社交网络中个体对群体影响力分析

顾亦然,马德营,孟繁荣
(南京邮电大学 自动化学院,江苏 南京 210023)

摘 要:随着互联网的快速发展,社交网络节点影响力的研究成为热点,影响力分析对理解网络谣言、信息、病毒传播等具有重要意义。当前研究多集中于如何评价和衡量影响力,少有从个体对群体角度研究影响力的作用机制。文中将影响力看作是一种能够在网络连边传递的能量,个体影响力的大小为传递能量的程度。基于该思想,提出了一种基于介数中心性和节点贡献度的个体对群体影响力算法。通过与介数中心性、k-shell 和 PageRank 算法的对比以及传染病传播模型的传播实验表明,该算法能够有效、准确地表征个体对群体的影响力;通过对个体对群体的影响力各区间段节点频数分布研究发现,其值呈现幂律分布,具有无标度现象,即少数节点具有较大的影响力,普通节点虽然数量较大,但其影响力普遍较小。

关键词:复杂网络;节点贡献度;群体影响力;病毒传播
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2018)12-0190-04
doi:10.3969/j.issn.1673-629X.2018.12.040

Analysis of Individual Influence on Groups in Social Network

GU Yi-ran, MA De-ying, MENG Fan-rong
(School of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: With the rapid development of the Internet, the study of the influence of social network nodes has become a hot spot. Influence analysis has great significance to the understanding of online rumors, information, virus transmission and so on. Current researches focus on how to evaluate and measure influence, and there is few mechanisms that quantify the impact from individual to group. We regard influence as a kind of power that can be transmitted along the network, and the impact of individual is the degree of transferring power. Based on this, we propose an algorithm of the influence of the individual on the group based on the centripetal centrism and node contribution. The comparison between the proposed algorithm with mediating centrality, k-shell and PageRank, and the propagation experiment of infectious disease transmission model show that it has effectiveness and accuracy to represent the influence of the individual to the group. Through the research of node frequency distribution on each interval segment of the influence of the individual on the group, its value shows power law distribution and has scale-free phenomenon, that is, a few nodes have great influence. Although the number of ordinary nodes is large but has small influence generally.

Key words: complex network; node contribution; group influence; virus transmission

1 概 述

影响力分析是复杂网络的重要研究内容,现实社会中的诸多系统都是以复杂网络(complex network)的形式存在[1]。比如社会网络、互联网、交通网络、生物网络、社交网络等。其中,社交网络用户影响力的研究成为当前的热点之一。

为了深入研究及分析社交网络的节点重要性,学者们从网络结构等方面对节点影响力进行了广泛研究[2-6]。其中度中心性[3]、PageRank[4]、k-shell[5]、介数中心性[6]等算法刻画了节点在网络拓扑结构上的重要程度。度中心性(degree centrality, DC)刻画网络的局部特性,无法从全局刻画网络特征;PageRank 算法认为节点的重要性取决于网络中指向该节点的节点数量和质量,容易陷入悬挂节点;k-shell 是将网络一层层分解找出不同层次不同影响力的节点,其度量重要性比较粗粒化。

收稿日期:2018-01-09 **修回日期**:2018-05-16 **网络出版时间**:2018-07-04
基金项目:教育部人文社会科学研究规划基金(15YJAZH016);江苏省普通高校研究生创新计划项目(SJZZ16_0151)
作者简介:顾亦然(1972-),女,博士,教授,研究方向为复杂网络理论与应用、嵌入式系统、通信网络等;马德营(1990-),男,硕士研究生,研究方向为网络影响力分析。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20180703.1511.032.html

此外,节点的重要性不仅体现在拓扑结构上,其社会关系、行为特性同样对邻居节点产生不可忽视的作用。Richardson 等[7]提出影响力的计算以及使其最大化是一算法问题。Wang 等[8]发现影响力的传播大多发生在社团内部。王金龙等[9]提出了相对权重影响力模型并对影响力传播路径进行分析。肖云鹏等[10]提出了基于节点动态行为的影响力传播模型,但其主要是刻画影响力强度而未定量分析影响力。

