居节点为节点15、17、34、35、37、41、44、62,可以看出,节点46的邻居节点的影响力在其他算法中没有节点

38的邻居节点高。但在网络结构上,节点46与其邻居节点的相似性比节点38的高,例如节点16与节点46有3个共同邻居节点而节点17与节点38只有两个。因此,本文算法认为节点38重要性低于节点46更为合理。

表3 Dolphins网络节点重要性排序

排名	度中心性		接近中心性		PageRank		LeaderRank		SRank	
	节点	DC	节点	CC	节点	PR	节点	LR	节点	SR
1	15	12	38	30.4	18	0.031 7	15	0.029 4	15	0.003 44
2	46	11	15	29.98	52	0.031 3	38	0.027 1	46	0.002 99
3	38	11	37	29.28	15	0.030 5	46	0.027 1	38	0.002 46
4	52	10	41	29.28	58	0.029 7	52	0.024 9	34	0.002 46
5	34	10	21	29.15	38	0.029 1	34	0.024 9	58	0.001 97
6	58	9	46	28.58	46	0.028 4	58	0.022 6	52	0.001 97
7	30	9	34	28.48	34	0.027 6	18	0.022 6	14	0.001 84
8	21	9	2	27.37	30	0.025 8	21	0.022 6	30	0.001 63
9	18	9	51	27.33	14	0.025 6	30	0.022 6	18	0.001 41
10	41	8	52	26.83	2	0.024 1	14	0.020 4	10	0.001 34

2.3 SIR传播验证

通常采用传染病模型评价节点重要性排序方法,本文采用SIR模型^[1-2,11]。SIR模型中,一个节点的传播能力被定义为该节点的平均传播范围,如果一个排序方法的结果使得网络流传播又快又广,则说明该重要节点排序方法优于其他方法。

定义 4 节点的传播重要度矩阵 S_{IR} 定义为多次传播仿真网络中I状态和R状态的节点占整个网络的比例的平均值。分别对网络中每一个节点进行传播仿真,即每一次仿真只有一个节点处于I状态,其他节点都为S状态,选取传播时间步为10,对每个节点进行100次传播仿真,因此节点 v_i 的传播重要度为:

$$S_{IR}(i) = \frac{1}{100} \sum_{j=1}^{100} \frac{n - n_{j-i-s}}{n} \quad (5)$$

式中, n_{j-i-s} 表示初始传播节点为节点 v_i 的第 j 次传播仿真中经过10步传播后网络中S状态的节点个数。

首先,选取Neural网络对其进行仿真,将本文算法和度中心性、接近中心性、PageRank和LeaderRank进行对比,得到每个算法的节点重要度和传播重要度关系如图3所示。由图3明显看出,PageRank、LeaderRank和SRank节点重要度和传播重要度的线性程度表现更好。换言之,本文算法在重要节点排序上要优于度中心性和接近中心性。图3c、3d、3e图中圆圈标注的部分中,LeaderRank、PageRank算法排名值高的其 S_{IR} 值低,而本文算法线

性程度好, 可知本文算法发现重要性低的节点, 比另外两种算法优秀; 此外, 图3中椭圆标注中, 本文算法在节点重要性高的部分其离散程度低于其他两种算法, 可知本文算法在发现网络重要节点更为有效。

其次选取Polblogs网络进行仿真, 将本文算法和度中心性、接近中心性、PageRank算法和LeaderRank算法进行对比, 得到每个算法的节点重要度和传播重要度关系如图4所示, PageRank、LeaderRank和SRank节点重要度和传播重要度的线性程度比其他算法表现更好。换言之, 本文算法在重要节点排序上要优于度中心性和接近中心性。

