基于中介中心度的微博影响力个体发现*

朱静宜
(浙江长征职业技术学院 计算机与信息技术系, 杭州 310023)

摘 要: 微博中重要影响力个体的发现有着极为重要的作用。中介中心度方法是发现网络中重要节点的有效方法,然而传统的中介中心度方法只适用于小规模的网络,对于海量的微博网络信息却无能为力。提出一种基于随机游走的中介中心度算法,该算法不仅能有效地应对海量的微博网络数据,而且其发现结果也明显优于相关的研究。

关键词: 微博; 影响力; 排名; 算法; 中介中心度
中图分类号: TP393 **文献标志码**: A **文章编号**: 1001-3695(2014)01-0131-03
doi:10.3969/j.issn.1001-3695.2014.01.031

Centrality based microblog influence entity discovery

ZHU Jing-yi
(*Dept. of Computer & Information Technology, Zhejiang Changzheng Vocational & Technical College, Hangzhou* 310023, *China*)

Abstract: Discovery of influential people is a very important research topic in microblog. Betweenness centrality is an efficient method to find important nodes in networks. However, traditional betweenness centrality algorithms are only fit in small size of networks, and not efficient in microblog of massive messages. This paper proposed a random walk based betweenness centrality algorithm, and this algorithm could deal with massive microblog messages, and was more efficient than related work.
Key words: microblog; influence; ranging; algorithm; betweenness centrality

微博客(microblog),简称微博,是一个基于用户之间的关注关系,实现信息分享、传播以及获取的平台。微博用户可以在个人电脑、移动电话、智能终端等设备上通过 Web 或 wap 访问微博服务。在微博中,用户可以即时地发布不多于 140 字的文字信息,该信息被称为用户的状态信息。国外最著名的微博网站是美国的 Twitter,国内有中国门户网站新浪推出的新浪微博,以及腾讯公司推出的腾讯微博。随着微博在网民中的日益火热,微博效应正在逐渐形成。在微博中,用户既可以作为信息发布者,在微博上发布内容供别人浏览,也可以作为观众,在微博上浏览感兴趣的信息。用户除了可以发布文字信息外,也可以发布图片、分享视频等。微博服务的最大优点是发布信息快速快,信息传播速度快。假设某用户有 200 万听众,那么他发布的信息瞬间就可以传播给这 200 万人。微博作为一种新的社会媒体,一经出现便吸引了大量的用户。当前,微博已经成为网民获取信息和表达情感的重要场所,因此也成为了社会热点事件的产生和讨论的重要场所[1]。在微博中,用户与用户间的关注关系作为网络中的节点和边是信息传播的基础,转发是信息传播的途径。微博信息在网络上经过重要用户的转发后,便会引起广泛的关注。这些重要的用户就是高影响力的用户。高影响力的用户发布或转发的信息是引起信息持续传播和形成更大传播规模的关键因素[2]。因此,识别和发现网络中高影响力的用户是微博中的重要研究内容之一。

本文将节点在网络中的中介中心度作为衡量节点影响力的标准,提出了一种基于随机游走的节点中介中心度算法。实验结果表明,本文提出的基于中介中心度的节点影响力排名算法明显优于传统的基于中介中心度的节点排名算法。

1 相关工作

网络中的影响力节点排名是社会网络研究领域的重要研究内容之一,其主要应用在 Web 网页排名和社交网络的用户影响力识别。在 Web 信息检索领域,网页作为节点,网页和网页之间的链接作为边。网页排名是按照网络的链接结构将网页在网络中的重要性进行排名。经典的网页排名方法有 PageRank[3]、HITS[4]、SALSA[5]等。PageRank 将网络看做一个有向图,节点的影响力与父节点的影响力成正比,与其父节点的子节点个数成反比。PageRank 假设网络中的影响力只包含一个属性,并且该影响力沿着一个方向在网络中流动。HITS 算法认为节点的影响力属性包含权威值(authority)和枢纽值(hub)。节点的权威值与其父节点的枢纽值成正比,与其父节点的子节点个数成正比;节点的枢纽值与其子节点的权威值成正比,与其子节点的父节点个数成正比。SALSA 算法类似于 HITS 算法,将网络划分为两个矩阵,即权威值矩阵和枢纽值矩阵,然后通过 PageRank 的迭代算法求出这两个矩阵的特征值。

在社交网络中,用户的影响力识别类似于 Web 信息检索中的网页排名。TunkRank[6]和 TwitterRank[7]将 Twitter 中的用户及用户间的关注关系构成一个有向图,然后应用 PageRank 算法的变体将用户排名。两者的区别在于,TwitterRank 既考虑了用户间的关注关系,又考虑了用户谈论的话题的相似性;而 TunkRank 则主要考虑用户对其关注者的影响。以上两种算法是基于 PageRank 算法的。IP 影响力算法[8]类似于 HITS 算

收稿日期: 2013-04-10; **修回日期**: 2013-05-24 **基金项目**: 浙江省自然科学基金资助项目(201200712)
作者简介:朱静宜(1982-),女,浙江杭州人,讲师,硕士,主要研究方向为计算机网络、社会网络挖掘等(gonggongis@163.com).