当前,在线社交网络发展迅速,社交平台用户之间或因为现实关系或共同的价值观等逐渐形成网络群体,在群体影响力的相互作用下,网络群体行为涌现的更加迅速。例如,在微博社交网络中,由于兴趣等关注点的不同,一个用户可能隶属于不同社团,又由于亲密关系的不同,一个用户又可以在不同社团中进行消息的传递,那么这样的节点实际上成为不同社团间信息传播的桥梁,具有控制网络信息传播的能力。其传播的影响力不仅仅是按照最短路径产生作用,亲密关系也成为消息传递,信息分享的一个重要因素。因此影响力的传递作用不仅和社交网络的拓扑结构有关还和社会关系的因素有关,如文献[11]发现节点贡献度与节点重要性及其影响力范围都存在一定关系。

因此,针对社交网络中桥节点这样的特殊位置节点在信息传播中的重要性,考虑到此类节点信息交流的繁忙程度,以介数中心性作为描述节点拓扑结构重要性的参数;考虑到亲密关系对传播的影响,以节点间的贡献度来描述节点亲密程度,提出了一种个体对群体影响力算法,以此刻画该节点对整个网络的影响力贡献情况,即个体对网络群体的影响力。另外还对计算结果进行定量分析,用以研究其内在规律。

2　相关概念及定义

2.1　介数中心性

介数中心性(betweenness centrality,BC)是以经过某一节点的最短路径数目来刻画节点重要性的指标。它体现了网络中节点对网络信息流动的影响力[12]。具体地,节点 v_i 的介数中心性如下:

$$\mathrm{BC}_i = \frac{2}{N^2 - 3N + 2} \sum_{s \neq i \neq t} \frac{n_{st}^i}{g_{st}} \tag{1}$$

其中,g_{st} 表示从节点 v_s 到节点 v_t 最短路径的数目;n_{st}^i 表示从节点 v_s 到节点 v_t 最短路径 g_{st} 中有 n_{st}^i 个经过节点 v_i。

2.2　节点间的贡献度

社交网络中节点之间的共同邻居越多说明节点间的关系越密切。其对彼此的影响力就不仅与自身所处的拓扑位置有关,还与共同邻居作用有关。现实生活中,节点 v_i 对 v_j 的影响程度和节点 v_j 对 v_i 的影响程度是有区别的,因此引入贡献度[11]这一概念。节点 v_i 对 v_j 的贡献度定义如下:

$$C_{ij} = \frac{|\Gamma_i \cap \Gamma_j| + 1}{|\Gamma_j| + 1} \tag{2}$$

其中,Γ_i 和 Γ_j 分别为节点 v_i 和 v_j 邻居节点的集合。

3　个体对群体影响力算法

文中将影响力看作是一种能够在节点之间进行传递的能量。那么,从节点 v_i 传递到 v_j 的能量多少,就是节点 v_i 对 v_j 的影响力。结合网络上刻画节点控制信息流动的指标介数中心性以及节点之间相互影响作用的贡献度概念,进一步描述个体间影响力的传递机制,建立个体对群体的影响力算法。

3.1　个体间影响力

正如上文所说,将影响力看作一种能够在节点之间流动的能量,该能量沿着节点连边进行扩散。如节点 v_i 传递到 v_j 的能量多少,就是节点 v_i 对 v_j 的影响力,可见,节点间的影响力刻画的是节点之间影响力传递能力的大小。

定义:个体间影响力 $f_n(i,j)$ 为沿最短路径从节点 v_i 开始扩散到节点 v_j 的影响力。

以计算 $f_n(i,j)$ 为例,计算过程如下:设从节点 v_i 到节点 v_j 的最短路径为 $v_i \to v_{x_1} \to v_{x_2} \cdots v_{x_{k-1}} \to v_j$,其距离为 d_{ij},则有:

$$f_n(i,j) = \alpha_{ix_1} * \mathrm{BC}_i * C_{ix_1} + \sum_{k=2}^{d_{ij}-1} \alpha_{ix_{k-1}} * \mathrm{BC}_{x_{k-1}} * C_{x_{k-1}x_k} + \alpha_{ix_k} * \mathrm{BC}_{x_k} * C_{x_kj} \tag{3}$$

其中,α 为根据路径远近设置的权重因子。

由于节点的影响力会因节点之间的距离远近产生差异,因此引入权重因子 α[13] 来分配计算权重。权重因子 α 与节点之间的距离 d_{ij} 存在以下关系:

$$\alpha_{ik} = \begin{cases} \dfrac{1}{d_{ik} + (d_{ij} - 1)} & d_{ik} > 1 \\ 1 - \sum_{k=2}^{N} \alpha_{ik} & d_{ik} = 1 \end{cases} \tag{4}$$