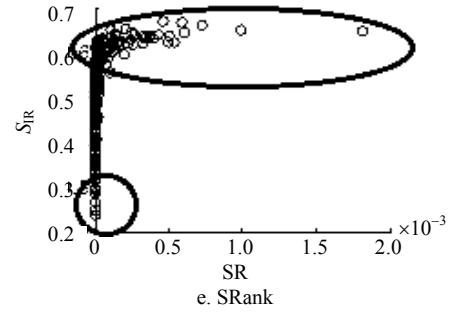
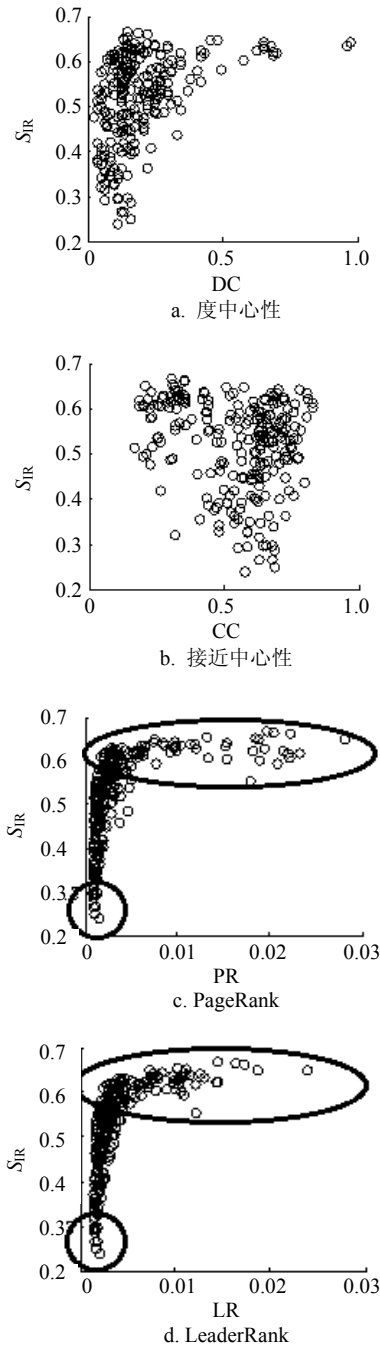
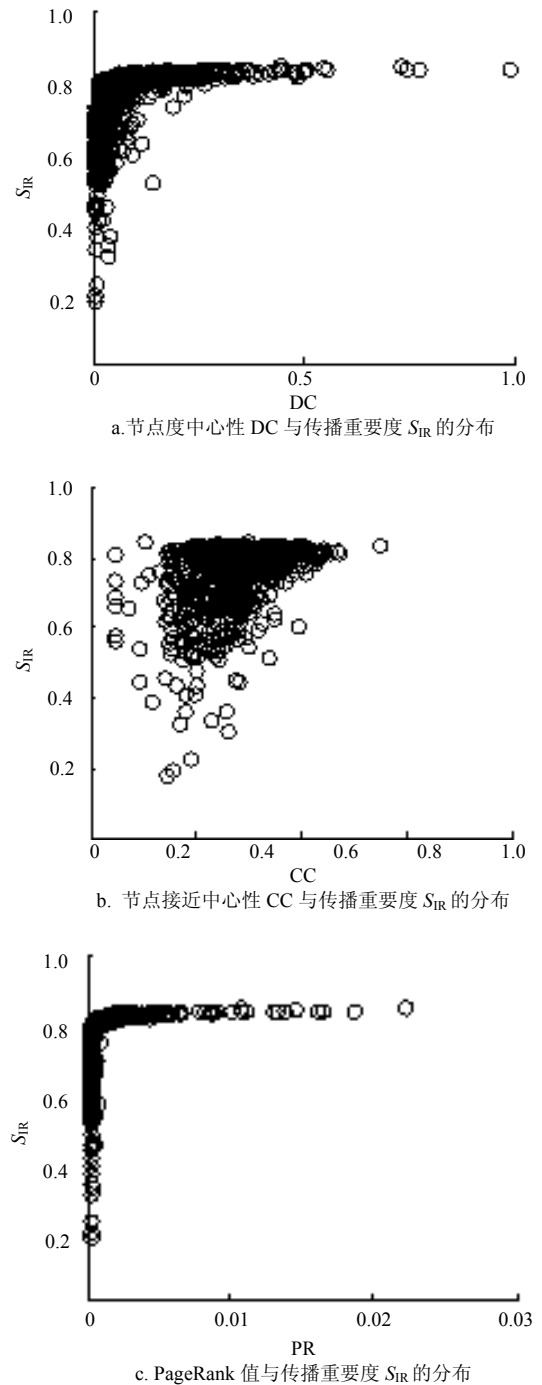