法,在网络的影响力属性中,将用户的影响力分为影响力(influence)和被动型(passivity)。在社交网络中,用户可以发言、回复和转发别人的发言,这些行为也可以构成社会网络图,如用户—发言图[9]和基于行为的影响力图[10]。根据这些社会网络图,仍然可以通过相关的定义将用户进行排名。

2 基于中介中心度的影响力排名

节点的中心度用来评估网络中节点的重要性,其作为研究社会网络的重要工具吸引了人们广泛的研究。在社会网络中,具有高中心度的节点往往被认为具有高的影响力。随着影响力的具体定义不同,学者们提出了许多中心度定义方法。本文提出中介中心度的评价方法来衡量节点在网络中的影响力。

在社会网络中,中心度高的点和边都占据非常重要的角色。由于社会网络中往往存在一些小组,这些小组内部节点之间的链接很紧密,小组与外界的链接很稀疏。当节点或边在连接这些小组的过程中起着非常重要的角色时,可以认定这些节点的中心性高。节点 i 的中介中心度 c_i 定义为

$$c_i = \sum_{j,k} \frac{b_{jik}}{b_{jk}} \tag{1}$$

其中:b_{jk}表示从节点 j 到 k 的最短路径的个数,b_{jik}表示从节点 j 到 k 的且经过节点 i 的最短路径的个数。

计算节点的中介中心度的最基本方法就是计算出所有的节点对之间的最短路径,这种方法需要 $\Theta(n^3)$ 的时间复杂性和 $\Theta(n^2)$ 的空间复杂性。为了降低该算法的复杂性,Brandes[11]基于单源最短路径的算法设计了一个快速的节点中心度计算方法。该算法的空间复杂性是 $O(n+m)$,时间复杂性是$O(nm)$和 $O(nm+n^2\log n)$。其中,n 为图中节点的个数,m 为图中边的个数。

在微博社会网络中,网路中的节点和边的数量巨大,因此上述方法并不能有效地计算出节点的中介中心度。本文应用随机游走的方法来计算节点的中介中心度。其主要思想是:在网络中进行大量的随机游走,根据这些随机游走计算出节点的中介中心度。其计算方法如下:

$$c_i^{rw} = \sum_{i \neq j \neq k} R_{jk}^{(i)} \tag{2}$$

其中:$R^{(i)}$为一个矩阵,其中的第(j,k)个元素 $R_{jk}^{(i)}$ 为某个随机游走从节点 j 到 k 的概率,并且该随机游走包含节点 i 作为中间节点。

基于随机游走的中介中心度的算法过程为:令网络中的每个节点 $u(u \in V)$作为初始节点,进行若干次 R 随机游走;在每一随机游走步中,以概率 ε 停止该随机游走,以概率 $1-\varepsilon$ 进行下一个的随机游走;在随机游走时,随机选取一条当前的节点的出边,并走向目标节点;重复上述随机游走过程,直到终止。算法终止后,根据式(2)计算出每个节点的中介中心度。算法的详细描述如下:

```
for i = 1 to R
  for each u ∈ V
    Let rw = u;
      While Random.next() > ε do
      rw.append(u.randomNeighbor())
    end while
  end for
end for
```

3 实验分析

3.1 评价标准

为了对提出算法的有效性进行验证,本文采用了信息检索的四个非常重要的指标:查准率(precision)、召回率(recall)、F-measure 和再现率(re-occur)来评价算法的有效性。这四个评价指标具体定义如下:

$$\text{precision} = \frac{TP}{TP+FP} \tag{3}$$

$$\text{recall} = \frac{TP}{TP+FN} \tag{4}$$

$$F\text{-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{5}$$

其中:$|TP|$、$|FP|$、$|FN|$分别表示 true positives、false positives 和 false negatives。

$$\text{re-occur} = \frac{\sum_{\forall u, u \in A_i,} \text{numOfOccur}(u, A_i, \bar{A_i})}{|A_i|} \tag{6}$$