其中,d_{ij} 为始末两节点间距离;d_{ik} 为初始节点 v_i 到节点 v_k 的最短距离,k 取正整数。

3.2　个体对群体影响力

如果说个体间的影响力刻画的是节点之间影响力传递能力的大小,那么个体对群体影响力就是刻画节点影响力的全局特性,即节点的影响力扩散对全网的影响程度。

定义:个体对群体影响力为节点 v_i 对群体 R 中所有其他节点的影响力之和,记为 F_{iR}。

在实际计算中,F_{iR} 从源节点依次计算至整个群

体，显然复杂度太高。在研究中发现，并非要计算至整个网络，原因有二：其一，根据微信以及 Facebook 最新数据分析，复杂在线网络两节点间平均距离为 3 左右；其二，由实际计算比对中发现，路径的距离 $d_{ij} > 3$ 时，路径权重 α 的大小对于计算结果排序并无明显影响。基于上述分析，将 F_{iR} 进一步定义如下：个体 v_i 对群体 R 的影响力 F_{iR} 为 v_i 对距源节点路径距离 $d_{ij} \leqslant 3$ 所有节点的影响力之和。公式为：

$$F_{iR} = \sum_{j \in V_i} f_n(i,j) \tag{5}$$

其中，V_i 表示距节点 v_i 路径距离 $d_{ij} \leqslant 3$ 的节点集合。

3.3 个体对群体影响力算法描述

综上所述，算法描述如下：

输入：$G = (V,E)$

输出：节点 v_i 对群体 R 的影响力

1. 通过广度优选搜索算法获取节点 v_i 到集合 V_i 中节点的路径与距离表 D_path

2. 计算网络节点间贡献度矩阵 C_n

3. 计算网络每节点 BC 值

4. 计算节点 v_i 到集合 V_i 中每个节点的影响力 $f_n(i,j)$

5. 由式 5 将每条路径影响力加和，求得 F_{iR}

6. END

4 仿真分析

文中研究的 F_{iR} 算法刻画的是节点影响力的全局特性，即节点的影响力扩散对全网的影响程度。显然，个体对群体的影响力是衡量节点重要性的指标之一。另外影响力大的节点往往是信息传播中的重要节点，信息传播能力强，传播至整个群体速度较快。基于这一思想，本节使用不同网络数据，通过经典的传染病模型进行仿真可以较为直观地分析节点信息传播情况，以验证算法的有效性。实验使用的网络分别为 Email 和 Zachary，两网络参数如表 1 所示。

表 1 网络参数

网络	节点数	连边数	平均度	网络直径	平均聚类系数	平均路径长度
Email	1 133	10 902	19.244	8	0.254	3.606
Zachary	34	78	4.59	5	0.57	2.41

首先使用 SI 传染病模型，对 Email 网络进行传播验证，其传播概率为 $\beta = 0.03$。

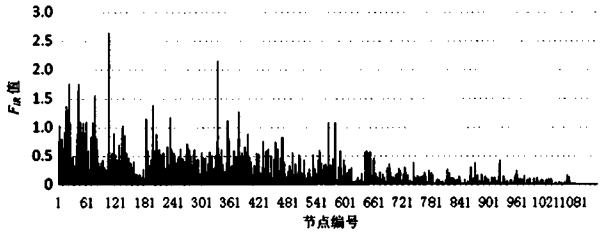
4.1 网络仿真分析

4.1.1 Email

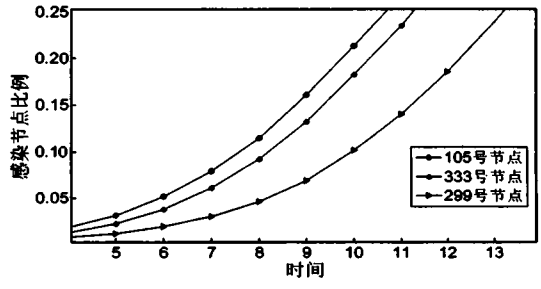
对 Email 网络进行 F_{iR}、BC、PageRank 和 k-shell 算法分析，结果如表 2 所示，其中 F_{iR} 的分析结果如图 1(a)所示。

