基于改进PageRank算法的微博用户影响力研究

□何 静 郭进利

[摘要]随着计算机技术的快速发展，微博已成为主流的在线社交网络平台。面对大规模的用户群，影响力成为衡量用户价值的一个重要指标。综合考虑微博用户关系网络特性和用户行为，提出了一种基于PageRank的微博用户影响力评价模型，用来评估用户影响力并对评估结果进行排序。实证分析结果表明，改进算法在微博应用中能够有效识别“僵尸”用户，客观地反映出用户的实际影响力。

[关键词]PageRank算法 微博 用户影响力 活跃度
[中图分类号]G206 [文献标识码]A [文章编号]1671-0029（2013）01（下）-0021-03

随着互联网的飞速发展，微博、博客、论坛等社交网络已成为人们生活工作中的一部分。微博作为新兴的即时通讯工具，具有即时发布、实时传播、多途径参与、简便易用等特点。据不完全统计，截至2011年12月，中国网民规模到达5.13亿，新浪微博注册用户数超过2.5亿，每天的信息量也突破1亿条。微博正在由最初的娱乐应用发展成为互联网的主流应用，并开始与社会现实广泛对接，并在信息传播中发挥着重要作用。

微博网络的发展过程可以看作是一个生长的复杂网络，而信息在微博网络中的传播过程又极具规律性。面对大规模的用户群，微博用户的影响力评估吸引了广大学者的研究。早期的研究者在对Tweeter的研究过程中，通常直接将粉丝数作为指标来衡量微博用户的影响力。Ye等人在对Tweeter用户的粉丝数、好友数、发布的消息数等数据进行研究后发现，用户的综合影响力与消息数量也呈现一定的相关性。Weng等人在对Tweeter用户的粉丝和关注数量进行研究后发现关注关系是高度对称的，且用户的综合影响力还与其粉丝的影响力有关。郭浩等人在对新浪微博进行研究后发现，在微博的特定话题中，参与程度较高、具有大量粉丝的用户未必是影响力高的用户。他们提出了一种计算用户影响力的算法，将用户的影响力分为直接影响力、级联影响力、绝对影响力和相对影响力，通过对用户的各种影响力指标的权重分配，最终得出综合影响力。原福永等人在对微博用户特性研究的基础上提出了用户活跃指数模型，用来评估单个用户在整个系统中的影响力，并据此计算用户活跃度。这些研究虽然给出了用户影响力的评估算法，但在一定程度上受人为因素对评价结果的影响，并不能真实地反映出用户的实际情况。

面对开放式社交网络的兴起和用户群体规模的不断扩大，僵尸用户和网络水军大量涌现，降低了网络中用户关注度的真实性。因此，用户价值排序研究对于信息传播显得尤为重要。本文在网页排序PageRank算法研究的基础上，结合微博的用户特性，提出了一种评估用户影响力的UIR（User Influence Rank）算法，并对计算结果进行排名。

一、PageRank算法原理

PageRank算法是1998 年由Sergey Brin 和Lawrence Page提出的基于链接分析的网页排序算法，其基本思想是通过统计网页被链接的次数来计算网页的重要性。它将页面重要性按一定的权重进行划分并传递到其所引用的页面，从而确定其所引用页面的能力值。假设页面A被页面B引用，即B认为A是重要的，就相当于B向A投了一票；如果A被多个页面引用，那这些页面都认为A是重要的，相当于多个页面都向A投了一票。一个网页的PageRank值可以用下式计算得到：

$$PR(p)=(1-\mathrm{d})+d\sum_{i=1}^{n}\frac{PR(T_i)}{C(T_i)}$$

其中，PR（p）表示网页P的PR值；Ti表示指向网页P的网页集；d为阻尼因子。

PageRank算法可以用P2P网络中的随机游走模型来解释，即用户从一个网页界面随机通过N次链接进入到目标网页，到达不同的页面也已看作相应的状态，这个过程可以看做是在图G上的随机游走。在实际的网络中，由于网页之间存在复杂的链入和链出关系，为了避免网页之间出现悬挂链接而导致稳态概率无法收敛的情况，引入了阻尼因子d（0<d<1），1－d表示用户从一个网页通过链接随机到达另一个网页的概率。按照PageRank算法，一个页面i的PR值只与页面i的入度和指向i的页面的PR值有关，也就是说网络的拓扑结构决定了节点的重要性。

二、基于微博用户的改进算法

1. 问题的提出

PageRank算法是用来衡量网络中网页重要性的经典算

法。该算法主要是通过分析网络节点的拓扑性质来获用户的重要性排名，其基本思想是将用户之间的链接关系看作一种投票行为。在微博用户关系网络中，微博用户可以看作网络节点，用户之间的粉丝与关注关系看作节点之间的连接边，每一个连接边都代表着一种关系。若用户A是B的粉丝，则意味着A给B投了一票；A的重要性越高，则传递到B的价值就越大。这种用户之间的连接关系与网页之间的链接相类似，因此微博用户的影响力评估也可以通过改进PageRank算法来获取。