图3 Neural网络传播仿真结果(传染率 $\alpha=0.3$, 免疫率 $\beta=0.1$), 其中横坐标为归一化后的节点重要度值



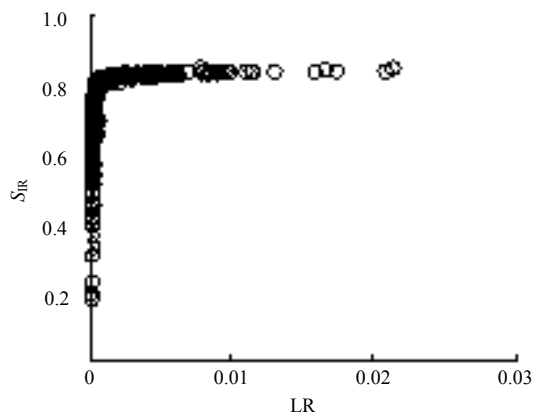
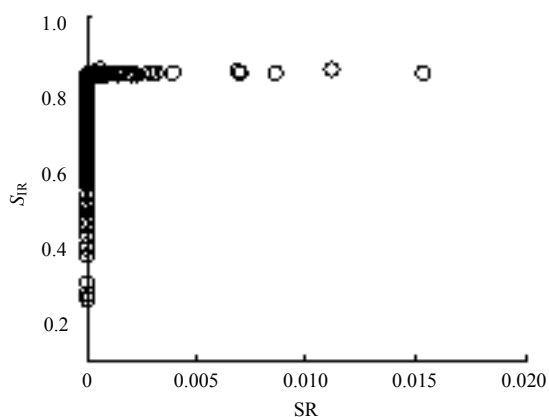
d. LeaderRank 值与传播重要度 S_{IR} 的分布e. SRank 值与传播重要度 S_{IR} 的分布

图4 Polblogs网络传播仿真结果(传染率 $\alpha=0.3$, 免疫率 $\beta=0.1$), 其中横坐标为归一化后的节点重要度值

定义 5 相关性系数 ρ 定义为节点重要度排序和传播重要度的相关性关系。 ρ 的绝对值越接近1, 两者的相关性越强, 节点重要度排序算法越好。 ρ 由斯皮尔曼等级相关系数^[12]计算, 定义为:

$$\rho = 1 - \frac{6 \sum_{i=1}^n (x_i - y_i)^2}{n(n^2 - 1)} \quad (6)$$

式中, x_i 、 y_i 分别为节点 v_i 在重要性排序算法中的排名和传播重要度中的排名。分别选取Nerual网络、Polblogs网络进行相关性分析。

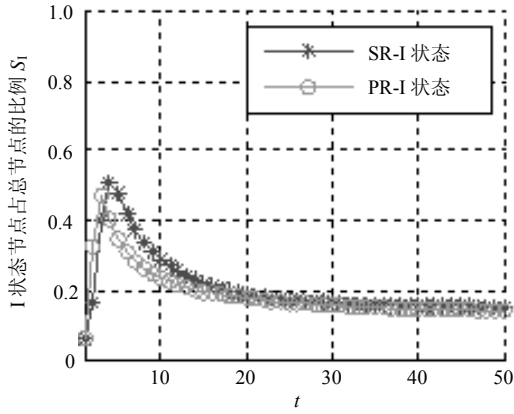
为了分析不同排名节点段的重要性排序结果, 本文分别选择PageRank、LeaderRank和SRank算法的重要性排序并对每个重要性排序算法的排名前10%、前20%以及所有节点分别进行考虑。相关性系数 ρ 对比结果如表4所示。由表4可知, 无论是整个网络的重要度排序还是节点重要性排名比较高的排序结果, 本文算法的相似性程度更高, 表明本文算法更优秀。

表4 节点重要性排序相关性系数 ρ 对比结果

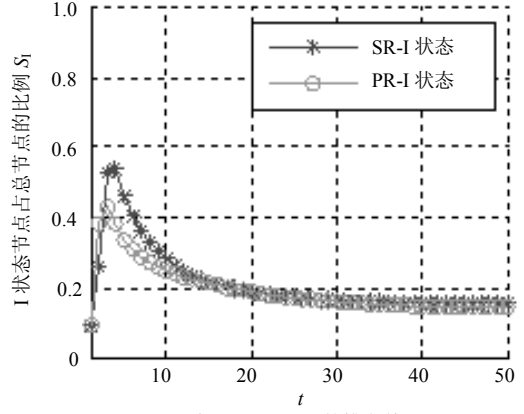
算法	Polblogs			neural		
	全部	前10%	前20%	全部	前10%	前20%
DC	0.823 1	0.993 6	0.977 4	0.444 7	0.950 1	0.876 0
CC	0.429 3	0.944 4	0.888 3	-0.181 8	0.828 2	0.690 0
PR	0.937 6	0.998 6	0.996 4	0.847 5	0.987 2	0.972 7
LR	0.960 9	0.997 7	0.993 6	0.878 4	0.977 6	0.956 3
SR	0.963 9	0.999 3	0.997 9	0.879 5	0.996 7	0.989 4