表 2 Email 网络各算法排名前三节点

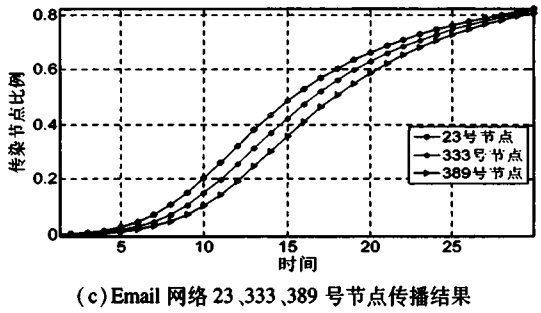
排名	F_{iR}	BC	PageRank	k-shell
1	105	333	105	299
2	333	105	23	389
3	23	23	333	434



(a) Email 网络各节点 F_{iR}



(b) Email 网络 105、333、299 号节点传播结果



(c) Email 网络 23、333、389 号节点传播结果

图 1 Email 网络部分节点传播图

根据表 2 排序结果，首先将四种算法排在第一位的 105、333、299 号节点进行传播仿真，结果如图 1(b)所示。可以看出，F_{iR} 计算出来的 105 号节点在传播初期明显优于其他影响力指标得出的节点；基于上述的思想将 F_{iR} 计算出来的 23、333 节点以及 k-shell 排序第二的 389 号节点进行传播仿真实验，结果如图 1(c)所示。从图中亦能看出，在整个传播仿真中 23 号节点较优于 333 号节点和 389 号节点。

综上可知，该算法能够较为准确地衡量个体对群体的影响力。

4.1.2 Zachary

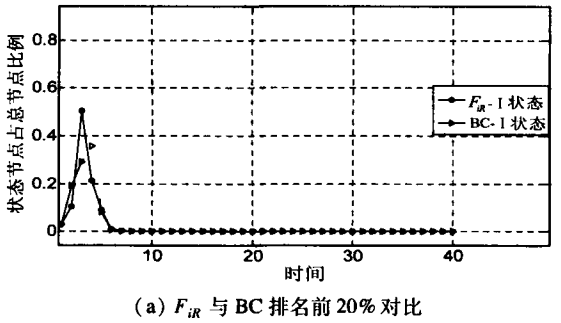
在 Zachary 网络中分别计算节点的 F_{iR}、BC、PageRank 和 k-shell，得到的结果如表 3 所示。

为了更全面地验证算法，本节传播仿真采用 SIR 传播。将文中算法排名前 20% 节点分别和其他三种算法得出的排名前 20% 节点进行传播比对，传播比对过程中去掉两算法中相同的节点，选取时间步为 40，传播 100 次取平均，结果如图 2 所示。由图 2(a)可

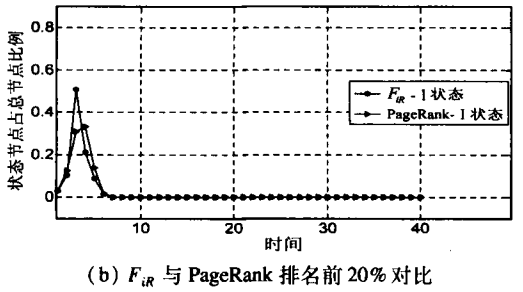
知,文中算法选取的节点不但感染峰值比例高于后者,而且到达峰值的时间也比 BC 早;图 2(b)中两者的差异更为明显;图 2(c)中两个最大感染比例差异虽没有图 2(a)、(b)大,但到达最大值的时间却比较滞后。通过上述分析对比,文中算法在 SIR 传播仿真中依然有较好表现。

表 3　Zachary 网络各算法排名前 20% 节点

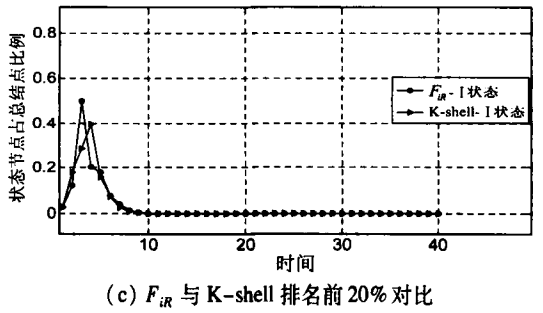
排名	F_{iR}	BC	PageRank	k-shell
1	1	1	34	1
2	34	34	1	2
3	3	33	33	3
4	32	3	3	4
5	20	32	2	8
6	9	9	32	9
7	14	2	4	14
8	12	14	24	31