其中:A_i表示采用算法 i 得到的计算结果;$|A_i|$表示结果中含有的节点个数;$\text{numOfOccur}(u, A_i, \bar{A_i})$表示算法 A_i得到的结果中的点 u 在除了 A_i算法以外的结果中的出现次数。该评价标准以某种算法的计算结果在其他算法的结果中出现的次数越多,该算法的性能越好。

3.2 实验环境及数据集

本实验的实验环境是个人 PC,操作系统是 Ubunbtu 10.0,实验平台为 Java,图处理工具包为 Jung 2.0。在对算法的有效性进行验证的过程中,本文采用了 Dianping 数据集[12]和 Gowalla 数据集[13]两个真实的数据集。

数据集基本信息如表 1 所示。

表 1 数据集基本信息

name	# nodes	# edges	average degree
Gowalla	196 591	1 900 654	1.744
Dianping	204 074	926 720	1.822

3.3 实验结果

在实验过程中,将本文提出的基于随机游走的节点中介中心度算法称为 RWBC,将传统的中介中心度算法称为 TBC,将改进的中介中心度算法称为 IBC。

由于 RWBC 采用的是基于随机游走的方法,而在随机游走过程中,随机游走的步长对算法的准确性起着非常重要的作用。本文分别考虑了随机游走的平均游走步长在 5、10、15、20 和 30 几种情况下算法的评价标准,结果如图 1 和 2 所示。可以看出,RWBC 算法随着随机游走步长的增加,算法的再现率增大;但是当步长增长到一定程度时(图中为 20),再现率的增加缓慢。因此,在后续实验中,本文采用平均步长为 20 的随机游走来计算节点的中介中心度。

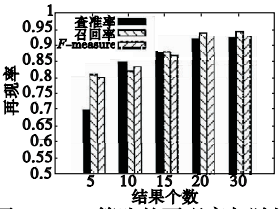

图1 RWBC算法的再现率与随机游走步长的关系(Dianping数据集)

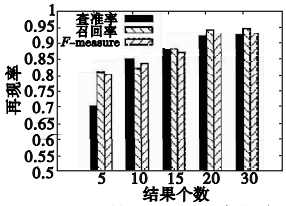

图2 PWBC算法的再现率与随机游走步长的关系(Gowalla数据集)

为了验证算法的有效性,本文分别比较了算法返回结果的查准率、召回率和 *F*-measure,比较结果如图 3 和 4 所示。

从图 3 和 4 可以看出,当用户在搜索影响力高的用户时,三种算法在性能上存在着差异,并且本文提出的基于随机游走的 RWBC 算法在查准率、召回率和 *F*-measure 三种评价标准上都优于其他两种算法。

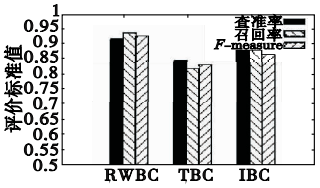

图3 算法的性能比较
(Dianping数据集)

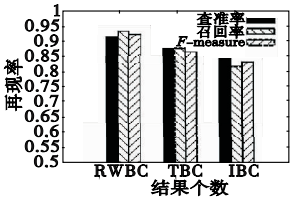

图4 算法的性能比较
(Gowalla数据集)

为了进一步比较算法的性能,本文对各种算法的再现率进行了比较。对算法返回的结果个数进行了限制,分别比较了算法在返回 10、20、50、100 和 150 个结果下各种算法的再现率,结果如图 5 和 6 所示。

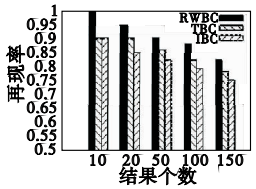

图5 算法的再现率比较
(Dianping数据集)

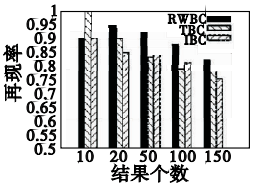

图6 算法的再现率比较
(Gowalla数据集)

从图 5 和 6 可以看出,三种算法的再现率都随着算法返回结果个数的增多而减小。然而在三种算法的比较中,本文提出的基于随机游走的节点中介中心度算法的再现率明显高于另外两种算法。

4 结束语

在微博中,用户可以实时地发布信息,这些信息经过重要用户转发后可以在网络中以爆炸式的方式传播,一个热点话题可以迅速蔓延到整个网络。为了监控微博网络中的话题传播,必须及时发现那些在网络中有重要影响力的个体。本文采用了节点中介中心度的思想对网络中的节点进行排名。区别于传统的中介中心度算法,本文提出了一种基于随机游走的算法。实验结果表明,本文提出的基于中介中心度的节点影响力排名算法明显优于传统的基于中介中心度的节点排名算法。

参考文献:

[1] 杨亮,林原,林鸿飞. 基于情感分布的微博热点事件发现[J]. 中文信息学报, 2012, 26(1): 84-90.