实际中尽管从图论的角度来看，微博关系网络和Web网络拥有相似的拓扑结构，但是其网络的形成机理和应用环境却大不相同。Web网络中页面之间只有单一的链接关系，在应用PageRank计算网页重要性时不需要考虑时间维度和网页自身特性等因素的影响。但是在微博关系网络中，一个微博用户的影响力不仅与其入度即粉丝数有关，还与用户认证、微博发布频率及传播深度等因素有关。因此，若直接利用PageRank算法来评估微博用户的影响力，单纯地对其好友的PR值进行叠加，就忽略了用户自身因素的影响，并不能客观地反映真实情况。

2. 微博用户影响力评价指标

通过对微博社区的用户行为特性进行分析，发现信息在微博上的传播过程是一个典型的级联传播。在微博网络中，用户与用户之间主要通过粉丝和关注关系进行信息的传播，若某用户发布一条信息，则该信息会沿着他的粉丝向外界层层传播。因此，用户的影响力可以定义为用户对于该网络的影响程度。

在对微博用户进行影响力的评估时，根据PageRank算法和用户微博用户特性，可加入新的评价指标来改进和优化算法。

（1）用户的粉丝数。用户的粉丝数代表着该用户节点的链接边数。粉丝数越多，意味着用户的接触面越广，则信息的传播范围就越广。

（2）用户的活跃度。用户的活跃程度主要与用户发布微博的频率、转发和评论好友的微博数有关，通常会影响到信息传播的速度与范围。

（3）用户微博的传播能力。用户微博的传播能力主要由用户的微博被转发和评论的次数来衡量。由于信息在微博中呈现出级联传播，因此传播的级数代表着用户的微博传播的深度。大量的样本分析得出，一条微博传播深度为0～6，平均传播深度值约为3.5。

（4）用户粉丝的影响力。信息在微博中的传播主要由粉丝来推动，因此粉丝的影响力也影响力到博主的影响力。

在微博社区中，信息传播的深度和广度与参与用户的影响力呈现出一定的正相关性。高影响力的用户通常会成为网络社区中的意见领袖和明星人物，对于推动和干预信息的传播尤为关键。在微博关系网络中，当一个新的节点用户进入到网络中时，他往往会选择那些具有较高知名度的用户，如娱乐明星、体育明星、企业家、新闻网站等，而这些用户往往会具有较高的影响力，成为网络中的中枢节点。

3. 微博用户影响力评价模型

在原有的网页排序的PageRank算法的基础上，针对微博用户关系网中用户影响力的评估，我们提出了改进的UIR（User Influence Rank）算法。算法的基本思想是：一个用户的影响力由他粉丝的影响力和自身的活跃度来综合度量。算法描述为

$$UIR(u)=(1-d)+d\sum_{v\in f(u)}Wuv\cdot UIR(v) \quad (1)$$

其中d为（0,1）区间上的衰减系数；f(u)为用户的粉丝集合；Wuv为用户u分配给用户v的UIR值的比例。

一个用户的活跃度Wu主要由以下因素确定：用户的粉丝数、发布的微博数（发布原创微博的频率）、转发和评论好友微博的次数、用户微博的传播能力。因此，用户u的活跃度可以表示为

$$\mathrm{Wu}=\sum\omega_{ij}\cdot A_{ij}+C \quad (2)$$

其中A_{ij}为用户i的第j个影响指标，w_{ij}为相应指标A_{ij}的权重；C为用户微博的传播层级数。各影响指标A_{ij}权重w_{ij} =（W_1，W_2，W_3……W_n）的确定可由层次分析法得到。用户的粉丝数F_N、转发和评论好友微博数Z_N和P_N可分别由用户的实际数据收集得到。用户发布微博频率$\Phi=\frac{N}{T_{end}-T_{first}}$。$T_{first}$表示第一条微博的发布时间，$T_{end}$表示最后一条微博的发布时间，N表示在$T_{first}$至$T_{end}$时间内发布的微博数量。那么用户u分配给用户v的UIR值的比例为

$$Wuv=\frac{\mathrm{Wu}}{\sum_{i=1}^{n}W} \quad (3)$$

其中，N是用户u的粉丝数，Wi为用户u的第i个粉丝的活跃度。为了保证算法的合理性和收敛性，我们假设最外层用户的初始UIR值为1。

根据微博关系网络图，我们可以按照（1）式给出的方法计算每个用户的UIR值。由于信息在微博中的传播方式为级联传播，因此通过有限次的反复迭代即可求得目标用户的影响力UIR值。该算法充分考虑了微博用户的粉丝数、微博发布频率、用户的活跃度等因素，并将这些影响因素按照一定的权重配比，因此可以排除僵尸用户对计算结果的干扰，实现对用户综合影响力的评估。