(a) F_{iR} 与 BC 排名前 20% 对比



(b) F_{iR} 与 PageRank 排名前 20% 对比

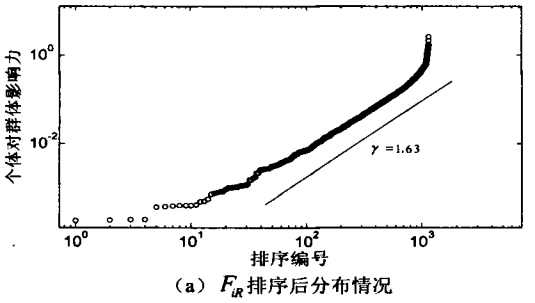


(c) F_{iR} 与 K-shell 排名前 20% 对比

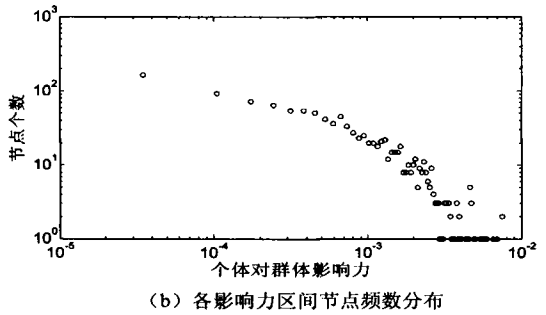
图 2　Zachary 网络传播仿真结果对比

4.2　定量分析

对算法计算结果的定量分析过程如下:首先将 Email 网络 F_{iR} 进行升序排列,如图 3(a)所示,分别对横纵坐标取对数,对曲线的走向进行分析发现,起始阶段含有较多的节点但是这部分节点影响力的变化范围却很小;在曲线顶部虽然节点数较少,但其影响力却明显快速增加。通过数值计算可以得到前 15.36% 的节点影响力值总和等于后 84.64% 的节点影响力值总和。由此可知,只需对前 15.36% 节点施加影响力便能很大程度上影响整个网络。当然进一步分析前 20% 节点对群体影响力占总影响力为 58.38%,并未出现或接近公众普遍认知的"二八效应"。同时,将 F_{iR} 值均分 150 个区间段,取每段中间值代替每段的数值,计算落在每段节点的频数,其频数分布符合幂率分布,分布曲线为 $y = 0.14 * x^{-0.70}$。将数值和频数取对数进行拟合,拟合的结果如图 3(b)所示,其斜率 $\gamma = -0.7$。可见,网络中少数节点的影响力远远超过大部分普通节点的影响力。



(a) F_{iR} 排序后分布情况



(b) 各影响力区间节点频数分布

图 3　F_{iR} 算法结果定量分析

5　结束语

由于节点影响力物理表征和刻画方法到目前为止并没有一个相对完整的体系,文中将影响力看作是一种能够在节点之间进行传递的能量,提出了 F_{iR} 的计算方法,并建立了算法模型,进行了 SI 和 SIR 传播仿真分析。结果表明,该算法在传播仿真中表现优异,能够准确地找出对群体影响力大的节点。进一步分析 F_{iR} 的分布情况得出影响力值的分布情况符合幂律分布,具有无标度现象。

参考文献:

[1] ADAMIC L A, HUBERMAN B A, BARABÁSI A L, et al. Power-law distribution of the world wide web[J]. Science, 2000, 287(5461): 2115.

[2] DING Fei, CHENG Hui, SI Xiameng, et al. Read and reply behaviors in a BBS social network[C]//2nd international conference on advanced computer control. Shenyang, China: IEEE, 2010: 571-576.

(下转第 199 页)

最后,动态地将任务提交给 Hadoop 集群,观察总利润的变化。从数据可以看出,该方法不仅接收到最多的候选任务,而且完成大部分任务,因此可以带来最大的利润。这说明所提出的方法也适用于动态提交的任务。

3　结束语

研究了 Hadoop 集群中的最大利润问题。为了使利润最大化,基于候选任务集的有效序列选择了一些高利润率的任务。此外,为了提高查找有效序列的效率,设计了一些修剪策略,并给出了相应的调度算法。实验结果表明,该算法的总收益非常接近理想的最大值,在不同的实验环境下明显优于相关的调度算法。

参考文献:

[1] 李玉丹,郑晓薇. Hadoop 下多模式并行分类算法及其应用研究[J]. 计算机工程,2014,40(12):45-49.