对各算法在线性虫神经网络、美国政治家博客网络的重要节点排序结果进行传播仿真。选取传播时间步为50, 观察网络中I状态节点个数 S_I 占网络的比例随时间的变化。针对本文算法所得排名前10%节点, 选取其中节点排名高于PageRank的节点作为传播节点; 同理, 针对PageRank所得排名前10%节点, 选取其中节点排名结果高于本文算法的节点作为传播节点。为了保证算法对比时, 不同方式选择传播节点个数相同, 传播节点个数较多的去除排名低的部分节点。其次, 分别选取本文算法排名前20%和全部节点进一步验证本文算法对比PageRank的有效性。本文算法与LeaderRank的对比亦如是。线性虫神经网络、美国政治家博客网络的传播仿真结果分别如图5、图6所示。

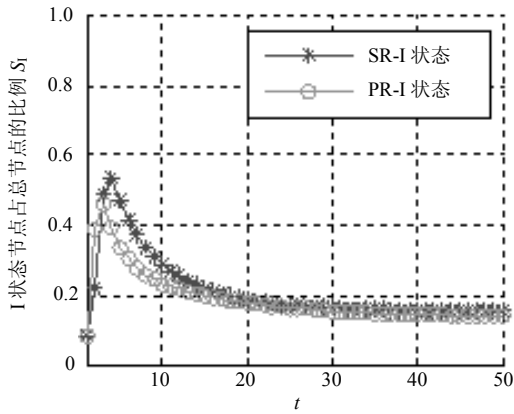
图5a~图5c为本文算法与PageRank的对比结果。图5a、图5b中, 本文算法的I状态比例 S_I 最大值超过0.5, 高于PageRank算法, 且前者增长时后者已经下降了, 表明本文算法的传播深度高于后者, 也证明了本文算法选取该网络的核心节点上表现比PageRank优越。图5c中, 两者的相差比例更加明显, 本文算法的斜率高于后者, 表明本文算法传播速度高于后者, 说明了本文算法在整个网络节点重要度排序要优于PageRank。同理, 由图5d~图5f本文算法与LeaderRank的对比结果可知, 本文算法I状态的最大值和斜率都高于后者, 说明本文算法无论是在发现该网络的核心节点上还是整个网络节点重要度排序上, 都要优于LeaderRank。由图6a~图6c本文算法与PageRank的对比结果可得, 本文算法的I状态比例的最大值超过后者, 且增长时其斜率高于后者, 表明本文算法的传播深度、深度高于后者, 本文算法无论是在发现该网络的核心节点上还是整个网络节点重要度排序上, 都要优于PageRank。同理, 由图6d~图6f本文算法与LeaderRank的对比结果可得, 本文算法要优于LeaderRank。