三、实证结果与分析

我们选取了Sina微博（http://weibo.com）上的微博用户数据作为实证研究的样本，运用爬虫软件收集了新浪微博中某微群1092名用户的数据，包括用户的粉丝数、发布的微博数及其时间、转发和评论好友的微博数等。

首先，运用层次分析法（AHP）得出用户活跃度的各影响因子的权重ωi=（0.267,0.483,0.250）。为了使算法收敛，根据微博关系网络特性取用户粉丝层级数为3，阻尼因

子d取经验值0.85。按照上述算法计算经过有限次迭代得到用户的影响力结果如表1：

表1 某微群前十位微博用户的影响力

序号	用户昵称	粉丝数	微博数	活跃度	UIR值
1	梦想方舟	10498	5363	4.82	8.94
2	数据挖掘PHP	12227	2168	2.53	8.37
3	fengyuncrw	5713	4611	4.73	7.58
4	Jee-e	911	7145	6.73	7.32
5	戴虎宁	3640	2340	3.56	7.21
6	mindy1	1180	5573	4.74	6.59
7	翟变变变	2076	973	1.03	6.18
8	pv-vq	1676	1415	1.98	5.99
9	扁骰子	880	1208	2.16	5.75
10	Ashleylv	670	1201	3.80	5.49

给出各用户的粉丝数、影响力和活跃度之间的关系，如图1：

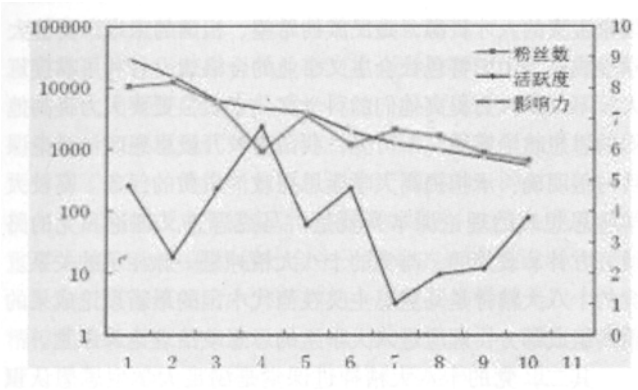

图1 双坐标下的微群用户粉丝数与用户影响力的对比

从实验结果可以看出，用户的影响力与用户的粉丝数和活跃度等因子之间并非呈现出一定的正相关性。粉丝数高的用户，其影响力不一定也高；同样活跃度高的用户的影响力也不一定高。但是相比用户的活跃度，用户的粉丝数对用户影响力的影响较大，其在改变的用户的影响力中起到了主导作用，而用户活跃度对用户影响力的作用则是相对较小的。这一结论与实际的微博应用中的情况基本相一致。在实际的微博中，存在着明星用户、普通用户和僵尸用户等等。对于明星用户和普通用户来说，其粉丝数和活跃度在一定程度上决定了他在局域网络中的影响力；而对于僵尸用户，由于其活跃度很小，计算得到的影响力也就非常小。因此，通过该算法基本上能够客观真实地反映微博用户的实际影响力。

四、结论

微博作为一种新兴的即时通讯工具，成为人们获取信息的重要媒介。本文在对PageRank算法分析的基础上提出了微博用户影响力的评价模型，并对微博用户进行了实际评估，得到用户的综合影响力能够客观地反映实际情况。这一评估算法对于商业营销具有很大的现实意义。

作为衡量微博用户的重要指标，用户的影响力表现为他在整个网络中的价值。这些影响力大的明星用户通常会对以他为中心的局域网络形成领导作用，从而推动信息的快速传播。面对发展迅速和竞争激烈的电子商务来说，通过微博等在线社交网络进行信息发布和商品推介宣传，较大影响力的用户无疑是最佳选择。此外，对于现实的人际关系网络来说，通过发掘用户的影响力和价值，还可以提高社会交往的质量和效果。

（作者单位：上海理工大学管理学院）

本文系国家自然科学基金项目（批准号：70871082）和上海市研究生创新基金项目（JWCXSL1202）。

参考文献

[1]中国互联网信息中心CNNC.第29次中国互联网络发展状况统计报告[R].2012（1）.

[2]郭浩，陆余良，王宇等.基于信息传播的微博用户影响力度量[J].山东大学学报（理学版），2012，47（5）.

[3]原福永，冯静，符茜茜.微博用户的影响力指数模型[J].现代图书情报技术，2012（6）.

[4]Kamvar S.Ex trapolation Methods for Accelerating PageRank computations[D].USA: Stanford University,2003.

[5]李稚楹，杨武，谢治军.PageRank算法研究综述[J].计算机科学，2011，38 （10A）.

[6]段庆锋，朱东华，汪雪锋.基于改进PageRank 算法的引文文献排序方法[J].情报理论与实践，2012，1（35）.

[7]詹圣君，邵雄凯，刘建舟.一种考虑用户行为的改进N—PageRank 算法[J].计算机技术与发展，2011，21（8）.

[8]李娜.企业微博营销策略研究[J].中州大学学报，2011.28（6）.

责任编辑：张硕