[2] 王静蕾. Hadoop 云计算框架中的分布式数据库 HBase 研究[J]. 商丘职业技术学院学报,2014,14(2):18-20.

[3] CHU Chengtao, KIM S K, LIN Yian, et al. Map-reduce for machine learning on multicore[C]//Proceedings of the 19th international conference on neural information processing systems. Canada: MIT Press, 2007.

[4] INZA I, LARRANAGA P, BLANCO R, et al. Filter versus wrapper gene selection approaches in DNA microarray domain[J]. Artificial Intelligence in Medicine, 2004, 31(2): 91-103.

[5] 向丽辉,缪　力,张大方. 压缩对 Hadoop 性能影响研究[J]. 计算机工程与科学,2015,37(2):207-212.

[6] 杨倩茹,黄梦醒,万　兵. 一种引入内存平衡的 Hadoop 平台作业调度算法[J]. 小型微型计算机系统,2014,35(12):2708-2012.

[7] 孙彦超,王兴芬. 基于 Hadoop 框架的 MapReduce 计算模式的优化设计[J]. 计算机科学,2014,41(11A):333-336.

[8] TRIPATHY B K, MITTAL D. Hadoop based uncertain possibilistic kernelized c-means algorithms for image segmentation and a comparative analysis[J]. Applied Soft Computing, 2016, 46: 886-923.

[9] GANESH S, BINU A. Statistical analysis to determine the performance of multiple beneficiaries of educational sector using Hadoop-Hive[C]// International conference on data science &engineering. Kochi, India: IEEE, 2014: 32-37.

[10] BERLINSKA J, DROZDOWSKI M. Scheduling divisible MapReduce computations[J]. Journal of Parallel and Distributed Computing, 2011, 71(3): 450-459.

[11] 李　洋,吕家恪. 基于 Hadoop 与 Storm 的日志实时处理系统研究[J]. 西南师范大学学报:自然科学版,2017,42(4):119-126.

[12] 梁俊荣. 基于 Hadoop 的图书馆复合大数据存储系统研究[J]. 现代情报,2017,37(2):63-67.

[13] 余　辉,黄永峰,胡　萍. 微博舆情的 Hadoop 存储和管理平台设计与实现[J]. 电子技术应用,2017,43(3):120-123.

[14] HO T K. The random subspace method for constructing decision forests[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(8): 832-844.

[15] 张建平,李　斌,刘学军,等. 基于 Hadoop 的异常传感数据时间序列检测[J]. 传感技术学报,2014,27(12):1659-1665.

(上接第 193 页)

[3] ALBERT R, JEONG H, BARABASI A L. The diameter of the world wide web[J]. Nature, 1999, 401: 130-131.

[4] BRIN S, PAGE L. The anatomy of a large-scale hypertextual Web search engine[C]//International conference on world wide web. Brisbane, Australia: Elsevier Science Publishers, 1998: 107-117.

[5] KITSAK M, GALLOS L K, HAVLIN S, et al. Identification of influential spreaders in complex networks[J]. Nature Physics, 2010, 6(11): 888-893.

[6] FREEMAN L C. A set of measures of centrality based on betweenness[J]. Sociometry, 1977, 40(1): 35-41.

[7] RICHARDSON M, DOMINGOS P. Mining knowledge-sharing sites for viral marketing[C]//Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining. Edmonton, Alberta, Canada: ACM, 2002: 61-70.

[8] WANG Yu, CONG Gao, SONG Guojie, et al. Community-based greedy algorithm for mining top-K influential nodes in mobile social networks[C]//16th ACM SIGKDD international conference on knowledge discovery and data mining. Washington, DC, USA: ACM, 2010: 1039-1048.

[9] 王金龙,刘方爱,朱振方. 一种基于用户相对权重的在线社交网络信息传播模型[J]. 物理学报,2015,64(5):63-73.

[10] 肖云鹏,李松阳,刘宴兵. 一种基于社交影响力和平均场理论的信息传播动力学模型[J]. 物理学报,2017,66(3):227-239.

[11] 顾亦然,王　兵,孟繁荣. 一种基于 K-Shell 的复杂网络重要节点发现算法[J]. 计算机技术与发展,2015,25(9):70-74.

[12] 汪小帆,李　翔,陈关荣. 网络科学导论[M]. 北京:高等教育出版社,2012:159-162.

[13] 张闪闪. 国内外作者贡献声明的贡献要素与贡献权重算法初探[J]. 图书情报工作,2016,60(1):125-134.