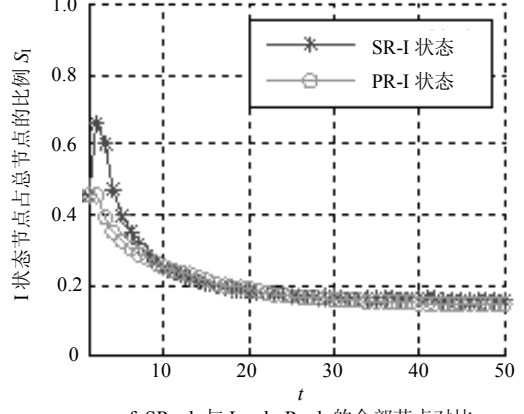
a. SRank 与 PageRank 的排名前 10%对比



e. SRank 与 LeaderRank 的排名前 20%对比

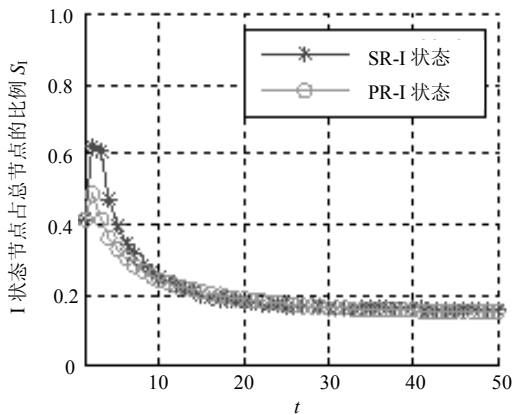


b. SRank 与 PageRank 的排名前 20%对比

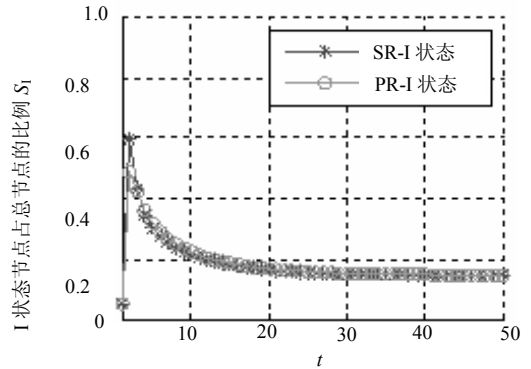


f. SRank 与 LeaderRank 的全部节点对比

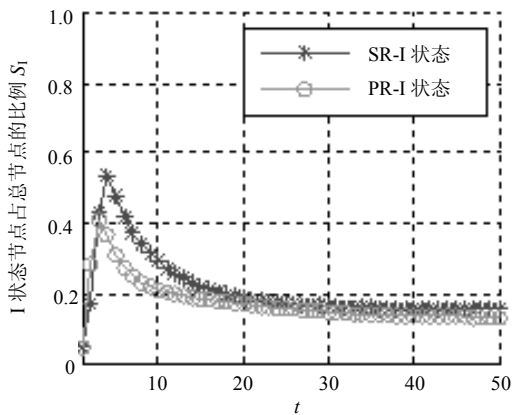
图5 Nerual网络传播仿真结果对比 (传染率 $\alpha=0.3$, 免疫率 $\beta=0.1$)