[2] 袁毅. 微博客信息传播结构、路径及其影响因素分析[J]. 图书情报工作, 2011, 55(12): 26-30.

[3] PAGE L, BRIN S, MOTWANI R, *et al*. The PageRank citation ranking: bringing order to the Web[EB/OL]. [2001-10-30]. http://ilpubs.stanford.edu:8090/422.

[4] KLEINBERG J M. Authoritative sources in a hyperlinked environment [J]. Journal of the ACM, 1999, 46(5): 604-632.

[5] LEMPEPEL R, MORAN S. The stochastic approach for link-structure analysis (SALSA) and the TKC effect[J]. Computer Networks, 2000, 33(1): 387-401.

[6] TUNKELANG D. A twitter analog to PageRank [R/OL]. (2009). http: thenoisychannel. con/2009/01/03/a-tweitter-analog-to-pagerank.

[7] WENG Jian-shu, LIM E P, JIANG Jing, *at al*. TwitterRank: finding topic-sensitive influential twitterers[C]//Proc of the 3rd ACM International Conference on Web Search and Data Mining. 2010.

[8] ROMERO D, GALUBA W, ASUR S, *et al*. Influence and passivity in social media[C]//Proc of Machine Learning and Knowledge Discovery in Databases. 2011: 18-33.

[9] KONG Shou-bin, FENG Ling. A tweet-centric approach for topic-specific author ranking in micro-blog[C]//Proc of the 7th International Conference on Advanced Data Mining and Applications-Volume Part I. 2011: 138-151.

[10] ZHANG Meng, SUN Cai-hong, LIU Wen-hui. Identifying influential users of micro-blogging services: a dynamic action-based network approach[C]//Proc of Pacific Asia Confernce on Information Systems. 2011.

[11] BRANDES U. A faster algorithm for betweenness centrality[J]. Journal of Mathematical Sociology, 2001, 25(1): 163-177.

[12] JIN Zhao-yan, SHI Dian-xi, WU Quan-yuan, *et al*. LBSNRank: personalized PageRank on location-based social networks[C]//Proc of ACM Confence Ubiquitous Computing. 2012: 980-987.

[13] RICHARDSON M, AGRAWAL R, DOMINGOS P. Trust management for the semantic Web[C]//Proc of the 2nd International Semantic Web Conference. 2003: 351-368.

(上接第 123 页)

[6] HO S L, YANG S, NI G, *et al*. A particle swarm optimization method with enhanced global search ability for design optimizations of electromagnetic devices [J]. IEEE Trans on Magnetics, 2006, 42(4): 1107-1110.

[7] XIE X, ZANG W, YANG Z. Dissipative swarm optimization[C]//Proc of IEEE Congress on Evolutionary Computation. [S.l.]: IEEE Press, 2002: 1456-1461.

[8] LIU B, WANG L, JIN H Y. Improved particle swarm optimization combined with chaos [J]. Chaos, Solitons & Fractals, 2005, 25(5): 1261-1271.

[9] 吕振肃,侯志荣. 自适应变异的粒子群优化算法[J]. 电子学报, 2004, 32(1): 416-420.

[10] MUKHOPADHYAY S, BANERJEE S. Global optimization of an optical chaotic system by chaotic multi swarm particle swarm optimization [J]. Expert Systems with Applications, 2012, 39(1): 917-924.

[11] 陈如清,俞金寿. 混沌粒子群混合优化算法的研究与应用[J]. 系统仿真学报, 2008, 20(3): 685-688.

[12] 刘玲,钟伟民,钱锋. 改进的混沌粒子群优化算法[J]. 华东理工大学学报, 2010, 36(2): 267-272.

[13] HE D, HE C, JIANG L G, *et al*. IEEE TENCON [C]//Proc of A Chaotic Map With Infinite Collapses. Kuala Lumpur: IEEE Press, 2000: 95-99.

[14] TAVAZOEI M S, HAERI M. Comparison of different one-dimensional maps as chaotic search pattern in chaos optimization algorithms[J]. Applied Mathematics and Computation, 2007, 187(2): 1076-1085.

[15] 徐文星,耿志强,朱群雄,等. 基于 SQP 局部搜索的混沌粒子群优化算法[J]. 控制与决策, 2012, 27(4): 557-561.

[16] 李祚勇,汪嘉杨,郭淳. PSO 算法优化 BP 网络的新方法及仿真实验[J]. 电子学报, 2008, 36(11): 2224-2228.