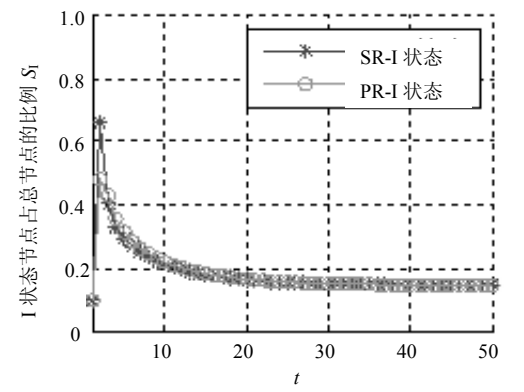
c. SRank 与 PageRank 的全部节点对比



a. SRank 与 PageRank 的排名前 10%对比



d. SRank 与 LeaderRank 的排名前 10%对比



b. SRank 与 LeaderRank 的排名前 20%对比

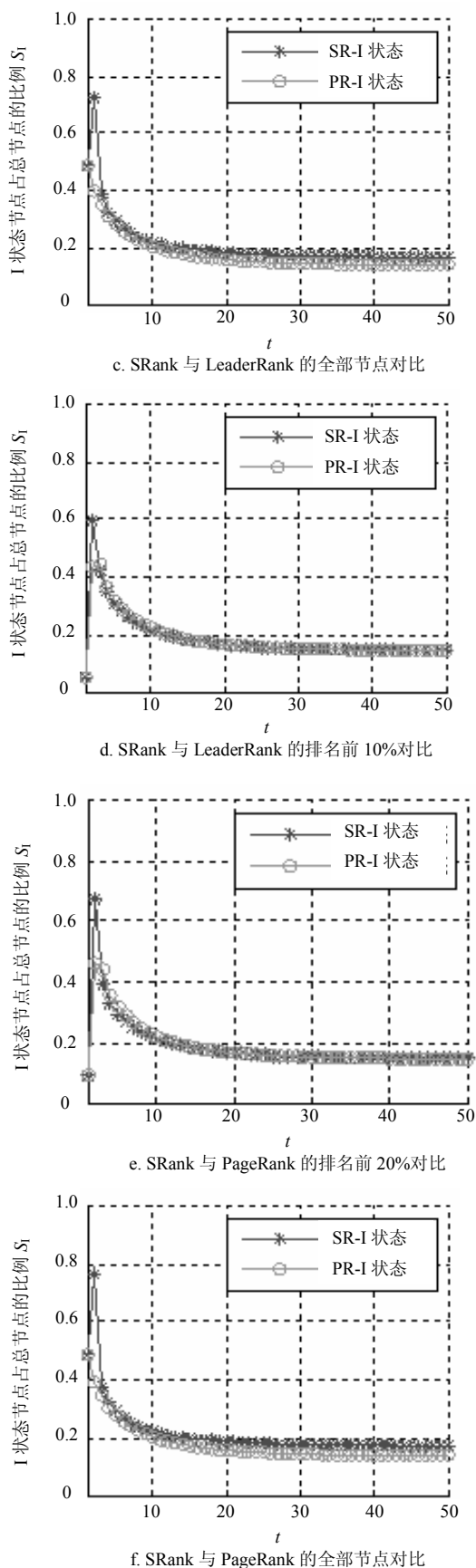


图6 Polblogs网络传播仿真结果对比
(传染率 $\alpha=0.3$, 免疫率 $\beta=0.1$)

3 结束语

本文主要工作是在LeaderRank算法的基础上引入了节点相似度的概念,综合考虑节点的局部特征和全局特征,提出了SRank节点重要性排序算法。对四种真实网络进行重要性排序,在通过分析网络结构来说明本文算法在无向网络的节点重要度排序的合理性同时,利用传播仿真成功验证了本文算法在有向网络节点重要度排序的有效性。

本文研究工作还存在一些不足。提出的算法主要针对于连通网络,对不连通网络还存在一定的不确定性。今后的工作需要更全面、科学地考虑不连通网络节点重要性的排序和验证方法。

参考文献

- [1] 周涛, 韩筱璞, 闫小勇, 等. 人类行为时空特性的统计力学[J]. 电子科技大学学报, 2013, 04(4): 481-540.
ZHOU Tao, HAN Xiao-pu, YAN Xiao-yong. Statistical mechanics on temporal and spatial activities of human[J]. Journal of University of Electronic Science and Technology of China, 2013, 42(4): 481-540.
- [2] 任晓龙, 吕琳媛. 网络重要节点排序方法综述[J]. 科学通报, 2014, 59(13): 1175-1197.
REN Xiao-long, LÜ Lin-yuan. Review of ranking nodes in complex networks[J]. Chin Sci Bull, 2014, 59(13): 1175-1197.
- [3] BONACICH P. Factoring and weighting approaches to status scores and clique identification[J]. Journal of Mathematical Sociology, 1972, 2(1): 113-120.
- [4] FREEMAN L C. Centrality in social networks conceptual clarification[J]. Social Networks, 1978, 1(3): 215-239.
- [5] BRIN S, PAGE L. The anatomy of a large-scale hypertextual web search engine[J]. Computer Networks & Isdn Systems, 1998, 30(98): 107-117.
- [6] LÜ L, ZHANG Y C, YEUNG C H, et al. Leaders in social networks, the delicious case[J]. Plos One, 2011, 6(6): e21202.
- [7] LI Q, ZHOU T, LÜ L. Identifying influential spreaders by weighted LeaderRank[J]. Physica A Statistical Mechanics & Its Applications, 2014, 404(24): 47-55.
- [8] 李鹏翔, 任玉晴, 席酉民. 网络节点(集)重要性的一种度量指标[J]. 系统工程, 2004, 22(4): 13-20.
LI Peng-xiang, REN Yu-qing, XI You-min. An importance measure of actors (set) within a network[J]. Systems Engineering, 2004, 22(4): 13-20.
- [9] DANGALCHEV C. Residual closeness in networks[J]. Physica A, 2006, 365: 556-564.
- [10] LIU Z, ZHANG Q M, LÜ L, et al. Link prediction in complex networks: a local naive Bayes model[J]. Epl, 2011, 96(4): 48007.
- [11] BONACICH P. Factoring and weighting approaches to status scores and clique identification[J]. J Math Sociol, 1972, 2: 113-120.
- [12] 马俊, 周刚, 许斌, 等. 一种基于话题传播的微博用户影响力分析方法[J]. 信息工程大学学报, 2013, 14(6): 735-742.
MA Jun, ZHOU Gang, XU Bin. User influence analysis in microblog based on topic diffusion[J]. Journal of Information Engineering University, 2013, 14(6): 735-742.

编辑 蒋晓